Protein Engineering vol.10 no.7 pp.743–746, 1997

FORUM

Residual colours: a proposal for aminochromography

William R.Taylor

Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Keywords: amino acid colouring/protein multiple sequence alignments

Aminochromatic blindness

'No doubt you are aware that the winds have colour. . . . There are four winds and eight sub-winds, each of which has its own colour. The wind from the east is a deep purple, the south a fine shining silver. The north wind is hard black and the west wind is amber. People in the old days had the power of perceiving these colours and could spend a day sitting quietly on a hillside watching the beauty of the winds, their fall and rise and changing hues, the magic of neighbouring winds when they inter-weaved like ribbons at a wedding. It was a better occupation than gazing at newspapers. The sub-winds had colours of indescribable delicacy, a reddish-yellow halfway between silver and purple, a greyish-green that was related equally to black and brown. What could be more exquisite than a countryside swept lightly by cool rain reddened by the south-west breeze.'

The gift of seeing these colours is very rare these days; indeed, the last recorded observation was now many years ago in a small isolated community in Ireland. [The lives (or more accurately, deaths) of these people, centering around the local constabulary, was recorded by Flann O'Brien in *The Third Policeman*, from which the opening quotation is taken. (Some neglected aspects of atomic theory, of undoubted relevance to protein structure, can also be found in this work).] Lamentable as the loss of these powers is, a matter of even graver concern is the growing sign that many people appear unable to see the colours of the amino acids.

In the days when protein sequence alignments were published only in black-and-white, much of this aminochromatic blindness passed unnoticed, but now, with plentiful coloured alignments appearing in the literature, it has become apparent that most people simply make up colours for the amino acids, seemingly at random and, worse still, fail to observe the rigid laws of colour combination when dealing with variable positions. Occasionally, some thought is applied to the problem (Gibson *et al.* 1994) but often such schemes are complex and arbitrary with respect to their choice of colours.

Although I, myself, do not have the gift of seeing the colours of the acids directly, I have managed (through late-night conversations with the officers of the local constabulary in Ballycarry) to elicit their true nature and feel bound to pass this information on to all those who would endeavour to illuminate their multiple sequence alignments.

The colour of the acids

Fixed points

All 20 amino acids have pure spectral colours (those seen in the rainbow). Given this fundamental fact, it takes only a prescribed cyclic ordering of the acids and two fixed points (say, North and East, as in the cardinal points of the compass) to establish the absolute colour of all. The cyclic ordering of the acids, while not completely unambiguous, is relatively clear and can be found in the convergence of a number of different arguments derived from the genetic code, relative mutability and basic physics and chemistry [See Taylor (1986) for a review]. This ordering is dominated by the two properties of amino acids (size and polarity) that are fundamental to the stability of protein structure. At opposite ends of the circle (say, North and South) lie the large and small acids, while orthogonal to this (say, East and West) lie the hydrophobic and hydrophilic acids.

To establish a unique orientation of the circle, both the N–S and the E–W axes must be assigned colours. One fixed point, that is seldom in dispute, is that cystine (cys, C) is yellow. This choice is inescapable, given the vivid pure yellow colour of sulphur. The second required orthogonal point is less obvious but hinges on the allocation of the amino acids with charged side groups. The negatively charged acidic groups (glutamate and aspartate) are often coloured red, following the conventional colour for oxygen in molecular models (originally from its visible manifestation in fire). Similarly, the positively charged basic groups (lysine and arginine) are coloured blue by extension from the blue colour of their charged nitrogen atoms (sky-blue to be exact). Unfortunately, some who have lingered over-long in physics laboratories reverse this convention and colour the positively charged acids red, following the colour of the positive terminal in electricity. This misconception forces black as the logical choice for the negatively charged residues, but as black is not a proper colour for an amino acid (which are all colours of the rainbow) this assignment can be rejected immediately.

