JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 5, 2009

Pp. 703–723
DOI: 10.1089/cmb.2008.0173

Interpretable Numerical Descriptors of Amino Acid Space

ALEXANDER G. GEORGIEV

ABSTRACT

Informative numerical representations of amino acid residues are essential for successful *in silico* modeling or establishing the structure–activity relationships of proteins. A straightforward approach is adopted here for representing more than 500 amino acid indices from the AAindex database by a set of uncorrelated scales, satisfying the VARIMAX criterion. Different measures are considered in order to demonstrate the improved interpretability of the current scales as compared to previously published ones. Performance is also addressed in a classification problem of G-protein coupled receptors, and is found to be similar or higher than the performance achieved by six other scale sets. Finally, a unique correspondence between numerical indices and mutation matrices is derived and discussed in light of the evolutionary conservation of amino acid properties. Conclusions from this study highlight the discord between ease of interpretation of amino acid scales and their relevance to protein structure conservation, as well as general considerations for designing custom scale sets.

Key words: algorithms, linear algebra, sequence analysis.

1. INTRODUCTION

PROTEINS ARE LINEAR POLYMERS of 20 species of L-amino acids, and owe their structure and function to the physico-chemical properties and specific order of their building blocks. Amino acids are traditionally represented by letter codes, which are less suitable for use in computational algorithms than numbers. Therefore, studies have aimed at mathematicizing the differences and similarities between amino acids by assigning numerical values to them, corresponding to various propensity scales. Ordering amino acids according to a certain measurable physico-chemical feature or propensity results in 20-dimensional vectors, or *amino acid indices* (Kidera et al., 1985a). Similar properties of amino acids are commonly assayed by more than one method, resulting in different indices associated with essentially the same feature.

In practical computational applications dealing with encoding protein sequence by numbers, a minimal set of indices is usually selected to avoid the ''curse of dimensionality'' (Bellman, 1961) arising from use of too many variables. Generally, these indices conform to the following rules: (i) they correspond closely to some measurable quality of single aminoacid residues; (ii) they are orthogonal (uncorrelated) so that the associated

Department of Biochemistry and Biophysics, Stockholm University, The Arrhenius Laboratories, Stockholm, Sweden.

property can be analyzed separately; and (iii) they are representative for the information available throughout the scientific literature.

Such sets of indices were derived in the past, notably by Sneath (1966), Kidera et al. (1985a), and Hellberg et al. (1986). Owing to their pioneering character, computation of these indices could not take advantage of data published during the last two decades. The recent rapid accumulation of structural data has made it possible to expand the statistics on structural features of individual amino acids, while novel approaches (e.g., ''biological'' hydrophobicity scale [Hessa et al., 2005]) have enriched the diversity of amino acid descriptors. Several more recent studies report scales derived to solve specific problems or designed for general use (Sandberg et al., 1998; Venkatarajan and Braun, 2001; Atchley et al., 2005; Opiyo and Moriyama, 2007). Sandberg et al. (1998) used principal components analysis (PCA) to further expand the *z-scales* originally derived from 29 physico-chemical descriptors by Hellberg et al. (1986). Multidimensional scaling (MDS) was employed by Venkatarajan and Braun (2001) to analyze 237 physico-chemical properties compiled from public databases. Atchley et al. (2005) performed factor analysis (FA), followed by PROMAX rotation, to compute five factors from 54 selected amino acid attributes. Most recently, Opiyo and Moriyama (2007) selected 12 physico-chemical properties, from which five principal components were derived by PCA.

The prevailing technique used to derive amino acid descriptor scales has been PCA (and similar techniques such as MDS or FA). These approaches are useful when the purpose of the analysis is dimensionality reduction; however, they are less useful in designing *interpretable* scales. Axes rotation is usually necessary in order to find the underlying *simple structure* in the data (Thurstone, 1938) and for improving the interpretability of principal components. While PROMAX rotation was used in one previous study (Atchley et al., 2005), only a small subset of properties was considered during the analysis, potentially explaining the poor interpretability of two of the resulting five scales.

The main purpose of the present study is to uncover the underlying structure and dimensionality of amino acid space by identifying as many *independent* scales as possible with good correlation to published scales. As a set of 20 naturally occurring amino acid residues can be completely described by the same number of scales, a maximum of 20 dimensions are considered initially; however, due to the similarities between amino acids, the number of meaningful dimensions is likely lower than this maximum. Dimensionality reduction based on the relative importance of the scales for protein structure conservation is also considered as a secondary task. Finally, the usefulness of the set of interpretable scales derived here in direct applications is illustrated by the presented classification analysis of G-protein coupled receptors (GPCRs).

2. METHODS

2.1. Source data—the AAindex

The AAindex is maintained at *www.genome.jp/aaindex* and has been described at various stages during its development (Nakai et al., 1988; Tomii and Kanehisa, 1996; Kawashima and Kanehisa, 2000). It comprises a collection of 544 amino acid (AA) indices (release 9.1, August 2006), extracted from published work on physical, chemical, statistical, and biological features of amino acids and residues. This database was used previously to develop amino acid descriptors for general application (Kidera et al., 1985a,b); since then, however, the collection has nearly tripled in size by the addition of a substantial amount of novel data.

The current dataset of 544 indices was pre-processed as follows. Incomplete indices (with fewer than 20 entries) were removed. The remaining 509 indices were centered around zero by subtracting their mean, and were rescaled to unit vector length.

2.2. Source data—G-protein coupled receptors

The 10.0 release (June 2006) of the collection of GPCRs at *www.gpcr.org/7tm/* (Horn et al., 2003) was used to retrieve 1882 sequences of Class A GPCRs. The classification into subfamilies from the same database was considered, with 16 subfamilies: amine, peptide, glycoprotein hormone, opsin, olfactory, prostanoid, nucleotide-like, cannabinoid, platelet-activating factor, gonadotropinreleasing hormone, thyrotropin-releasing hormone, melatonin, lysosphingolipid, leukotriene, viral receptors, and orphan receptors. Upon further evaluation, the orphan family and the viral family were excluded, since they deviated significantly in the computational models.

2.3. Scale interpretation

Utilizing protein sequence statistics, a single entry in an amino acid scale can be viewed as the logarithm of the odds ratio of an amino acid to be found within a certain structural region (such as a β-sheet or a membrane spanning helix), and the frequency of occurrence of the amino acid multiplied by the fraction of residues in the dataset present in the structure in question, as follows

$$s_{x,i} = log(q_{x,i}) = log\frac{p_{x,i}}{f_x f_i}, \tag{1}$$

where f_x is the observed frequency of occurrence of amino acid x in the database, f_i is the fraction of residues characterized by structure i, and $p_{x,i}$ is the joint frequency of occurrence of amino acid x in structure i calculated over the entire dataset. This definition of amino acid scales is analogous to the definitions used in substitution matrices (Henikoff and Henikoff, 1992; Eddy, 2004); however, the natural logarithm is considered here instead of base 2 logarithm, leading to a multiplicative constant difference without further implications on the data presented here.

The expression *amino acid space*, previously used by Kidera et al. (1985a), was also employed here to refer to a multidimensional Euclidean space where amino acids determine the axes, while the scales from various published studies are points or vectors. This is in contrast to *index space*, where amino acids are represented as points with coordinates determined by indices corresponding to measurable physico-chemical or statistical properties.

2.4. Correlation

The main tool used throughout this work for comparing two indices was the *Pearson product-moment correlation coefficient* (referred to as *correlation* further in the text), calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y}, \tag{2}$$

where x and y are the indices being compared, $\bar{x}$ and $\bar{y}$ their arithmetic means, and σ_x and σ_y are the standard deviations. For normalized data with zero mean and unit standard deviation, the above equation reduces to $r_{xy} = \sum x_i y_i / (n-1)$. Whenever the sign of the correlation is unimportant, the squared correlation r^2 was used, which has the important advantage of being additive.

2.5. Principal components analysis

PCA (Pearson, 1901) is a matrix decomposition technique used mainly for dimensionality reduction. It involves a linear transformation of the data matrix to a coordinate system of orthogonal vectors (called *principal components*), which consecutively capture the maximum remaining variance in the data. Technically, it is often performed through eigen-decomposition of the covariance matrix.

2.6. VARIMAX rotation

The VARIMAX criterion was developed by Kaiser (1958) as a solution to the *simple structure* described by Thurstone (1947). It aims at maximizing the variance V of squared factor loadings q^2 from a $n \times r$ (*columns* by *rows*) matrix $Q = [q_{ij}]$, calculated as:

$$V = \sum_{j=1}^{r}\left\{\frac{1}{n}\sum_{i=1}^{n} q_{ij}^4 - \frac{1}{n^2}\left(\sum_{i=1}^{n} q_{ij}^2\right)^2\right\} \tag{3}$$

The VARIMAX criterion is commonly used for factor rotation following factor analysis; among its advantages is the invariance of the solution with different composition of the test set (Kaiser, 1958).

2.7. Reduced-redundancy set selection

In order to reduce the redundancy of the AA index and thus decrease the bias caused by overrepresented properties, the following selection procedure was performed. All indices were compared pairwise in

terms of squared correlation, and one of each two most correlated with each other indices was removed from the dataset. The procedure was repeated and more indices removed sequentially, until 50 indices were left.

2.8. Consensus scales

Individual indices were selected from the AAindex based on *keywords* present in their description in order to form *subsets* of indices corresponding to these keywords. The subsets of indices collected in this way were further processed by PCA, and the loadings on the first principal component of this analysis were used as a scale representative of the feature associated with the keyword. For example, the word ''alpha'' was used to select scales potentially associated with α-helicity and include them in the ''alpha'' subset. The intersection of this subset with another one described by either ''helix'' or ''helicity'' gave the final *alpha-helicity* subset. Similarly, the *hydrophobicity* subset was obtained by uniting scales with ''hydrophobicity,'' ''hydrophillicity,'' ''hydropathy,'' or ''membrane'' in their description. The remaining subsets were constructed analogously (Table 1).

