

Amino acid names and parlor games: from trivial names to a one-letter code, amino acid names have strained students' memories. Is a more rational nomenclature possible?

M. Saffran

2331 Hempstead Road, Toledo, OH 43606, USA

Abstract

This paper explores the origins of the names and single-letter abbreviations of the 20 amino acids found in proteins. Knowing the background of the nomenclature may help the student to remember the amino acid names and abbreviations. © 1998 IUBMB. Published by Elsevier Science Ltd. All rights reserved

1. Introduction

The English alphabet of 26 letters contains several letters with almost identical sound values, e.g. c and s, c and k and q. If these duplications are removed, the alphabet consists of 23 different letters and sounds. The Hebrew alphabet contains 23 letters; if the sound duplications are removed, the alphabet contains 21 letters. The Greek alphabet has 24 letters and 23 sounds. Most other alphabets cluster around the 20 to 30 letter size. Is the clustering a coincidence, or is it driven by a basic biologic force? The proteins that make up much of our body are constructed from simpler substances, the amino acids. There are 20 different amino acids in proteins. (I wonder if there is any relationship between the letters in so many alphabets and the number of amino acids.)

All amino acids contain two identical chemical groups, the amino group ($-\text{NH}_2$) and an acid group ($-\text{COOH}$). In addition, each amino acid has a characteristic side chain or R group which provides the individuality of each amino acid and determines its chemical properties. The arrangement of different amino acids in a protein molecule gives the protein its characteristics: its ability to provide shape to a cell or body structure, its ability to carry out its task of providing energy or of fighting off infections, and so on. The 20 amino acids seem to have the attributes of letters of an alphabet, which form the protein words, which carry the message.

2. The names of the amino acids

During the nineteenth century, chemists, mostly in Germany, purified and analyzed the then new family of amino acids. Proteins were too complex to be studied by the techniques of the nineteenth century, but heating the proteins in acid or alkali separated the amino acids from each other. The amino acids could now be isolated and studied by available methods. By the beginning of the twentieth century, 18 amino acids were identified as components of proteins. As each amino acid was described, the discoverer gave it a name depending upon its source, its properties, or its chemical structure. Except for the suffix *-ine*, for (am)-*ine*, in the name of most amino acids, there was no systematic naming of the amino acids. As a result students are faced with the task of memorizing the names of the amino acids as an exercise in rote learning. Also, the chemists of the nineteenth century did not know that boiling in acid destroyed two amino acids: these were not identified until the first and third decades of the twentieth century.

A good dictionary provides information about the naming of the amino acids. My source is *The American Heritage Dictionary of the English Language*, edited by William Morris, published in 1970 by the American Heritage Publishing Co. and Houghton Mifflin Co., Boston, MA. Glycine, the smallest amino acid, was so

named because it was sweet tasting. The prefixes, gly- or glu-, are derived from the Greek *γλυκερος*, meaning sweet. The amino acid, arginine, was isolated as crystals resembling white, shiny silver (*αργινοεισ* in Greek). Aspartic acid and asparagine were isolated from asparagus, while glutamic acid and glutamine were named after their source, the wheat protein, gluten. Histidine was isolated from tissues (cf. histology, the study of tissues), and lysine from tissues undergoing lysis, or liquefaction. Alanine was so-called because of the mistaken impression that it contained an aldehyde group (Al-a-nine). Tyrosine was first isolated from cheese (*τυρος* = Greek for cheese). Serine was isolated from ser(ic)ine, a protein that adheres to silk (*sericus* is the Latin adjective, silky). Cysteine was isolated from urine, which comes from the urinary bladder, or cyst.

A few of the amino acids were named for their resemblance to other chemicals or to their chemical components. Methionine is a compound name composed of Me(methyl)+Thio(sulfur)+N+ine. Phenylalanine is indeed a phenyl group (benzene ring)+alanine. Proline is named after the P(yr)ROL(ring)+ine. Threonine, discovered in the 1930s, was named after its resemblance to the sugar, threose. Valine was named after VAL(eric acid)+ine.

Tryptophan(e) was discovered in the early years of the twentieth century and was named after the enzymes that were used instead of hot hydrochloric acid to break down the parent proteins: TRYP(sin) and pep(T)ic enzymes+OPHANE, a made-up suffix, as in cellophane. Its discovery was celebrated in verse in *Brighter Biochemistry*, an informal publication of the Biochemistry Department at Cambridge University, where tryptophane was discovered:

In our laboratory, there are two classes.
Some are genii, others asses.
The former discovered tryptophane,
the latter flushed it down the drain.

3. Rational nomenclature

If the amino acids had been discovered in the twentieth, rather than the twenty-first century, a more rational nomenclature might have evolved. For example, alanine, with its chain of three carbon atoms, might have been named as a derivative of propane, the hydrocarbon with a three-carbon chain: alanine = 2-amino propanoic acid. In fact, the synonym for alanine in the Merck Index of chemical compounds is just that! Similarly, leucine can be named as a derivative of the five-carbon chain, pentane, as 2-amino, 4-methyl-pentanoic acid.

Another system of nomenclature can consider all amino acids, except glycine, as derivatives of alanine. I

have already mentioned phenylalanine as an example. Serine would then be 3-hydroxyalanine, and leucine, 3-dimethylmethylalanine, etc. But usage has enshrined the original amino acid names in the literature and the text books and change would be too disruptive.