The cardinal colours

Given the fixed points, the colours of the acids themselves (referred to below as *acidic colours* because of their sharp clarity), although now ranked, are still not fixed in their relative spacing. If only the three primary colours (red, green and blue) are considered, the number of acids in the range red–blue can, at most, be a quarter of the total (red = d, E, Q, N, K, r = blue, with d and r contributing one half to this sector), rather than the third required for even spacing around the red–green–blue wheel. This arithmetic inconvenience, however, can be overcome by declaring yellow as a 'primary' colour, giving four equidistant cardinal colours. Four acids can now be assigned to cardinal colours with four lying between, giving equal spacings for the 20 acids around this four-colour circle (Table I). The resulting

Table I. The colours of the amino acids

Amino acid			Colour			R	G	B
Glutamine	(Q)	Gln	RBR	Reddish-bluey-red	(magenta)	1.0	0.0	0.8
Glutamate	(E)	Glu	BRR	Blueish-reddey-red	(violet)	1.0	0.0	0.4
Aspartate	(D)	Asp	RRR	Reddish-reddey-red	(red)	1.0	0.0	0.0
Serine	(S)	Ser	YRR	Yellowish-reddey-red	(scarlet)	1.0	0.2	0.0
Threonine	(T)	Thr	RYR	Reddish-yellowey-red	(vermillion)	1.0	0.4	0.0
Glycine	(G)	Gly	YRY	Yellowish-reddey-yellow	(orange)	1.0	0.6	0.0
Proline	(P)	Pro	RYY	Reddish-yellowey-yellow	(tangerine)	1.0	0.8	0.0
Cystine	(C)	Cys	YYY	Yellowish-yellowey-yellow	(yellow)	1.0	1.0	0.0
Alanine	(A)	Ala	GYY	Greenish-yellowey-yellow	(lemon)	0.8	1.0	0.0
Valine	(V)	Val	YGY	Yellowish-greeney-yellow	(lemon-lime)	0.6	1.0	0.0
Isoleucine	(I)	Ile	GYG	Greenish-yellowey-green	(lime)	0.4	1.0	0.0
Leucine	(L)	Leu	YGG	Yellowish-greeney-green	(grass)	0.2	1.0	0.0
Methionine	(M)	Met	GGG	Greenish-greeney-green	(green)	0.0	1.0	0.0
Phenylalanine	(F)	Phe	BGG	Blueish-greeney-green	(emerald)	0.0	1.0	0.4
Tyrosine	(Y)	Tyr	GBG	Greenish-blueey-green	(turquoise)	0.0	1.0	0.8
Tryptophan	(W)	Trp	BGB	Blueish-greeney-blue	(cyan)	0.0	0.8	1.0
Histidine	(H)	His	GBB	Greenish-blueey-blue	(peacock)	0.0	0.4	1.0
Arginine	(R)	Arg	BBB	Blueish-bluey-blue	(blue)	0.0	0.0	1.0
Lysine	(K)	Lys	RBB	Reddish-bluey-blue	(indigo)	0.4	0.0	1.0
Asparagine	(N)	Asn	BRB	Blueish-reddey-blue	(purple)	0.8	0.0	1.0
Glutamine	(Q)	Gln	RBR	Reddish-bluey-red	(magenta)	1.0	0.0	0.8
Glutamate	(E)	Glu	BRR	Blueish-reddey-red	(violet)	1.0	0.0	0.4
Aspartate	(D)	Asp	RRR	Reddish-reddey-red	(red)	1.0	0.0	0.0
Serine	(S)	Ser	YRR	Yellowish-reddey-red	(scarlet)	1.0	0.2	0.0
Threonine	(T)	Thr	RYR	Reddish-yellowey-red	(vermillion)	1.0	0.4	0.0

Each amino acid (left) is assigned a systematic colour, Xish-Xey-X, where X represents at most two colours combined in proportions, Xish $<$ Xey $<$ X. A trivial name is also given (in parentheses) along with the RGB (red, green, blue) weights. (N.B., magenta and cyan, which are normally the complements of green and red, have been slightly displaced from their conventional assignments.) The order is circular, as indicated by repeated entries above and below the lines.

assignment has some convenient groupings that reflect well a number of the more important properties:

- hydrophobic amino acids are green (GYY–BGB);
- aromatic amino acids are greeney blue (BGG–GBB);
- amino acids found in loops are red and orange (RRR–RYY);
- large polar acids are purple and blue (GBB–RBR).

Within these groups there is little scope for change that does not separate some closely linked pair or triplet (ST, LIV, FYW, RK, NQ, DE). Greatest ambiguity is found in the orange acids, where any permutation of the four acids would be acceptable. That chosen in Table I keeps the oxygen-containing S and T close to red. An alignment of flavodoxin sequences is shown in Figure 1a with each residue coloured using its individual acidic colour.

