Protein Engineering vol.11 no.3 pp.153–169, 1998

Which effective property of amino acids is best preserved by the genetic code?

Georges Trinquier[1] and Yves-Henri Sanejouand

Laboratoire de Physique Quantique, IRSAMC-CNRS,
Université Paul-Sabatier, 31062 Toulouse Cedex, France

[1]To whom correspondence should be addressed

Simple procedures are proposed to quantify how much an effective property embodied in a given ranking of the twenty amino acids can be affected by random point mutations at nucleotide bases. As expected, of the various orderings tested, rankings based on most hydrophobicity scales exhibit low scores, thus offering better immunity towards such single-base mutations. This, however, occurs to different extents and the method allows sharp discriminations between the scales. Hydrophobicity scales based on global properties such as spatial environment data of proteins residues, or mutation matrices of amino acid replacements, generally behave better than those based on pure physicochemical properties of isolated residues. An averaged scale built from the available hydrophobicity scales exhibits one of the most favorable scores. A systematic search for the best amino acid order has been carried out across all possible scales. Optimized scales are characterized by the existence of a clustering scheme into three zones, within which permutations are more or less tolerated, depending on the zone and on the summation procedure used in the score calculation. The first cluster corresponds to the hydrophobic side, and includes the ten amino acids WMCFILVGRS. Next follows the ATP triad. The third cluster coincides with the hydrophilic side and includes, in the last seven positions, the amino acids EDKNQHY. Interpretation of these optimized scales in terms of codon positions in the genetic code further suggests a clustering scheme composed of four groups, WMCFILV-GRS-ATP-EDKNQHY, emphasizing the role of the second base as the main driving parameter. As a consequence, the conserved character of the genetic code is better reflected when it is displayed in UGCA ordering rather than in the commonly used UCAG ordering. The present a priori classification of the amino acids could find potential use in protein sequence homology and structure prediction.

Keywords: amino acid effective property/genetic code/hydrophobicity scales

Introduction

With its degenerate character, the genetic code is redundant in that many punctual mutations at the third base of the codons often do not change the encoded residue. It also offers a certain resistance of the hydrophobicity property towards random mutations. For the typically hydrophobic amino acids of the first column (phenylalanine, leucine, isoleucine, valine) a change in the first codon base will maintain the hydrophobic character of the mutated residue. This also holds true to some extent for the polar amino acids associated with the codons of the third column and constitutes the well-known second-base effect on hydrophobicity. The genetic code is, therefore, not neutral towards this property, and one can wonder whether there exists another amino acid characteristic—physicochemical or more general—that could be better conserved throughout random nucleotide mutations. The aim of the present work is to try to outline, through simple models, such an effective property.

After its decipherment and the settling of its quasi-universal character, the genetic code has been the object of much attention. The following characteristics have been particularly scrutinized:

(i) Its origin and evolution (Epstein, 1966; Crick, 1968; Orgel, 1972; Jukes, 1973; Brack and Orgel, 1975; Wong, 1975; Eigen and Schuster, 1978; Marlborough, 1980; Sukhodolets, 1989; Konecny *et al.*, 1993, 1995; Béland and Allen, 1994; Delarue, 1995; Jimenez-Sanchez, 1995; Hartman, 1995; Wetzel, 1995; Demongeot and Besson, 1996; Di Giulio, 1996);
(ii) Its symmetric and antisymmetric properties (Hasegawa and Miyata, 1980; Chipens *et al.*, 1989; Chipens, 1991);
(iii) Its degree of optimization (Seybold, 1976);
(iv) The links between the codon disposition and various amino acid properties, including the hydrophobic or polar character (Woese *et al.*, 1966; Volkenstein, 1966; Wolfenden *et al.*, 1979; Siemion, 1995), the frequency of occurrence in proteins (King and Jukes, 1969, Siemion, 1994b), the biosynthetic pathways (Taylor and Coates, 1989), and other geometrical (Siemion, 1994a, 1995), physicochemical (Grantham, 1974; Sjöström and Wold, 1985), or chemical (Siemion and Stefanowics, 1992) properties.

Whereas the issue of random or better-optimized codes has been addressed (Wong, 1980; Cullmann and Labouygues, 1983; Haig and Hurst, 1991), attempts to determine a synthetic amino acid property—through a unique optimized ranking—best suited to the code as it presently exists, have yielded few, if any, answers. More recently, a neural network model of the genetic code provided a mapping of the codons and a grouping of the amino acids which shows an interesting resemblance to a hydrophobicity scale (Tolstrup *et al.*, 1994). The present treatment is related to that approach in that it will provide a clustering and ranking of the amino acids without introducing any '*a priori* relationship between the nucleotides or amino acids'.

The procedure proposed in this work, considering amino acid properties which are based on a ranking of the twenty amino acids, allows quantification of the resistance, towards nucleotide mutations. This treatment will be applied, first, to known hydrophobicity scales, and next will be used to determine—*ab initio* from the genetic code—which scale is associated with the most stable effective property.

The paper is organized as follows. First, we will recall some definitions and briefly review various hydrophobicity scales. Secondly, a method section will describe our score evaluation

procedures. Thirdly, we will present the results obtained with the three possible score options. Finally, these results will be discussed and interpreted in terms of the structure and function of the genetic code.

Hydrophobicity

The residue property which has appeared to be an essential driving force in protein folding, and eventually in its structure and function, is hydrophobicity (for a review on this topic, see Rose and Wolfenden, 1993). Hydrophobicity can be defined in several ways, but its complex and general character is recognized. For instance, more than fifteen years ago, Charton and Charton (1982) wrote: 'no single hydrophobicity parameter can represent the complete range of amino acid behavior. There is no special phenomenon denoted by hydrophobicity in amino acids'. It is by reference to this operational character that we shall use the term *effective property*, as used by physicists whenever known effects are better understood than the underlying physics. According to the context in which hydrophobicity (or hydrophilicity or hydropathy) has been introduced, it has thus received various definitions. As a tendency to exhibit less affinity for water, it can be estimated from physicochemical measurements on isolated amino acids. As a tendency to occupy the core of proteins, it can be estimated from statistical properties of residues in the three-dimensional structures of proteins. In this case, it carries more complex information than a single physical or chemical property. Accordingly, hydrophobicity scales basically divide into two classes:

(i) The first one is based on experimental measurements of chemical behavior or physicochemical properties of isolated amino acids such as solubility in water, partition between water and an organic solvent, chromatographic migration, or effects on surface tension.
(ii) The second one is more 'operational', and is based on the *known* environmental characteristics of residues in proteins, such as their solvent accessibility or their inclination to occupy the core of proteins. The corresponding measurements are based on the position of amino acids in the tertiary structures of proteins, as observed from crystallographic or NMR data. This has been made possible because of the increasing number of three-dimensional structures now available in databases. A variant of this class is based on mutation data, which handles even more global information and is particularly relevant to the present approach (Sweet and Eisenberg, 1983; Cornette *et al.*, 1987). Let us illustrate with a few examples how such scales may differ.

Some rankings of the amino acids according to selected hydrophobicity scales are reported in Table I, top (see the appendix for details on the origins and descriptions of the scales). There is a wide diversity of position for many amino acids. To make it more conspicuous, one can count their occurrence at each position along the scale, deduce a mean position, and construct a mean scale which averages the 43 selected scales. This is done in Table II, which also displays for each amino acid the standard deviations and the number of observed ranks. Some interesting points are suggested by this treatment. There is a good consensus to assign hydrophobic character to I, F, L, V and M; to assign hydrophilic character to N, Q, E, D and K; and to locate A or T at mid-scale. There is more divergence when ranking the remaining amino acids, in particular C, P and R, as can be gauged from their standard deviations. Cysteine is the only amino acid that can be found either as the most hydrophobic one (in six scales), or as the most hydrophilic one (in Aboderin's scale). Proline is encountered in as many as 16 different positions. Arginine can be encountered in the most hydrophilic position (20), or at mid-scale (positions 12 and 13). Similarly, tryptophan can occupy the most hydrophobic rank (position 1), or it can be located at mid-scale (position 9). Although our set of hydrophobicity scales is rather limited for reliable statistical analysis, and being aware that some of these have a composite character (see Appendix), nevertheless the scale resulting from the mean ordering of Table II will prove to possess interesting properties.

Method

The fundamental operation consists of allowing *all* the possible mutations at each base position for each codon of the twenty amino acids arranged in a given order, and then seeing how far the mutated residue is removed from the current position in the scale. The extent of change in the property associated with a given scale brought by the punctual mutations is measured by the summation of these simple distances. The corresponding score is thus defined as:

$$S = \sum_{AA=1}^{20} \sum_{cod=1}^{ncod} \sum_{bp=1}^{3} \sum_{mn=1}^{3} |x_{AA} - x_M|$$

in which AA are the amino acids, cod their associated codons, bp the base positions on the codon, mn the mutated nucleotides, and x_{AA} and x_M the ranking positions on the scale of the current and mutated amino acid, respectively.

Similar distance evaluations have been used before (Alff-Steinberger, 1969; Wong, 1980; Cullmann and Labouygues, 1983; Perlwitz *et al.*, 1988; Di Giulio, 1989b; Haig and Hurst, 1991), but in different or more restrictive contexts, and never in a complete-scanning strategy as in the present case. The procedure can be illustrated in an alphabetic ordering of the amino acids (Figure 1). On this scale, the first residue is alanine, A. One possible codon associated with this residue A is GCU. The three possible mutations on the first base will yield UCU, CCU and ACU, corresponding to mutated residues serine, S, proline, P, and threonine, T, respectively. On our scale, where position 1 is assigned to A and position 20 to Y (tyrosine), S occupies the position 16, P the position 13 and T the position 17. The A–S distance can thus be appraised as the difference 16–1=15, the A–P distance as 13–1=12, and the A–T distance as 17–1=16. The three mutated residues will thus introduce penalties of 15, 12 and 16, respectively. A synthetic index reflecting the resistance of a certain property, embodied in the ordering, to random nucleotide punctual mutations can be generated by summing up such increments over all mutations at all three positions of each codon, and for each codon associated with all the residues in the given scale.

The reader can redo the processing for other codons, as partly exemplified in Figure 1, keeping in mind that the distances are always expressed in absolute values. For sake of simplicity, if the mutated codon is associated with the same residue, the increment counts zero, and if the mutated codon is a stop codon, the mutation is not taken into account.

Intentionally, the model is kept as simple as possible in order to allow for a better interpretation of the results. Yet, several strategies for summing up the increments are possible. One can choose, as suggested above, to keep the summation

