Opinion

On the evolution of the standard amino-acid alphabet

Yi Lu and Stephen Freeland

Address: Department of Biological Sciences, University of Maryland, Baltimore County, Baltimore, MD 21250, USA.

Correspondence: Stephen Freeland. Email: freeland@umbc.edu

Published: 1 February 2006

Genome ***Biology*** 2006, **7:**102 (doi:10.1186/gb-2006-7-1-102)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/1/102

Abstract

Although one standard amino-acid 'alphabet' is used by most organisms on Earth, the evolutionary cause(s) and significance of this alphabet remain elusive. Fresh insights into the origin of the alphabet are now emerging from disciplines as diverse as astrobiology, biochemical engineering and bioinformatics.

At the root of biology there are a handful of biochemical standards, the ubiquity of which tempts us to take them for granted. One is the standard 'alphabet' of 20 encoded amino acids, shared by organisms that diverged as early as *Escherichia coli* and human beings. But numerous lines of evidence, from abiotic chemistry to protein engineering, combine to indicate that this alphabet could potentially have consisted of fewer, more, or just plain different amino acids. So why have these 20 become the standard alphabet?

Extensive scientific research has explored both the order by which amino acids entered the primordial genetic code and the ways in which variations of the alphabet affect the structure and function of proteins. But knowing the history of the alphabet's formation and appreciating the high tolerance of protein structures for alternative constituents merely highlights the deeper question of the alphabet's cause. New research, from synthetic biology [1,2], genomic analysis [3] and computational biochemistry [4,5], is shedding new light on the question. Greater understanding in this area would potentially help scientific adventures as diverse as the search for extraterrestrial life and the drive to improve standard bioinformatic procedures such as homology detection and protein-structure prediction.

Why ask why?

Given the phenotypic diversity that has evolved, the revelation in the 20th century of a highly conserved biochemical framework beneath that diversity was remarkable. This uniformity - which goes from the structure of DNA, via the 'central dogma' of molecular biology that 'genes make RNA make proteins', to the codon assignments of the standard genetic code - spurred the scientific revolution that has carried us into the post-genomic era.

But behind the biochemical canon lie the deeper questions of why life is built this way, including the question of why proteins are constructed using a standard alphabet of exactly these 20 amino acids. Although recent publications have considered similar questions for nucleic acids [6], nucleotides [7] and even ribose [8], the cause(s) of the amino-acid alphabet have not been fully and directly addressed in more than two decades [9]. Indeed, most authors have considered the amino-acid alphabet as a mere sub-component of a multifaceted phenomenon - the genetic code [10,11]. But understanding whether the amino-acid alphabet reflects some independent logic of its own would provide valuable input on two very different research fronts.

In one direction, as astrobiology turns skywards to search for extraterrestrial life [12], it behoves us to ask what exactly we are looking for. Should we anticipate a more or less universal biochemistry? Pace [13] and Benner *et al.* [14] have each considered this question, only to reach opposite conclusions. Without a quantitative framework for these analyses, it is hard to evaluate who has the stronger argument. At worst, the current absence of such a framework seems to

encourage the specter of pseudo-scientific claims of a mysterious 'external force' directing the natural world that continues to haunt American popular culture [15].

With feet firmly back on Earth, a deeper understanding of amino-acid biochemistry is also of major importance to the emerging field of bioinformatics. In particular, protein sequence alignment (which underpins homology searching, phylogenetic reconstruction and even protein-structure prediction) is built up essentially from a quantitative model of amino-acid similarity. Increasingly, researchers are seeking further improvements here by replacing generalized, global models of observed amino-acid substitution patterns (models, such as PAM [16] and BLOSUM [17], that apply to all proteins in all organisms) with specialized models, such as those used for particular protein families [18,19] or for genomes that have evolved under unusual mutation biases or selection regimes [20-22]. Discovering in detail how the amino-acid alphabet evolved (developing its 'quantitative etiology') could make it possible to unify such models into a common theoretical framework derived from biophysical considerations.

