J Mol Model (2001) 7:445–453
DOI 10.1007/s00894-001-0058-5

ORIGINAL PAPER

**Mathura S. Venkatarajan · Werner Braun**

# New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties

Received: 15 August 2001 / Accepted: 18 October 2001 / Published online: 8 December 2001


**Abstract** We derive new quantitative descriptors for the 20 naturally occurring amino acids based on multidimensional scaling of 237 physical–chemical properties. We show that a five-dimensional property space can be constructed such that the amino acids are in a similar spatial distribution to that in the original high-dimensional property space. Properties that correlate well with the five major components were hydrophobicity, size, preferences for amino acids to occur in α-helices, number of degenerate triplet codons and the frequency of occurrence of amino acid residues in β-strands. Distances computed for pairs of amino acids in the five-dimensional property space are highly correlated with corresponding scores from similarity matrices derived from sequence and 3D structure comparison. We used the five-dimensional property distances to cluster the amino acids in groups depending on a cutoff distance. These groups define a reduced amino acid alphabet for protein folding studies. Our descriptors should provide a quantitative means to identify property motifs in sequences of protein families. Electronic supplementary material to this paper can be obtained by using the Springer Link server located at http://dx.doi.org/10.1007/s00894-001-0058-5.

**Keywords** Multidimensional scaling · Amino acid · Substitution matrices · BLOSUM · PAM · Physical–chemical properties · Cluster analysis

Electronic supplementary material to this paper can be obtained by using the Springer Link server located at http://dx.doi.org/10.1007/s00894-001-0058-5.

M.S. Venkatarajan · W. Braun (✉)
Sealy Center for Structural Biology,
Department of Human Biological Chemistry and Genetics,
301 University Boulevard,
The University of Texas Medical Branch,
Galveston, TX 77555-1157, USA
e-mail: werner@newton.utmb.edu
Tel.: +1 409 747-6810, Fax: +1 409 747-6850

## Introduction

Homologous proteins share a common fold, even when the overall sequence identity is less than 10%. [1] The physical–chemical properties of the less conserved residues still encode the information necessary for folding. Hydrophobicity or charge is to a high degree conserved in structurally equivalent positions among evolutionarily related proteins, even when the individual amino acid residues are different. [2, 3, 4, 5] Profile methods [6, 7] such as PSI-Blast, [8, 9] and hidden Markov models [10, 11, 12] rely implicitly on conserved patterns of amino acids to detect homologous proteins. Protein threading methods rely even less on sequence identity and detect the homologous folds primarily based on the structure-forming properties of areas of the protein sequence. [13, 14, 15, 16, 17, 18, 19] While these methods have shown extraordinary success in recognizing targets with similar folds and low sequence identity, results from recent CASP competitions [8] indicate that there are opportunities for improvements.

We suggest here a new method to summarize information about physical–chemical properties that should prove useful in identifying protein homologues on the basis of property-based motifs. We used multidimensional scaling of 237 physical–chemical properties to derive quantitative descriptors for all 20 naturally occurring amino acids. We were able to reproduce the main variations of all properties for the 20 amino acids through five quantitative descriptors. Multidimensional scaling is a general classification approach to reconstructing the geometrical configuration of large point sets in lower dimensions. [20, 21, 22] This approach has been used in distance geometry to calculate protein structures from NMR data [23] and to classify 3D conformations of proteins, [24, 25] but our method of reducing the large redundancy in physical–chemical properties by multidimensional scaling is novel. While the physical meaning of each descriptor can be correlated with individual properties, the five descriptors cannot simply be replaced by five individual properties. A similarity measure based on our

descriptors correlated well with similarity indices derived from substitution matrices such as PAM250, [26] BLOSUM62, [27] sub-structural [28] and in particular Gonnet. [29] These quantitative descriptors can characterize motifs in protein families (work in progress), and may aid in improving alignments for protein modeling. The information has been incorporated into our MASIA program and is available for use at our website http://www.scsb.utmb.edu/masia. [30, 31, 32]

## Materials and methods

### Property normalization

Each property was normalized such that the standard deviation is 1 and the average is 0. Standard deviation was calculated using the "biased" method. The normalization ensures that all properties are expressed as dimensionless numbers.

$$S_\alpha(i) = \frac{(P_\alpha(i) - \overline{P_\alpha})}{\sigma_{P\alpha}} \tag{1}$$

$$\sigma_{P\alpha} = \sqrt{\frac{20\sum_{i=1}^{20} P_\alpha(i)^2 - \left(\sum_{i=1}^{20} P_\alpha(i)\right)^2}{400}} \tag{2}$$

where $S$ is normalized property values, $\alpha$ is the index of the property and $i$ stands for the amino acid. $P$ is the property value, $\overline{P}_\alpha$ and $\sigma_{P\alpha}$ are the average and the standard deviation of property $\alpha$.