Table I. Some scales and their calculated scores

Denomination	N	D	Residue sequences	Scores		
				total	averaged	weighted
Hydrophobicity scales						
Zimmerman	p	2	LYFPIVKCMHRAEDTWSGNQ	2644	914.8	43.7
Jones	p	2	WIFPYLVMKCHAREDGNSTQ	2530	889.8	41.1
Wolfenden	p	20	GLIVAFCMTSWPYQKNEHDR	2434	791.8	39.4
Bull	p	4	LIYFWVMQNGHRDAEKSCTP	2372	828.8	38.5
von Heijne		1	FILVWAMGTSYQCNPHKEDR	2338	793.4	38.2
Levitt	(p)	32	WFYLIVPMCHATGQNSDERK	2302	814.2	36.9
Frömmel		1	WFYMILPVHCKTRAEQSNGD	2280	830.8	37.6
Rekker	p	20	WFILYVMPCRAKGDEHTSNQ	2280	801.5	37.0
Krigbaum	(p)	4	WFIMYKCVLPATSHDGENQR	2276	834.4	37.9
Efremov		2	IFLCMVWYHAGTNRDQSPEK	2268	776.0	37.6
Nishikawa		1	CVILWMYFAHGRTQSEPNKD	2264	799.5	38.1
Guy		1	FMCVLIWHYTAGNSPDEQKR	2240	780.8	35.7
Meirovitch		48	FIVWMLCHYARNTPGDSEQK	2234	778.4	37.6
Guy 2		1	FCIMLVWHYATGNSPQEDRK	2230	780.6	36.2
Eisenberg		1	IFVLWMAGCYPTSHENQDKR	2230	755.6	36.6
Robson		160	IWFCYLMVHPRQGTNEKASD	2224	766.8	36.6
Cornette 5		1	FILVYMHCWQARETGSPNDK	2214	781.0	37.9
Janin		8	CIVLFMGAWYSHTPDNEQRK	2214	754.7	35.7
Fauchère	p	1	IFVLWMAGCYPTSHQNEDKR	2214	745.6	36.3
Cornette 4		1	FILVMHYCWAQRTGESPNKD	2210	773.9	37.6
Chothia		8	IVFCLMAGWTSPEHDYNQKR	2208	738.4	35.5
Krigbaum 2	(p)	1	CIMFWVYLATSHDGQENRPK	2206	763.6	35.7
Cornette 2		1	LIVFMCYRWHAGSNETPQKD	2204	755.4	37.4
Kyte	(p)	24	IVLFCMAGTSWYPHDNEQKR	2198	727.3	35.4
Chothia 2		1	IVFLMACGWSTEPHYDNQKR	2194	724.8	35.8
Cornette 7		1	LFIMVYCHWRAGQTESPNKD	2190	749.3	36.9
Eisenberg 2		2	WMFILYVPAHTCSGQKNEDR	2186	787.4	35.1
Cornette		1	CIVLWFMYAGHTRSQNPEDK	2164	767.8	36.5
Ponnuswamy		2	VCIMLFYWHAGTRQSPENDK	2164	764.3	36.7
Cornette 3		2	LVIFMCYRWHAGNSPTEQKD	2164	742.6	36.9
Janin 2		48	CIVFLMWGASHTPYNDEQRK	2164	741.8	34.7
Hopp	(p)	384	WFYILVMCAHTPGQNSRKDE	2160	781.8	35.9
Olsen		1	IVMFLCAGTSPWEHDYNQKR	2146	719.2	34.7
Cornette 6		1	LIFVMYCHWRAGSTQENPKD	2140	740.4	36.3
Wertz 2		8	WCILVFHMYASRNGTEDPQK	2134	763.4	35.3
Engelman	(p)	1	FMILVCWATGSPYHQNEKDR	2102	705.8	34.0
Aboderin	p	1	LFIWMVYAPGTSREHQKDNC	2070	769.0	34.1
Wertz		8	FWCILMVHYASRNGTEDPQK	2060	732.3	34.2
Meek	p	8	WIFLYPMVTSRNAKGHQCDE	2034	762.7	33.6
Sanejouand		2	IWFLCVMYAHTGSPRQNDEK	2026	731.8	34.3
Rose		144	CFIVMLWHYAGTSRPNDEQK	1994	725.0	33.4
Miyazawa		2	FMILCWVYAHTGRSPQNEDK	1994	716.3	34.1
Sweet	m	2	FYILMVWCTAPSRHGKQNED	1884	681.0	32.4
Miscellaneous scales						
Alphabetic		1	ACDEFGHIKLMNPQRSTVWY	2680	939.4	44.2
pK_1'	p	2	HFDCPNRQKEYSMVAGLIWT	2610	908.4	42.5
Isoelectric points	p	1	DECNFQYSMWVGLAITPHKR	2498	916.5	42.9
Occurrence in proteins		1	LASGVETKIDRPNQFYMHCW	2480	881.7	39.8
Molar masses	p	2	GASPVTCILDNEQKMHFRYW	2444	843.6	38.3
pK_2'	p	8	DEHCNKRYQFSMWGLVTIAP	2266	783.6	38.0
Mean above hydr. scales		1	IFLVMWCYAHGTPSRNQEDK	2020	718.8	34.6
Singular scores						
Mean 100000 sampling				2744	952.3	46.0
Minimum total score [a]		144	WMCFILVGRSTAPYNKDEHQ	1600	597.9	27.0
Minimum averaged score		1	MIFLVWCGRSTAPYNKDEHQ	1634	581.5	28.1
Minimum weighted score		1	WCMFILVGRSATPEDKNQHY	1602	599.3	26.4
Maximum total score[a]		8	RLANCETYQMWFDIGHKPVS	3784	1199.5	61.1
Maximum averaged score		1	RNAFEMHCTVWPYQIGKDSL	3682	1232.6	61.7
Maximum weighted score		1	LANRECTYMQWFGHIPDKVS	3734	1205.8	62.8

The origins of the scales are given in the appendix, with the corresponding references. The column N (for nature) points the scales built from physicochemical experimental data on amino acid, either entirely, p, or partly (p), or built from mutation matrix data, m. The other scales are built from protein structural data, averaged or not. The column D (for degeneracy, or degrees of freedom) indicates the number of equivalent sequences, due to residue equivalency, among which that giving the best score has been selected. See the appendix for more details.
[a]Only one sequence of the degenerate set is given.

Table II. Occurrence of residues at each sequence rank for the hydrophobicity scales listed in Table I

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Mean	SD	Occupied ranks
I	9	13	14	3	3	1															2.6	1.2	6
F	10	11	9	7	1	4		1													2.9	1.7	7
L	7	1	5	14	9	5		1	1												4.0	1.8	8
V	1	6	7	7	3	11	4	4													4.7	2.0	8
M		4	2	5	10	9	8	4	1												5.5	1.8	8
W	9	3		2	7	1	7	1	9		2	1				1					5.9	3.6	11
C	6	3	2	3	3	4	6	5	4	2		1	1					2		1	6.7	4.5	14
Y		2	4		6	3	7	4	6	3	1	1	2	1	1	2					7.9	3.6	14
A					1	2	5	3	8	9	9	2	1	2				1			9.8	2.4	11
H						1	2	7	4	9	4	2		9	1	3		1			11.1	2.9	11
G	1						1	8		4	4	8	4	5	4	2		1	1		11.7	3.3	12
T									7	3	7	10	4	4	3	2	1		2		12.3	2.5	10
P				2		1	2	2	1	2	4	4	3	3	6	1	6	4	1	1	13.1	4.0	16
S										7	6	2	8	5	2	5	5	2	1		13.6	2.6	10
R								2		3	4	6	7	2	1		1	1	4	12	15.0	4.0	11
N									1			2	6	2	4	8	9	8	3		15.8	2.2	9
Q								1		1	1	2	1	5	6	6	3	9	5	3	16.0	2.7	12
E												1	4	2	7	6	11	7	3	2	16.3	1.9	9
D													2	2	7	4	4	5	10	9	17.5	2.1	8
K						1	1		1		1	1		1	1	3	3	2	13	15	17.6	3.5	12

over all the synonymous codons that the degeneracy of the genetic code permits for a given residue, or one could take the average over the codon degeneracy of the residue concerned. Both strategies make sense. The former one, leading to *total* scores, may seem spurious since it exaggerates the role of highly degenerate residues. Besides its conceptual simplicity, it has nonetheless several advantages, such as producing integer scores, to allow many exact degeneracies; and mirroring, in some way, the importance such residues were conferred by evolution. The latter one, leading to *averaged* scores, treats all residues on an equal footing and may seem more natural and reasonable, although we know synonymous codons are not encountered with equal frequencies (Médigue *et al*., 1993). Both types of summations lead to distinct results, but they exhibit many common features. A third strategy has also been tried, which consists in further weighting the averaged score of each residue by its mean rate of occurrence in proteins. Such a rate can be measured using sequence databases for instance. This will lead to *weighted* scores. Interestingly this reflects many conclusions obtained from the total scores. The detailed results obtained from each of these scoring options are now examined.

Results

Total scores

In the total summation, the procedure involves a multiple loop running over 61 codons, 3 base positions and 3 mutated bases. Since the three stop codons UAA, UAG and UGA can be reached from $3\times3\times3=27$ different mutations, and since 4 of these correspond to the amber/ochre and ochre/opal intraconversions, $27–4=23$ mutations lead to a stop codon. Under these conditions, the computational process will select $(61\times3\times3)–(23)=526$ mutations. Such a summation happens to give only even integers. Typically, the score associated with the alphabetic ordering illustrated in Figure 1 is calculated at 2680. The total scores obtained by our selected hydrophobicity scales are listed, in decreasing order, in Table I, top. For the sake of comparison, a few disparate scales are also reported in the same table.

A sampling of 100 000 random scales led to a bell-shaped distribution of the scores centered around 2744, with a standard deviation of 202 (Figure 2). Scales built from molar masses, isoelectric points, pK'_1, mean occurrence in proteins, or, of course, alphabetic order, all give poor scores, i.e. not that different from the random mean scale (only pK'_2 does a little better). Scales associated with hydrophobicity all exhibit much better scores, in the range 2000–2300, thus distant from the random average by more than two standard deviations (see Figure 2 and Table I). Another way of illustrating the differentiating power of our total scoring is to plot it along the coordinate of the 526 mutations. This is done in Figure 3 for two typical scale samples, a random one, and a hydrophobic one. The divergence of both skeins is conspicuous.

Here also, some disparity between the hydrophobicity scales can be observed. The scales built from Zimmerman's or Jones' indices, based on solubility data, exhibit high scores of 2644 and 2530 respectively, which is hardly better than a random ordering score. On the other hand, nine scales happen to give scores lower than 2138, which makes more than three standard deviations apart from the mean random score. Expressing the scores by their difference to the random value, in standard deviation units, in fact provides more meaningful indices. Such Z-scores are given in Table III, which can therefore be seen as a measure, for the property embodied in the given scale, of the inertia of the response of the genetic code to DNA mutations.

The best resistance to nucleotide mutations is obtained with Sweet's scale, the scale based on the mutation matrix of amino acid replacement amongst related proteins. This scale already conveys some residue mutational information, so it is natural that it offers the highest passivity towards nucleotide mutations. All scales built from similar mutation-matrix data should behave so, as we could check on that labelled SWEIG in Cornette *et al.* (1987) (nearly identical to Sweet's scale, this

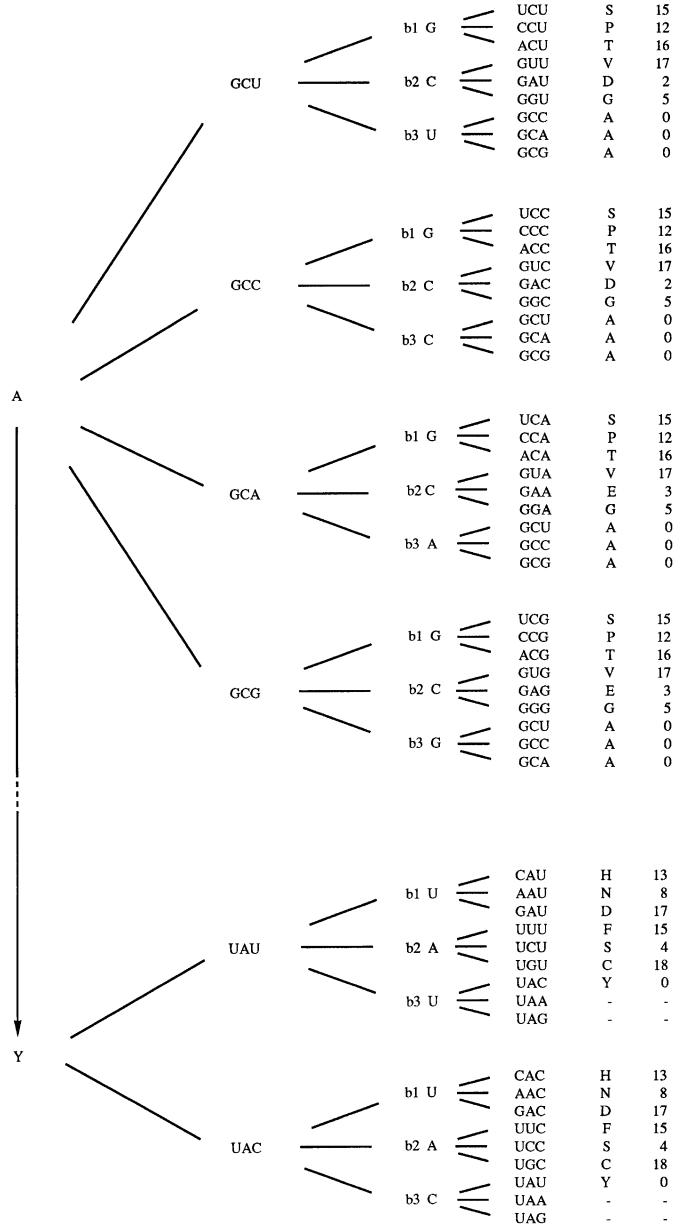

Fig. 1. The procedure used to get total scores, as illustrated by an alphabetic ordering of the amino acids.

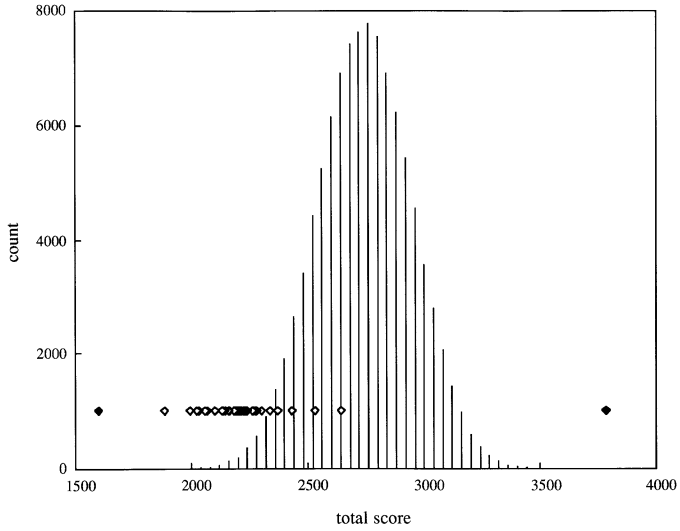

Fig. 2. Total-score distribution for a sampling of 100 000 random scales. The dots mark the various hydrophobicity scales (empty) and the extreme possible scores (filled).

scale is not included in Table I). Most favorable scores are obtained with scales in which residues are ranked according to their inclination to occupy the interior of proteins. The best ones are those of Miyazawa, Rose, Sanejouand, Wertz, Engelman (note however that some scales, like Rose's, have many degrees of freedom, which favors them spuriously, whereas others, like Engelman's, are unique). The tendency to occupy the protein core embodies more than a simple physicochemical character of residues, so it is not surprising that scales obtained from such phenomenological analysis, and therefore carrying more global information, prove to be more effective. Two scales based on chromatography data have remarkably low scores: (i) Aboderin's scale, built from mobilities of the amino acids on chromatography paper, and (ii) Meek's scale, built from retention times of peptides in high-pressure liquid chromatography. The other scales based on residue properties have either a fair score, such as the popular Kyte and Doolittle scale which is based in part on water affinity of the amino acid side chains, or a rather high score.