In fact, these two seemingly very different research frontiers, exobiology and bioinformatics, meet at several unexpected junctures. For example, some researchers interpret recent insights into the variation and distribution of protein folds as clues that the particular protein families that we find populating our biosphere were as inevitable to evolution as inorganic crystal structures are to physics [23]. This fascinating idea is of equal relevance to drug design and protein-structure prediction as it is to exobiology. Its proponents have so far, however, failed to consider the role of the amino-acid alphabet from which protein folds are constructed. If the standard alphabet were different, what would the impact be on protein evolution? Analysis of protein-space fold suggests that the answer is not trivial [3-5]. Encouragingly, emerging technologies such as chemoinformatics are opening up new approaches to the exploration of amino-acid etiology, more cheaply and rapidly than anything that has been done before. The time is ripe to reassess what we know and thus to highlight directions for future investigation.

Could alternative alphabets have been encoded?

In seeking a justification for the 20 amino acids we have, we imply that other alphabets were possible. Is this really the case? Early explanations for the size and content of the standard alphabet worked from the very premise that what we see today was somehow an inevitable outcome (see [24] for a review). But as scientific progress undermined these flawed ideas, only one argument against alternative alphabets retained its plausibility. This was the general evolutionary observation that as organisms evolve an increasing complexity, emerging characters can easily become 'locked in' by subsequent evolutionary innovations that are adaptive only in relation to these early characters. Perhaps, then, the first amino acids to enter the code, for whatever reason, were frozen into evolutionary history by a proteome (and hence metabolism) built from them?

Until recently, it did indeed appear that the potential for proteomic disruption was preventing any natural turnover of the standard amino-acid alphabet. Even the discovery of a widely distributed, 21st 'encoded' amino acid - selenocysteine (Sec) - appeared to support this view, once it was realized that significant extra molecular machinery is required for selenocysteine translation. Specifically, there is no explicit selenocysteine aminoacyl-tRNA synthetase that charges an appropriate tRNA; rather, serine aminoacyl-tRNA synthetase charges $tRNA^{Sec}$ with (canonical) serine [25]. Enzymes then modify the serine into selenocysteine *in situ* while it is attached to the tRNA. Furthermore, a *cis*-encoded mRNA secondary structure downstream of the relevant codon is required to pause translation long enough for special elongation factors to supervise the incorporation of selenocysteine (reviewed in [26]). All in all, one might view this as *prime facie* evidence that that the standard amino-acid alphabet is hard to change.

Biochemical engineering has, however, steadily built up a contrasting picture of flexibility that suggests that a rethink is in order. To start with, something close to 100 non-standard amino acids have been successfully incorporated into various 'natural' protein structures [27,28]. The biochemistry of protein folds does not therefore tightly restrict the contents of the alphabet - although it remains to be seen whether different alphabets could enable fundamentally different folds. Nor is the alphabet directly and obviously limited by constraints of the translational machinery, as several studies have introduced 'unnatural' amino acids into the genetic code [1,2] through rational modification of appropriate tRNAs and the aminoacyl-tRNA synthetase molecules that charge them (see [29] and references therein).

Most directly of all, the discovery of a 22nd encoded amino acid, pyrrolysine, shows that the alphabet can grow and change naturally, not just in the laboratory. Like the 20 standard amino acids, pyrrolysine has its own aminoacyl-tRNA synthetase and its translation requires no unusual *cis* or *trans* elements (see [29] for an overview). Viewed in this light, the special decoding arrangements for selenocysteine, including its *in situ* modification from seryl-tRNA into selenocysteinyl-tRNA, can be interpreted as exactly the sort of evolutionary intermediate that might be expected to arise during alphabet expansion under natural selection, as a way of minimizing disruption to preexisting coded protein products. Indeed, the knowledge that *in situ* tRNA modification is exactly how two of the standard amino acids (glutamine and asparagine) are coded in many microorganisms adds credibility to this interpretation (see [30] and references therein) and sits well with theories for the origin of the standard alphabet.

Where did the standard alphabet come from?