Colour combination

As the winds inter-weave to produce ‘colours of indescribable delicacy’, so too do the acids as they mingle in a multiple sequence alignment. In contrast to the wind, however, with the aid of Table I, the colours of any combination of amino acids can be described exactly.

The delicate colours

The simplest approach is to average the acidic colours (using the RGB values as specified in Table I). So, for example, a position with an equal mix of Asp and Arg would be a purplish colour. With the addition of a little hydrophobicity at the same position, the average would shift towards the centre of the colour wheel, producing a less saturated lilac shade of purple. Figure 1b shows the example alignment coloured in this way. This simple combination recipe has, however, one unsatisfactory feature: a mixture of amino acids can give the same colour as a completely conserved position. Given the importance of conservation in multiple sequence analysis, this lack of discrimination cannot be ignored.

Accretions of darkness

Colours can be described either as the mixture of red, green and blue (the RGB values) or, alternatively, by hue, saturation and intensity (HSI). The colours of the acids themselves are all at full intensity and full saturation (rainbow colours) while the delicate colours have full intensity but variable saturations (a mix of all acids would be white). This leaves the third component free to encode conservation: full intensity for total conservation and black for all different acids. Other schemes can easily be imagined, such as the mean pairwise similarity or weights derived from relative sequence similarity. A bonus from this method is that gaps in an alignment should obviously darken the colour: implying that gaps are simply black amino acids, as might naturally be expected. Figure 1c shows the example alignment with variable positions darkened by subtracting $(n - 1)/20$ from each colour component (where n is the number of different acids seen in the column being coloured). This reduction was applied to each residue in the column being darkened.

