BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(12)

On the utility of alternative amino acid scripts

Darren R Flower

Aston Pharmacy School, Life and Health Sciences, Aston University, Aston Triangle, Birmingham, B4 7ET, UK; Email: D.R.Flower@aston.ac.uk; Phone: +44 (0)121 204 5182.

Received May 29, 2012; Accepted June 03, 2012; Published June 28, 2012

Abstract:
In this work we propose the hypothesis that replacing the current system of representing the chemical entities known as amino acids using Latin letters with one of several possible alternative symbolic representations will bring significant benefits to the human construction, modification, and analysis of multiple protein sequence alignments. We propose ways in which this might be done without prescribing the choice of actual scripts used. Specifically we propose and explore three ways to encode amino acid texts using novel symbolic alphabets free from precedents. Primary orthographic encoding is the direct substitution of a new alphabet for the standard, Latin-based amino acid code. Secondary encoding imposes static residue groupings onto the orthography of the alphabet by manipulating the shape and/or orientation of amino acid symbols. Tertiary encoding renders each residue as a composite symbol; each such symbol thus representing several alternative amino acid groupings simultaneously. We also propose that the use of a new group-focussed alphabet will free the colouring of amino acid residues often used as a tool to facilitate the representation or construction of multiple alignments for other purposes, possibly to indicate dynamic properties of an alignment such as position-wise residue conservation.

Key words: Atom pair, CDK-2, Similarity searching, Molecular similarity

Background:
Science long ago christened the standard twenty protein-making - or biogenic - amino acids; their full, informal names arising through historical happenstance. The amino acids were discovered during a hundred and thirty year period between 1806 and 1935. Asparagine was the first to be discovered; threonine the last. Two further biogenic amino acids were identified more recently: selenocysteine, the 21st amino acid, in 1986, and the 22nd, Pyrrolysine, in 2002. The three-letter and one-letter codes used universally in bioinformatics derive, in part at least, from these long names: thus "Glycine" becomes "Gly" (3-letter code) becomes "G" (1-letter code), and so on for other residues.

As one-letter codes, exclusively written using ASCII characters derived primarily - though not exclusively – from letters used in classical Latin, the twenty biogenic amino acids form an alphabet, from which the protein sequences comprising the proteome are constructed. IUPAC introduced the one-letter code for the 20 amino acids in 1968, complementing the earlier three-letter code. The IUPAC nomenclature evolved from an original proposal formulated during the 1950s by Frantisek Sorm (1913-1980). When selecting letters to represent different amino acids, Sorm omitted B, O, U, J, X, and Z. At this time, Sorm's coding was not widely used, and many thought it might spell out obscene words and offensive phrases.

The present amino acid nomenclature, either the three-letter or one-letter codes, is no more than an arbitrary historical accident. Science might easily have chosen a very different set of letters. We could map Latin letters to an arbitrary choice of Amino Acids. The IUPAC one-letter code offers one alternative but there are more. Another might choose to compare the frequency of letters in English, or other languages, to the frequency of the different amino acids. There are 26!/6! different ways to map the 26 letters in the English alphabet to the twenty different amino acids. 26!/6! works out to be 5601270293425078272000000. Listing all such possibilities, writing one such permutation every second, would take 200000 times longer than the estimated age of the universe. Allowing such free substitution,

or rather permutation, of letters representing each amino acid, we can quickly identify much longer words within naturally-occurring proteins sequences than those unearthed by previous efforts **[1, 2].** These studies assumed the conventional one-letter coding, finding HIDALGISM or ANNIDAVATE. Allowing free permutation, words such as DICHLORODIPHENYLTRICHLOROETHANE or CYCLOTRIMETHYLENETRINITRAMINE appear. They obviously map to utterly different looking sequences using the conventional coding, but the underlying pattern of permutation is the same.