The biggest single clue to understanding the origin of the standard amino-acid alphabet comes from our understanding of the prebiotic chemistry of Earth (see, for example [31]) and space (see, for example [32]), which suggests that amino acids were likely to have been obvious commodities that primordial life has exploited. The standard amino-acid alphabet is no mere passive reflection of chemistry, however: any correlation between an amino acid's likely prebiotic abundance and its presence within the standard alphabet is weak [9]. Moreover, even the most optimistic assessment admits that lysine, arginine and histidine have never been observed in simulation experiments or in meteorites [33]. In other words, it is clear that not all prebiotically synthesized amino acids ended up in the standard alphabet, and equally clear that not all members of the standard alphabet were prebiotically synthesized (Figure 1).

The latter observation has received the most attention to date, stimulating theories that at least some of the 20 standard amino acids originated as biosynthetic modifications of the others (Figure 1). In particular, Wong [34] extensively developed the idea that the order in which amino acids were added to the alphabet can be seen from the metabolic pathways by which amino acids are biosynthesized in present-day organisms.

But even a 'consensus order' [35] derived from many different precise models of alphabet expansion cannot explain the current situation fully, because all organisms biosynthetically derive amino acids that are definitely not incorporated into the genetic code. Of these, ornithine, citrulline and homoserine are three of the most ubiquitous, although for many lineages the total number is undoubtedly in the hundreds, if not the thousands [36]. Moreover, post-translational modification introduces many further amino acids into proteins without them ever being 'coded' in any meaningful sense [37]. Of course the term 'amino acid' describes the infinite series of molecular structures that contain both an amino and a carboxyl (acid) group, many of which could plausibly be biosynthesized by the right protein machinery. And let us not forget that within the standard alphabet, proline does not meet even these minimal requirements because it is an amino acid in which a cyclic side chain binds back to the 'backbone' nitrogen, generating a C=NH group where the amino acids have the NH_2 group.

At a deeper level, it is not entirely clear why early evolutionary expansion of the alphabet should have occurred at all. Experimental and theoretical analyses of amino-acid alphabet size (see [38,39], respectively, and references therein) suggest that a much smaller amino-acid alphabet might be sufficient to produce most of the fold structures that have been observed. Such hypothetical alphabets are much more plausible starting points, given the amino acids that are thought to have been generated by prebiotic chemistry. So we need to ask again, why have these 20 amino acids been used in the code?

Evolutionary causes for the size and contents of the alphabet

To date, only one publication from 1981 has offered detailed, case-by-case, feature-by-feature justifications for the members of the standard amino-acid alphabet [9], "... on the basis of the availability in the primitive ocean, function in proteins, the stability of the amino acid and its peptides, stability to racemization, and stability on the transfer RNA". The specific explanations given for individual members of the standard alphabet in this work [10] were all strictly qualitative, however, and they are hard to assess, beyond being plausible. At best, then, we have some good ideas for the themes involved in amino-acid alphabet selection. At worst, we have untestable explanations that critics could dismiss as 'adaptive storytelling'. One pointed example is that the dismissal of β-amino acids on the grounds that they could not support stable secondary structures [9], turns out to be incorrect [40].

Contrasting with these specific arguments, others have certainly suggested general, adaptive criteria, although often only as brief comments within work of a different primary focus. Among the most common is that the amino-acid alphabet was somehow selected for its biochemical diversity: for example, Szathmary [41] suggests that "proteins provided a greater catalytic versatility than nucleic acids (20 versus 4 building blocks)". But simulations of protein evolution consistently indicate a high degree of functional redundancy in the standard alphabet (see, for example, [42,43]), suggesting that diversity alone is not a good explanation. Also, at an intuitive level, the presence of the very similar amino acids valine, leucine and isoleucine suggests that biochemical diversity is hardly maximized in the standard alphabet. Another possible explanatory factor derives from the observation that bulky amino acids, such as phenylalanine and tyrosine, are used much less within 'natural' proteins than simple and small alternatives [44,45]. Perhaps entry into the standard alphabet was restricted to the smallest and cheapest amino acids that could form a functional protein library following simple, economic principles?