### Representation of properties

Each amino acid $i$ is represented as a vector $\underline{S}(i)$ in a 237-dimensional continuous space, where the components $S_\alpha(i)$ are the normalized property values. The scalar product $Q_{ij}$ between two vectors $\underline{S}(i)$ and $\underline{S}(j)$, where $j$ is another index for an amino acid, is given by

$$Q_{ij} = \underline{S}(i) \bullet \underline{S}(j) \tag{3}$$

$$Q_{ij} = \sum_{\alpha=1}^{237} S_\alpha(i).S_\alpha(j) \tag{4}$$

The positive symmetric 20×20 matrix $Q$ consists of the scalar products of the property vectors **$\underline{S}$(i)** and **$\underline{S}$(j)**, where $i$=1...20 and $j$=1...20.

### Calculation of eigenvectors and eigenvalues

Eigenvectors $E$ and eigenvalues $\lambda$ of the matrix $Q$ were computed using the JACOBI and EIGSRT subroutines provided in Numerical Recipes. [33]

$$Q.E = \lambda E \tag{5}$$

As $Q$ is of order 20, we will have 20 eigenvectors and eigenvalues $\lambda$ and the smallest eigenvalue $\lambda_{20}$ is equal to 0 due to normalization of the properties. The subroutine JACOBI implements a Jacobian transformation of a symmetric matrix and returns the eigenvectors and values of the $Q$ matrix. The eigenvalues and their corresponding eigenvectors are indexed in decreasing order of the eigenvalues.

### Calculation of distances in the property space

If $\mu$ represents the index of eigenvalue and eigenvector, then the elements of the $Q$ matrix can be equated to eigenvalues and eigenvectors as:

$$Q_{ij} = \sum_{\mu=1}^{20} \lambda_\mu E_i^\mu . E_j^\mu \tag{6}$$

The first five significant eigenvalues were selected for the representation of amino acids, thus $Q_{ij}$ can be written as:

$$Q_{ij} \approx \sum_{\mu=1}^{5} \lambda_\mu E_i^\mu . E_j^\mu \tag{7}$$

Each amino acid can be represented as a vector in the five-dimensional Euclidean space (a.k.a. Eigen subspace) with each dimension perpendicular to each other. The co-ordinates of the $i$th amino acid can be written as:

$$\sqrt{\lambda}_{\mu=1} E_i^{\mu=1}, \sqrt{\lambda}_{\mu=2} E_i^{\mu=2}, \sqrt{\lambda}_{\mu=3} E_i^{\mu=3}, \sqrt{\lambda}_{\mu=4} E_i^{\mu=4}, \sqrt{\lambda}_{\mu=5} E_i^{\mu=5} \tag{8}$$

The distance between the $i$th and $j$th amino acids is given by

$$d_{ij} = \sqrt{\sum_{\mu=1}^{5} \left(\sqrt{\lambda}_\mu E_i^\mu - \sqrt{\lambda}_\mu E_j^\mu\right)^2} \tag{9}$$

Distances computed between amino acids in the five-dimensional Eigen sub-space constitute the property distance matrix (PDM). Small distance values between two amino acids indicate they are similar in all of their 237 physical-chemical properties.

### Generation of normally distributed random numbers

Random numbers with normal (Gaussian) distribution were generated using the Box–Muller method and implemented using the GASDEV subroutine from Numerical Recipes. [33] The distribution of these random numbers in multiple dimensions will approximate a spherical surface.

### Calculation of correlation coefficient

Pearson's correlation coefficient between pairs of quantities $(x_i,y_i)$ is given by the formula

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad (10)$$

### Cluster analysis

In a first approach amino acids were clustered according to their property distances below a given threshold. A single cluster core, formed by the closest amino acid pair in the five-dimensional Eigenspace, is identified first. Distances from the centroid of this core to all other amino acids are calculated, and the amino acid with the least distance to the centroid is added, provided the distance is below the threshold. A new centroid is computed each time an amino acid is added. The procedure is repeated until there are no other amino acids closer than the threshold to the centroid of the cluster. The whole procedure is repeated with the remaining amino acids, which are not yet part of a cluster. As a second approach we used the hierarchical clustering program KITSCH of the phylogenetic package PHYLIP version 3.2.3. [34] The PDM between all amino acids was used as the input to KITSCH. The program produces a best tree that has similar distances computed from the tree and the distance matrix using the Fitsch–Margolish algorithm. [35]

## Results

### Quantitative representation of amino acids in five dimensions

A comprehensive list of 237 physical–chemical properties was compiled from the public databases SWISS-PROT [36] and dbGET. [37] Only those properties of the side chains with numerical values for all amino acids were considered. We included experimentally determined properties, such as different scales of hydrophobicities derived from transfer free energy values, volumes, molecular weight, or helix-coil equilibrium constants, and statistically derived quantities, such as propensity values for secondary structures. A complete list of all 237 included properties is provided as electronic supplementary material.