Apart from the storage of sequence data, the one-letter amino acid code is key to the manual construction of pair-wise and multiple protein sequence alignments **[3]**. Likewise, the one-letter code plays a pivotal role in the human visualisation of alignments and thus their interpretation in terms of function, evolution, and structure, as well as the identification of characteristic conserved motifs. As the parsing of the complex patterns inherent within large multiple alignments is often extremely difficult, colours were long ago introduced to ease the analytical process. There are several standard conventions for colouring residues **[3]**, and attempts have been made to optimise these **[4, 5].** Generally, these colours are used to encode the grouping of amino scids according to their physical properties. As such different properties vary differently between different amino acid residues, there are thus many, many ways to group them. Using Stirling numbers of the second kind allows us to quickly identify the total number of different ways to group the 20 amino acids: over 51724 billion. However, using colour alone to show groupings, limits us to but a single choice of clustering. If we could find another way to represent the many different choices of grouping, colour could usefully take on other roles when we seek to interpret and understanding sequence conservation as revealed by multiple alignments.

Use of the one-letter IUPAC code is now so prevalent as to be near universal; as bioinformaticians, we cannot easily imagine using any other nomenclature. Yet there are arguments for doing precisely that; and in this paper, we begin to investigate this idea. Rather than suggest a definitive alternative, we introduce several strategies for renaming the amino acids, and explore how these alternative orthographies can combine synergistically with colour potentially to enhance visual analysis of multiple protein sequence alignments.

Methodology:

Rationale

To aid the construction and interpretation of multiple sequence alignments, we explore alternative orthographic symbols for the twenty standard biogenic amino acids (**Figure 1)**.

Primary orthographic encoding

We introduce first the direct one-to-one substitution of an alternative alphabet for the twenty conventional Latin letters. The choice of an alternative orthography is wholly arbitrary: there are potentially an infinite number of different alternatives to the traditional IUPAC code. We could use non-alphanumeric characters – such as @#$% – already available in extant fonts. Or choose all twenty letters from a single other language; for example, the many beautiful letter symbols of the Tibetan alphabet. Or we might choose one letter from each of twenty different languages; or several letters from each of several languages; or design a wholly-new set of previously unknown letter symbols.

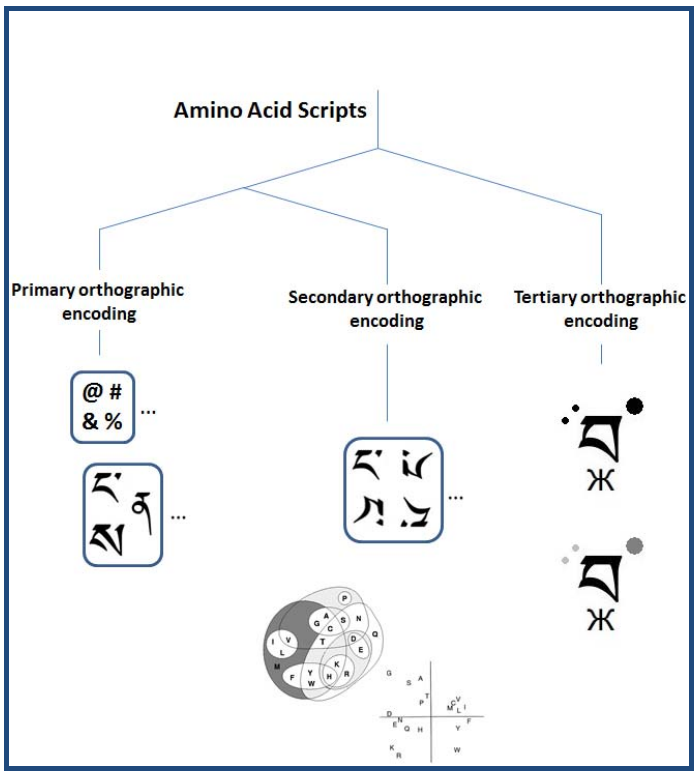

Figure 1: The schematic representation embodied in this figure attempts to capture the quintessence of the proposed alternative scripts. Primary orthographic encoding is a literal alternative with a straight substitution of one alphabet for the standard, Latin-based amino acid code. As examples of what might be possible, we show here non-alphanumeric characters and letters culled from an aesthetically pleasing but little used language. Secondary encoding attempts to impose static amino acid groupings onto the orthography of the alphabet, either by using letter rotation or by encoding similarity in residue physical properties as the similarity of shape between letters. This captures either the explicit categorisation of amino acids into defined groups or the implicit groupings more usually represented using a principal component plot. Tertiary encoding renders each amino acid as a composite symbol; each such symbol thus representing several groupings simultaneously. As indicated, orthographically-encoded groupings can be effectively augmented by colouring each element differently.