Remarkably, the residue ordering resulting from the averaging of all our selected hydrophobicity scales gives a low score of 2020—the fourth best score. Such averaging may have therefore cancelled both the bias of definitions, the arbitrariness of selected data and the errors dues to experimental uncertainties.

A systematic search for the lower scores was then undertaken. For this typical combinatorial-optimization problem, to be run over a scale space of $20! \approx 2 \times 10^{18}$, we used Monte Carlo simulated annealing algorithms (see Appendix for details). We got a degenerate set of 144 scales with a minimum score of 1600, i.e. over five standard deviations away from the average. On the left side, the side of higher hydrophobicity in our convention, these scales have in common the ten residues WMCFILVGRS. Then at positions 11, 12 and 13, one finds the three residues TAP in any order. Last, a hydrophilic part involves the seven residues occupying the positions 14 to 20. Out of the $7! = 5040$ possible arrangements, only 24 arrangements (0.5%) are allowed, each one combining with all the $3! = 6$ possible arrangements of the preceding middle cluster, thus leading to the total 144 sequences (24×6). Bearing in mind the unicity of the first cluster, the degeneracy of the second one, and the variability of the third one, these 144 best-score scales can be summarized as:

WMCFILVGRS TAP YNKDEHQ

Let us examine each of these three clusters. The hydrophobic cluster includes amino acids which are generally accepted to have hydrophobic character. It begins with the two non-degenerate residues tryptophan and methionine, WM. In many scales, tryptophan is actually located on the hydrophobic side, but at different positions. Methionine is also considered to be hydrophobic, but its typical position on the scales is 5 or 6. Next follows cysteine C. The positioning of the two sulfur-containing residues M and C next to each other is encountered in seven of our hydrophobicity scales. Then follows the classical hydrophobic set FILV, in an ordering also found in many scales. Next one finds glycine G, which is encountered

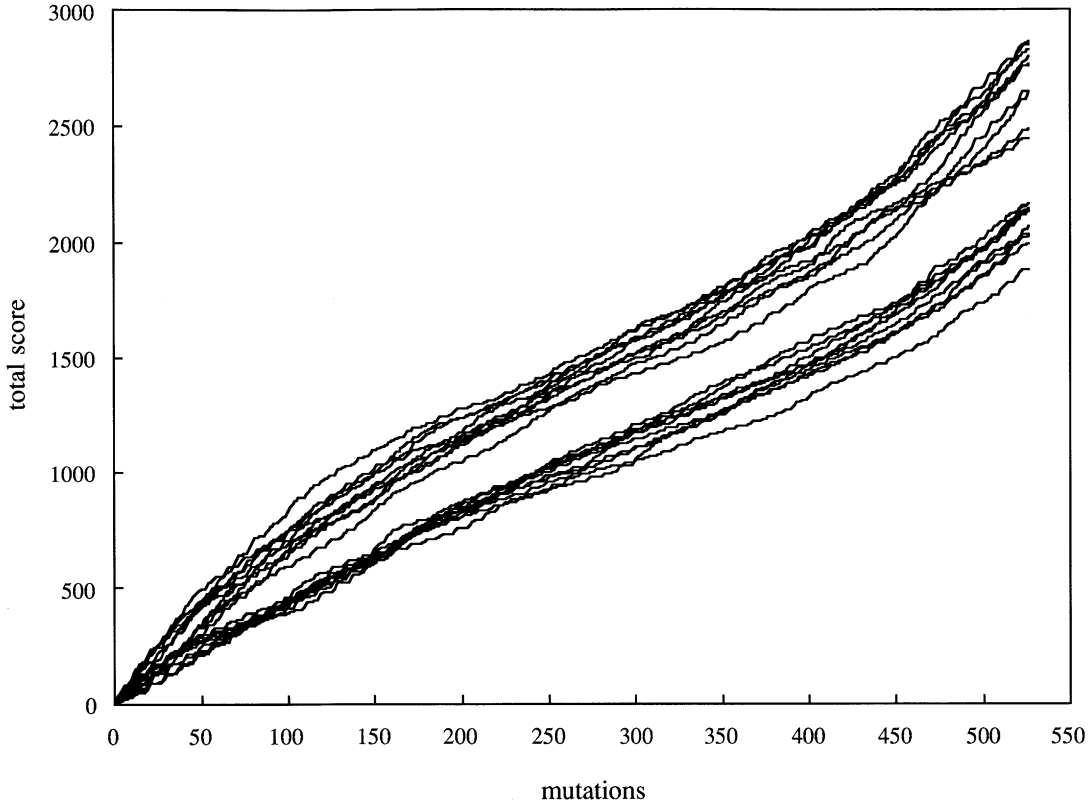

Fig. 3. Total scores along the quadruple loop of mutations for ten typical random scales (top), and ten typical hydrophobicity scales (bottom).

ordinarily at mid-scale, with a preference for positions 8, as here, or position 12. The cluster ends with two six-codons residues, arginine and serine RS. Serine, which bears an OH bond on its side-chain, is generally located either at mid-scale, or on the hydrophilic side. More unexpected is finding arginine in position 9, most scales locate it clearly on the hydrophilic side: it occupies position 20 in as many as twelve scales, and incidentally is never met in position 9.

The middle cluster, corresponding to positions 11–13, includes threonine, alanine and proline, TAP, with strict equivalency of the six possible arrangements. These residues all correspond to codons with cytosine as second base. The remaining amino acid of this second column of the genetic code, serine, has a degeneracy different from T, A or P. In the examined hydrophobicity scales, alanine, A, is located in the middle of the scale, with a mean position of 9.8, and a maximum occurrence at positions 10 and 11. Threonine, T, is also well centered around position 12, with a mean position of 12.3. Proline, P, occupies as many as sixteen different positions, but its mean position is 13.1. Although there is not a TAP cluster at positions 11–13 in the mean scale, the condition is nevertheless nearly realized, since A, T, and P occupy mean positions 10, 12 and 13, respectively, and absolute positions 9, 12, and 13, respectively. This is probably the reason for the low score obtained by the mean scale. The best scale (Sweet) has also a TAP bundle at positions 9–11. This condition is however not sufficient for a good score, as can be checked in Table II.

The last cluster constitutes the hydrophilic zone and includes residues given as hydrophilic by most scales, putting aside tyrosine, and, to a lesser extent, histidine. It comprises the charged amino acids: glutamic acid, aspartic acid and their amide forms, glutamine and asparagine, EDQN, the positively charged residues histidine and lysine, HK, and tyrosine, Y. These seven residues constitute the third column of the genetic code, again emphasizing the peculiar role of the second codon position. Note that the remaining positively charged residue, arginine, is not included in this zone, but it has a higher codon degeneracy. As mentioned above, arginine, R, is considered to be hydrophilic according to most scales. It occupies the hydrophilic end, position 20, in twelve cases, and has also a maximum occurrence at mid-scale, at positions 12 or 13. Tyrosine, on the other hand, is generally given as rather hydrophobic (see Table II). Histidine is more equivocal, located both at 8–10 and at 14.

The 5040 possible arrangements in the hydrophylic zone have been explored. For each of them, the TAP degeneracy still remains, and the scores now range from 1600 to 1680. Only four-manifold scores are allowed, and overlapping effects generate non-obvious degeneracies. These scores may, of course, be outclassed by those arising from reorganizations in the remaining zones of the scale. All the possible permutations in the first zone have also been explored. The corresponding configurational space involves $10!=3\,628\,800$ arrangements. Again, the TAP degeneracy still operates since any arrangement of these three residues leads to a same score. These scores range from 1600 to 2316, according to a bell-shaped distribution, with a mean value of 1991, and a standard deviation of 114. The constraint imposed by such a clustering can be visualized by plotting on the same picture this distribution and the global sampling. Such a diagram of occurrence, or score degeneracy, not shown here, exhibits no strict symmetry, with a remarkable alternation between neighbor scores.

Averaged scores

When the scores for each residue are averaged over its synonymous codons, a large part of the degeneracy is removed.

Table III. Results of Table I expressed in Z-scores

	Total	Averaged	Weighted
Zimmerman	0.5	0.6	0.7
Jones	1.1	1.0	1.5
Wolfenden	1.5	2.7	2.0
Bull	1.8	2.1	2.3
von Heijne	2.0	2.7	2.4
Levitt	2.2	2.3	2.8
Frömmel	2.3	2.0	2.6
Rekker	2.3	2.5	2.8
Krigbaum	2.3	2.0	2.5
Efremov	2.4	2.9	2.6
Nishikawa	2.4	2.6	2.4
Guy	2.5	2.9	3.2
Meirovitch	2.5	2.9	2.6
Guy 2	2.6	2.9	3.0
Eisenberg	2.6	3.3	2.9
Robson	2.6	3.1	2.9
Cornette 5	2.6	2.9	2.5
Janin	2.6	3.3	3.2
Fauchère	2.6	3.5	3.0
Cornette 4	2.7	3.0	2.6
Chothia	2.7	3.6	3.2
Krigbaum 2	2.7	3.2	3.2
Cornette 2	2.7	3.3	2.6
Kyte	2.7	3.8	3.3
Chothia 2	2.7	3.8	3.1
Cornette 7	2.8	3.4	2.8
Eisenberg 2	2.8	2.8	3.4
Cornette	2.9	3.1	2.9
Ponnuswamy	2.9	3.1	2.9
Cornette 3	2.9	3.5	2.8
Janin 2	2.9	3.5	3.5
Hopp	2.9	2.8	3.1
Olsen	3.0	3.9	3.5
Cornette 6	3.0	3.5	3.0
Wertz 2	3.0	3.2	3.3
Engelman	3.2	4.1	3.7
Aboderin	3.3	3.1	3.7
Wertz	3.4	3.7	3.6
Meek	3.5	3.2	3.8
Sanejouand	3.6	3.7	3.6
Rose	3.7	3.8	3.9
Miyazawa	3.7	3.9	3.7
Sweet	4.3	4.5	4.2
Mean scale	3.6	3.9	3.5
Best allowed	5.7	6.2	6.0
Worst allowed	–5.2	–4.7	–5.2

Since there are, on average, 3.05 codons per residue, the averaged scores are therefore expected to be around one third of the total scores. A sampling over 100 000 random scales led to a bell-shaped distribution centered around 952.3, with a standard deviation of 60.0 (Figure 4). Although most of the trends discussed above are still valid, the relative scores for the scales of Table I are somewhat different. Sweet's scale still remains the most favored scale, whereas Engelman's scale now has the next lowest score. Four other scales are also particularly favored by the averaged mode of scoring, these are: Chothia, Chothia 2 (14-step jump), Kyte (13-step jump) and Olsen (7-step jump). Again, the mean scale built from the hydrophobic set would have the fourth lowest score.

The search for the most favorable arrangements led to scales exhibiting both common and different features from those of best total score. The division into three clusters, as above, is preserved. The hydrophobic cluster is modified in its seven first positions and now ranks as MIFLVWCGRS. The TAP degeneracy is lowered as follows. The TAP arrangement is always preferred, then the TPA or ATP arrangements contribute to a 2/3 increase in the score, next the APT or PTA arrangements contribute a further 4/3 increase, and last the PAT arrangement still bring a 2/3 increase in the score. The removal of the 6-fold TAP degeneracy therefore results in a 1/2/2/1 symmetrical splitting of total width 2.67. In the hydrophilic cluster, of the 24 previously favorable configurations, one is now preferred, YNKDEHQ, the other ones being higher by up to 10. Changing YNK into KNY, for instance, only penalizes by 1/3, while changing NKDE by DENK penalizes by 4/3. These effects can be summarized by plotting the scores of the former degenerate 24 arrangements followed by the fixed TAP arrangement (Figure 5, bottom). If the six arrangements within the TAP cluster are further taken into account, the picture becomes more intricate, with larger dispersion and higher average values (Figure 5, top).

Permutations within the hydrophobic cluster increase the score to different extents. The change FL→LF increases the score by 1/2, the change GR→RG does it by 3/4, while the permutation VLF→FLV results in an increase of 2/3. These increments appear to be additive. The ten lowest-score scales identified are listed in Table IV.

The preferred averaged orderings for both TAP and hydrophilic clusters are common with one of the 144 preferred total-score configurations, and the end of their hydrophobic clusters is also common. Taking these two facts into consideration, the absolute minimum in the averaged summation model appears to differ from that in the total summation model in only the first seven residues:

lowest total: WMCFILV GRSTAPYNKDEHQ
lowest averaged: MIFLVWC GRSTAPYNKDEHQ

To assess the incidence of each amino acid on the quality of a scale, we have examined the contributions of each residue to the averaged scores. As an illustration, the individual contributions of each amino acid to the averaged scores are plotted in Figure 6 for three pairs of scales: (a) the two extreme hydrophobicity scales of our set (Sweet and Zimmerman); (b) the best hydrophobicity scale (Sweet) and the best possible minimum; (c) the scale of Engelman and that of Sanejouand, which gave rise to an inversion with respect to their total scores. Proline in position 4 and glutamine in position 19 appear to penalize the Zimmerman scale particularly. The best hydrophobicity scale roughly follows the best possible minimum scale, although it could do better for aspartic acid, asparagine and tyrosine. Arginine and phenylalanine would favor the scale of Sanejouand over that of Engelman. However, the latter overcompensates by the improved behavior of histidine, isoleucine, asparagine and tryptophan.