Each amino acid is represented as a vector in the 237-dimensional space of normalized properties with mean value of zero and standard deviation 1. Our multidimensional scaling approach reveals the high redundancy of the property values (see Materials and methods for details). The computational approach and justification for reduction to a lower dimensional space follows closely the practice of embedding in distance geometry. [23] The distribution of the eigenvalues of the *Q* matrix

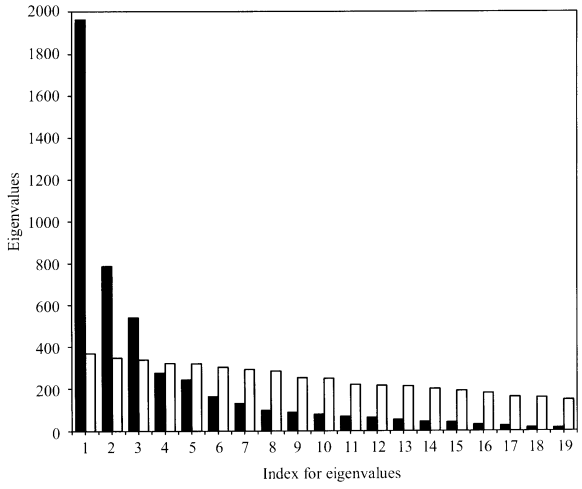


**Fig. 1** Distribution of the eigenvalues of the components, computed from 237 normalized properties (*black bars*). Indices for the components are sorted according to decreasing order of the eigenvalues. As a control, eigenvalues of a matrix derived from 237 random uniformly distributed vectors are shown in *white bars*

(Fig. 1), containing the scalar products between all pairs of the 237-dimensional amino acid vectors, rapidly decreases from the largest value $\lambda_1$=1962 to $\lambda_{19}$=16. As a control we generated 237 random vectors with a uniform distribution of property values, and calculated the distribution of eigenvalues for these "random" property values. In the random case, the eigenvalues are all almost equal as expected. In contrast, the rapid decrease of the eigenvalues derived from the 237 physical–chemical properties shows a large anisotropy of the distribution of the property values. This anisotropy is a consequence of the large redundancy in the sets of property values. This suggests that the number of properties can be reduced while retaining approximately the same distribution of amino acids in the property space. The eigenvalues rapidly decrease within the five largest eigenvalues and are then substantially smaller than the values calculated from the "random" property set. We then compared distances in the original property space with those regenerated from a subset of *n* eigenvectors, varying *n* systematically from 2 to 20 (Fig. 2a). The correlation coefficient between the original and regenerated distances is more than 99% for *n*=5, and approaches 1 very rapidly. We therefore chose the first five eigenvalues and eigenvectors to calculate five-dimensional numerical descriptors of the amino acids. The individual distances in the original property space and in the subspace using the first five eigenvectors are highly correlated (Fig. 2b). The five numerical descriptors for each amino acid are given in Table 1.

### Physical–chemical meaning of the numerical descriptors

Two-dimensional representations of the 20 amino acids in the plots E1 versus the eigenvectors E2 to E5 provide

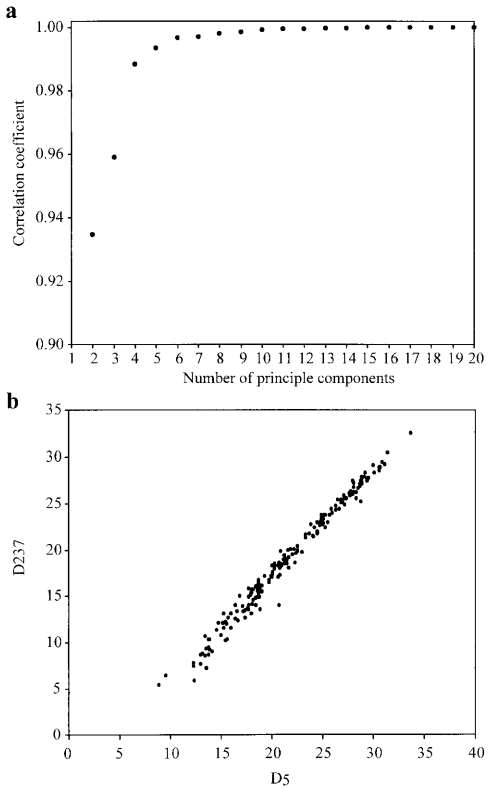


**Fig. 2a,b** Comparison of distances between amino acids in the original property space and component subspaces. **a** Linear correlation coefficient between distances of all amino acid pairs in the original 237-dimensional property space and in subspaces of *n* components with *n* varying from 2 to 20. **b** Correlation plot between distances of amino acid pairs in the five-dimensional space formed by the first five components ($D_5$) and distances in the original property space ($D_{237}$). The linear correlation coefficient is 0.992