Table 1. List of Indices Used to Derive Consensus Natural Scales by PCA

Property	*Entries from the AAindex related to the amino acid property*					
α-Helicity	BURA740101	CHOP780201	GEIM800102	GEIM800104	ISOY800101	LEVM780101
	LEVM780104	MAXF760101	NAGK730101	PALJ810101	PALJ810102	PALJ810107
	PALJ810108	PALJ810109	PONP800104	PRAM900102	QIAN880101	QIAN880102
	QIAN880103	QIAN880104	QIAN880105	QIAN880106	QIAN880107	QIAN880110
	QIAN880111	QIAN880112	QIAN880113	ROBB760101	TANS770101	
β-Sheet propensity	CHAM830102	CHOP780202	CHOP780208	CHOP780209	CRAJ730102	KANM800102
	KANM800104	LEVM780102	LEVM780105	PALJ810103	PALJ810104	PALJ810110
	PALJ810111	PALJ810112	PONP800105	PRAM900103	QIAN880114	QIAN880115
	QIAN880116	QIAN880117	QIAN880118	QIAN880119	QIAN880120	QIAN880121
	QIAN880122	QIAN880123	QIAN880124	QIAN880125	QIAN880126	
Size	FASG760101	BIGC670101	BULH740102	COHE430101	FAUJ880103	GOLD730102
	GRAR740103	KRIW790103	TSAJ990101	TSAJ990102	HARY940101	PONJ960101
	FAUJ880104	DAWD720101	ZIMJ680102			
Hydrophobicity	DESM900102	NAKH900109	NAKH900110	NAKH900111	NAKH900112	MONM990101
	MONM990201	CEDJ970103	PUNT030101	PUNT030102	ARGP820101	ARGP820102
	ARGP820103	DESM900101	EISD840101	NAKH920101	NAKH920102	NAKH920103
	NAKH920104	NAKH920105	NAKH920106	NAKH920107	NAKH920108	VHEG790101
	WIMW960101	MITS020101	JURD980101	CIDH920101	CIDH920102	CIDH920103
	CIDH920104	CIDH920105	EISD860102	FAUJ830101	GOLD730101	HOPT810101
	JOND750101	KYTJ820101	LEVM760101	MANP780101	PONP800101	PONP800102
	PONP800103	PONP800104	PONP800105	PONP800106	PRAM900101	ROSM880101
	ROSM880102	ROSM880103	SWER830101	ZIMJ680101	NADH010101	NADH010102
	NADH010103	NADH010104	NADH010105	NADH010106	NADH010107	PONP930101
	WILM950101	WILM950102	WILM950103	WILM950104	KUHL950101	WOLR790101
	KIDA850101	COWR900101	BLAS910101	CASG920101	ENGD860101	FASG890101
	BIGC670101	BULH740101	BULH740102	CHAM820101	CHAM820102	GOLD730102
	KANM800101	KANM800102	KANM800103	KANM800104	LAWE840101	MEIH800101
	MEIH800102	MEIH800103	NAKH900101	NAKH900102	NAKH900103	NAKH900104
	NAKH900105	NAKH900106	NAKH900107	NAKH900108	NAKH900113	PARJ860101
	PLIV810101	PONP800107	PONP800108	RACS770101	RACS770102	RACS770103
	ROSG850101	ROSG850102	VENT840101	BASU050101	BASU050102	BASU050103
	JACR890101	CORJ870101	CORJ870102	CORJ870103	CORJ870104	CORJ870105
	CORJ870106	CORJ870107	CORJ870108			
Composition	FUKS010110	DAYM780101	JUNJ780101	JUKT750101	CEDJ970101	JOND920101
	CEDJ970102	NAKH900101	CEDJ970104	FUKS010112	KUMS000101	KUMS000102
	FUKS010111	CEDJ970105	NAKH920101	NAKH920106	NAKH920102	CEDJ970103
	NAKH900109	NAKH920103	NAKH920104	NAKH900102	NAKH920107	

2.9. Interpretability criterion

The following criterion was used to compare interpretability of different derivative indices: the squared correlations of the index in question to all 509 AA indices in the dataset were computed and sorted in decreasing order. The third highest value was used as a robust measure of interpretability (RMI). Choosing the second highest correlation, or a lower by rank value, or using an average of several of the highest correlations, gave similar results.

2.10. Distance matrices and AA scales conversions

The inner product of two amino acid vectors in index space was used as a similarity measure between the amino acids.

To recalculate a substitution matrix from a set of amino acid indices, I used the following equation:

$$M = AB(AB)^T = AB^2A^T \tag{4}$$

where M is the substitution matrix, A is a matrix with normalized (unit length) indices as columns, and B is a diagonal matrix with the weights (relative importance) of each index. If we set $B^2 = \Lambda$, the above Equation 4 becomes equivalent to the spectral (eigen) decomposition of a symmetric matrix:

$$M = A\Lambda A^T \tag{5}$$

where M is the substitution matrix, A is the matrix of eigenvectors, and Λ is a diagonal matrix with the eigenvalues. The above equations provide a unique correspondence between substitution matrices and sets of weighted amino acid scales (see Appendix).

The results from pairwise comparisons of protein sequences are invariant upon addition of a constant to the distance matrix. Thus, eigenvectors produced upon matrix decomposition should also be independent of the sum of all matrix elements. In addition, negative eigenvalues are meaningless in the context of distance matrices, since they penalize the conservation of identical residues. In order to make sure that the distance matrices such as BLOSUM62 used in this study are positive semidefinite, they were pre-processed by projecting all eigenvectors on the plane perpendicular to the ''all-ones'' vector before eigen decomposition.

To calculate the weight b of an existing scale a from a given substitution matrix M, the *Rayleigh quotient* was used, which provides an approximation to an eigenvalue of a matrix from an approximate eigenvector. Thus, the weight of an individual scale $\vec{a}_i$, is given by

$$b_i = \sqrt{\frac{a_i^T \cdot M \cdot a_i}{a_i^T \cdot a_i}} \tag{6}$$

2.11. Auto and cross-correlation of protein sequence features

The auto-cross correlation method of representing protein sequences (Cornette et al., 1987; Wold et al., 1993) has proven useful for comparing protein sequences of varying length. The method consists of (i) translating each residue from the primary sequence into one to several amino acid descriptors, and (ii) calculating the correlations between the sequence and a lagged (shifted) copy of it, as follows:

$$ACC_{j,k,l} = \sum_{i}^{n-l} \frac{s_{j,i} \times s_{k,i+l}}{n-l}, \tag{7}$$

where i is the amino acid position in the protein, n is the number of residues in the sequence, l is the lag between the two correlated residues, j and k are the index identifiers, and $s_{m,n}$ is the value from the m^{th} scale corresponding to the n^{th} residue in the protein. For auto covariances $j = k$, while for cross covariances $j \neq k$. The *ACC* vector is used in subsequent calculations in place of the original protein sequence.

2.12. Partial least-squares discriminant analysis

Partial least-squares (PLS) is a multivariate statistical regression method capable of dealing with a large number of variables in data sets suffering from noise and missing values (Wold et al., 1984). The PLS

algorithm relates two data matrices X and Y by maximizing the *covariance* between the response variable Y and a linear combination of the original variables $t = Xw$, where t is the score vector and w is a weight vector (Burnham et al., 1999). PLS–discriminant analysis (PLS-DA) is used as a method to discriminate between classes by using *Boolean* variables in Y to describe class assignment.

2.13. Software

PCA and PLS-DA were performed with SIMCA-P 11.0 (Umetrics AB, Umeå, Sweden). The R computational environment (R Development Core Team, 2006) was used for the rest of the statistical analysis. The R package *seqinr* was used to extract AAindex database information.

3. RESULTS

3.1. Consensus natural scales

To facilitate interpretation of the amino acid indices, a set of *consensus indices* with clear and unambiguous meanings was derived. Six ''natural'' features of amino acids were considered: *hydrophobicity*, *α-helix propensity*, *β-sheet propensity, bulkiness, charge*, and *composition* (frequency of occurrence in protein sequences). The consensus scales were derived by PCA of selected subsets of scales from the AAindex, according to the procedure described in Methods (Table 1). For each separate PCA, one to a few principal components were determined to be statistically significant, with the first one explaining most of the variation in the set (>50%), thereby demonstrating that the selection criterion was adequate and that indeed closely related scales were combined to produce the ''natural'' scales. The potential charge of acidic and basic side chains, which is important for electrostatic interactions of proteins and is an intuitive biochemical descriptor, was represented by a single scale, pI (ZIMJ680104) (Zimmerman et al., 1968). The complete set of indices is presented in Figure 1A.

Upon comparison between the natural indices (Fig. 1B), it became evident that some of the features used to describe amino acids are not mutually independent. For example, *hydrophobicity* and *β-sheet propensity*

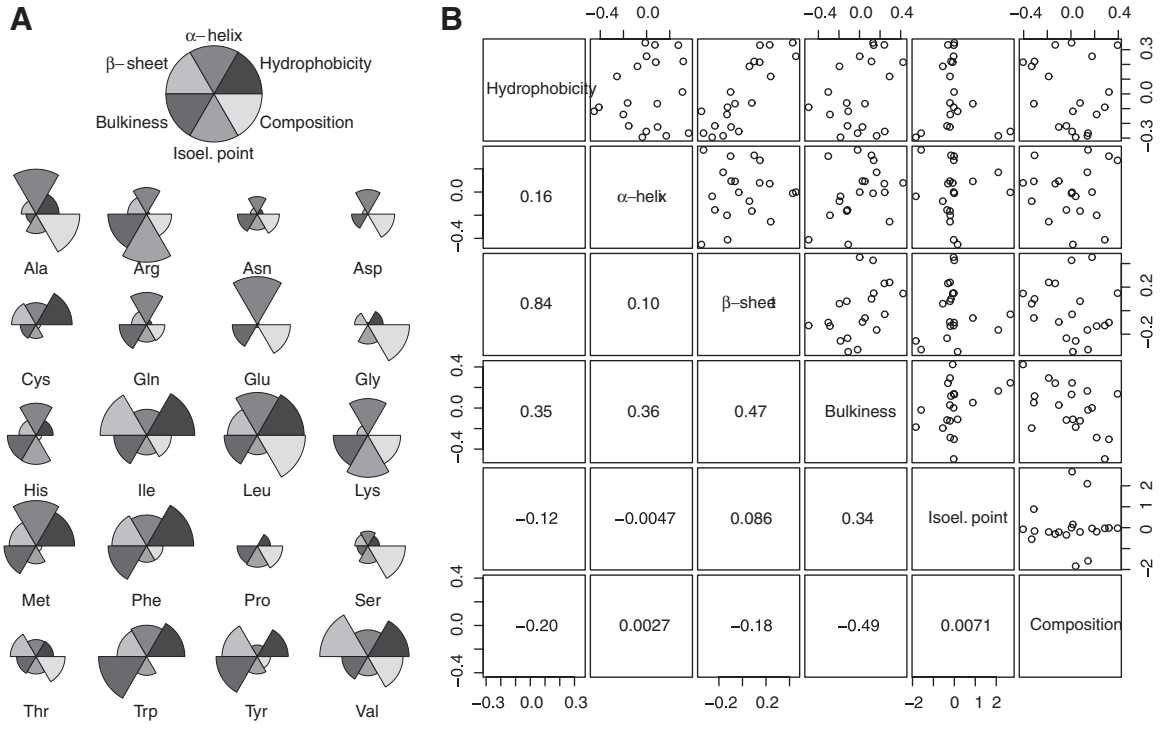


FIG. 1. (**A**) Consensus natural properties of the 20 amino acids. Radii correspond linearly to the magnitude of the properties, with the most negative values represented by zero and the most positive by unit radius. (**B**) Pairwise correlation of the natural indices (lower triangle) and scatter bi-plots (upper triangle).

are highly correlated ($r = 0.84$), consistent with the frequent occurrence of hydrophobic amino acid residues in β-sheets. The measure of side chain size or *bulkiness* is correlated with all other indices, making it difficult to separate its effect on protein structure from that of the other characteristics (Fig. 1B).