Engelman's scale (also known as GES scale; Engelman *et al.*, 1986) was already noticed to 'best-fit' the grouping pattern obtained from a neural network model of the genetic code (Tolstrup *et al.*, 1994). In our averaged-score classification, putting aside Sweet's mutational scale, the GES scale is better by far than the other scales. The difference between this scale and the minimum scale is plotted, again for each residue, in Figure 6d (solid line), together with Sweet's scale for comparison (dashed line). As noticed by Tolstrup *et al.*, (1994), the main 'deficiency' of the GES scale is the location of arginine in position 20, at the hydrophilic end. Similarly, the Sweet scale mainly suffers from the position of tyrosine in the hydrophobic end, at position 2 (see Table I and Figure 6d).

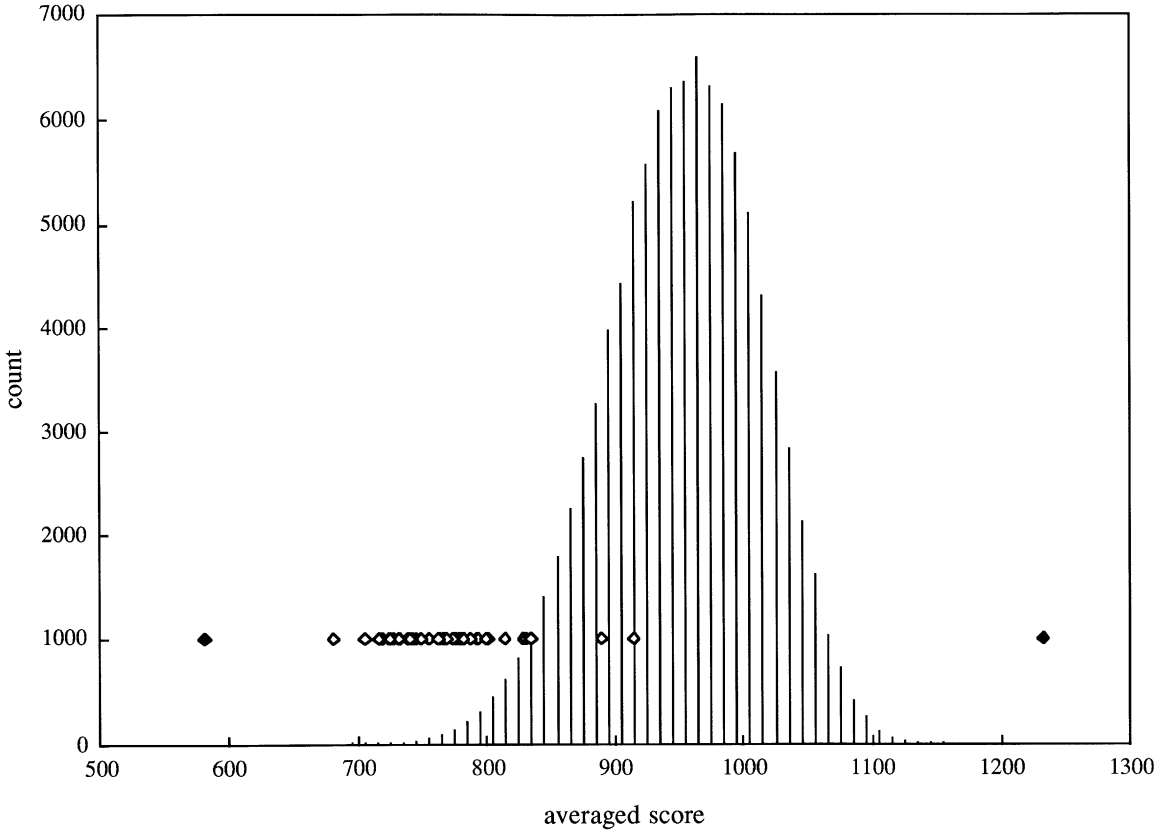

Fig. 4. Averaged-score distribution for a sampling of 100000 random scales. The dots mark the various hydrophobicity scales (empty) and the extreme possible scores (filled).

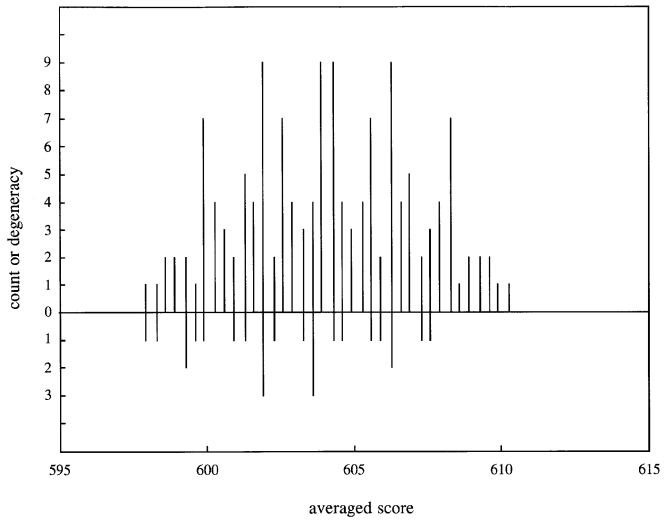

Fig. 5. Averaged-score histogram of the preferred total-score orderings. Bottom: the middle-zone is kept in its more advantageous TAP configuration (24 sequences); top: all six middle-zone arrangements are taken into account (144 sequences).

Incidentally, minor misfits can also induce notable responses in the scoring, as illustrated by tryptophan in the Engelman scale, and asparagine in the Sweet scale. These residues occupy positions that are not far from their position in the optimized scale, however they bring the second contributions to the respective scores. Disregarding arginine in the GES scale, the optimal clusters are particularly well reproduced in this scale. Following the hydrophobic cluster FMILVCW, a set ATGSP occupies the middle of the scale, preceding the hydrophilic set YHQNEKD. The intermediate class GSRTAP can be seen, to some extent, as an autonomous cluster which separates the hydrophobic and hydrophilic groups. Were arginine in position 8–13, the clustering scheme adopted in the Engelman scale would be perfect. This suggests testing how much the score is improved when arginine is allowed to occupy other positions in this scale.

The averaged scores obtained when arginine is located at each of the twenty positions in the Engelman scale are plotted in Figure 7. Location at the hydrophilic end (position 20) is indeed the worse location. For most other positions, it would score much better than any other scales, including Sweet'. This result holds true for total scores as well as weighted scores. With the best location of arginine in position 12, one gets Z-scores of 4.9, 5.3 and 4.7 for total, averaged, and weighted scores, respectively, which approaches the best allowed Z-scores (see Table III, bottom). The U shape of the curve further indicates that residue, R, has some freedom of displacement within an intermediate zone, running from position 8 to position 13, without significant penalty. The consistency of the Engelman scale with the internal structure of the genetic code has been suggested to originate from the fact that it avoids the common shortcoming of not specifically reflecting 'the circumstances in which amino acids appear in proteins' (Tolstrup *et al.* 1994). Actually, in the GES scale, the free energies of water–oil transfer are computed for amino acid residues engaged in an alpha helix, and not for isolated individual amino acids (Engelman *et al.* 1986).

In the Sweet scale, besides a favorable grouping of TAPSR in positions 9–13, there are two major misplacements: tyrosine

Table IV. The ten lower-score arrangements found with the averaged and weighted calculations

Averaged score		Weighted score	
MIFLVWCGRS ATP KNYDEHQ	582.50	WCMFILVRGS PAT EDKNQHY	26.442
MIVLFWCGRS TAP KNYDEHQ	582.50	WCMFILVRGS TAP EDKNQHY	26.440
MIFLVWCGRS TPA KNYDEHQ	582.50	WCMFLIVGRS APT EDKNQHY	26.440
MIFLVWCRGS TAP YNKDEHQ	582.25	WCMFLIVGRS ATP EDKNQHY	26.430
MIFLVWCGRS ATP YNKDEHQ	582.17	WCMFILVRGS APT EDKNQHY	26.430
MIVLFWCGRS TAP YNKDEHQ	582.17	WCMFILVRGS ATP EDKNQHY	26.420
MIFLVWCGRS TPA YNKDEHQ	582.17	WCMFILVGRS PAT EDKNQHY	26.413
MILFVWCGRS TAP YNKDEHQ	582.00	WCMFILVGRS TAP EDKNQHY	26.411
MIFLVWCGRS TAP KNYDEHQ	581.83	WCMFILVGRS APT EDKNQHY	26.401
MIFLVWCGRS TAP YNKDEHQ	581.50	WCMFILVGRS ATP EDKNQHY	26.391

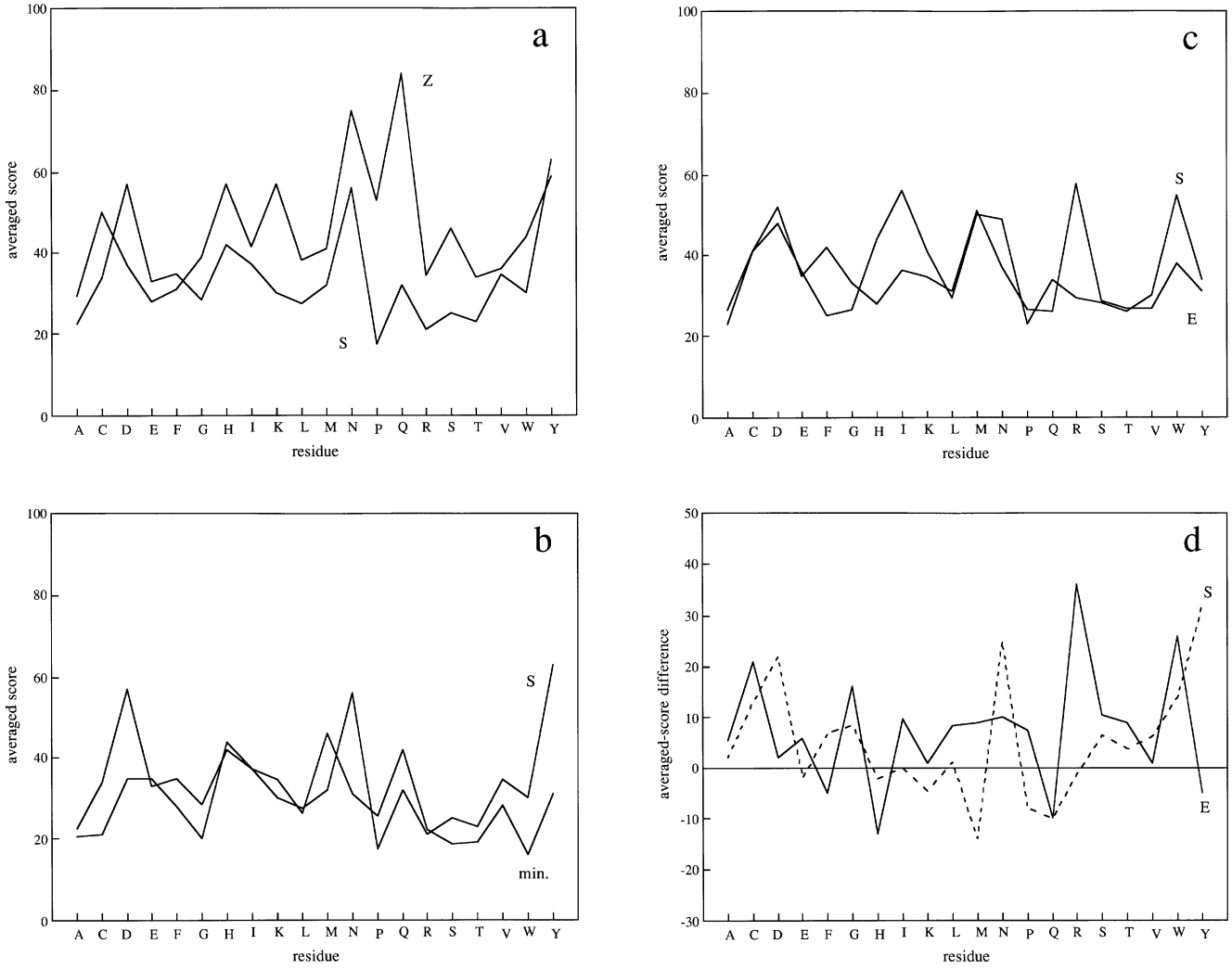

Fig. 6. Contributions of each amino acid to the averaged scores for some scales. (**a**) lowest hydrophobic score (S) versus higher hydrophobic score (Z). (**b**) Lowest hydrophobic (S) versus lowest possible score (min). (**c**) Scale of Sanejouand (S) versus scale of Engelman (E). (**d**) Difference between Engelman (solid line) or Sweet (dashed line) scales and the lowest possible scale.

lies in the hydrophobic cluster (position 2), and glycine in the hydrophilic one (position 15). It turns out that these deviations are less severe for the scoring than was the arginine position in the GES scale.