**Table 1** Components E1 to E5 of 237 physical–chemical properties for each amino acid

| Eigenvalue (λ) | Eigenvector[a] | | | | |
|---|---|---|---|---|---|
| | E1 1961.504 | E2 788.200 | E3 539.776 | E4 276.624 | E5 244.106 |
| A | 0.008 | 0.134 | –0.475 | –0.039 | 0.181 |
| R | 0.171 | –0.361 | 0.107 | –0.258 | –0.364 |
| N | 0.255 | 0.038 | 0.117 | 0.118 | –0.055 |
| D | 0.303 | –0.057 | –0.014 | 0.225 | 0.156 |
| C | –0.132 | 0.174 | 0.070 | 0.565 | –0.374 |
| Q | 0.149 | –0.184 | –0.030 | 0.035 | –0.112 |
| E | 0.221 | –0.280 | –0.315 | 0.157 | 0.303 |
| G | 0.218 | 0.562 | –0.024 | 0.018 | 0.106 |
| H | 0.023 | –0.177 | 0.041 | 0.280 | –0.021 |
| I | –0.353 | 0.071 | –0.088 | –0.195 | –0.107 |
| L | –0.267 | 0.018 | –0.265 | –0.274 | 0.206 |
| K | 0.243 | –0.339 | –0.044 | –0.325 | –0.027 |
| M | –0.239 | –0.141 | –0.155 | 0.321 | 0.077 |
| F | –0.329 | –0.023 | 0.072 | –0.002 | 0.208 |
| P | 0.173 | 0.286 | 0.407 | –0.215 | 0.384 |
| S | 0.199 | 0.238 | –0.015 | –0.068 | –0.196 |
| T | 0.068 | 0.147 | –0.015 | –0.132 | –0.274 |
| W | –0.296 | –0.186 | 0.389 | 0.083 | 0.297 |
| Y | –0.141 | –0.057 | 0.425 | –0.096 | –0.091 |
| V | –0.274 | 0.136 | –0.187 | –0.196 | –0.299 |

[a] The numerical descriptors for each amino acid *i* are calculated by $\sqrt{\lambda_\mu} E_i^\mu$ for the five eigenvectors $E^\mu, \mu = 1..5$.

insights into the physical meaning of the descriptors (Fig. 3). The hydrophobic/hydrophilic character is the main determinant of the distribution of amino acids along E1. Hydrophobic residues I, F, L, W, M, C and Y are grouped towards negative values along E1, while the polar and charged residues have positive values. Small residues, G, P and S cluster at positive values along E2 and large residues R, K, E and W at negative values. Residues with a high helix-forming propensity (A, L, E) are grouped at the bottom of Fig. 3b, and helix-breakers are at the top.

The physical meaning of the components E4 and E5 are not immediately obvious from the two-dimensional representation; we therefore calculated linear correlation coefficients between all 237 properties and the five numerical descriptors E1 to E5 (Table 2). These data confirm the high correlation of E1 to E3 for hydrophobicity, size and helical propensity. The correlation coefficients for E4 are highest for partial specific volumes, relative abundance of amino acids, and the number of codons. The β-strand forming propensity seems the dominant factor for E5. The highest correlation coefficients for E2 to E5 are significantly different from 1, and while deviations from a linear correlation between E1 and hydrophobicity, and E2 and size (represented by side chain length) are small (Fig. 4a, b), significant differences exist for residues W and G for the correlation between descriptor E3 and the α-helix propensity, and for K between E4 and the number of codons (Fig. 4c, d). Thus the five components cannot simply be replaced by five individual properties.

## Cluster analysis of the amino acids

The five-dimensional descriptors represent an exhaustive large set of physical–chemical properties. The Euclidian distances between the five-dimensional descriptors, PDM, can be used to group the amino acids in different clusters according to this large set of properties. In our first approach we clustered similar amino acids according to their property distances with an increasing distance threshold from 9.5 to 20, shown in bold face in Table 3. For small distance thresholds only the large hydrophobic residues I, V, L, F and the polar and charged residues S, T, Q, N and D form clusters. These clusters grow as the threshold is increased until they form only two groups comprising hydrophobic and hydrophilic residues. The small amino acids A, G, C and P are not part of any cluster until large values of the threshold. This cluster analysis can be used to rationally define a reduced alphabet for the amino acid sequences, e.g. a seven letter code with a cut-off value of 12.5.