3.2. Natural versus derivative scales

Five sets of scales derived in previous studies (Kidera et al., 1985a,b; Sandberg et al., 1998; Atchley et al., 2005; Opiyo and Moriyama, 2007; Venkatarajan and Braun, 2001) were compared to the *natural indices*. Pearson correlations were used as a measure for similarity between the natural indices, on one hand, and the five sets of derivative scales, on the other (Fig. 2). Most of the latter sets had at least one scale with a good match to *hydrophobicity*. Further, the scales derived by Kidera et al. (1985a) correlated well with α-helicity (factor 1, $r = -0.93$), bulkiness (factor 2, $r = 0.95$), β-sheet propensity (factor 3, $r = 0.72$) and composition (factor 6, $r = -0.70$), but notably none of these scales represented pI. Of the scales derived by Sandberg et al. (1998) (the first three of which are a linear transformation of the z-scales of Hellberg et al. [1986]), scale 2 had a more pronounced match to bulkiness ($r = 0.76$), while the rest appeared as a combination of several natural indices and α-helicity was not represented at all. Similar was the situation with the scales derived by Venkatarajan and Braun (2001), which provided a better representation of structural features but lacked a good match for charge/pI. Three of the scales derived by Atchley et al. (2005) matched well hydrophobicity ($r = -0.90$), α-helicity ($r = -0.96$), and composition ($r = 0.92$); however, the remaining two scales did not correlate with any of the natural indices, and *bulkiness* was not represented well. Finally, the set generated by Opiyo and Moriyama (2007) was the only one lacking a clear match for *hydrophobicity* and modeled poorly compositional statistics. Notably, none of the examined derivative sets represented all of the consensus natural scales.

3.3. Principal components analysis of the complete AAindex

Initially, the dataset of 509 AA indices was analyzed by PCA (Fig. 3). This method is robust, tolerates multicollinearity, and produces a unique solution. A common problem with PCA and related techniques is determining the number of significant components to be retained. The software SIMCA-P identified the first three components as significant via cross-validation. In contrast, according to the rule for retaining factors with corresponding eigenvalues greater than one, the first five components should be kept. The fraction explained variation suggested not more than eight factors to account for 83% of the variation. Finally, comparing the eigenvalues to the corresponding obtained by similar analysis of an equally sized random set, as done previously by Opiyo and Moriyama (2007), one would consider four components as significant (more likely than chance).

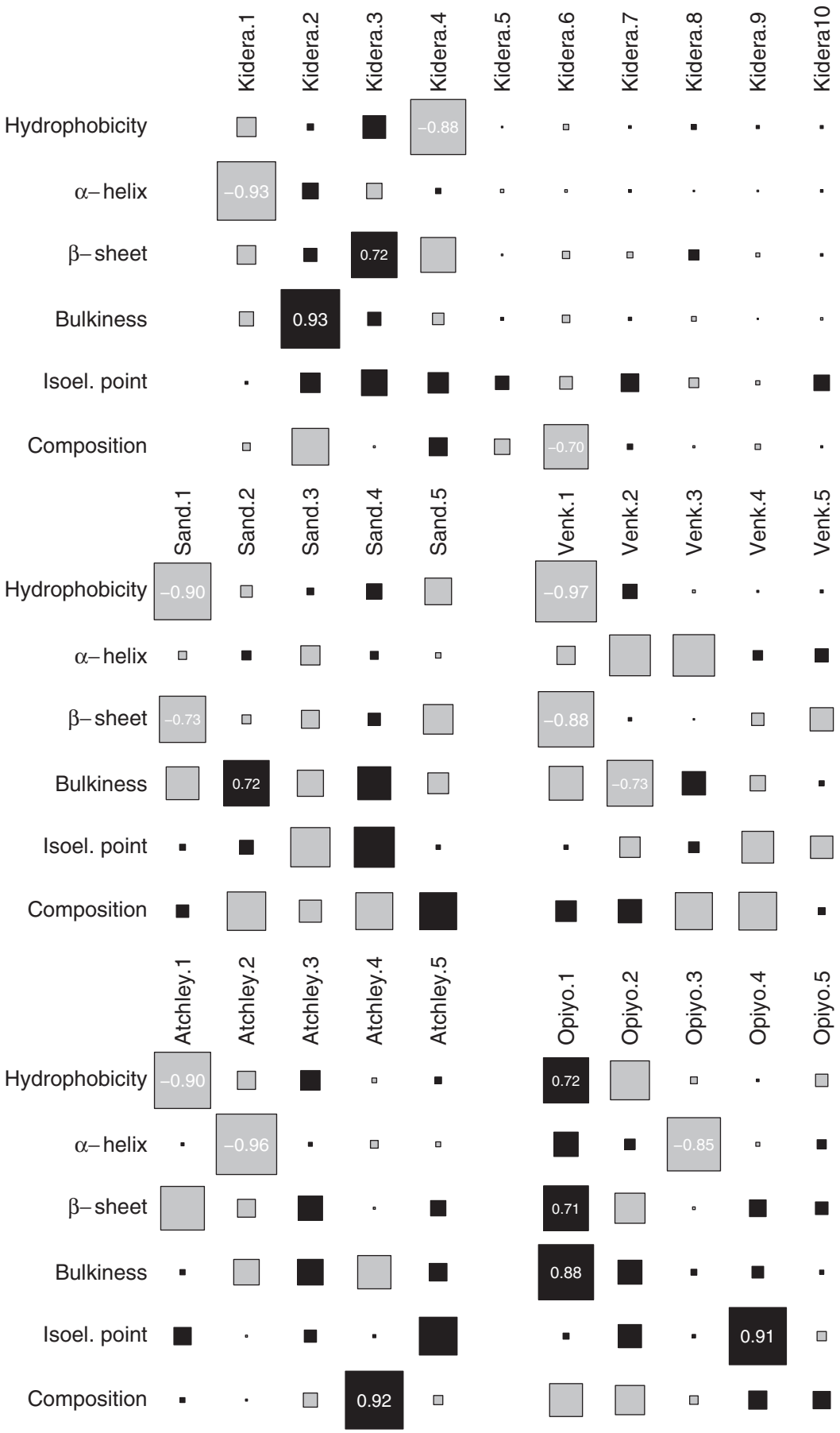
3.4. Minimum factors

To obtain a better idea of the dimensionality of the dataset, I employed two approaches. First, I searched for combinations of AA indices which are orthogonal to each other, by iteratively selecting those with mutual squared correlation below a threshold. I found 73 sets of 10 scales with as little as $r^2 = 0.04$ maximum correlation between any two of them, suggesting that a similar number of dimensions of AA space was needed to accommodate them all. Therefore, any dimensionality reduction approach ending up with a lower number of dimensions would be discarding a significant part of the variation represented by these ''naturally orthogonal'' scales.

Further, I selected a subset of the entire original AAindex, in which the duplicated or closely related scales were removed by an iterative procedure until 50 scales with no more than 50% correlation between them were left (Table 2). This dataset of reduced redundancy was also subjected to PCA. During the analysis, no less than 12 factors were needed in order to explain 80% of the variation in the subset. Therefore, preserving the diversity of amino acid features present in the AAindex reqired a higher number of dimensions than has been considered previously.

3.5. Maximum factors

In contrast to common dimensionality reduction approaches, where one aims to minimize the number of factors, I considered the maximum possible number of scales that could be potentially useful in distinguishing



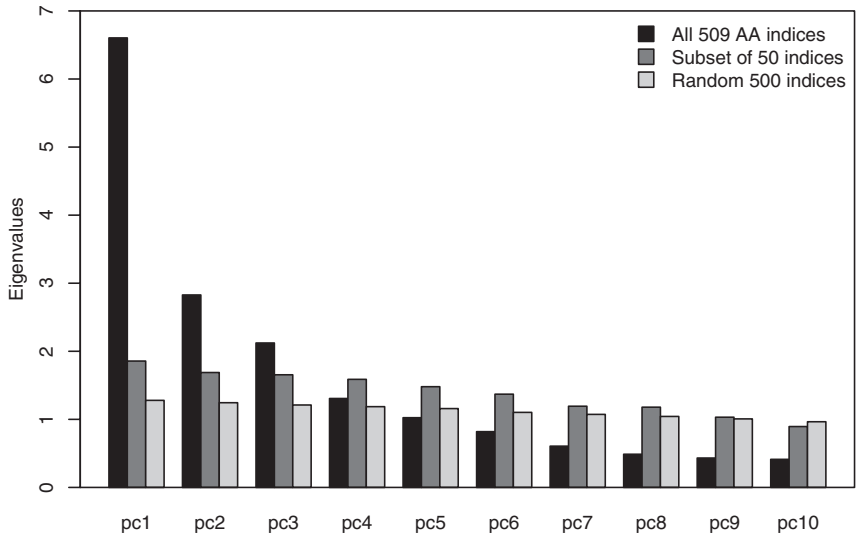


FIG. 3. The first 10 components of PCA analysis of 509 complete AA indices (black), a subset of 50 non-redundant indices (dark gray), and 500 randomly generated indices (light gray).

amino acids based on their features. *Amino acid space* has 20 dimensions ($\mathbb{R}^{20}$), equal to the number of amino acids. During the pre-processing, scales are centered around zero, equivalent to projecting them onto a 19-dimensional hyperplane perpendicular to the ''all-ones'' vector. Scaling the indices to a unit length further positions them along a hypersphere; thus, 18 of the dimensions are independent, while the 19^{th} is uniquely defined by the remaining 18.