Weighted scores

If the principle of least change in a decisive property affects all living organisms, the entire set of proteins should be protected against nucleotide blind mutations. Since each residue has a distinct frequency of occurrence in the various proteins of the various organisms, a global optimization, in a Darwinian frame, would suggest the residue contributions be weighted by their occurrence frequencies. In other words, the penalty in our scoring system should be proportional to the importance of the residue, in turn reflected by its global abundance in proteins. Such a weighted score can be obtained by multiplying the contribution of each residue, averaged over its possible codons, by its mean occurrence in proteins. Technically, this

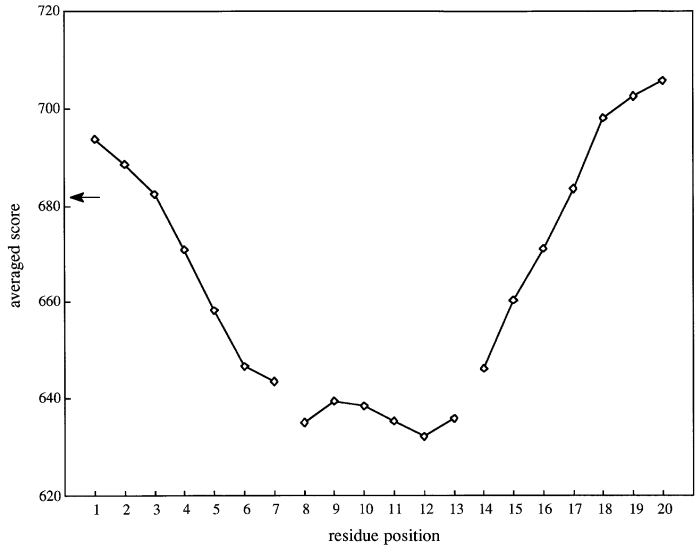

Fig. 7. Averaged score for the twenty positions of arginine in the Engelman scale. The arrow marks the score for the best hydrophobicity scale (Sweet). The discontinuities underline the existence of a neutral intermediate zone.

Table V. Amino acid composition of the sequence database OWL, release 26.0

Residue	%	Codon degeneracy
L	9.1	6
A	7.5	4
S	7.4	6
G	7.1	4
V	6.5	4
E	6.3	2
T	6.0	4
K	5.8	2
I	5.5	3
D	5.2	2
R	5.2	6
P	5.1	4
N	4.6	2
Q	4.1	2
F	3.9	2
Y	3.3	2
M	2.8	1
H	2.2	2
C	1.8	2
W	1.3	1

is done in the third loop of our process, where each residue contribution is multiplied by its abundance fraction. The weighted scores are, therefore, expected to be around 1/20 of the averaged scores. The amino acid composition of proteins is of course variable (Reeck, 1976), but a protein sequence database such as, for instance, the OWL database, may provide a reasonable clue for mean residue abundance. From the amino acid composition of this database, we extracted the percentages given in Table V, and used them as our residue weightings. Due to the experimental origin of these relative abundances, the weighting now introduces in our procedure some empirical data. We could check that the ensuing uncertainties, not present in the preceding scoring modes, do not affect the results too much, which happen to be quite stable towards variations of up to one unit in these percentages.

The main effect of this weighting on the hydrophobicity scales is to bring it back close to the total-score rating of the scales, as can be seen in Tables I and III. A sampling over 100 000 random scales led to a bell-shaped distribution centered at 46.0, with a standard deviation of 3.2. For the hydrophobicity scales, the scores are here again at several standard deviations from this mean value (Figure 8). Typically, the mean hydrophobicity scale is at 3.5 standards deviations.

The search for the optimal scale came close to the results obtained with the total score. The first hydrophobic cluster is nearly identical with the total-score one, now beginning with WCM, instead of WMC. Since residue occurrences are totally non-degenerate, this engenders a complete breaking of degeneracy in all three zones. What still justifies the clustering into the three zones is the remarkable constancy and additivity of the score penalty associated with a given permutation within a zone, regardless of which configuration is adopted in the other zones. The degeneracy breaking in the TAP cluster thus induces a splitting of 0.08 for all arrangements in the hydrophobic zone and in the hydrophilic zone. No matter how residues are arranged within these zones, the six TAP relative scores will be invariably ordered as:

TPA 0.079
PTA 0.072
PAT 0.022
TAP 0.020
APT 0.010
ATP 0.

Similarly, elemental permutations within the hydrophobic and hydrophilic zones yield to a set of unvarying and transposable relative increments, which happen to be further additive within the zones, as exemplified below for the ten lowest permutations within the hydrophobe and hydrophile zones.

WCFMLIVRGS	0.143	DEKNQYH	0.131
WMCFILVGRS	0.120	EDQKNHY	0.118
WCFMLIVGRS	0.114	DEKNYQH	0.115
WCFMILVRGS	0.105	EDKQNHY	0.115
WCMIFLVRGS	0.102	EDKNYHQ	0.113
WCFMILVGRS	0.076	EDNKQHY	0.105
WCMIFLVGRS	0.073	KNEDQHY	0.092
WCMFLIVRGS	0.067	EDKNQYH	0.067
WCMFLIVGRS	0.039	DEKNQHY	0.064
WCMFILVRGS	0.029	EDKNYQH	0.052
WCMFILVGRS	0.	EDKNQHY	0.

Given these values, and with a TAP width of 0.08, overlaps between intra-zone distributions soon occur in the spectrum, but it is easy to generate a global ranking from these sub-rankings—say the 600 lower-score arrangements from two subsets of 10 sub-arrangements. The ten lowest sequences found this way are given in Table IV.

The effects of the clustering constraint are illustrated in Figures 9–11. Plotted in Figure 9, besides the full random histogram, is the histogram for the sequences corresponding to all possible permutations within the TAP zone, and all possible permutations within the hydrophilic zone (this makes 3!x7!=30204 sequences). Not unexpectedly, the constraints at hand—three clusters, the first of these keeping its most favorable configuration—have a tremendous effect on improving the score. A comparison of such histograms for two typical arrangements in the hydrophobic zone (they only differ in WCM vs WMC) is plotted in Figure 10, showing the extent of intricacy of both distributions. Figure 11 addresses a

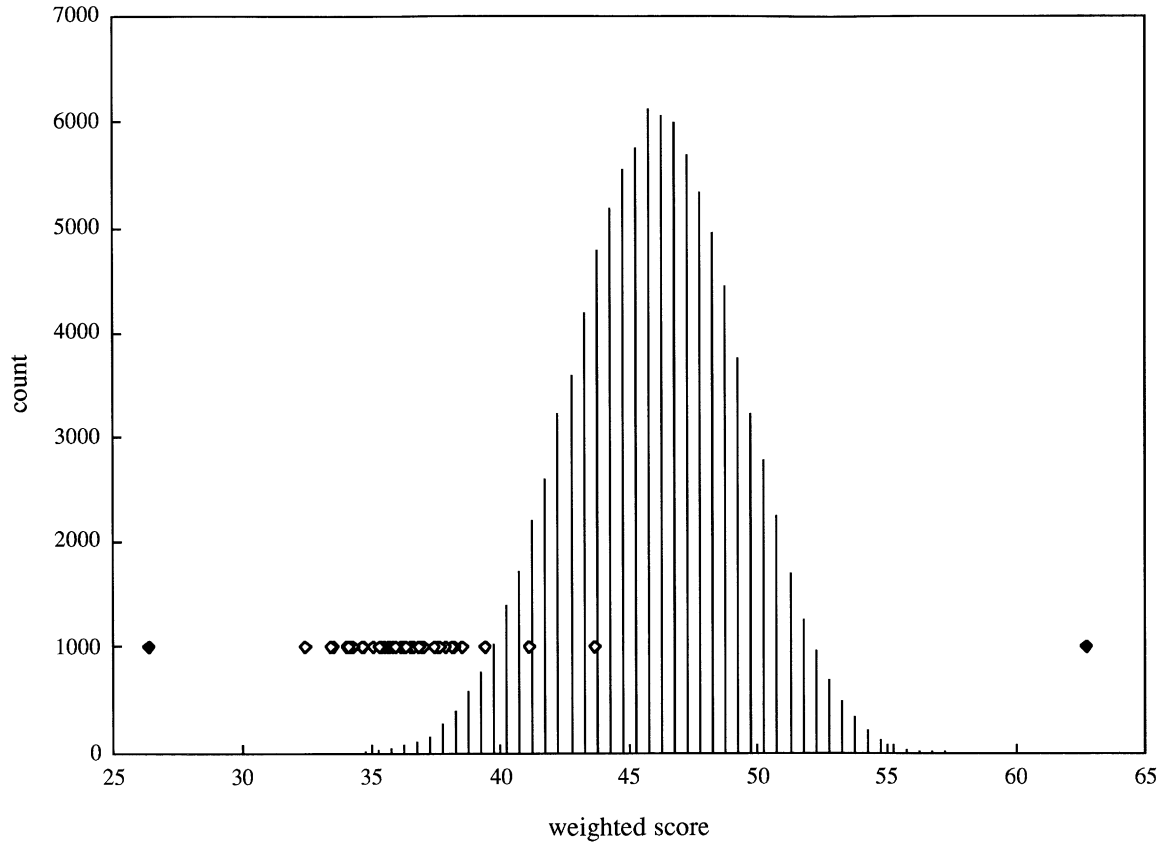

Fig. 8. Weighted-score distribution for a sampling of 100 000 random scales. The dots mark the various hydrophobicity scales (empty) and the extreme possible scores (filled).

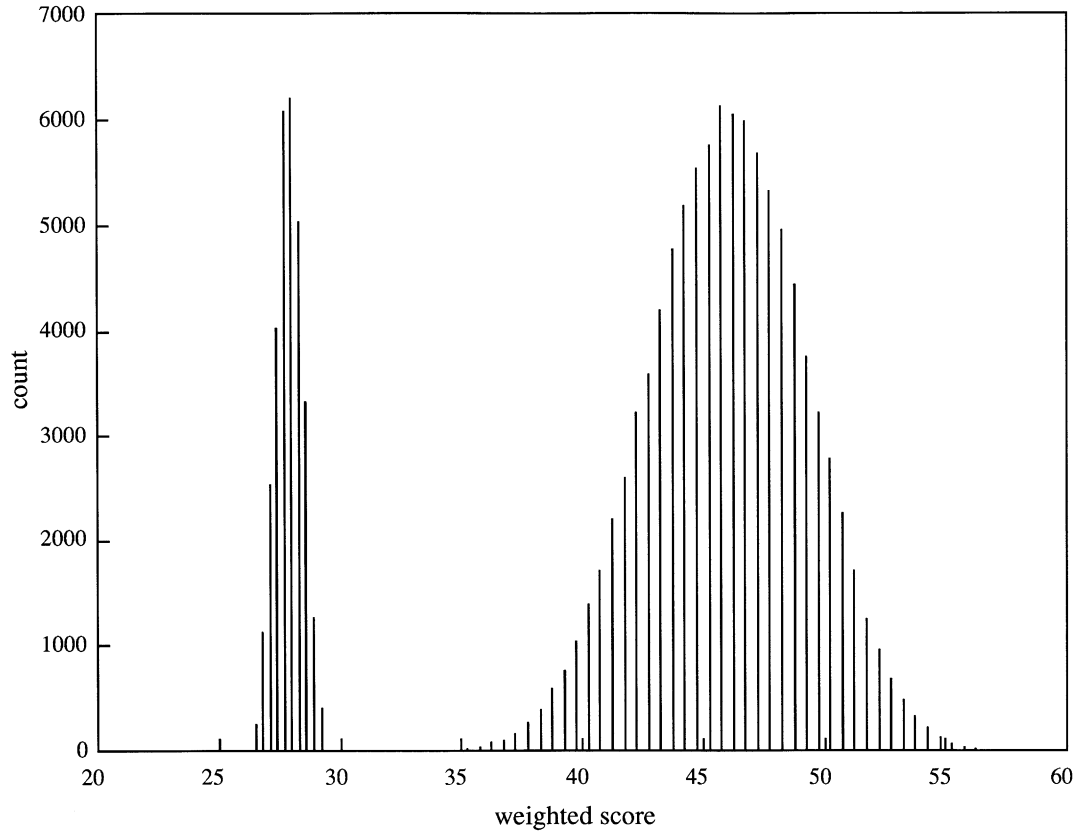

Fig. 9. Left: histogram of weighted scores for all permutations within 11–13 and within 14–20, keeping positions 1–10 in its preferred arrangement. Right: random sampling histogram.