Of course, many other adaptive criteria can easily be formulated; the question is how we can render such speculations as testable science. In principle, statistical analysis would allow us to test whether the standard amino-acid alphabet forms a non-random collection against the background of plausible alternatives, provided we have reliable, quantitative metrics of important biophysical properties (for example, size, charge and hydrophobicity) for all the relevant molecules. A wealth of such data already exists for the 20 standard amino acids: indeed, the AAIndex database [46,47] has collated many of these into a free online resource. These data do

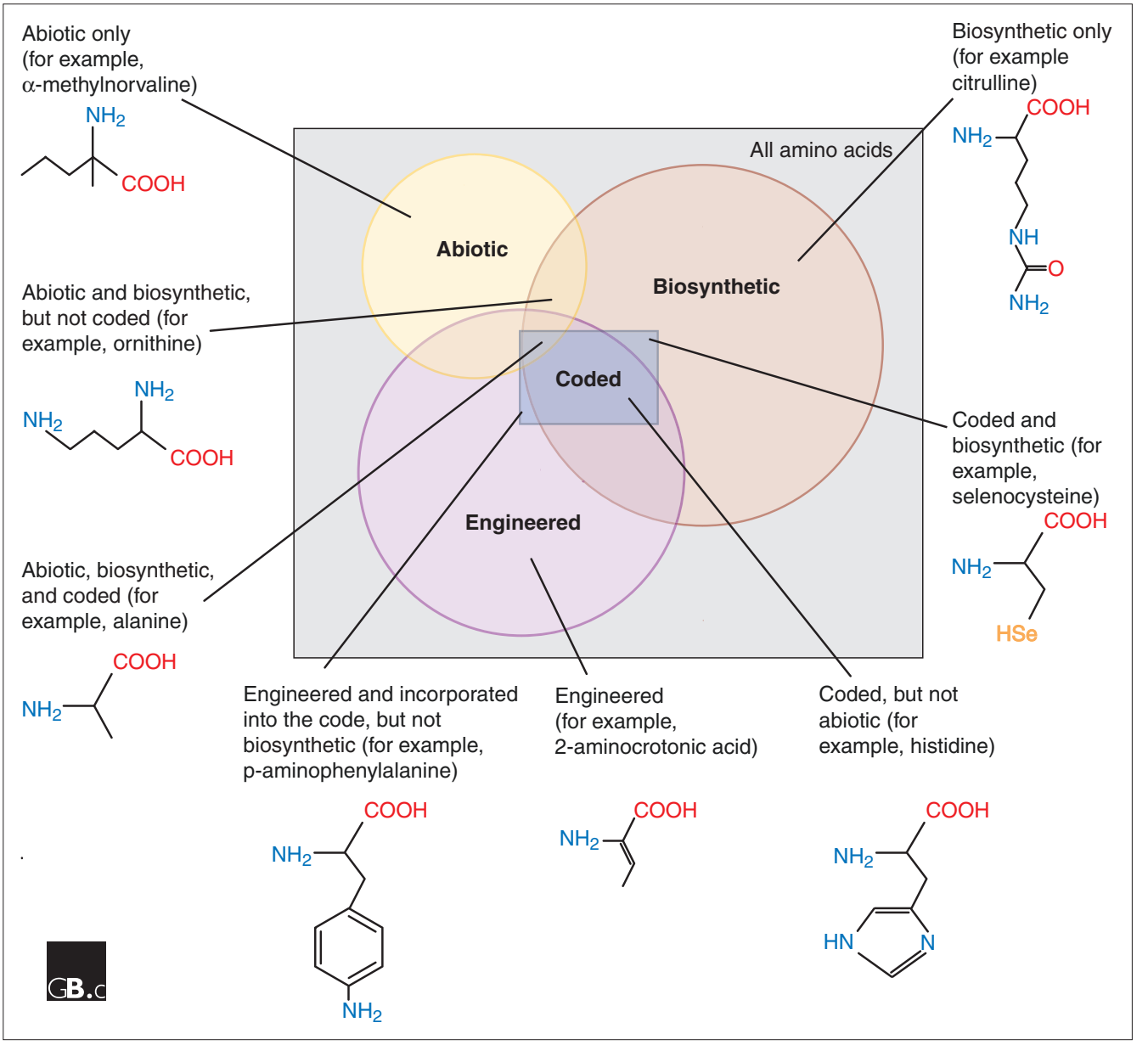

Figure 1
A Venn diagram showing different categories of amino acids: abiotic, approximately 80 amino acids which were probably produced by abiotic synthesis before life evolved (see, for example, [53]); biosynthetic, approximately 900 amino acids which are produced by natural biosynthetic pathways [54,55]; and engineered, at least 118 amino acids which have been experimentally engineered and placed into proteins by biomedical research projects [56]. The group of coded amino acids includes the standard amino-acid alphabet of 20 coded amino acids and the coded and biosynthetic amino acids selenocysteine [26] and pyrrolysine [29], as well as at least 30 engineered amino acids which have been cotranslationally incorporated into proteins [28]. One example is shown for each region of the Venn diagram. At least some of the 20 coded amino acids are thought to have originated as biosynthetic modifications of the others. The diagram shows that the 20 coded amino acids of the standard amino-acid alphabet are a small subset of what was chemically and/or biologically possible.

not, however, extend to non-standard amino acids, for the simple reason that synthesis of a molecule and analysis of its biophysical properties is a slow and expensive endeavor, even for a small molecule. For the hundreds of biosynthetically available alternatives, let alone the thousands that are biochemically plausible, such constraints are prohibitive.

New technologies to address old questions

It is in the analysis of the properties of hundreds of compounds that emerging technologies seem set to open new research possibilities. Specifically, the explosive growth in computational power and sophistication that biologists encounter through bioinformatics extends into chemical