As was shown in the previous section, the number of unique uncorrelated amino acid indices approaches the maximum number of possible orthogonal scales. In factor analysis, it is known that underfactoring usually prevents the detection of structure while over-factoring is not a problem. This provided additional motivation to analyse amino acid space without initial reduction of the number of dimensions.

3.6. VARIMAX analysis

Axes rotation is commonly performed following a solution for a starting loading matrix of the eigenvectors in factor analysis and related methods. Several criteria for optimal factor rotation exist, of which the most popular is VARIMAX (Kaiser, 1958). The VARIMAX criterion maximizes the variance of factor loadings. Similarly to PCA, it has a unique solution and is numerically stable. A two-dimensional example of applying PCA and VARIMAX to the same set of random vectors is shown on Figure 4.

In order to explore the full dimensionality of amino acid space, derivation of a starting matrix was omitted. Using built-in functions from the R statistical package (R Development Core Team, 2006), a solution was found for VARIMAX rotation of 20 arbitrary orthogonal axes, optimally describing the entire AAindex dataset. In the resulting rotation matrix (Table 3), 19 vectors were centered around zero, while the 20^{th} had equal values along all axes due to the pre-centering of the data.

3.7. Interpretation of the PCA and VARIMAX derived scales

The scales calculated by PCA or VARIMAX were in turn compared to the natural consensus indices. As shown (Fig. 5A), the first PCA scale correlated well with *hydrophobicity* ($r = 0.95$), the second was inversely correlated to *α-helicity* ($r = -0.75$), and the third was correlated to *composition* ($r = 0.85$). None of the other

FIG. 2. *Hinton* diagrams comparing five sets of scales to the consensus natural indices. Positive correlations are displayed by black squares, negative ones by gray. The side of the squares is proportional to the magnitude of the correlation. Scale sets are indicated by the (abbreviated) name of the first author of the corresponding publication: Kidera (Kidera et al., 1985), Sand (Sandberg et al., 1998), Venk (Venkatarajan and Braun, 2001), Atchley (Atchley et al., 2005), Opiyo (Opiyo and Moriyama, 2007).

Table 2. Subset of 50 AA Indices with Reduced Redundancy

AAindex	*Description (see www.genome.jp/aaindex for complete citations)*
BUNA790102	alpha-CH chemical shifts (Bundi-Wuthrich, 1979)
BUNA790103	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)
CHAM830102	Residuals from the best correlation of the Chou-Fasman parameter of β-sheet (Charton-Charton, 1983)
CHAM830103	The number of atoms in the side chain labelled $1+1$ (Charton-Charton, 1983)
CHAM830104	The number of atoms in the side chain labelled $2+1$ (Charton-Charton, 1983)
CHOP780215	Frequency of the 4th residue in turn (Chou-Fasman, 1978b)
FINA910103	Helix termination parameter at position j-2,j-1,j (Finkelstein et al., 1991)
ISOY800107	Normalized relative frequency of double bend (Isogai et al., 1980)
JOND750102	pK (-COOH) (Jones, 1975)
JOND920102	Relative mutability (Jones et al., 1992)
KARP850103	Flexibility parameter for two rigid neighbors (Karplus-Schulz, 1985)
KHAG800101	The Kerr-constant increments (Khanarian-Moore, 1980)
MAXF760103	Normalized frequency of zeta R (Maxfield-Scheraga, 1976)
OOBM770105	Short and medium range non-bonded energy per residue (Oobatake-Ooi, 1977)
OOBM850103	Optimized transfer energy parameter (Oobatake et al., 1985)
PALJ810107	Normalized frequency of alpha-helix in all-alpha class (Palau et al., 1981)
PRAM820101	Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)
PRAM820102	Slope in regression analysis x 1.0E1 (Prabhakaran-Ponnuswamy, 1982)
QIAN880101	Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988)
QIAN880102	Weights for alpha-helix at the window position of -5 (Qian-Sejnowski, 1988)
QIAN880117	Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)
QIAN880128	Weights for coil at the window position of -5 (Qian-Sejnowski, 1988)
QIAN880129	Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
QIAN880138	Weights for coil at the window position of 5 (Qian-Sejnowski, 1988)
RACS820103	Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga, 1982)
RACS820106	Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga, 1982)
RACS820107	Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga, 1982)
RICJ880101	Relative preference value at N'' (Richardson-Richardson, 1988)
RICJ880105	Relative preference value at N2 (Richardson-Richardson, 1988)
RICJ880106	Relative preference value at N3 (Richardson-Richardson, 1988)
RICJ880114	Relative preference value at C1 (Richardson-Richardson, 1988)
RICJ880116	Relative preference value at C' (Richardson-Richardson, 1988)
RICJ880117	Relative preference value at C'' (Richardson-Richardson, 1988)
ROBB760107	Information measure for extended without H-bond (Robson-Suzuki, 1976)
TANS770102	Normalized frequency of isolated helix (Tanaka-Scheraga, 1977)
TANS770107	Normalized frequency of left-handed helix (Tanaka-Scheraga, 1977)
TANS770108	Normalized frequency of zeta R (Tanaka-Scheraga, 1977)
VASM830101	Relative population of conformational state A (Vasquez et al., 1983)
VELV850101	Electron-ion interaction potential (Veljkovic et al., 1985)
WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)
AURR980105	Normalized positional residue frequency at helix termini Nc (Aurora-Rose, 1998)
AURR980118	Normalized positional residue frequency at helix termini C″ (Aurora-Rose, 1998)
AURR980120	Normalized positional residue frequency at helix termini C4′ (Aurora-Rose, 1998)
MITS020101	Amphiphilicity index (Mitaku et al., 2002)
COSI940101	Electron-ion interaction potential values (Cosic, 1994)
WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al., 1995)
WILM950103	Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O (Wilce et al., 1995)
WILM950104	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H2O (Wilce et al., 1995)
GEOR030103	Linker propensity from 2-linker dataset (George-Heringa, 2003)
GEOR030107	Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa, 2003)

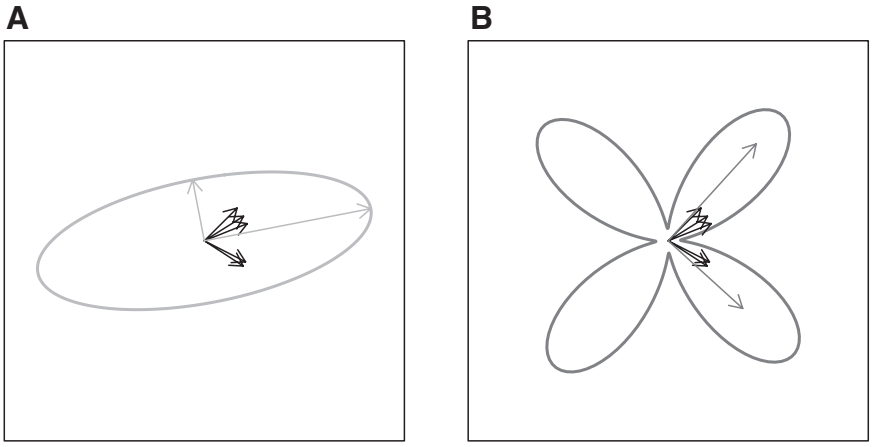


FIG. 4. Example of applying PCA or VARIMAX to the same set of vectors (black). Gray arrows represent the two principal components (**A**); dark gray arrows show the direction of the axes after VARIMAX rotation (**B**). The contours show the result of transforming a unit circle with the correlation matrix of the five vectors in the dataset (A) or the relative value of the VARIMAX criterion as computed by Equation 3 (B).

scales matched the natural indices; however, two indices were partially explained by more than one scale, *isoelectric point* (PCA scales 5 and 7) and *bulkiness* (scales 1–3).

In contrast, the VARIMAX scales, designed to provide good interpretability, showed excellent correlation with the natural indices (Fig. 5B). VARIMAX index 1 matched *hydrophobicity* ($r = 0.98$) and VARIMAX index 2 matched *α-helicity* ($r = 0.98$). VARIMAX index 3 matched *bulkiness* ($r = 0.88$), index 4 matched *composition* ($r = 0.94$), and index 7 matched *isoelectric point* ($r = 0.87$). These correlations, which were substantially higher those for most of the previous scales, allowed for an unambiguous interpretation of some of the VARIMAX scales.

Additional AAindex collection entries with highest correlation to the VARIMAX scales (Table 4) provided information about those VARIMAX scales which were not explained equally well by the natural consensus indices. Index 5 was found to correlate best with the frequency of amino acid occurrence in interdomain linkers (George and Heringa, 2002), and could thus be described as a measure of *linker propensity*. Index 6 was close to several measures of *C-terminal helix capping* structural preferences (Doig and Baldwin, 1995; Aurora and Rose, 1998). Index 8 could be explained as a supplementary *β-sheet* measure, noting that the natural *β-sheet propensity* is already explained partly by the *hydrophobicity* scale, or index 1.

3.8. Interpretability of amino acid scales compared

The VARIMAX scales demonstrated a feasible solution for a set of independent scales closely related to the natural indices. However, the natural indices are the consensus of several similar indices selected from the entire AAindex, and are potentially biased by the subjective selection of key words. Therefore, in order to address objectively the ease of interpretation of any derivative amino acid scale, I defined RMI based on comparing the scale in question with the dataset of published AA indices. The RMI shows to what extent the meaning any scale can be revealed by its correlation with any of the 509 complete indices from the AAindex. It can be applied without any prior assumptions about the nature of the scale in question.