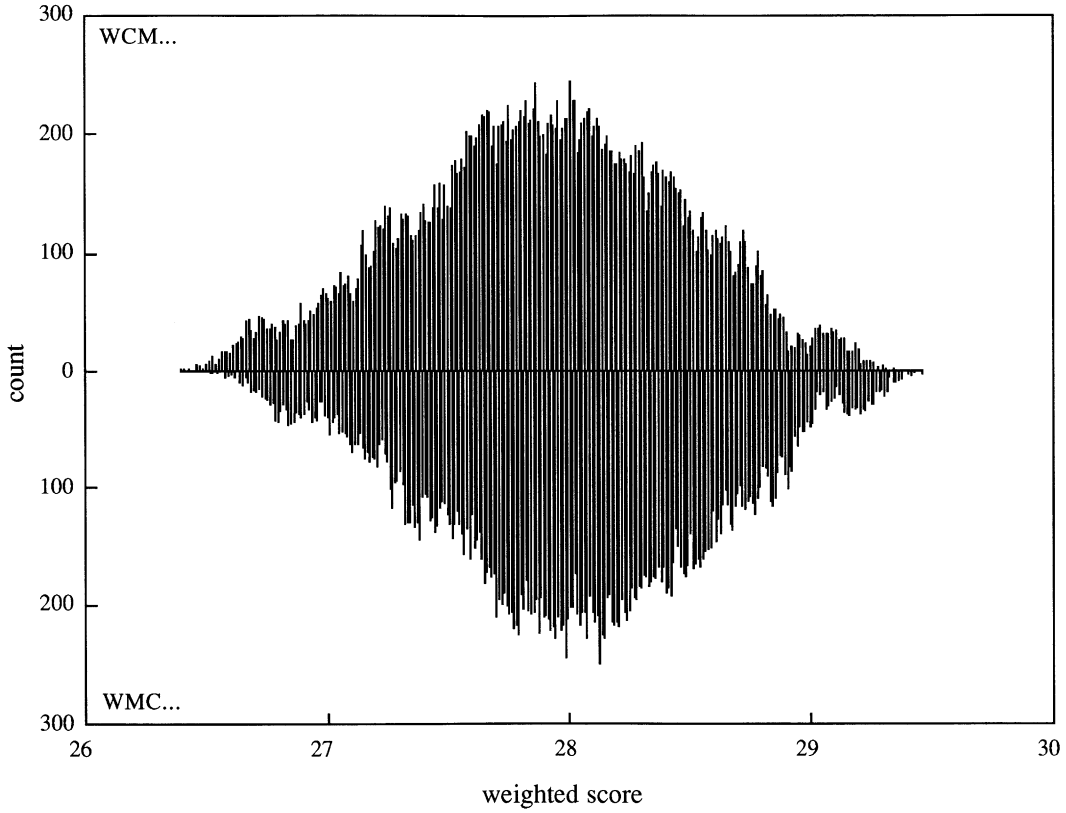

Fig. 10. Histograms of weighted scores for all permutations within 11–13 and within 14–20. Top: best arrangement in 1–10 (WCM...). Bottom: second best arrangement in 1–10 (WMC...).

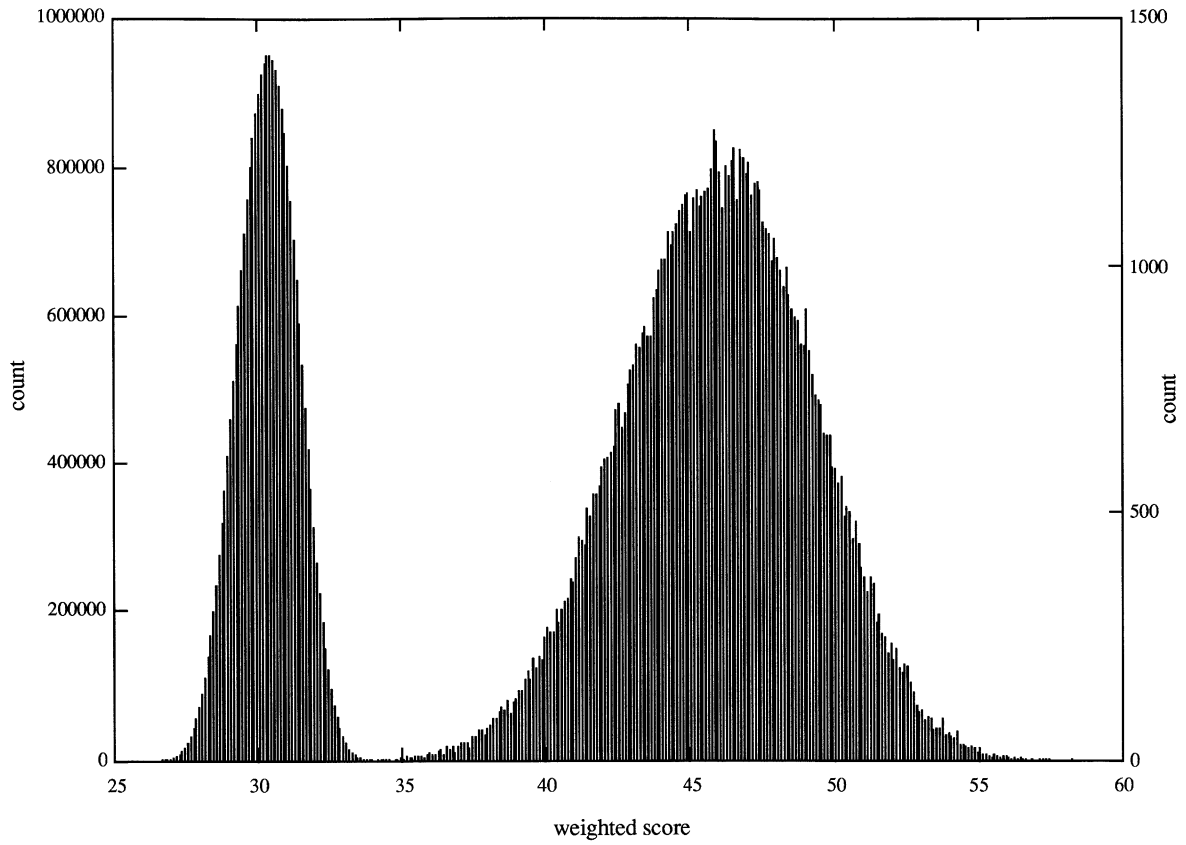

Fig. 11. Histogram of weighted scores for all possible arrangements within the hydrophobic zone (positions 1–7), and within the hydrophilic zone (positions 14–20), keeping the intermediate zones (positions 8–13) in the preferred arrangement (GRSATP). Histogram for all-shuffled sampling is given on the right.

Table VI. A table of the genetic code, simplified for our purposes

1st nucl.	2nd nucl.				3rd nucl.
	U	C	A	G	
U	F	S	Y	C	U
	F	S	Y	C	C
	L	S	–	–	A
	L	S	–	W	G
C	L	P	H	R	U
	L	P	H	R	C
	L	P	Q	R	A
	L	P	Q	R	G
A	I	T	N	S	U
	I	T	N	S	C
	I	T	K	R	A
	M	T	K	R	G
G	V	A	D	G	U
	V	A	D	G	C
	V	A	E	G	A
	V	A	E	G	G

slightly different constraint, corresponding to a shuffling within positions 1–7 (we shall see that this also defines a relevant cluster), and within positions 14–20, the remaining positions, 8–13, being kept in the favorable arrangement, GRSATP. 25 401 600 configurations (=7!x7!) were thus scanned for this purpose. This single constraint also appears to carry a huge part of the score lowering.

It may seem puzzling that the weighting of averaged scores leads back to one of the total-score scales. This probably originates from the fact that the mean occurrence of residues is not totally independent of their codon degeneracy, as noticed long ago (Goldberg and Wittes, 1966; McKay, 1967), and as illustrated in Table V. In a sense, our results suggest that the weighting of the contribution of a residue by its occurrence rate in proteins restores the importance given to it by evolution through the degeneracy of its codons. We therefore tend to consider the following preferred weighted-score arrangement as highly significant:

WCMFILVGRS ATP EDKNQHY

The position of tyrosine at the extreme hydrophilic end, shown here and in 25% of the total-score best scales, notably contrasts with its currently admitted character (see Table II).

Discussion

Interpretation from codon distribution

Let us now try to understand our optimized scales in terms of the structure of the genetic code, a simplified table of which is given in Table VI. As mentioned, the three-zone clustering obtained with total scores could be related, to a good extent, to the distribution of the codons over the four columns of the table of the genetic code. When optimized with averaged scores, where all residues are treated as if they were associated with a codon behaving as the average of all the synonymous codons, the preferred ranking exactly coincides with a grouping of the residues according to their second base. This suggests clustering the scale as follows, according to the four columns of the genetic code:

MIFLV WCGRS TAP YNKDEHQ
11111 44444 222 3333333

Note that serine, S, is located here quite logically, since it belongs to both the second and fourth columns. Such a perfect correlation no longer holds for the scales optimized with the total or weighted scores, but in all cases the end of the hydrophobic cluster, GRS, is conserved. This suggests dividing the scales into four clusters, each one being related to the four possibilities for the second base. When applied to our four typical optimized scales (averaged, weighted, and those of the total ones that best match these), the result is the new clustering :

MIFLVWC GRS TAP YNKDEHQ
WMCFILV GRS TAP YNKDEHQ
WMCFILV GRS ATP EDKNQHY
WCMFILV GRS ATP EDKNQHY

This clustering discloses the following relationships:

(i) The residues whose codons have uracile as second base (FLIMV) are all included in the first zone.
(ii) The second zone includes residues whose codons have guanine as second base (RSG).
(iii) The third zone includes residues whose codons have cytosine as second base (PTA).
(iv) The residues whose codons have adenine as second base (YHQNKDE) are all included in the fourth zone.

This apparent regularity is temperated by some non-obvious trends. First, one of the residues with cytosine as second base belongs to the second zone instead of the third one (S). Grouping serine with arginine can however be understood since these two residues have in common both the highest degeneracy (6) and the guanine as second base. Next, and most important, the first zone also includes residues whose codons have guanine as second base (CW). The grouping in a common hydrophobic cluster of tryptophan and cysteine, belonging to the guanine column, with the residues of the uracile column, may also result from an interplay between the second-base column and the degeneracies. However, these results are all the more sensible since these two residues, W and C, are clearly located on the hydrophobic side in most hydrophobicity scales. Furthermore, as many as eleven hydrophobicity scales begin with the hydrophobic set WMCFILV, of course in various orderings, as reflected by the beginning of the mean hydrophobicity scale: IFLVMWC.

The global clustering of the residues, therefore, emanates from the nature of the second base, modulated by some conditions on the codon degeneracies. The interplay between both parameters can ultimately be traced to the following occurrence rules:

(i) The hydrophobic class incorporates the residues with uracile as second base, without condition on the degeneracy, and those with guanine as second base on condition that degeneracy is lower than or equal to 2.
(ii) The first intermediate class incorporates the residues with guanine as second base, on condition that degeneracy is higher than or equal to 4.
(iii) The second intermediate class incorporates the residues with cytosine as second base, on condition that degeneracy is equal to 4.
(iv) The hydrophilic zone incorporates the residues with adenine as second codon base, without condition on the degeneracy.

Table VII. Second nucleotide and degeneracy for the codons associated with the residues ordered according to the four favorable scales (averaged, total, total and weighted, respectively). The nucleotide labels 1–4 correspond to uracile, cytosine, adenine and guanine respectively

	M	I	F	L	V	W	C	G	R	S	T	A	P	Y	N	K	D	E	H	Q
2nd nucl.	1	1	1	1	1	4	4	4	4	4	2	2	2	3	3	3	3	3	3	3
Codon deg.	1	3	2	6	4	1	2	4	6	6	4	4	4	2	2	2	2	2	2	2
	W	M	C	F	I	L	V	G	R	S	T	A	P	Y	N	K	D	E	H	Q
2nd nucl.	4	1	4	1	1	1	1	4	4	4	2	2	2	3	3	3	3	3	3	3
Codon deg.	1	1	2	2	3	6	4	4	6	6	4	4	4	2	2	2	2	2	2	2
	W	M	C	F	I	L	V	G	R	S	A	T	P	E	D	K	N	Q	H	Y
2nd nucl.	4	1	4	1	1	1	1	4	4	4	2	2	2	3	3	3	3	3	3	3
Codon deg.	1	1	2	2	3	6	4	4	6	6	4	4	4	2	2	2	2	2	2	2
	W	C	M	F	I	L	V	G	R	S	A	T	P	E	D	K	N	Q	H	Y
2nd nucl.	4	4	1	1	1	1	1	4	4	4	2	2	2	3	3	3	3	3	3	3
Codon deg.	1	2	1	2	3	6	4	4	6	6	4	4	4	2	2	2	2	2	2	2

The optimized scaling finally tries to keep a compromise for the regularity in the grouping of these two parameters along the scale. In our typical scales, the two factors are put in regard in Table VII. The last twelve positions, 8–20, are common to all cases. The groupings happen to be perfect for both factors in the last two clusters (11–13, 14–20). The scales differ in the first seven ranks. In the averaged ranking, the organization is matchless for the second bases. The total score rather favors a regular arrangement of the degeneracies now organized as 11 22 36 44 66 444 2222222. The weighted scale brings about some compromise, destroying part of this degeneracy order, and restoring a better second-base clustering.

A two-entry table of the codons, corresponding to the number of amino acids associated with both a given second base and a given degeneracy, is proposed in Figure 12. To emphasize the logic underlying the disposition of the four classes, the usual order of the columns of the genetic code was changed. Interestingly, such a UCAG→UGCA reordering of the genetic code seems to reveal an underlying symmetry of its structure (see Table VIII). The degeneracy of the third codon base is almost identical for codons with complementary first base (i.e. G/C and A/U). This is likely to be meaningful since the two exceptions to this rule, codon AUA coding for I instead of M and codon UGA coding for 'Stop' instead of W, are not universal. In the mitochondrial genetic code of vertebrates, arthropods, mollusca, ascidians, nematodes or yeasts, AUA actually codes for M and UGA for W (Osawa *et al.*, 1992). Such an alternative display of the genetic code had been proposed thirty years ago in its reverse ACGU ordering as best reflecting some statistical regularities (Volkenstein, 1966). The present study suggests it actually carries more conservative information than the commonly used one.