The hierarchical clustering program KITSCH of the phylogenetic package PHYLIP version 3.2.3 [34] yields a

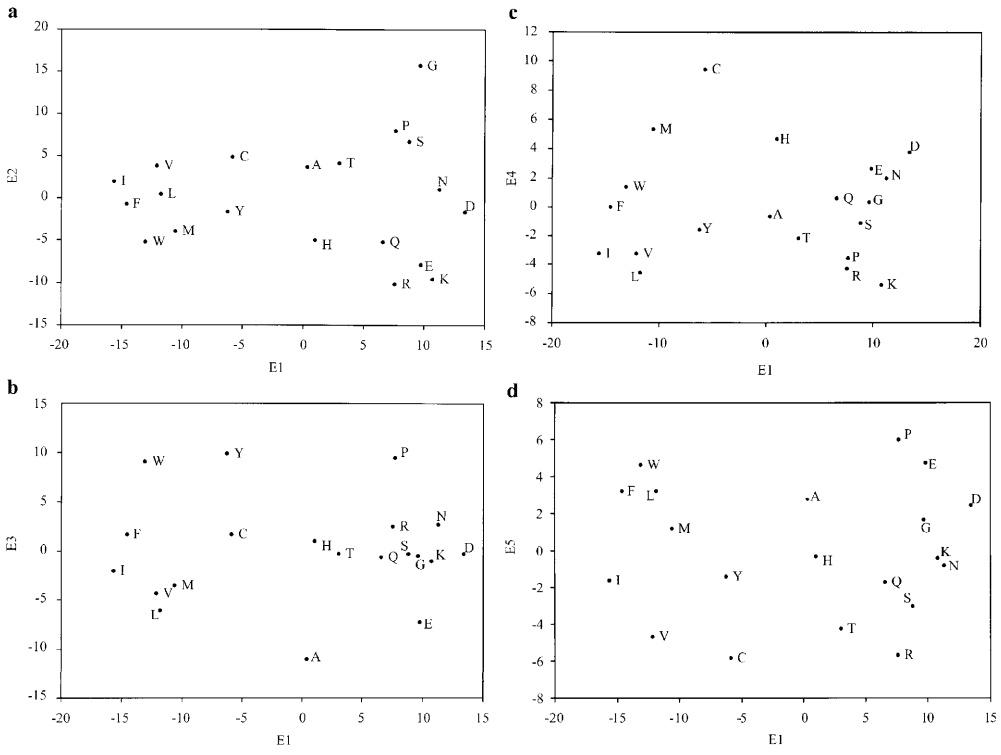


**Fig. 3a–d** Distribution of amino acids along the principal components E1, E2, E3, E4 and E5. **a** Two-dimensional plot E2 versus E1. The E1 and E2 axis correspond to hydrophobicity/hydrophilicity and molecular size of amino acids. **b** Two-dimensional plot E3 versus E1. The E3 component corresponds approximately to helical-propensity. **c** Two-dimensional plot E4 versus E1. E4 is related to the relative abundance of amino acids. **d** Two-dimensional plot E5 versus E1. E5 shows a weak correlation to β-strand propensity