realms ('chemoinformatics'), particularly in the form of algorithms to predict the shape and properties of user-defined molecules (see, for example, [48]). Although accurate predictions remain elusive for macromolecules such as proteins [49], there have been steady improvements in the prediction of structure (see, for example, [50]) and biophysical properties (see, for example, [51]) of smaller molecules. This, then, offers a relatively quick and low-cost approach to exploring the chemically possible amino acids. Theoretical predictions must be developed with caution, under the guidance of empirical data; this challenge is easily met when considering amino acids, however, because the experimentally derived metrics of the 20 standard amino acids offer a natural 'control group' for testing the accuracy of computational predictions.

Thus, the computational infrastructure of 21st-century biochemistry puts us within reach of asking what, if any, properties of the standard amino-acid alphabet distinguish its contents from the vast array of prebiotically and biosynthetically plausible alternatives - and for only a modest investment of time and money. It is possible that this cornerstone of biochemistry will defy all attempts at logical explanation, leaving us to conclude that the emergence of the standard amino-acid alphabet was an entirely arbitrary outcome. It would certainly match one school of evolutionary thinking [52] if it was discovered that the whole of life is in fact built upon meaningless accidents of chemistry and history.

What is important is that we can now see ways to ask such questions with scientific rigor. Indeed, as this and other questions of biochemical etiology become amenable to rigorous scientific inquiry, the life sciences will be contributing directly to cosmology: there are few biological questions deeper than asking to what extent life (either our kind of life or indeed any kind of life) was implicit within the physics of this universe.

Acknowledgements

This work was funded in part by NASA Exobiology award NNG04GJ72G and NSF award DBI 0317349. We thank Gang Wu, Blasej Bulka, Wen Zhu and Nick Keulmann for insights and comments that improved this article.