Comparison with literature

Using a neural network, Tolstrup *et al.* (1994) have obtained a grouping of the residues along the edges of a square, broadly corresponding to the four middle bases. Running through such a circular pattern, one can range the residues along a scale, and gather them into groups. The grouping used by these authors opposes the first column (uracile, hydrophobic) to the third one (adenine, hydrophilic), linked by an intermediate zone of the amino acids with guanine (fourth column) or cytosine (second column) as second base: IMVFL WCRSGPAT DENKHQY. The circular pattern allows other clustering schemes as well, and one of them happens to fit our partition

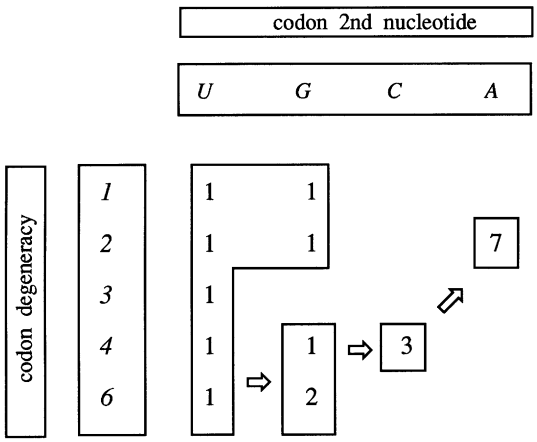

Fig. 12. Amino acid clustering as a function of the codon degeneracies and codon second bases. For convenience, serine is included in the guanine column.

Table VIII. The display of the genetic code which best reflects the presently studied effective property

1st nucl.	2nd nucl.				3rd nucl.
	U	G	C	A	
U	F	C	S	Y	U
	L	W	S	–	G
	F	C	S	Y	C
	L	–	S	–	A
G	V	G	A	D	U
	V	G	A	E	G
	V	G	A	D	C
	V	G	A	E	A
C	L	R	P	H	U
	L	R	P	Q	G
	L	R	P	H	C
	L	R	P	Q	A
A	I	S	T	N	U
	M	R	T	K	G
	I	S	T	N	C
	I	R	T	K	A

scheme perfectly: IMVFLWC RSG PAT DENKHQY. The ordering within each group is not the same as ours, but there are some common features such as the position in 6,7 of WC, as in our averaged-score scale, and the ending position of Y, as in our total-score and weighted-score scales.

Earlier attempts to devise relationships between the twenty amino acids from the genetic code, include the 'similarity alphabet', from a Gray-code model (Swanson, 1984), and the dendrogram, from cluster analysis of angles between vectors (Di Giulio, 1989a). The corresponding diagrams exhibit some features that are common with our grouping patterns, but they also display many significant differences.

Maximum scores

Contrary to the minimum-entropy arrangements of Table VII, the scales that maximize the scores exhibit a maximum alternance in the second bases and degeneracies. Eight degenerate arrangements maximize the total score, corresponding to the following scale

RLAN CE TYQMW FD IG H K P VS

with three allowed permutations between C and E, F and D, and H and P. The unique scales associated with the averaged or weighted maximum scores differs somewhat from this arrangement (Table I, bottom), but basically preserves the maximum entropy distribution of our two parameters. In particular, two of the six-codon residues, LRS, are always located on the borders. Would a crucial protein property be closer to the virtual property associated with these scales, single-base mutations would be prone to be lethal, and the genetic code would be self-destructive instead of self-protective.

Conclusion and prospects

In tackling the question of amino acid usage and the genetic code, we have tried to keep the model as simple as possible. Some refinements can be considered, such as the use, in the distance calculation, of continuous scales, instead of our discrete one-step function. Since this tends to favor confined clusters, additional constraints must therefore be introduced in the optimization process, which partly destroys the external information-free character of the model. More interesting extensions could be the taking into account of stop codons, and the study of altered codes such as the one operating in mitochondria.

Our distance-based scoring procedure to measure the resistance of an effective property towards single-base mutation, led to residue-ranking scales optimized for this virtual property. These scales, which are *ab initio* from the genetic code, more or less reflect the available hydrophobicity scales. Due to non-uniformities in the number of synonymous codons and their use, and in the frequency of residues in proteins, three different scoring schemes were implemented, which all led to a clustering of the residues into broadly three classes. The scales begin with a hydrophobic cluster including the classical hydrophobic residues of the first column of the genetic code, plus tryptophan and cysteine, both belonging to the fourth column. It ends with a hydrophilic cluster containing the classical hydrophilic residues of the third column, including tyrosine. These groups are connected by the residues belonging to the fourth and second columns, with serine logically joining these two subgroups. The absolute ranking within these clusters may vary according to the mode of score summation. We propose to retain the rankings obtained with the averaged-score calculation, and the averaged-and-weighted-score calculation, both sharing many common features with the 144 degenerate scales obtained from the total-score evaluation. These are respectively:

MIFLVWC GRSTAP YNKDEHQ
WCMFILV GRSATP EDKNQHY

Most strategies for predicting protein folds from primary sequences make use of hydrophobicity scales or preliminary-defined hydrophobic clusters. Hydrophobic sets used in two recent promising methods (Huang *et al.*, 1995; Srinivasan and Rose, 1995) happen to coincide exactly with our seven-residue hydrophobic class. The present derivations may provide some support for the choice of such hydrophobic sets, and suggest a refinement to the models by introducing polar or amphipatic sets defined from the present 'natural' effective property. This should hold everytime dichotomies between residue properties are introduced, either in the context of fold prediction (Gaboriaud, 1987; Sun *et al.*, 1995), lattice models (Li *et al.*, 1996) or in statistic explorations of primary sequences, such as random walk trajectories (Pande *et al.*, 1994). The present scales can also be used in sequence homology, for protein families sharing little sequence identity. Preliminary results in this field are encouraging.

Since we cannot define *ex nihilo* an absolute and critical residue property, the present work does not settle the question as to whether the code is optimized or not. It however provides some clues. Proteins are critical for life, and their folding is critical for fulfilling the functions they are assigned. The following residue characteristics: hydrophobicity, tendency to form residue–residue interactions through hydrophobic link, hydrogen bond, salt bridge, or disulfide bridge, tendency to prefer water–residue hydrogen bonds, all concur to produce for a given sequence, a fold and hence a function. In the light of what is presently known about these properties, in particular through the hydrophobicity scales considered as a whole, the present study supports the idea that the code as it has evolved is optimized in such a way that the machinery of life can best resist random nucleotide mutations.

Acknowledgements

Our interest in the links between hydrophobicity and the genetic code was aroused from discussions with Professor Jean Durup. We are grateful to Professor Philippe Dessen and the Centre de Bioinformatique CNRS-INSERM de Villejuif for giving us access to the sequence database facilities.

References

Aboderin,A.A. (1971) *Int. J. Biochem.*, **2**, 537–544.
Alff-Steinberger,C. (1969) *Proc. Natl Acad. Sci. USA*, **64**, 584–591.
Béland,P. and Allen,T.F.H. (1994) *J. Theor. Biol.*, **170**, 359–365.
Brack,A. and Orgel,L.E. (1975) *Nature*, **256**, 383–387.
Bull,H.B. and Breese,K. (1974) *Arch. Biochem. Biophys.*, **161**, 665–670.
Charton,M. and Charton,B.I. (1982) *J. Theor. Biol.*, **99**, 629–644.
Chipens,G.I. (1991) *Zh. Evol. Biokhim. Fisiol.*, **27**, 522–529.
Chipens,G.I., Gnilomedeva,L.E., Ievinya,N.G., Kudryavtsev,O.E., Rudzish,R.V. and Sklyarova,S. N. (1989) *Zh. Evol. Biokhim. Fisiol.*, **25**, 654–663.
Chothia,C. (1976) *J. Mol. Biol.*, **105**, 1–14.
Cornette,J.L., Cease,K.B., Margalit,H., Spouge,J.L., Berzofsky,J.A. and DeLisi,C. (1987) *J. Mol. Biol.*, **195**, 659–685.
Crick,F.H.C. (1968) *J. Mol. Biol.*, **38**, 367–379.
Cullmann,G. and Labouygues,J.M. (1983) *Biosystems*, **16**, 9–29.
Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*. vol. 5, suppl. 3. National Biomedical Research Foundation, Silver Spring, MD, pp. 345–352.
Degli Espositi,M., Crimi,M. and Venturoli,G. (1990) *Eur. J. Biochem.*, **190**, 207–219.

Delarue,M. (1995) *J. Mol. Evol.*, **41**, 703–711.
Demongeot,J. and Besson,J. (1996) *C. R. Acad. Sci. Paris*, **319**, 443–451.
Di Giulio,M. (1989a) *J. Mol. Evol.*, **29**, 191–201.
Di Giulio,M. (1989b) *J. Mol. Evol.*, **29**, 288–293.
Di Giulio,M. (1996) *Orig. Life Evol. Biosphere*, **26**, 589–609.
Efremov,R.G. and Alix,A.J.P. (1993) *J. Biomolec. Struct. Dyn.*, **11**, 483–507.
Eigen,M. and Schuster,P. (1978) *Naturwissenschaften*, **65**, 341–369.
Eisenberg,D. and McLachlan,A.D. (1986) *Nature*, **319**, 99–203.
Eisenberg,D., Weiss,R.M., Terwilliger,T.C. and Wilcox,W. (1982) *Faraday Symp. Chem. Soc.*, **17**, 109–120.
Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) *J. Mol. Biol.*, **179**, 125–142.
Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.
Epstein,C.J. (1966) *Nature*, **210**, 25–28.
Fauchère,J. and Pliska,V. (1983) *Eur. J. Med. Chem.*, **18**, 369–375.
Frömmel,C. (1984) *J. Theor. Biol.*, **111**, 247–260.
Gaboriaud,C., Bissery,V., Benchetrit,T. and Mornon,J.P. (1987) *FEBS Lett.*, **224**, 149–155.
Goldberg,A.L. and Wittes,R.E. (1966) *Science*, **153**, 420–424.
Grantham,R. (1974) *Science*, **185**, 862–864.
Guy,H.R. (1985) *Biophys. J.*, **47**, 61–70.
Haig,D. and Hurst,D. (1991) *J. Mol. Evol.*, **33**, 412–417.
Hartman,H. (1995) *J. Mol. Evol.*, **40**, 541–544.
Hasegawa,M. and Miyata,T. (1980) *Orig. Life.*, **10**, 265–270.
Hopp,T.P. (1986) *J. Immunol. Methods.*, **88**, 1–18.
Hopp,T. and Woods,K.R. (1981) *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.
Huang,E.S., Subbiah,S. and Levitt,M. (1995) *J. Mol. Biol.*, **252**, 709–720.
Janin,J. (1979) *Nature*, **277**, 491–492.
Jimenez-Sanchez,A. (1995) *J. Mol. Evol.*, **41**, 712–716.
Jones,D.D. (1975) *J. Theor. Biol.*, **50**, 167–183.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) *Nature*, **358**, 86–89.
Jukes,T.H. (1973) *Nature*, **246**, 22–26.
Karlin,S., Zuker,M. and Brocchieri,L. (1994) *J. Mol. Biol.*, **239**, 227–248.
King,J.L. and Jukes,T.H. (1969) *Science*, **164**, 788–798.
Konecny,J., Eckert,M., Schöniger,M. and Hofacker,G.L. (1993) *J. Mol. Evol.*, **36**, 407–416.
Konecny,J., Schöniger,M. and Hofacker,G.L. (1995) *J. Theor. Biol.*, **173**, 263–270.
Krigbaum,W.R. and Komoriya,A. (1979) *Biochim. Biophys. Acta.*, **576**, 204–228.
Kuntz,I.D. (1971) *J. Amer. Chem. Soc.*, **93**, 514–516.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Ladunga,I. and Smith,R.F. (1997) *Protein Engng*, **10**, 187–196.
Levitt,M. (1976) *J. Mol. Biol.*, **104**, 59–107.
Li,H., Helling,R., Tang,C. and Wingreen,N. (1996) *Science*, **273**, 666–669.
Lüthy,R. and Eisenberg,D. (1991) In Gribskov,M. and Devereux,J. (eds), *Sequence Analysis Primer*. Stockton Press, New York, pp. 61–87.
McKay,A.L. (1967) *Nature*, **216**, 159–160.
Marlborough,D.I. (1980) *Orig. Life.*, **10**, 3–14.
Médigue,C., Viari,A., Hénaut,A. and Danchin,A. (1993) *Microbiol. Rev.*, **57**, 623–654.
Meek,J.L. (1980) *Proc. Natl Acad. Sci. USA*, **77**, 1632–1636.
Meirovitch,H., Rackovsky,S. and Scheraga,H.A. (1980) *Macromolecules*, **13**, 1398–1405.
Meister,A. (1965) *Biochemistry of the amino acids*, second edition, volume I. Academic Press, New York, p. 28.
Miyazawa,S. and Jernigan,R.L. (1985) *Macromolecules*, **18**, 534–552.
Nakai,K., Kidera,A. and Kanehisa,M. (1988) *Protein Engng*, **2**, 93–100.
Nishikawa,K. and Ooi,T. (1980) *Int. J. Pept. Protein Res.*, **16**, 19–32.
Nozaki,Y. and Tanford,C. (1971) *J. Biol. Chem.*, **246**, 2211–2217.
Olsen,K.W. (1980) *Biochim. Biophys. Acta.*, **622**, 259–267.
Orgel,L.E. (1972) *Isr. J. Chem.*, **10**, 287–292.
Osawa,S., Jukes,T.H., Watnabe,K. and Muto,A. (1992) *Microbiol. Rev.*, **56**, 229–264.
Pande,V.S., Grosberg,A.Y. and Tanaka,T. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 12972–12975.
Perlwitz,M.D., Burks,C. and Waterman,M.S. (1988) *Advan. Appl. Math.*, **9**, 7–21.
Pixner,P., Heiden,W., Merx,H., Moeckel,G., Möller,A. and Brickmann,J. (1994) *J. Chem. Inf. Comput. Sci.*, **34**, 1309–1319.
Ponnuswamy,P.K., Prabhakaran,M. and Manalavan,P. (1980) *Biochim. Biophys. Acta.*, **623**, 301–316.
Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in Fortran. The Art of Scientific Computing*, second edition. Cambridge University Press, pp. 436–448.
Reeck,G. (1976) In Fasman, G. D. (ed.), *Handbook of Biochemistry and Molecular Biology*, 3rd Edition. CRC Press, Cleveland, pp. 504–520.
Rekker,R.F. (1977) *The Hydrophobic Fragmental Constant*. Elsevier Scientific Publishing Company, Amsterdam.
Robson,B. and Osguthorpe,D.J. (1978) *J. Mol. Biol.*, **132**, 19–51.
Rodionov,M.A. and Galaktionov,S.G. (1992) *Mol. Biol.*, **36**, 777–783.
Rose,G.D. and Wolfenden,R. (1993) *Annu. Rev. Biomol. Struct.*, **22**, 381–415.
Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985a) *Science*, **229**, 834–838.
Rose,G.D., Gierasch,L.M. and Smith,J.A. (1985b) *Advan. Protein Chem.*, **37**, 1–109.
Seybold,P.G. (1976) *Int. J. Quantum Chem. Quantum Biology Symp.*, **3**, 39–43.
Siemion,I.Z. (1994a) *BioSystems*, **32**, 25–35.
Siemion,I.Z. (1994b) *BioSystems*, **32**, 163–170.
Siemion,I.Z. (1995) *Amino Acids*, **8**, 1–13.
Siemion,I.Z. and Stefanowicz, P. (1992) *BioSystems*, **27**, 77–84.
Siemion,I.Z., Siemion,P.J. and Krajewski,K. (1995) *BioSystems*, **36**, 231–238.
Sjöström,M. and Wold,S. (1985) *J. Mol. Evol.*, **22**, 272–277.
Srinivasan,R. and Rose,G.D. (1995) *Proteins*, **22**, 81–99.
Sukhodolets,V.V. (1989) *J. Theor. Biol.*, **141**, 379–389.
Sun,S., Thomas,P.D. and Dill,K.A. (1995) *Protein Engng*, **8**, 769–778.
Swanson,R. (1984) *Bull. Math. Biol.*, **46**, 187–203.
Sweet,R.M. and Eisenberg,D. (1983) *J. Mol. Biol.*, **171**, 479–488.
Taylor,F.J.R. and Coates,D. (1989) *Biosystems*, **22**, 177–187.
Tolstrup,N., Toftgard,J., Engelbrecht,J. and Brunak,S. (1994) *J. Mol. Biol.*, **243**, 816–820.
von Heijne,G. and Blomberg,C. (1979) *Eur. J. Biochem.*, **97**, 175–181.
Wertz,D.H. and Scheraga,H.A. (1978) *Macromolecules*, **11**, 9–15.
Wetzel,R. (1995) *J. Mol. Evol.*, **40**, 545–550.
Woese,C.R., Dugre,D.H., Dugre,S.A., Kondo,M. and Saxinger,W.C. (1966) *Cold Spring Harbor Symp. Quant. Biol.*, **31**, 723–736.
Wolfenden,R., Cullis,P.M. and Southgate,C.C.F. (1979) *Science*, **206**, 575–577.
Wolfenden,R., Andersson,L., Cullis,P.M. and Southgate,C.C.F. (1981) *Biochemistry*, **20**, 849–855.
Volkenstein,M.V. (1966) *Biochim. Biophys. Acta.*, **119**, 421–424.
Wong,J.T. (1975) *Proc. Natl Acad. Sci. USA*, **75**, 1909–1912.
Wong,J.T. (1980) *Proc. Natl Acad. Sci. USA*, **77**, 1083–1086.
Zimmerman,J.M., Eliezer,N. and Simha,R. (1968) *J. Theor. Biol.*, **21**, 170–201.