I calculated the RMI of both the VARIMAX derived and PCA derived scales, and compared it to that of a randomly generated set of scales and to the five published sets of scales addressed earlier (Fig. 6). As expected, individual scales with high correlation to consensus natural indices were also characterized by high RMI. In addition, two more of the VARIMAX derived axes had high RMI scores ($RMI_{A5} = 0.66$; $RMI_{A6} = 0.64$), consistent with the identification of those as measures of *linker* and *C-cap* propensities above. In comparison, the PCA derived scales had in general lower RMI, and only five of them were above the threshold determined by a simulation of the RMI of 10 000 random vectors at the 99% confidence level. Of the previously published sets analyzed here, only the one by Kidera et al. (1985a) contained six scales with better interpretability than random, however with lower RMI values than corresponding VARIMAX

Table 3. Orthonormal Axes Obtained after varimax Rotation of the Entire Set of 509 Complete AA Indices

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
Ala	0.57	3.37	−3.66	2.34	−1.07	−0.40	1.23	−2.32	−2.01	1.31	−1.14	0.19	1.66	4.39	0.18	−2.60	1.49	0.46	−4.22
Arg	−2.80	0.31	2.84	0.25	0.20	−0.37	3.81	0.98	2.43	−0.99	−4.90	2.09	−3.08	0.82	1.32	0.69	−2.62	−1.49	−2.57
Asn	−2.02	−1.92	0.04	−0.65	1.61	2.08	0.40	−2.47	−0.07	7.02	1.32	−2.44	0.37	−0.89	3.13	0.79	−1.54	−1.71	−0.25
Asp	−2.46	−0.66	−0.57	0.14	0.75	0.24	−5.15	−1.17	0.73	1.50	1.51	5.61	−3.85	1.28	−1.98	0.05	0.90	1.38	−0.03
Cys	2.66	−1.52	−3.29	−3.77	2.96	−2.23	0.44	−3.49	2.22	−3.78	1.98	−0.43	−1.03	0.93	1.43	1.45	−1.15	−1.64	−1.05
Gln	−2.54	1.82	−0.82	−1.85	0.09	−0.60	0.25	2.11	−1.92	−1.67	0.70	−0.27	−0.99	−1.56	6.22	−0.18	2.72	4.35	0.92
Glu	−3.08	3.45	0.05	0.62	−0.49	−0.00	−5.66	−0.11	1.49	−2.26	−1.62	−3.97	2.30	−0.06	−0.35	1.51	−2.29	−1.47	0.15
Gly	0.15	−3.49	−2.97	2.06	0.70	7.47	0.41	1.62	−0.47	−2.90	−0.98	−0.62	−0.11	0.15	−0.53	0.35	0.30	0.32	0.05
His	−0.39	1.00	−0.63	−3.49	0.05	0.41	1.61	−0.60	3.55	1.52	−2.28	−3.12	−1.45	−0.77	−4.18	−2.91	3.37	1.87	2.17
Ile	3.10	0.37	0.26	1.04	−0.05	−1.18	−0.21	3.45	0.86	1.98	0.89	−1.67	−1.02	−1.21	−1.78	5.71	1.54	2.11	−4.18
Leu	2.72	1.88	1.92	5.33	0.08	0.09	0.27	−4.06	0.43	−1.20	0.67	−0.29	−2.47	−4.79	0.80	−1.43	0.63	−0.24	1.01
Lys	−3.89	1.47	1.95	1.17	0.53	0.10	4.01	−0.01	−0.26	−1.66	5.86	−0.06	1.38	1.78	−2.71	1.62	0.96	−1.09	1.36
Met	1.89	3.88	−1.57	−3.58	−2.55	2.07	0.84	1.85	−2.05	0.78	1.53	2.44	−0.26	−3.09	−1.39	−1.02	−4.32	−1.34	0.09
Phe	3.12	0.68	2.40	−0.35	−0.88	1.62	−0.15	−0.41	4.20	0.73	−0.56	3.54	5.25	1.73	2.14	1.10	0.68	1.46	2.33
Pro	−0.58	−4.33	−0.02	−0.21	−8.31	−1.82	−0.12	−1.18	0.00	−0.66	0.64	−0.92	−0.37	0.17	0.36	0.08	0.16	−0.34	0.04
Ser	−1.10	−2.05	−2.19	1.36	1.78	−3.36	1.39	−1.21	−2.83	0.39	−2.92	1.27	2.86	−1.88	−2.42	1.75	−2.77	3.36	2.67
Thr	−0.65	−1.60	−1.39	0.63	1.35	−2.45	−0.65	3.43	0.34	0.24	−0.53	1.91	2.66	−3.07	0.20	−2.20	3.73	−5.46	−0.73
Trp	1.89	−0.09	4.21	−2.77	0.72	0.86	−1.07	−1.66	−5.87	−0.66	−2.49	−0.30	−0.50	1.64	−0.72	1.75	2.73	−2.20	0.90
Tyr	0.79	−2.62	4.11	−0.63	1.89	−0.53	−1.30	1.31	−0.56	−0.95	1.91	−1.26	1.57	0.20	−0.76	−5.19	−2.56	2.87	−3.43
Val	2.64	0.03	−0.67	2.34	0.64	−2.01	−0.33	3.93	−0.21	1.27	0.43	−1.71	−2.93	4.22	1.06	−1.31	−1.97	−1.21	4.77

The axes (column vectors) have been scaled to length of 10 units. The axis corresponding to the "all-ones" vectors has been removed; the remaining 19 are shown.

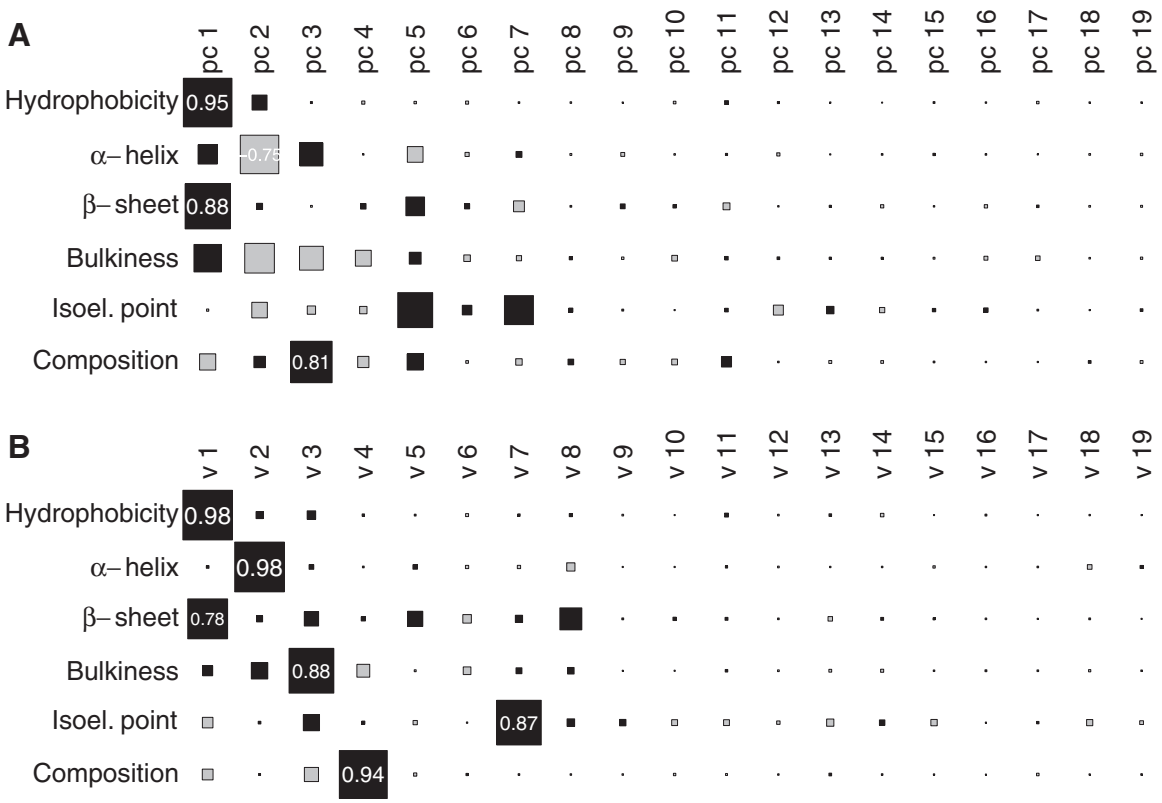


FIG. 5. Hinton diagrams comparing PCA (**A**) and VARIMAX (**B**) derived scales in terms of correlation to the consensus natural indices.

scales in all cases except for α-helicity. The other four sets had three or four interpretable scales ($RMI > 0.41$), and all had at least one scale that correlated so little with any of the AA indices as to be difficult to interpret (Fig. 6).

3.9. Scales derived from a distance matrix

The existence of a unique correspondence between distance matrices and amino acid indices presented above is an interesting observation in itself. Equation 4 was used to compute the indices corresponding to the widely used BLOSUM62 matrix (Henikoff and Henikoff, 1992), which has become a *de facto* standard for protein alignments (Eddy, 2004). The resulting indices (Table 5) describe amino acid properties important for retaining selection pressure on individual residues in the proteins from the blocks database which had served during the computation of BLOSUM62.

To find out how these indices correspond to the *consensus natural scales*, I examined the correlation between the two sets (Fig. 7). The first BLOSUM62-derived index showed strong correlation with *hydrophobicity* ($r^2 = 0.94$). The remaining indices were less than 50% correlated to the various natural scales, with higher correlations observed to pI (0.40), bulkiness (0.30), and α-helicity (0.24). The top three correlated indices from the AAindex collection provide some further insight in the interpretation of the BLOSUM62-derived indices (Table 6).

Using Equation 6, I obtained the weights to be assigned to the VARIMAX-derived axes in order to use them to reproduce the BLOSUM62 matrix (Table 7 and Fig. 8). The difference matrix reconstituted according to Equation 4 was highly correlated with the original BLOSUM62 ($r = 0.90$); this correlation was higher than the one between BLOSUM62 and another widely used matrix considered to be equivalent in the amount of information it provides, PAM160 ($r = 0.88$). The root mean squared deviation (RMSD) between the reconstituted BLOSUM and BLOSUM62 was 1.28. Using only seven of the VARIMAX scales yielded an approximation to BLOSUM62 with correlation coefficient $r = 0.80$ and RMSD = 1.58.