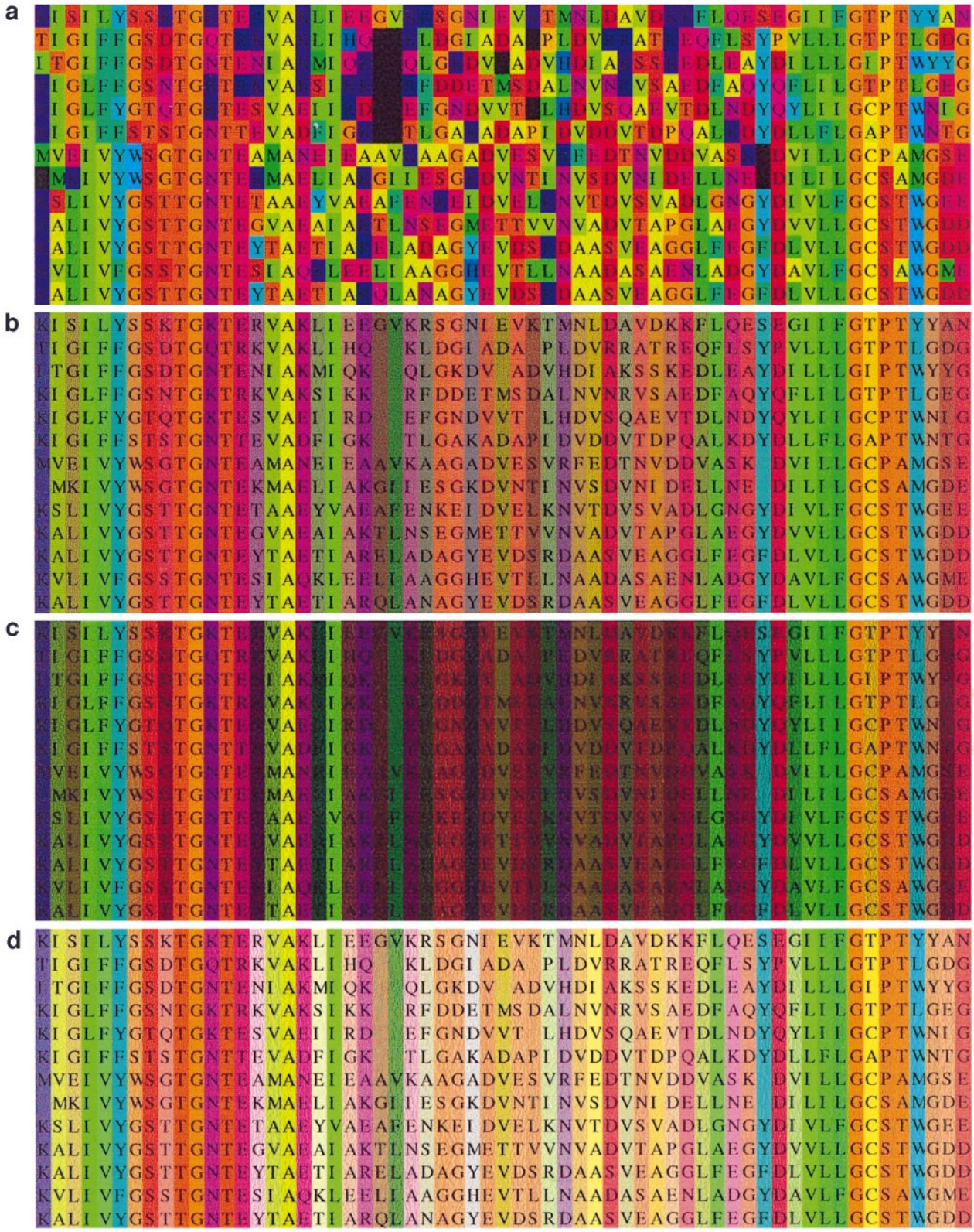

Fig. 1. Coloured alignments: (**a**) individual residues coloured by pure acidic colours; (**b**) each column coloured by the average acidic colours; (**c**) as (b) but with variable positions darkened; (**d**) as (b) but with variable positions lightened. See text for details.

With this added dimension, all colours attain meaning: a ‘greyish-green equally related to black and brown’ might derive from a slightly aromatic hydrophobic position shifted towards white, possibly by Pro or Gly, and darkened into grey-green by the dark-orange (brown) Gly and Pro component aided by black gaps.

The only colour that can never be reached is white, as any combination that produces it has variation and is therefore darkened to grey.

Half-way between silver and purple

Further dimensions in colour become available in the more subtle properties of translucency, reflectivity and specularity (the last term is often described by those in the motor trade as 'metallic finish'.) Such properties are already used to colour protein structures in a variety of aesthetically motivated schemes, none of which yet bear any relationship to the nature of the underlying amino acids but clearly have great potential.

The limitations of paper

For printing on paper, the lack of pure white is an advantage as positions cannot disappear completely, however, against the white background, the most obvious positions will be the less important dark unconserved and gapped regions. For this printing medium it may be best to lose some information and 'darken' the colours using white. A mixed red and blue position will still attain a lilac shade (indicating that the colour does not derive from a single acid) and all that will be lost is some discrimination in the more delicate shades. Figure 1d shows the example alignment faded by adding white to the variable positions. This was achieved by adding $(n - 1)/20$ to each position in a column (rather than subtracting when darkening, as described above).

With the demise of paper-based publishing, such problems will not be a concern for much longer.

Applications and developments

Alignment evaluation function

The most perfect sequence alignments when coloured by the current scheme will be those with the brightest, purest colours, having few browns or greys or other 'muddy' colours. If an alignment does not attain this ideal, the sequences can always be realigned to improve matters. Applying this logic recursively, it is clear that complete mutliple alignments can be evolved to minimize muddy colours. This approach to alignment may have the interesting property that, since a gap is treated simply as a black amino acid, and is a prime source of greys and browns, there may be no need to have a separate gap penalty. The simultaneous elimination of Dayhoff-like matrices and the gap penalty (from which all diseases of the alignment emanate) can only be a good thing, and will be pursued elsewhere.

Protein structure colouring

The natural resting place for a coloured alignment is on a protein structure. This might be the structure of a protein whose sequence lies in the alignment or, using the new powerful method of threading, might be a proposed relationship. In this context, it is inconvenient to display the alignment directly on the structure, so placing the burden of capturing the nature of each position more directly on to the colouring scheme. With the power of the current scheme, little reference to the source alignment will be necessary, and without lifting one's eyes from the protein structure, it will be easy to follow the flux of colours along the chain.

With the power of perceiving these colours one could spend a day sitting quietly on a computer watching the beauty of the acids, their fall and rise and changing hues, the magic of neighbouring positions when they inter-weave like ribbons at a wedding.—Surely, a better occupation than gazing at scientific papers.

References

Gibson,T.J., Hyvönen,M., Musacchio,A., Saraste,M. and Birney,E. (1994) *Trends Biochem. Sci*., **19**, 349–353.

Taylor,W.R. (1986) *J. Theor. Biol*., **119**, 205–218.

Received March 17, 1997; accepted April 9, 1997