Secondary orthographic encoding

In contrast to the one-to-one substitution described above, we could encode similarity of different amino acids one-to-another via the visual similarity of the letter symbols themselves. There are several ways to achieve this. For example, the four aliphatic hydrophobic amino acids LIVM could be represented by the same letter rotated successively by 90 degrees. Or we could more subtly alter the shape of letters to capture their continuous, quantitative - rather than categoric - similarity, as is more often represented by projection of high dimensional data

into the plane **[6],** designing a novel amino acid orthography for that purpose.

Tertiary orthographic encoding

The previous encodings presume a single character to encode each amino acid. However, we could instead create symbols which combine several distinct, independent elements to form each character. Each element would encode a different and distinct grouping of the amino acids. Thus, each symbol would simultaneously capture several different aspects of the complex, multi-dimensional similarity between the amino acids comprising the protein-making alphabet.

Colour

Combining these three encodings with residue colouring opens up several interesting and potentially useful enhancements to multiple alignment construction and interpretation. While combining colour with primary orthographic encoding allows us to reproduce conventional alignment processes, combination of colour with secondary encoding allows us simultaneously to capture two kinds of grouping: one in the form of the symbols used, the other in the colours adopted. We can extend this idea with tertiary orthographic encoding, where we can combine the many groupings implicit in the multi-component symbols with an extra group represented by colours. Colouring different elements differently within each symbol can effectively double the number of groupings represented. When combined with secondary or tertiary orthographic encoding, colour can be used to represent dynamic information deriving from the alignment, rather than static, pre-imposed information about different physical property-based amino acid groupings implicit within the symbols used.

Discussion:

We have described three encodings of increasing sophistication able to represent amino acid sequences within single and multiple alignments, using new symbols free from precedent and associations. While this idea is, to the best of our knowledge, novel, it is not without thematic precedent. Multiple alignments often use non-alphanumeric symbols, such as hashes, to represent, say, conservation within columns. The program Joy **[7],** represented the different structural state of residues within multiple structural alignments by using italics, bold, or underlining to indicate the presence of hydrogen bonds or highly accessible residues.

Our suggested encodings offer three options: the direct substitution of essentially random symbols for the conventional one-letter code, building the similarity apparent in amino acid properties into the graphical similarity of a new set of amino acid symbols, and combining several orthographic elements to create new symbols that better capture the multidimensional nature of amino acid similarity. Any of these alternate scripts would, or at least should, be free of precedent in extant written languages. Pre-existing linguistic associations with so-called written natural language, as is the case with the current IUPAC code based on familiar Latin letters, precludes the ready identification of important functional patterns in sequence alignments: real biological patterns are hidden beneath linguistic ones. When constructing or interpreting multiple protein sequence alignments, we often see words or pseudo-words in the amino acid texts; much as Scrabble players are reputed to easily differentiate real from contrived words, even if such words are unknown to them. Thus freeing the human analysis of multiple alignments from psychological preconceptions should facilitate deeper and more perspicacious understanding of structure-function relations currently lying unseen in single sequences and multiple sequence alignments.

As we have said, both secondary and tertiary orthographic encoding allows for the direct representation of amino grouping and similarity between amino acid residues. The secondary encoding envisaged above attempts to impose a particular static amino acid grouping directly onto the orthography of the amino acid alphabet. It can do this by using, say, letter rotation to indicate a group of similar amino acids – ILVM for example - and thus captures the explicit categorisation of amino acids into defined group. Alternatively, we can do so by encoding the implicit similarity of residue physical properties as the similarity in shape between different symbols. This captures the implicit groupings more usually represented using a principal component plot or other data reduction technique. Tertiary encoding allows us to capture several different ways of grouping amino acids simultaneously by combining several elements in the one symbol.