Table 4. AA Indices with Highest Correlation to the varimax-Derived Scales

V	*AAindex*	*r*	*Description*
1	NADH010102	0.969	Hydropathy scale based on self-information values in the two-state model (9% accessibility) (Naderi-Manesh et al., 2001)
1	BIOV880101	0.968	Information value for accessibility; average fraction 35% (Biou et al., 1988)
1	ROSG850102	0.965	Mean fractional area loss (Rose et al., 1985)
2	PALJ810102	0.982	Normalized frequency of alpha-helix from CF (Palau et al., 1981)
2	KANM800101	0.979	Average relative probability of helix (Kanehisa-Tsong, 1980)
2	ISOY800101	0.978	Normalized relative frequency of alpha-helix (Isogai et al., 1980)
3	PONJ960101	0.893	Average volumes of residues (Pontius et al., 1996)
3	TSAJ990102	0.888	Volumes not including the crystallographic waters using the ProtOr (Tsai et al., 1999)
3	FAUJ880103	0.886	Normalized van der Waals volume (Fauchere et al., 1988)
4	NAKH900101	0.954	AA composition of total proteins (Nakashima et al., 1990)
4	JOND920101	0.954	Relative frequency of occurrence (Jones et al., 1992)
4	CEDJ970102	0.953	Composition of amino acids in anchored proteins (percent) (Cedano et al., 1997)
5	BUNA790101	0.852	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
5	FINA910102	−0.851	Helix initiation parameter at position i,i + 1,i + 2 (Finkelstein et al., 1991)
5	AURR980119	−0.812	Normalized positional residue frequency at helix termini C″ (Aurora-Rose, 1998)
6	AURR980117	0.811	Normalized positional residue frequency at helix termini C′ (Aurora-Rose, 1998)
6	FAUJ880107	−0.803	N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988)
6	RACS820106	0.799	Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga, 1982)
7	KLEP840101	0.932	Net charge (Klein et al., 1984)
7	ZIMJ680104	0.875	Isoelectric point (Zimmerman et al., 1968)
7	FINA910103	0.806	Helix termination parameter at position j-2,j-1,j (Finkelstein et al., 1991)
8	QIAN880117	0.753	Weights for beta-sheet at the window position of −3 (Qian-Sejnowski, 1988)
8	QIAN880118	0.567	Weights for beta-sheet at the window position of −2 (Qian-Sejnowski, 1988)
8	PALJ810110	0.556	Normalized frequency of beta-sheet in all-beta class (Palau et al., 1981)
9	BUNA790103	0.646	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)
9	JOND750102	−0.628	pK (-COOH) (Jones, 1975)
9	FASG760105	−0.605	pK-C (Fasman, 1976)
10	MAXF760103	0.783	Normalized frequency of zeta R (Maxfield-Scheraga, 1976)
10	DAYM780201	0.618	Relative mutability (Dayhoff et al., 1978b)
10	WERD780102	0.607	Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga, 1978)

The weights calculated above demonstrated that the importance of the different amino acid features presented by the scales from an evolutionary perspective was different from what could be inferred by the fraction of explained variation of the AAindex dataset. Most important was axis 1 (weight 4.95), corresponding to hydrophobicity; however, second was axis 3 (weight 3.70) despite the fact that axis 2 explained a greater proportion of the variation in the AAindex. Several axes had very similar weights between 2.77 and 2.99 (axes 2, 4–7, 9), and the variation in general was less than that of the original BLOSUM62 eigenvalues. The scaled VARIMAX axes are shown in Table 7, and a comparison of their weights to BLOSUM62 eigenvalues is shown on Figure 8.

3.10. Performance of amino acid indices: classification of GPCRs

The importance of finely tuned scales for structure prediction has been demonstrated by Cornette et al. (1987), who used the performance in detecting helix amphipaticity as a criterion for choosing an optimal hydrophobicity scale. To make sure that the improved interpretability of the VARIMAX axes did not come at the expense of decreased performance, I compared them to earlier published scales in a real-life computational biology classification problem. Protein classification experiments were done using the ACC transform of protein sequences. In order focus on the performance of the scales rather than on the test data, I selected a well studied object, the classification of G-protein coupled receptors (GPCRs) which was recently addressed by Lapinsh et al. (2002), and compared various scales used previously to the VARIMAX scales derived here.

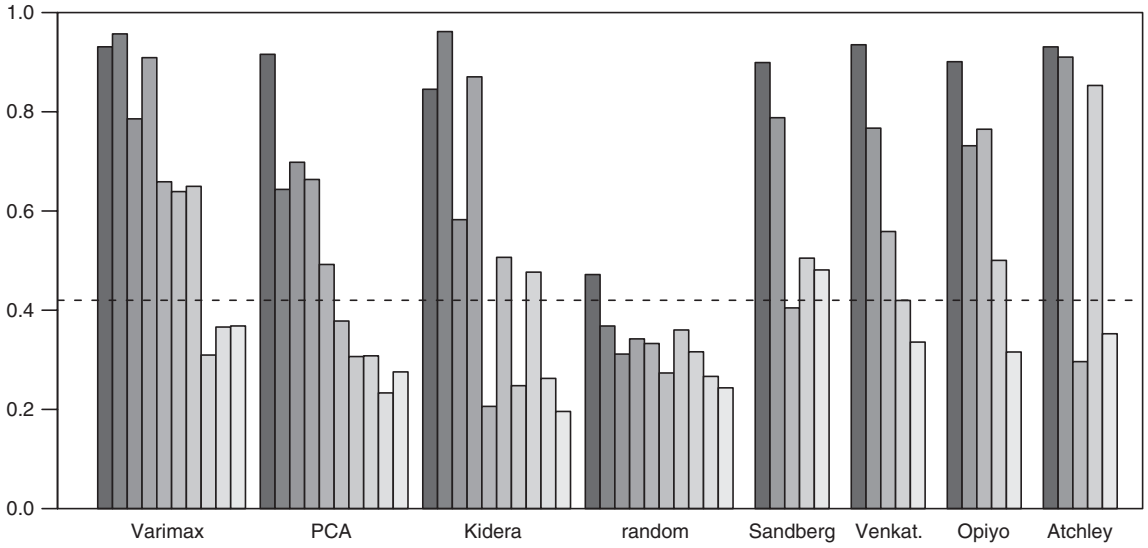


FIG. 6. Scale interpretability presented as the squared correlations to the third closest scale from the AAindex.

The GPCR sequences were translated into numerical vectors using the ACC transform separately with each one of the amino acid index sets. Five descriptor scales were used in each case; scales with a close match among the natural indices were selected from the set by Kidera et al. (1985a) (numbers 1, 2, 3, 4, 6) as well as from the VARIMAX scales (numbers 1, 2, 3, 4, 7). Forty lags were used in each case, similarly to previous studies using 40–80 lags (Lapinsh et al., 2002). This resulted in a total of 1000 variables and 14 sub-classes.

PLS-DA models were built of each dataset. Up to 40 principal components for each model were retained. The PLS-DA model performance was evaluated by comparing the predictive power (Q^2_{cum}) of each model obtained with different scales by cross validation (Fig. 9). Most of the sets of scales being compared demonstrated similar performance. Notably, the best predictive power ($Q^2_{cum} = 0.84$ for a 40-component model) was calculated with the BLOSUM62 set. The rest of the scale sets scored close in all models, with the exception of the Atchley scales leading to models with the lowest predictive power.

TABLE 5. THE FIRST 10 BLOSUM62-DERIVED INDICES

	B1	*B2*	*B3*	*B4*	*B5*	*B6*	*B7*	*B8*	*B9*	*B10*
Ala	0.077	−0.916	0.526	0.004	0.240	0.190	0.656	−0.047	1.357	0.333
Arg	1.014	0.189	−0.860	−0.609	1.277	0.195	0.661	0.175	−0.219	−0.520
Asn	1.511	0.215	−0.046	1.009	0.120	0.834	−0.033	−0.570	−1.200	−0.139
Asp	1.551	0.005	0.323	0.493	−0.991	0.010	−1.615	0.526	−0.150	−0.282
Cys	−1.084	−1.112	1.562	0.814	1.828	−1.048	−0.742	0.379	−0.121	−0.102
Gln	1.094	0.296	−0.871	−0.718	0.500	−0.080	−0.442	0.202	0.384	0.667
Glu	1.477	0.229	−0.670	−0.355	−0.284	−0.075	−1.014	0.363	0.769	0.298
Gly	0.849	0.174	1.726	0.093	−0.548	1.186	1.213	0.874	0.009	0.242
His	0.716	1.548	−0.802	1.547	0.350	−0.785	0.655	−0.076	−0.186	0.990
Ile	−1.462	−1.126	−0.761	0.382	−0.599	0.276	−0.132	0.198	−0.216	0.207
Leu	−1.406	−0.856	−0.879	−0.172	0.032	0.344	0.109	0.146	−0.436	−0.021
Lys	1.135	−0.039	−0.802	−0.849	0.819	0.097	0.213	0.129	0.176	−0.850
Met	−0.963	−0.585	−0.972	−0.528	0.236	0.365	0.062	0.208	−0.560	0.361
Phe	−1.619	1.007	−0.311	0.623	−0.549	0.290	−0.021	0.098	0.433	−1.288
Pro	0.883	−0.675	0.382	−0.869	−1.243	−2.023	0.845	−0.352	−0.421	−0.298
Ser	0.844	−0.448	0.423	0.317	0.200	0.541	0.009	−0.797	0.624	−0.129
Thr	0.188	−0.733	0.178	−0.012	0.022	0.378	−0.304	−1.958	0.149	0.063
Trp	−1.577	2.281	1.166	−1.610	0.122	0.239	−0.542	−0.398	−0.349	0.499
Tyr	−1.142	1.740	−0.582	0.747	−0.119	−0.475	0.241	−0.251	0.713	−0.251
Val	−1.127	−1.227	−0.633	0.064	−0.596	0.158	0.014	0.016	0.251	0.607

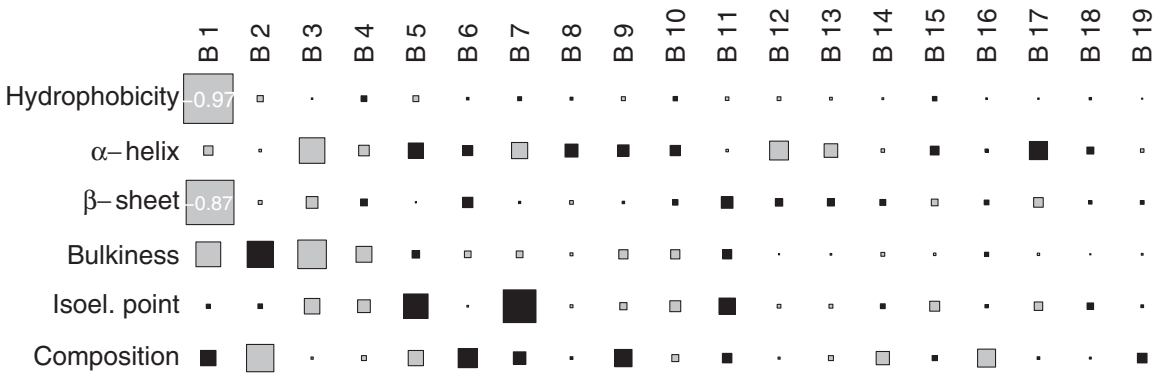


FIG. 7. Hinton diagram showing the correlation between 19 scales, derived by decomposition of the BLOSUM62 substitution matrix, and the consensus natural indices.