Received May 28, 1997; revised November 26, 1997; accepted December 1, 1997

Appendix

Hydrophobicity scales

Many of the hydrophobicity scales corresponding to the ranks listed in Table I were selected from the work by Cornette *et al.* (1987), in which 38 literature scales are reviewed, and eight new ones proposed. Whenever several amino acids were assigned the same hydrophobicity index, the ordering giving the more advantageous averaged score has been chosen (total and averaged scores most often led to similar results anyway). This degree of freedom cleverly gives advantage to the scales that exhibit many degenerate residues, especially when there are greater than two degeneracies. This is why the scales including too many degenerate values, such as that of Kuntz (1971) have been discarded. Such opimization may have more or less incidence on the scores. In the scale of Sweet, for instance, the score for KG ordering is larger than the selected GK ordering by as much as 34 (total score) or 10 (averaged score). On the other hand, in the scale of Efremov, the score for the WV arrangement is only at +2 (total) or +0.6 (averaged) above that for VW. The scales in which two or more amino acids were lacking have also been discarded. When only one amino acid was lacking, as in the scales of Rekker (R), Wolfenden (P), or Robson (G), its ranking has been optimized. For the three afore-mentioned cases, the preferred locations occur at positions 10, 12 and 13, respectively. The total number of degree of freedoms, or degeneracy, is indicated in the column labelled D in Table I.

More complete reviews on hydrophobicity scales can be found in Rose *et al.*, 1985b; Hopp, 1986; Cornette *et al.*, 1987; Degli Esposito *et al.*, 1990; Rodionov and Galaktionov, 1992; Rose and Wolfenden, 1993. Exhaustive analyses of physicochemical and biochemical properties of amino acids have been done by Nakai *et al.* (1988), and Ladunga and Smith (1997). For a 'behavioral' analysis of selected amino acids, see Karlin *et al.*, 1994. For hydrophobicity in organic compounds in general, see Pixner *et al.*, 1994. See below, in alphabetic order, a brief description and the references to selected hydrophobicity scales. The reader is referred to the article by Cornette *et al.* (1987) for more details on most of them.

Aboderin: mobilities of amino acids on chromatography paper (Aboderin, 1971).

Bull: effects of amino acids on the surface tension of water (Bull and Breese, 1974).
Chothia: proportion of residues that are buried in the native structures of a set of proteins (Chothia, 1976).
Chothia 2: transfer energies computed from the above scale (Cornette *et al.*, 1987).
Cornette, and Cornette 1–7: the scales labelled NNEIG, PRIFT, PRILS, ALTFT, ALTLS, TOTFT, and TOTLS, respectively, in the paper by Cornette *et al.* (1987). These are composite scales built from elaborate statistical treatments of literature scales.
Efremov: three-dimensional molecular hydrophobicity potentials, extracted from experimental geometrical data on proteins (Efremov and Alix, 1993).
Eisenberg: average of five scales: Nozaki and Tanford (1971), Wolfenden, Chothia, Janin, and von Heijne (Eisenberg *et al.*, 1982; 1984).
Eisenberg 2: combines atomic solvent-accessible areas with atomic solvation parameters fitting the scale of Fauchère (Eisenberg and MacLachlan, 1986).
Engelman: Semitheoretical approach, taking into account side chains attached to an α-helical framework (Engelman *et al.*, 1986).
Fauchère: octanol/water distribution measurements on the acetyl-amino acid amides side-chain analogs (Fauchère and Pliska, 1983).
Frömmel: apolar accessible surface areas (Frömmel, 1984).
Guy: experimental geometrical data on proteins (Guy, 1985).
Guy 2: combines the scale defined above with three others: Ponnuswamy, Meirovitch, and Wertz (Cornette *et al.*, 1987).
Hopp: adjustment of Levitt's scale in view of antigenic determinants identification (Hopp and Woods, 1981).
Janin: molar fractions of buried to accessible residues, from experimental geometrical data on proteins (Janin, 1979).
Janin 2: transfer energies computed from the above scale (Cornette *et al.*, 1987).
Jones: Zimmerman's scale adjusted to other experimental data (Jones, 1975).
Krigbaum, and Krigbaum 2: combine solubility data on residues and crystallographic structural data on proteins (Krigbaum and Komoriya, 1979).
Kyte: combines Wolfenden's scale, Chothia's scale, and estimates based on the constituent parts of the side chains (Kyte and Doolittle, 1982).
Levitt: combine the data of Nozaki and Tanford (1971) on the free energy of transfer of amino acid side-chains and backbone peptide units from water to ethanol and dioxane, from the solubilities of the amino acids in those media, with data on accessible surface areas (Levitt, 1976).
Meek: retention coefficients for amino acids, computed from the retention times of peptides in high-pressure liquid chromatography (Meek, 1980).
Meirovitch: Cα-protein center distances from experimental geometrical data (Meirovitch *et al.*, 1980).
Miyazawa: effective contact energies, from protein geometrical data (Miyazawa and Jernigan, 1985).
Nishikawa: contact number of residues in proteins, from experimental Cα distances matrices (Nishikawa and Ooi, 1980).
Olsen: average internal preferences (Olsen, 1980).
Ponnuswamy: surrounding hydrophobicity, obtained by combining Jones' values with geometrical proximity data from a set of proteins (Ponnuswamy *et al.*, 1980).
Rekker: fragmental constants from measured partition values of organic molecules (Rekker, 1977).
Robson: experimental geometrical data on proteins (Robson and Osguthorpe, 1978).
Rose: mean fractional area loss by each amino acid in a set of proteins (Rose *et al.*, 1985a).
Sanejouand: this scale was built from three-dimensional structural data on proteins (Sanejouand, unpublished work), on the same principle as that of Nishikawa and Ooi (1980). It measures the mean number of heavy atoms neighboring a residue within 10 Å, as *averaged* over *all* side-chain heavy atoms, in the 102 proteins of the non-redundant set of Jones *et al.* (1992). Doing so led to the following mean number of neighbors. I: 177, W: 175, F: 175, L: 171, C: 170, V: 169, M: 165, Y: 155, A :146, H :145, T: 137, G: 135, S: 131, P: 128, R: 126, Q: 125, N: 122, D: 115, E: 109, K: 103.
Sweet: optimal matching hydrophobicity, evaluated from the mutation matrices of Dayhoff *et al.* (1978) (Sweet and Eisenberg, 1983).
von Heijne: free energy of transfer of a single residue in a polypeptide from random coil in aqueous phase to helix form in non-polar environment, combined with data of accessible surface area of Chothia (1976) (von Heijne and Blomberg, 1979).
Wertz: inside-vs-outside classification of residues in a set of proteins (Wertz and Scheraga, 1978).
Wertz 2: transfer energies computed from the above scale (Cornette *et al.*, 1987).
Wolfenden: distribution of amino acid side-chains between dilute aqueous solutions and the vapor phase (Wolfenden *et al.*, 1981).
Zimmerman: solubility data (Zimmerman *et al.*, 1968).

Miscellaneous scales

The alphabetic scale is given here as an example of random scale. Incidentally, it is used in Figure 1 to illustrate the scoring procedure. The scales of pK'_1, pK'_2, and pI, given in increasing order, were taken from Meister (1965). Molar masses correspond to molecular weights at pH 7 (Lüthy and Eisenberg, 1991). The residue occurrence scale, given in decreasing order, reflects the amino acid composition of the protein sequence database OWL, release 26.0 (June 1995), which included 105 990 sequences, and 32 623 718 residues. We could check that this composition is fairly stable over the successive versions (only some tenths of a percent changes for each residue fraction). Comparison with a smaller and older database, such as PIR, release 26, (1990, 25 814 sequences, 7 348 950 residues) also results in changes of at most a few tenths of a percent for each residue fraction. This stability would suggest that, in spite of various compositional biases, such databases provide a reasonable clue for the mean occurrence of residues in proteins.

Scoring and simulation details

In our scoring system, the ranking direction of the 20 amino acids is, of course, arbitrary. Scores remain unchanged if scales are transposed leftside right. In other words, a scale is a vector, but its direction is arbitrary, so that it is by pure convention that our scales run from hydrophobic character, or else, at left, to hydrophilic character, at right. Seeking for the extreme scores over a configurational space of $20! \approx 2.10^{18}$ is a typical combinatorial optimization problem, like that of the travelling salesman, for which simulated annealing algorithms have been devised (for an introduction, see e.g. Press *et al.*, 1992). Careful descent in temperature from any starting point in principle ensures the best chance of reaching the desired extremum. However, there is no guarantee at all of finding the unique lowest (or highest) configuration in the full space. Strictly speaking, there is therefore no absolute assurance of finding, the extreme-score scales. We tried to put all chances on our side by first running hundreds of explorations from random or non-random scales. We then deduced, from such results, various subsets of scales prone to have potential favorable scores. These subsets were sometimes exhaustively scanned within a clustering constraint, otherwise they were taken as starting points for further searches. A typical Monte Carlo simulated-annealing exploration starts with temperatures corresponding to up to +100 points above the initial total score, and scanned several thousands of random permutations for each of the –1 point descent in temperature. For averaged and weighted scores, these parameters are adjusted accordingly. The latter case, with its total lack of configuration degeneracy, required the most thorough explorations.