In the current work, we have purposely not attempted to prescribe the precise choice of orthography, but merely to initiate a discussion, however brief, around the subject. The choice is open. We could use letters from widely written languages, such as Arabic or Hebrew, but this would represent only a minor improvement over the conventional Latin letters. Alternatively, we could utilise little-used but aesthetically-pleasing scripts such as those of the Tibetan alphabet. Better still perhaps would be to use newly designed characters entirely free of extant connotation or psychological preconceptions. A new symbolic representation would allow us to design sets of characters whose internal graphical similarity might adroitly echo the similarity in physical properties across the twenty biogenic amino acids. Optimising such representations to achieve this is best left as a collaborative exercise for bioinformaticians, font designers, and psychologists.

We have also purposely not attempted to implement this idea, yet instantiation of such scripts would, in the modern age, be relatively straightforward. There are now many so-called open source fonts (http://openfontlibrary.org/), and innumerable word processing programs able to use them; thus the mechanics of designing one or more wholly new fonts would be facile. There are also many programs which create and/or display multiple sequence alignments, and again adding the option of using such new amino acid-specific fonts should prove equally straightforward.

When combined with different orthographic encoding, colour can add additional information. This may be static information, such as a physical property-based residue grouping, or dynamic information, derived from the alignment. Dynamic information changes as the amino acid alignment changes, and could include position-wise sequence conservation; window-averaged - rather than positional - physical properties, of which hydrophobicity is the most obvious; or an alignment reliability score **[8].** Alternatively, we could use colour to encode entirely different data: for example, functional data, such as post-

translational modification sites, or structural data, such the location of secondary structure elements – helices or sheets - or the relative surface accessibility of residues.

Conclusion:

The sequences comprising strings of amino acids that form proteins constitute perhaps the most natural language we know; a language common to all humanity: a language shared irrespective of individual creed, ethnicity, culture, or political affiliation. Thus it seems appropriate to promulgate the idea that this language should have its own unique symbolic representation without precedent in any pre-existing written language. Currently, history, and the dominance of English, has dictated that the amino acids are written using the Latin letters used by the languages of Western Europe. The replacement with an alternate and hitherto unknown script will it is to be hoped free the representation of protein sequences from all literary connotation, racist euro-centricity, and psychological preconceptions of any kind.

We have deliberately avoided prescribing a new alphabet, leaving that to others, but have indicated the broad routes we might take, the many potential advantages to be gained, and the significant benefit of the whole undertaking. The wide implementation of this concept has much to recommend it, and may lead to the development of much more sophisticated software for protein sequence alignment and display, and thus to much more complete, meaningful, and far-sighted interpretations of the complex and befuddling patterns that such protein alignments contain.

References:

[1] Gonnet GH *et al*. *Nature*. 1993 **361**: 121 [PMID: 8421517]
[2] Jones D, *Nature*. 1993 **361**: 694 [PMID: 8441464]
[3] Parry-Smith *et al*. *Gene*. 1998 **221**: GC57 [PMID: 9852962]
[4] Taylor WR, *Protein Eng*. 1997 **10**: 743 [PMID: 9342138]
[5] Lin K *et al. J Theor Biol*. 2002 **216**: 361 [PMID: 12183124]
[6] Livingstone CD *et al. Comput Appl Biosci*. 1993 **9**: 745 [PMID: 8143162]
[7] Mizuguchi K *et al*. *Bioinformatics*. 1998 **14**: 617 [PMID: 9730927]
[8] Mevissen HT *et al*. *Protein Eng*. 1996 **9:** 127 [PMID: 9005433]

Edited by P Kangueane
Citation: **Flower,** Bioinformation 8(12): 539-542 (2012)