4. DISCUSSION

Representation of amino acid properties by numerical scales has been employed in the past with two major tasks in mind: (i) the supervised exploration of physico-chemical properties along the sequence of proteins as aids in determining features important for protein structure and function; and (ii) the unsupervised learning and prediction of protein functional or structural class belonging. The first task is enhanced by using numerical representations that can be interpreted as measurable physico-chemical properties of the individual amino acids. In contrast, the success of the second task is dependent on finding out what features have been subjected to direct evolutionary pressure, without the implied interpretability of those features on the level of individual residue properties.

As noted previously, PCA is a technique which captures the greatest variance in the data in orthogonal principal components. Thus, it relies on the idea that the dataset is normally distributed and that its variance reflects the actual importance of the amino acid features studied. However, since the AAindex is an artificial collection, it is more likely to be biased towards the investigators' interests rather than provide an optimal representation for the structure of *amino acid space*. This bias is also present in studies where a subjective selection of the indices or physico-chemical properties is used (Sandberg et al., 1998; Venkatarajan and Braun, 2001; Atchley et al., 2005; Opiyo and Moriyama, 2007). Thus, one less obvious problem with using PCA or multidimensional scaling for reduction of the data dimensionality is the assumption that a high fraction of explained variation in the retained dimensions is sufficient to justify the reduction. In fact, due to the strong bias in compiling the source data, it is very likely that less-well represented features are removed in favor of overrepresented ones. In support of this, it was demonstrated here that a reduced dataset with lower redundancy could not be represented sufficiently well by less than 12 independent principal components.

A direct conclusion from the above observation is that, for an exhaustive representation of AA space by derivative scales, it is important that the starting dataset used to derive those scales spans the entire vector space. In fact, a selected subset has been used during the derivation of several scales thus leading to underrepresentation of particular subspaces. For example, Kidera et al. (1985a) excluded scales deviating from normal distribution, thereby preventing charge-related features from entering their final solution. Sandberg et al. (1998) developed their scale for the study of short (unstructured) peptides, thus excluding structure-related features such as α-helical propensities.

The use of a different approach in the current study, namely the VARIMAX criterion instead of PCA, led to amino acid scales with dramatically increased ease of interpretation. They clearly correspond to well studied physical, chemical, and statistical measures—hydrophobicity, α-helix forming propensity, residue volume, frequency of occurrence, and net charge. Moreover, this was achieved with only a minor decrease in the amount of explained information present in the complete dataset as compared to the theoretically optimal method PCA (not shown). In addition to enhanced interpretability, the VARIMAX scales compared favourably to the other scales investigated in terms of performance in the task of Class A GPCR subfamily classification.

In addition to developing a new potentially useful set of amino acid scales using the AAindex database, the current study demonstrates a method to decompose the BLOSUM62 substitution matrix into scales analogous to single amino acid features. It should be noted that eigenvectors of distance matrices have been

TABLE 6. AA INDICES WITH HIGHEST CORRELATION TO THE BLOSUM62-DERIVED SCALES

BI	*AAindex*	*r*	*Description*
1	ZHOH040103	−0.962	Buriability (Zhou-Zhou, 2004)
	MIYS990104	0.962	Optimized relative partition energies—method C (Miyazawa-Jernigan, 1999)
	GRAR740102	0.958	Polarity (Grantham, 1974)
2	SNEP660103	0.774	Principal component III (Sneath, 1966)
	CHAM830105	0.749	The number of atoms in the side chain labeled 3 + 1 (Charton-Charton, 1983)
	MITS020101	0.743	Amphiphilicity index (Mitaku et al., 2002)
3	RICJ880112	−0.771	Relative preference value at C3 (Richardson-Richardson, 1988)
	AURR980114	−0.746	Normalized positional residue frequency at helix termini C2 (Aurora-Rose, 1998)
	GEOR030108	−0.740	Linker propensity from helical (annotated by DSSP) dataset (George-Heringa, 2003)
4	BUNA790103	0.686	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)
	RACS820107	0.578	Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga, 1982)
	RACS820102	−0.574	Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga, 1982)
5	FASG760104	−0.784	pK-N (Fasman, 1976)
	RICJ880108	0.695	Relative preference value at N5 (Richardson-Richardson, 1988)
	RICJ880113	0.672	Relative preference value at C2 (Richardson-Richardson, 1988)
6	ROBB760107	−0.757	Information measure for extended without H-bond (Robson-Suzuki, 1976)
	BUNA790101	0.708	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)
	FINA910102	−0.692	Helix initiation parameter at position i,i + 1,i + 2 (Finkelstein et al., 1991)
7	NADH010107	−0.731	Hydropathy scale based on self-information values in the two-state model (50% accessibility) (Naderi-Manesh et al., 2001)
	FAUJ880112	−0.673	Negative charge (Fauchere et al., 1988)
	RICJ880106	−0.666	Relative preference value at N3 (Richardson-Richardson, 1988)
8	RICJ880117	−0.544	Relative preference value at C″ (Richardson-Richardson, 1988)
	QIAN880115	−0.541	Weights for beta-sheet at the window position of −5 (Qian-Sejnowski, 1988)
	SNEP660104	−0.525	Principal component IV (Sneath, 1966)
9	KUMS000103	0.545	Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al., 2000)
	KUMS000104	0.481	Distribution of amino acid residues in the alpha-helices in mesophilic proteins (Kumar et al., 2000)
	RACS820106	−0.472	Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga, 1982)
10	KARP850103	−0.456	Flexibility parameter for two rigid neighbors (Karplus-Schulz, 1985)
	CRAJ730101	0.431	Normalized frequency of middle helix (Crawford et al., 1973)
	WERD780103	−0.421	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)

used previously to measure *hydrophobicity* (Cornette et al., 1987). The mathematically justified relation between substitution matrices and amino acid indices used here is different than that proposed by others (Kidera et al., 1985b), with the advantage of providing a simple and unambiguous method of conversion between the two types of measures.

Comparing, on one hand, the scales in the AAindex database and derivatives therein (such as the VARIMAX scales presented here), and on the other hand, the scales derived by spectral decomposition of a mutation matrix, brings forward two interesting observations. The first one is that the scales derived from BLOSUM62, with one notable exception, have very low interpretability. The exception is the first eigenvector of BLOSUM62, which correlates >94%, and thus is clearly related, to *hydrophobicity/hydrophilicity*. This is a direct demonstration of the well known importance of residue hydrophobicity for protein structure/function and in

Table 7. Weighted varimax Scales (First 10) with Weights Determined from Their Rayleigh Quotient for blosum62

	WV1	*WV2*	*WV3*	*WV4*	*WV5*	*WV6*	*WV7*	*WV8*	*WV9*	*WV10*
Ala	0.281	0.962	−1.352	0.674	−0.320	−0.117	0.349	−0.530	−0.558	0.334
Arg	−1.386	0.089	1.050	0.073	0.059	−0.107	1.085	0.224	0.672	−0.251
Asn	−0.999	−0.548	0.014	−0.186	0.482	0.605	0.113	−0.565	−0.020	1.785
Asp	−1.219	−0.187	−0.213	0.039	0.223	0.071	−1.466	−0.268	0.202	0.381
Cys	1.315	−0.432	−1.217	−1.082	0.886	−0.649	0.125	−0.797	0.616	−0.962
Gln	−1.256	0.518	−0.304	−0.531	0.026	−0.174	0.071	0.483	−0.532	−0.425
Glu	−1.526	0.984	0.018	0.178	−0.147	−0.000	−1.613	−0.025	0.413	−0.575
Gly	0.076	−0.993	−1.100	0.593	0.208	2.178	0.117	0.371	−0.131	−0.738
His	−0.195	0.284	−0.234	−1.002	0.016	0.121	0.458	−0.137	0.984	0.386
Ile	1.535	0.106	0.097	0.298	−0.013	−0.345	−0.061	0.789	0.238	0.503
Leu	1.344	0.535	0.711	1.531	0.025	0.027	0.077	−0.928	0.119	−0.306
Lys	−1.925	0.419	0.720	0.337	0.159	0.029	1.142	−0.002	−0.073	−0.422
Met	0.937	1.106	−0.579	−1.030	−0.763	0.604	0.238	0.423	−0.567	0.198
Phe	1.542	0.194	0.887	−0.100	−0.263	0.473	−0.043	−0.093	1.163	0.186
Pro	−0.289	−1.234	−0.006	−0.060	−2.482	−0.530	−0.033	−0.269	0.000	−0.168
Ser	−0.543	−0.585	−0.809	0.391	0.531	−0.981	0.396	−0.276	−0.783	0.098
Thr	−0.324	−0.455	−0.514	0.180	0.404	−0.714	−0.187	0.783	0.093	0.062
Trp	0.934	−0.025	1.556	−0.795	0.216	0.252	−0.304	−0.379	−1.626	−0.169
Tyr	0.389	−0.745	1.520	−0.180	0.566	−0.155	−0.369	0.299	−0.155	−0.241
Val	1.308	0.008	−0.246	0.672	0.190	−0.587	−0.093	0.898	−0.057	0.322

support of the relevance of Equation 4. Other examples exist where hydrophobicity was derived by purely statistical means from protein sequence/structure databases rather than by direct measurements of residue properties (Cornette et al., 1987). The remaining scales obtained from blosum62 are difficult to interpret in terms of relatedness to single measurable amino acid properties. This suggests that there may exist properties important for protein sequence conservation that have received little attention in studies targeting individual residue features. Interestingly, Equation 6 provides a direct way to calculate the relative importance of any AAindex for sequence conservation from blosum62 or any other substitutional matrix, which can be used to rank the relevance of indices to protein sequence conservation.