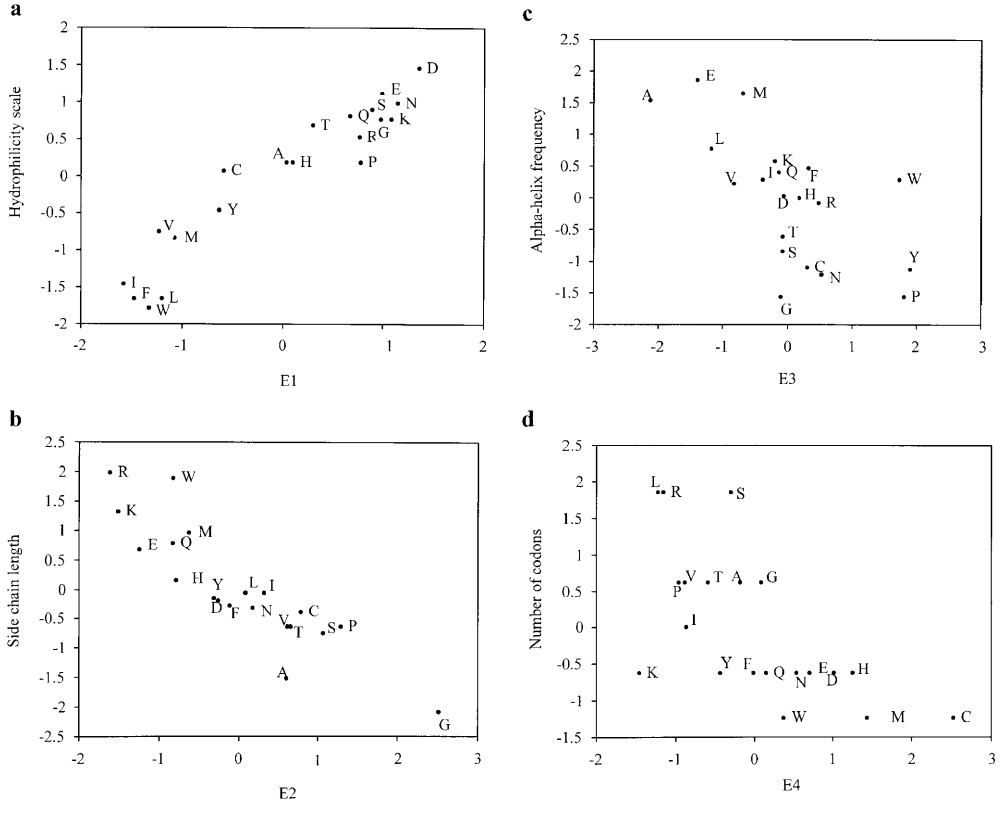


**Fig. 4a–d** Correlation between components (horizontal axis) and high correlating individual properties (vertical axis). **a** Correlation plot between hydrophilicity scale [57] and E1 (|*r*|=0.95). **b** Correlation plot between side chain length of amino acids [64] and E2 (|*r*|=0.89). **c** Correlation plot between alpha-helical frequency [74] of amino acids and E3 (|*r*|=0.67). **d** Correlation plot between the number of codons coding for each of the amino acids and E4 (|*r*|=0.65)

tree of clusters very similar to our first approach (Fig. 5). The separation in seven clusters at the cut-off distance of 12.5 in Table 3 is exactly matched by the hierarchical clustering method (dotted line I in Fig. 5), and the GP cluster is found at the slightly higher cut-off value 14.5 (dotted line II in Fig. 5). The fine grain groupings IV-LFM, WY, RK and ST are found by both procedures. The only major difference is that Q and H, which did not cluster at small threshold values in our first procedure, are grouped early in the hierarchical method.

**Table 2** Individual physical–chemical properties with high correlation to each of the five components E1 to E5

| Eigen-axes | Property name and reference | $\|r\|$[a] |
|---|---|---|
| E1 | 14 Å contact number [50] | 0.971 |
| | Effective partition energy [39] | 0.967 |
| | Average reduced distance for $C_\alpha$ [56] | 0.953 |
| | Hydrophilicity scale derived from HPLC [57] | 0.950 |
| | Hydrophobicity scale [58] | 0.950 |
| | Partition co-efficient [59] | 0.946 |
| | Information value for accessibility [60] | 0.942 |
| | Average surrounding hydrophobicity [61] | 0.928 |
| | Long range non-bonded energy per atom [62] | 0.924 |
| | Hydrophobic parameter pi [63] | 0.923 |
| E2 | STERIMOL length of the side chain [64] | 0.889 |
| | Distance between $C_\alpha$ and centroid of side chain [65] | 0.881 |
| | Radius of gyration of side chain [65] | 0.876 |
| | Residue accessible surface area [66] | 0.822 |
| | Entropy of formation [67] | 0.815 |
| | Absolute entropy [67] | 0.791 |
| | Molecular weight [68] | 0.783 |
| | Side chain torsion angle phi [65] | 0.746 |
| | Principal component II [69] | 0.742 |
| | Average volume of buried residue [70] | 0.742 |
| E3 | Normalized frequency of turn [71] | 0.747 |
| | Conformational parameter for β-turn [72] | 0.714 |
| | Information measure for loop [73] | 0.713 |
| | Information measure for turn [73] | 0.706 |
| | Normalized frequency of α-helix [74] | 0.674 |
| | Average relative probability of helix [75] | 0.674 |
| | Information measure for α-helix [73] | 0.662 |
| | Normalized frequency of α-helix [76] | 0.656 |
| | Normalized relative frequency of α-helix [77] | 0.644 |
| | Normalized frequency of α-helix [78] | 0.640 |
| E4 | Partial specific volume [79] | 0.658 |
| | Number of codon(s) | 0.651 |
| | Amino acid composition of total proteins [80] | 0.637 |
| | Amino acid composition in SWISSPROT [36] | 0.623 |
| | Amino acid composition [81] | 0.610 |
| | Apparent partial specific volume [82] | 0.606 |
| | Relative frequency of occurrence [83] | 0.604 |
| | Composition [38] | 0.594 |
| | Amino acid composition of total proteins [80] | 0.547 |
| | Amino acid distribution [84] | 0.538 |
| E5 | Frequency of extended structure [85] | 0.560 |
| | Free energy in β-strand region [86] | 0.534 |
| | Beta-strand indices [87] | 0.529 |
| | Free energy in β-strand region [86] | 0.513 |
| | Information measure for pleated-sheet [73] | 0.487 |
| | Frequency of β-sheet [76] | 0.484 |
| | Retention co-efficient in TFA [88] | 0.479 |
| | Information measure for extended structure [73] | 0.473 |
| | Net charge [89] | 0.425 |
| | Normalized frequency of extended structure [90] | 0.423 |