References

1. Hahn ME, Muir TW: **Manipulating proteins with chemistry: a cross-section of chemical biology.** *Trends Biochem Sci* 2005, **30:**26-34.
2. Benner SA, Sismour AM: **Synthetic biology.** *Nat Rev Genet* 2005, **6:**533-543.
3. Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E: **Protein structure and evolutionary history determine sequence space topology.** *Genome Res* 2005, **15:**385-392.
4. Bastolla U, Roman HE, Vendruscolo M: **Neutral evolution of model proteins: diffusion in sequence space and overdispersion.** *J Theor Biol* 1999, **200:**49-64.
5. Cellmer T, Bratko D, Prausnitz JM, Blanch H: **Protein-folding landscapes in multichain systems.** *Proc Natl Acad Sci USA* 2005, **102:**11692-11697.
6. Szathmary E: **Why are there four letters in the genetic alphabet?** *Nat Rev Genet* 2003, **4:**995-1001.
7. Eschenmoser A: **Chemical etiology of nucleic acid structure.** *Science* 1999, **284:**2118-2124.
8. Schoning K, Scholz P, Guntha S, Wu X, Krishnamurthy R, Eschenmoser A: **Chemical etiology of nucleic acid structure: the alpha-threofuranosyl-(3'→2') oligonucleotide system.** *Science* 2000, **290:**1347-1351.
9. Weber AL, Miller SL: **Reasons for the occurrence of the twenty coded protein amino acids.** *J Mol Evol* 1981, **17:**273-284.
10. Knight RD, Freeland SJ, Landweber LF: **Selection, history and chemistry: the three faces of the genetic code.** *Trends Biochem Sci* 1999, **24:**241-247.
11. Wong JT-F: **Coevolution theory of the genetic code at age thirty.** *BioEssays* 2005, **27:**416-25
12. Bada JL: **Astronomy: a field with a life of its own.** *Science* 2005, **307:**46.
13. Pace NR: **The universal nature of biochemistry.** *Proc Natl Acad Sci USA* 2001, **98:**805-808.
14. Benner SA, Ricardo A, Carrigan MA: **Is there a common chemical model for life in the universe?** *Curr Opin Chem Biol* 2004, **8:**672-689.
15. Lynch M: **Simple evolutionary pathways to complex proteins.** *Protein Sci* 2005, **14:**2217-2225.
16. Dayhoff MO, Schwartz RM, Orcutt BC: *Atlas of Protein Sequence and Structure*. Washington DC: National Biomedical Research Foundation; 1978.
17. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89:**10915-10919.
18. Vilim RB, Cunningham RM, Lu B, Kheradpour P, Stevens FJ: **Fold-specific substitution matrices for protein classification.** *Bioinformatics* 2004, **20:**847-853.
19. Teodorescu O, Galor T, Pillardy J, Elber R: **Enriching the sequence substitution matrix by structural information.** *Proteins* 2004, **54:**41-48.
20. Yu YK, Altschul SF: **The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions.** *Bioinformatics* 2005, **21:**902-911.
21. Bastien O, Roy S, Marechal E: **Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions.** *C R Biol* 2005, **328:**445-453.
22. Pacholczyk M, Kimmel M: **Analysis of differences in amino acid substitution patterns, using multilevel G-tests.** *C R Biol* 2005, **328:**632-641.
23. Denton MJ, Marshall CJ, Legge M: **The protein folds as platonic forms: new support for the pre-Darwinian conception of evolution by natural law.** *J Theor Biol* 2002, **219:**325-342.
24. Hayes B: **The invention of the genetic code.** *Am Sci* 1998, **86:**8-14.
25. Small-Howard AL, Berry MJ: **Unique features of selenocysteine incorporation function within the context of general eukaryotic translational processes.** *Biochem Soc Trans* 2005, **33:**1493-1497.
26. Hatfield DL, Gladyshev VN: **How selenium has altered our understanding of the genetic code.** *Mol Cell Biol* 2002, **22:**3565-3576.
27. Hendrickson TL, de Crecy-Lagard V, Schimmel P: **Incorporation of nonnatural amino acids into proteins.** *Annu Rev Biochem* 2004, **73:**147-176.
28. Xie J, Schultz PG: **Adding amino acids to the genetic repertoire.** *Curr Opin Chem Biol* 2005, **9:**548-554.
29. Zhang Y, Baranov PV, Atkins JF, Gladyshev VN: **Pyrrolysine and selenocysteine use dissimilar decoding strategies.** *J Biol Chem* 2005, **280:**20740-20751.
30. Wong JT: **On the formation of Asp-tRNA(Asn) by aspartyl-tRNA synthetases.** *Bioessays* 2005, **27:**1309.
31. Miller SL: **A production of amino acids under possible primitive earth conditions.** *Science* 1953, **117:**528-529.
32. Shock EL: **Astrobiology: seeds of life?** *Nature* 2002, **416:**380-381.
33. Miller SL: **Current status of the prebiotic synthesis of small molecules.** *Chem Scr* 1986, **26B:**5-11.
34. Wong JT: **A co-evolution theory of the genetic code.** *Proc Natl Acad Sci USA* 1975, **72:**1909-1912.
35. Trifonov EN: **Consensus temporal order of amino acids and evolution of the triplet code.** *Gene* 2000, **261:**139-151.
36. Bell EA, John DI: **Amino acids.** In *Organic Chemistry. Series 2, Volume 6. Amino Acids, Peptides and Related Compounds*. Edited by Rydon HN. London: Butterworths; 1976, 1-32.
37. Baumann M, Meri S: **Techniques for studying protein heterogeneity and post-translational modifications.** *Expert Rev Proteomics* 2004, **1:**207-217.