The second observation is that the blosum62 derived scales perform better than any other set tested in the classification of Class A GPCRs, as measured by the PLS-DA models' predictive ability. This may be due to

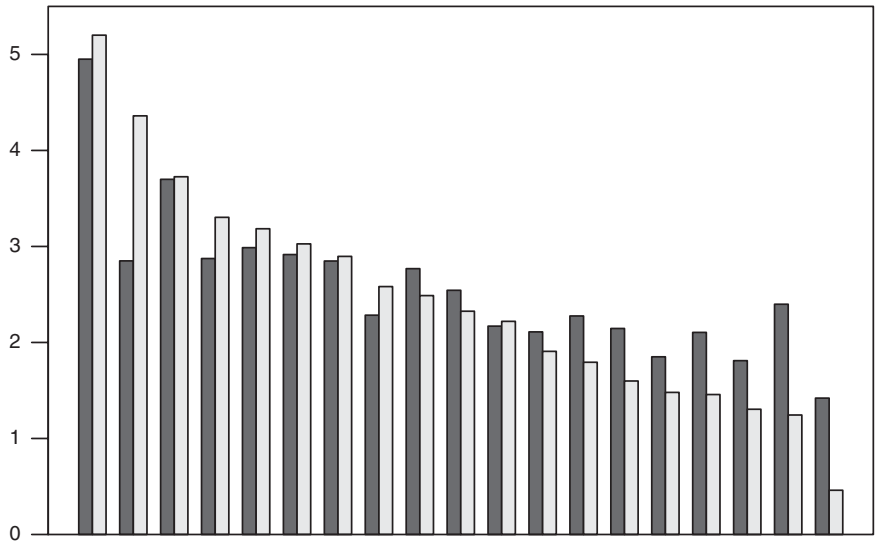


FIG. 8. Axes weights calculated from the blosum62 matrix (dark gray). The square root of the blosum62 eigenvalues shown for comparison (light gray).

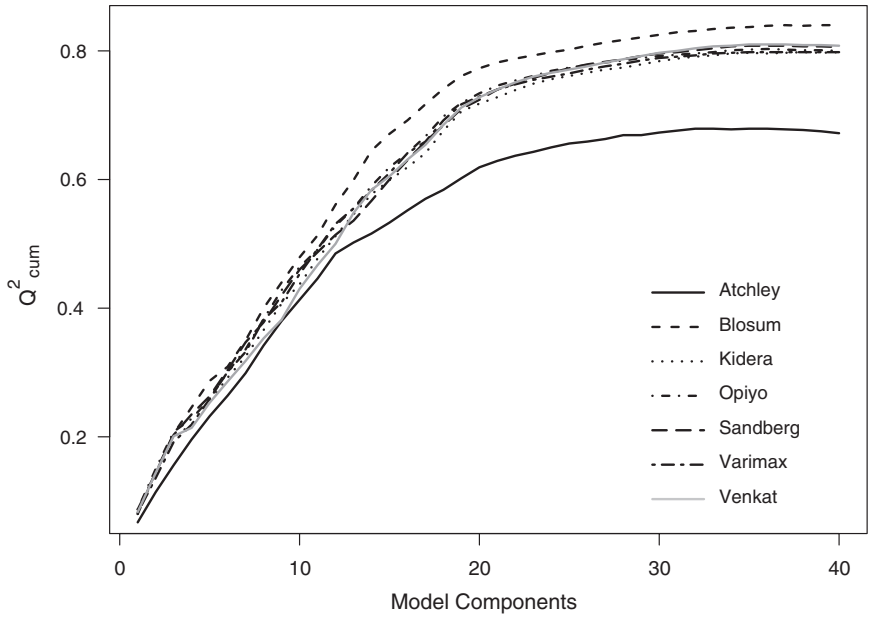


FIG. 9. Performance of sets of scales in PLS-DA models of GPCRs, assessed by the predictive power Q^2_{cum}.

the analogies between the ACC approach to represent sequences when the BLOSUM62 scales are used, and the widely used dynamic programming approaches to biological sequence comparison based on similarity matrices and pioneered by Needleman and Wunsch (1970).

There is an infinite number of ways to represent numerically amino acids, but our means of interpreting those representations are limited. The current study has revealed that the dimensionality of protein index space is higher than previously shown, with at least six to seven independent indices containing information relevant to protein structure. It is conceivable that future studies may bring forth novel aspects of amino acid space and thus present new, uncorrelated to the current indices. For the purpose of computational characterization of proteins and peptides, suitable combinations of several indices corresponding to features of interest can be selected.

Importantly, the low interpretability of scales derived from BLOSUM62 compared to the excellent performance of those scales in a bioinformatics modeling application indicated that current research, aimed at ranking amino acid residues, has yet to uncover some salient properties important for conservation of protein sequence. Conclusions from the current work suggest that, for the two different tasks, *exploration* and *modeling* as defined above, different scales may be appropriate to use. While the VARIMAX scales are suitable for exploratory analyses, the BLOSUM62 scales seem to be a better choice for unsupervised learning and modeling applications.

5. APPENDIX: CONVERSION BETWEEN AMINO ACID SCALES AND SUBSTITUTION MATRICES

The substitution of one amino acid residue with another in conserved protein domains is an indication of higher propensity of the two residues to reside within a similar structural context. In other words, substitutions between two amino acids will be less likely to disrupt protein structure if the two amino acids have similar impact on the local secondary structure. I reasoned that substitution frequency of two amino acids X and Y can be defined by the probabilities of occurrence of each of these amino acids within the same structural context, as follows:

$$p_{xy} = \frac{p_{x,i}p_{y,i}}{f_i} + \frac{(f_x - p_{x,i})(f_y - P_{y,i})}{1 - f_i}, \tag{8}$$

where $p_{x,i}$ is the joint probability of occurrence of amino acid x in structure i, and f_i is the frequency of occurrence of structure i. Dividing both sides by $f_x f_y$ gives

$$q_{xy} = \frac{p_{i|x}p_{i|y}}{f_i} + \frac{(1-p_{i|x})(1-P_{i|y})}{1-f_i}, \tag{9}$$

where $p_{i|x} = p_{x,i}/f_x$ is the conditional probability of occurrence of structure i calculated over all X residues. With several *independently* occurring structural features, the cumulative q_{xy} ratio is given by the product:

$$q_{xy} = \prod_{i=1}^{n} [\frac{p_{i|x}p_{i|y}}{f_i} + \frac{(1-p_{i|x})(1-P_{i|y})}{1-f_i}], \tag{10}$$

This leads to

$$q_{xy} = \prod_{i=1}^{n} [\frac{p_{i|x}p_{i|y} + f_i - f_i p_{i|x} - f_i P_{i|y}}{f_i(1-f_i)}], \tag{11}$$

and after substituting back $p_{x,i} = f_x \times p_{(i|x)}$ and rearranging

$$q_{xy} = \prod_{i=1}^{n} [\frac{f_i}{1-f_i}(\frac{p_{x,i}}{f_x f_i} - 1)(\frac{P_{y,i}}{f_y f_i} - 1) + 1], \tag{12}$$

Further, by taking the natural logarithms of the two sides of the equation and using the first order approximation $log(x+1) \approx x$, we obtain

$$log(q_{xy}) \approx \sum_{i=1}^{n} [\frac{f_i}{1-f_i}(log\frac{p_{x,i}}{f_x f_i})(log\frac{p_{y,i}}{f_y f_i})] \tag{13}$$

which, using the definition of amino acid scales (Equation 1) is identical with

$$s_{xy} = \sum_{i=1}^{n} (\frac{f_i}{1-f_i} \times s_{x,i} \times s_{y,i}) \tag{14}$$

Substituting $f_i/(1-f_i)$ for the elements of Λ and $s_{x,i}$ for the elements of A, we obtain Equation 5.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Atchley, W.R., Zhao, J., Fernandes, A.D., et al. 2005. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. USA* 102, 6395–6400.

Aurora, R., and Rose, G.D. 1998. Helix capping. *Protein Sci.* 7, 21–38.

Bellman, R.E. 1961. *Adaptive Control Processes.* Princeton University Press, Princeton, N.J.

Burnham, A.J., MacGregor, J.F., and Viveros, R. 1999. Latent variable multivariate regression modelling. *Chemometrics Intell. Lab. Sys.* 48, 167–180.

Cornette, J.L., Cease, K.B., Margalit, H., et al. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* 195, 659–685.

Doig, A.J., and Baldwin, R.L. 1995. N- and C-capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci.* 4, 1325–1336.

Eddy, S.R. 2004. Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.* 22, 1035–106.

George, R.A., and Heringa, J. 2002. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng.* 15, 871–879.

Hellberg, S., Sjöström, M., and Wold, S. 1986. The prediction of bradykinin potentiating potency of pentapeptides. An example of a peptide quantitative structure–activity relationship. *Acta Chem. Scand. B* 40, 135–140.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89, 10915–10919.

Hessa, T., Kim, H., Bihlmaier, K., et al. 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* 433, 377–381.

Horn, F., Bettler, E., Oliveira, L., et al. 2003. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.* 31, 294–297.

Kaiser, H. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 187–200.

Kawashima, S., and Kanehisa, M. 2000. AAindex: amino acid index database. *Nucleic Acids Res.* 28, 374.

Kidera, A., Konishi, Y., Oka, M., et al. 1985a. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.* 4, 23–55.

Kidera, A., Konishi, Y., Ooi, T., et al. 1985. Relation between sequence similarity and structural similarity in proteins. Role of important properties of amino acids. *Protein J.* 4, 265–297.

Lapinsh, M., Gutcaits, A., Prusis, P. et al. 2002. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci.* 11, 795–805.

Nakai, K., Kidera, A., and Kanehisa, M. 1988. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* 2, 93–100.

Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.

Opiyo, S.O., and Moriyama, E.N., 2007. Protein family classification with partial least squares. *J. Proteome Res.* 6, 846–853.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Phil. Mag. J.* 6, 559–572.

R Development Core Team. 2006. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna.

Sandberg, M., Eriksson, L., Jonsson, J., et al. 1998. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* 41, 2481–2491.

Sneath, P.H. 1966. Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* 12, 157–195.

Thurstone, L. 1938. A new rotational method in factor analysis. *Psychometrika* 3, 199–218.

Thurstone, L.L. 1947. *Multiple-Factor Analysis: A Development and Expansion of The Vectors of Mind.* University of Chicago Press, Chicago.

Tomii, K., and Kanehisa, M. 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* 9, 27–36.

Venkatarajan, M., and Braun, W. 2001. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. *J. Mol. Modeling* 7, 445–453.

Wold, S., Ruhe, A., Wold, H., et al. 1984. The collinearity problem in linear regression. the partial least squares approach to general inverses. *SIAM J. Sci. Stat. Comput.* 5, 735–743.

Wold, S., Sjöström, M., Sandrberg, M., et al. 1993. DNA and peptide sequences and chemical prosesses multivariately modelled by principal component analysis and partial least squares projections to latent structures. *Anal. Chim. Acta* 277, 239–253.

Zimmerman, J.M., Eliezer, N., and Simha, R. 1968. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* 21, 170–201.

Address reprint requests to:
Dr. Alexander G. Georgiev
Department of Biochemistry and Biophysics
Stockholm University
The Arrhenius Laboratories
106 91 Stockholm, Sweden

E-mail: alge@dbb.su.se