[a] Ten individual properties with highest correlation to each principal components E1 to E5 are given. Absolute value of the linear correlation coefficient calculated between the components and the properties

## Comparison between property distances and similarity scores derived from substitution matrices

Amino acids in close proximity in descriptor space have similar physical–chemical properties. Substitutions of amino acids in related proteins are usually to amino acids with similar properties. We thus anticipate that the substitution frequencies of amino acids in protein families should be inversely related to our property distances. To test this hypothesis we correlated our property-based distances with all major substitution matrices. [26, 27, 29, 38, 39, 40] We also included substitution matrices derived from 3D structure comparisons, as they are usually more sensitive to detect distantly related proteins. [41] Recently an amino acid similarity matrix, sub-structural matrix (SSM), based on the structural similarity comparison of non-homologous proteins [28] and a structure derived similarity matrix were published. [42]

Linear correlation coefficients between the off-diagonal elements of the substitution matrices and our PDM of the 5D descriptors are shown in Table 4. High absolute correlation coefficients of about 0.8 were observed for

**Table 3** Cluster analysis of amino acids according to their property distance

| Distance cut-off[a] | Amino acid clusters[b] |
|---|---|
| 9.5 | **IVLF** M W Y C A G P E **RK** H **ST QDN** |
| 10.0 | **IVLF** M W Y C A G P E **RK** H **STQDN** |
| 10.5 | **IVLFM WY** C A G P E **RK** H **STQDN** |
| 11.0 | **IVLFM WY** C A G P E **RK HSTQDN** |
| 12.5 | **IVLFM WY** C A G P **ERKHSTQDN** |
| 14.5 | **IVLFMWY** C A **GP ERKHSTQDN** |
| 15.0 | **IVLFMWYCHT** A **GP ERKSQDN** |
| 18.0 | **IVLFMWYCHTAERKSQDNP** G |
| 20.0 | **IVLFMWYCHTAERKSQDNPG** |

[a] Euclidean distance in the five-dimensional eigenspace
[b] Clusters obtained below distance cut-off are shown in bold

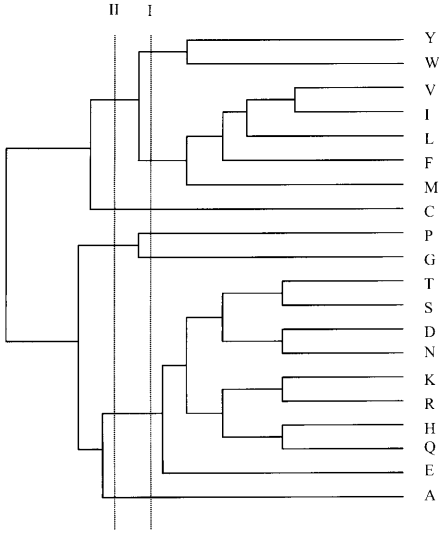


**Fig. 5** Hierarchical clustering of amino acids using the PDM as input to the program KITSCH of the phylogenetic package PHYLIP. Amino acids connected to the same node are closely located in the property space. This best fit binary tree (standard deviation of 21.12 and sum of squares of 16.86) was selected from 2,196 trees generated by the program. A vertical dotted line divides amino acids into groups that correspond to those in Table 3. Vertical bar I separates amino acids that approximates our clusters at a cutoff distance of 12.5 and II corresponds to a cutoff of 14.5