4. Abbreviations

As methods evolved to study the amino acid composition and sequence in proteins, writing out the full names of the amino acids became tedious. A joint nomenclature commission of the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Biochemistry (IUB), chaired by H.B.F. Dixon from the Department of Biochemistry in Cambridge, examined the names of the amino acids and their derivatives, along with the abbreviation of the amino acid names to the first three letters of their names. The shortened names became very popular. Soon, however, complications arose. To avoid overloading the proof-reading skills of the scientific world, Asn replaced the earlier Asp-NH₂ for asparagine and Glu.NH₂ became Gln for glutamine. The easily confused pair, Tyr for tyrosine and Try, for tryptophan, became Tyr (unchanged) and Trp. Ileu for isoleucine was simplified to Ile. Soon, the three-letter abbreviations became tiresome to write, and a one-letter abbreviation system was proposed by a group of Czech biochemists. A 20-letter amino acid alphabet was designed and adopted by the IUPAC-IUB Joint Commission on Biochemical Nomenclature. The deliberations and decisions of the Joint Commission were published in about 10 of the major biochemical journals in five languages (see Ref. [1] for details).

5. The one-letter system

The simple approach, using the first letter of the name of the amino acid, only sufficed for those compounds with unique first letters: C could only stand for cysteine, H for histidine, M for methionine, and V for valine. But alanine, arginine, aspartic acid, and asparagine could not all be represented by the first letter A! Glutamic acid, glutamine, and glycine shared G. Leucine and lysine begin with L. Threonine, tyrosine and tryptophan all begin with T. The simplest of the amino acids with shared initial letters was assigned that letter. Alanine became A, glycine became G, leucine became L, and threonine became T. For the others, the members of the Nomenclature Committee of the Biochemical Society apparently applied their skills at the parlor game of charades, in which the participants guess the name of the

person or the object with the aid of visual clues. For example, touching the nose and cupping the ear is the clue for 'sounds like nose' or *knows*. And so arginine was abbreviated to R, because arginine sounds like 'R'-ginine. Aspartic acid was abbreviated to D, because the amino acid's name sounds like aspar-D-ic acid. Similarly, glutamine became Q, sounds like Qlutamine, and phenylalanine was abbreviated to F, for F-enylalanine. Asparagine was abbreviated to N for asparagi-N-c.

So why is glutamic acid represented by E? Because aspartic acid is D, and glutamic acid is larger than aspartic acid by one carbon atom, hence glutamic acid is the next letter of the alphabet, E. While Leucine is L, the other L is lysine. Because L is taken, for leucine, lysine became K, next to L. Tyrosine is represented by Y for t-Y-rosine. Finally, tryptophan is abbreviated to W because the shape of the letter W reminded the committee of the shape of the ring structure of atoms in the side chain of tryptophan.

Most students are unaware of the ingenuity that entered into these abbreviations, and they continue to commit the list to memory. Dexter Moore [2] has proposed an elaborate system of informative acronyms to couple amino acid names, as represented by the first letters, with their chemical characteristics and their metabolic fates. Later, B.H. Nicholson [3] suggested a modification of Moore's scheme in which the amino acids were represented by their official one-letter abbreviations. I tried to use these mnemonics in lectures on the amino acids to medical students and found that the acronyms helped only as far as the next examination. In view of the push to replace rote memory with understanding, a better way might be to demonstrate very early in a course the importance of biochemistry, and indeed in all the life sciences, of knowing the structures of the amino acids and their characteristics as intimately as knowing one's family and friends. Therefore, I began the amino acid lectures by asking the students to write a list of the names and characteristics (e.g. gender, age, height, weight, personality, generosity, etc.) of 20 relatives and friends. This exercise demonstrated that committing such a list of details to memory is not an insurmountable task and that matching the 20 amino acids and their characteristics can be done easily and quickly.

6. Nomenclature problems continue

The apparent coincidence of 20 amino acids and 20+ letters in major alphabets has come full circle with the assignment of 20 letters to the amino acids. Naming newly discovered molecules is still a hit or miss activity. Some new growth factors were named after their first source, e.g. PDGF for platelet-derived growth factor, from the platelets of the blood. Others are named for their chemical and biological resemblance to well-established hormones, e.g. insulin-like growth factors I and II. Still others are named after their biological activity, e.g. maganin (Hebrew for shield) for a protective protein, and erythropoietin, for its ability to enhance the production of erythrocytes or red blood cells. Confusion reigns when several groups of investigators almost simultaneously isolate the same agent from several sources. Each laboratory then gives the agent a name based upon the source or biological activity or even after the university in which the laboratory is located (e.g. tuftsin, after Tufts University in Boston). Eventually the realization that several groups have discovered the same substance results in a shakedown of the names into the one that is most publicized, or most convenient or easiest to remember. The other terms wither and disappear.

Perhaps it is time to apply the ingenuity used in devising the single letter code for amino acids to a more rational approach to naming newly discovered molecules.

Acknowledgements

I am indebted to H.B.F. Dixon for clarification of the details of the deliberations and recommendations of the JCBN.

References

- [1] IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN), Nomenclature and symbolism for amino acids and peptides, recommendations 1983, *Biochem. J.* 219 (1994) 345–373.
- [2] D.S. Moore, Biochemically informative acronyms for the twenty common amino acids, *Biochem. Educ.* 15 (1987) 74–76.
- [3] B.H. Nicholson, Amino acid acronyms (letter to the editor), *Biochem. Educ.* 16 (1988) 49.