38. Walter KU, Vamvaca K, Hilvert D: **An active enzyme constructed from a 9-amino acid alphabet.** *J Biol Chem* 2005, **280:**37742-37746.
39. Fan K, Wang W: **What is the minimum number of letters required to fold a protein?** *J Mol Biol* 2003, **328:**921-926.
40. Koyak MJ, Cheng RP: **Design and synthesis of biologically active β-peptides.** *Meth Mol Biol*, in press.
41. Szathmary E: **The origin of the genetic code: amino acids as cofactors in an RNA world.** *Trends Genet* 1999, **15:**223-229.
42. Taverna DM, Goldstein RA: **Why are proteins so robust to site mutations?** *J Mol Biol* 2002, **315:**479-484.
43. Xu YO, Hall RW, Goldstein RA, Pollock DD: **Divergence, recombination and retention of functionality during protein evolution.** *Hum Genomics* 2005, **2:**158-167.
44. Dufton MJ: **Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins?** *J Theor Biol* 1997, **187:**165-173.
45. Akashi H, Gojobori T: **Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*.** *Proc Natl Acad Sci USA* 2002, **99:**3695-3700.
46. **AAindex** [http://www.genome.jp/aaindex/]
47. Kawashima S, Ogata H, Kanehisa M: **AAindex: amino acid index database.** *Nucleic Acids Res* 1999, **27:**368-369.
48. Gonzalbes R, Doucet JP, Derouin F: **Application of topological descriptors in QSAR and drug design: history and new trends.** *Curr Drug Targets Infect Disord* 2002, **2:**93-102.
49. Ginalski K, Grishin NV, Godzik A, Rychlewski L: **Practical lessons from protein structure prediction.** *Nucleic Acids Res* 2005, **33:**1874-1891.
50. Ponce MY, Castillo Garit JA, Nodarse D: **Linear indices of the 'macromolecular graph's nucleotide adjacency matrix' as a promising approach for bioinformatics studies. Part 1: prediction of paromycin's affinity constant with HIV-1 psi-RNA packaging region.** *Bioorg Med Chem* 2005, **13:**3397-3404.
51. Eros D, Keri G, Kovesdi I, Szantai-Kis C, Meszaros G, Orfi L: **Comparison of predictive ability of water solubility QSPR models generated by MLR, PLS and ANN methods.** *Mini Rev Med Chem* 2004, **4:**167-177.
52. Gould SJ: *Wonderful Life: The Burgess Shale and the Nature of History*. New York: WW Norton; 1990.
53. Cronin JR, Pizzarello S: **Amino acids in meteorites.** *Adv Space Res* 1983, **3:**5-18.
54. Uy R, Wold F: **Posttranslational covalent modification of proteins.** *Science* 1977, **198:**890-896.
55. Fowden L: **Plant amino acid research in retrospect: from Chinball to Singh.** *Amino Acids* 2001, **20:**217-224.
56. Khosla MC, Cohn WE: **Structures and symbols for synthetic amino acids incorporated into synthetic polypeptides.** In *Handbook of Biochemistry and Molecular Biology. Volume I*. Edited by Fasman GD. Baton Rouge, FL: CRC; 1976, 96-108.