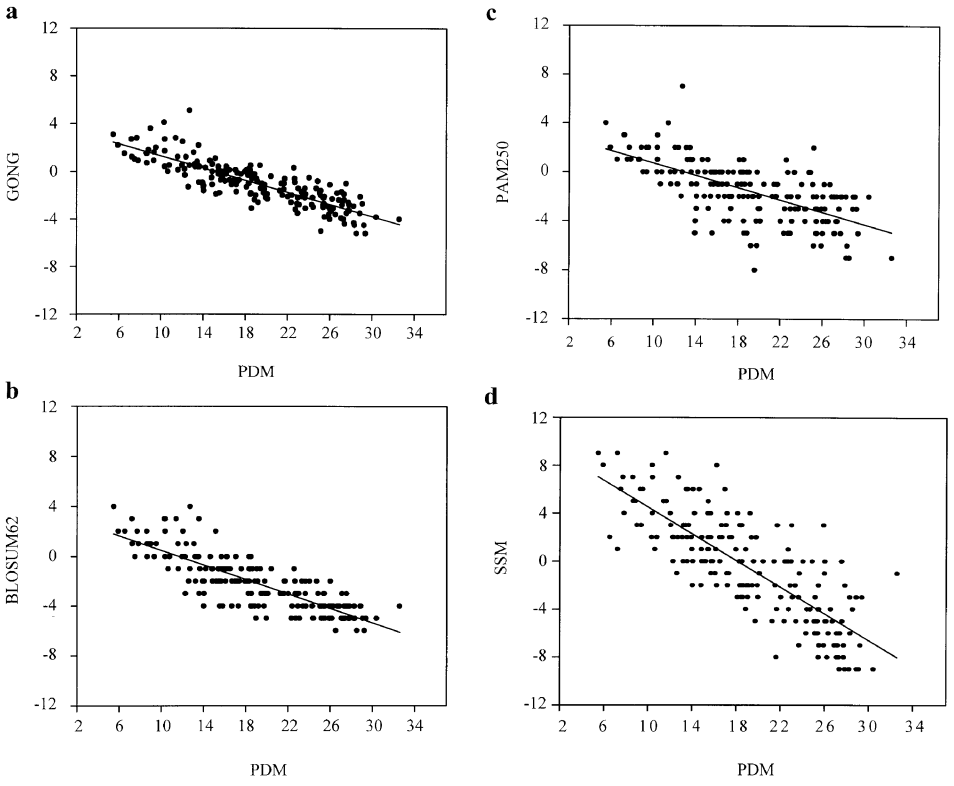


**Fig. 6a–d** Correlation plots between substitution matrices based on sequence and 3D structural analysis and our PDM. Diagonal values were excluded from the comparison. **a** GONG matrix [29] (correlation coefficient $r$=–0.848). **b** BLOSUM62 matrix [27] ($r$=–0.822). **c** PAM250 matrix [26] ($r$=–0.653). **d** Sub-structural matrix (SSM), a 3D structure-based comparison matrix [28] ($r$=–0.801)

**Table 4** Comparison of property distances and similarity scores derived from substitution matrices

| Basis of the scoring matrix | Matrix name and reference | $r$[a] |
|---|---|---|
| Sequence based | A composite log-odds matrix [29] | –0.848 |
| | Log-odds scoring matrix in 74-100 PAM [91] | –0.840 |
| | BLOSUM62 Substitution matrix [27] | –0.822 |
| | The PAM250-PET91 matrix [83] | –0.716 |
| | PAM250 The log odds matrix [26] | –0.653 |
| Structure based | Sub-structural matrix [28] | –0.801 |
| | Homologous structure derived matrix [42] | –0.793 |
| | Structure derived matrix [42] | –0.778 |
| | Structure-based amino acid scoring table [41] | –0.731 |
| | Secondary structural similarity matrix [92] | –0.709 |
| | Structure-based comparison table [93] | –0.692 |

[a] Linear correlation coefficient calculated between off-diagonal values of the PDM and corresponding values in substitution matrices

all major substitution matrices, with the highest absolute value of 0.85 between our property-based distances and the Gonnet matrix. All recent substitution matrices derived from 3D structure comparison show high correlation coefficients. Correlation plots between more familiar matrices like PAM250, [26] BLOSUM62 [27] and sub-structural matrices [28] are shown in Fig. 6.

## Discussion

### Meaning of the five-dimensional descriptors

Our multidimensional scaling method reduces a large pool of meaningful physical–chemical properties to a small set of five quantitative descriptors for amino acids. Five components of the properties were sufficient to reproduce the distances in the complete property space, a measure of the similarity of amino acids. The first and second components are dominated by hydrophobicity/hydrophilicity and amino acid size, respectively. However, one cannot simply replace all five components by individual properties, as several linear combinations of properties contribute to the components. We found similar values for the descriptors by repeating the analysis with smaller subsets of properties if we include at least one of the general property types (hydrophobicity, size, or secondary structure propensities) in the property list.

Our goal is to develop a sensitive motif search, based on physical properties of amino acids rather than on sequence identity. In general it is difficult to decide a priori which of the many properties one should use. Our quantitative descriptors represent a precise spatial relation of all amino acids with respect to many physical–chemical properties. They should be useful to identify related proteins in genome projects. [43, 44, 45, 46]

Previous studies have addressed the question of property-based similarity searches. [5, 6, 47] All of these studies used a small subset of a priori chosen individual properties. Otsuka and co-workers [48] used four properties to investigate similarity relationships between DNA and RNA polymerases. Grigoriev and Kim have used five physical properties along with secondary structures to represent proteins. [5] Their proximity correlation matrix method identified hydrophobicity as the property most strongly correlated within a family of proteins with similar folds.

A recent study to characterize disordered regions of proteins by a large number of physical–chemical properties ranks hydrophobicity as the major factor. [49] Interestingly the same property, the 14 Å contact number for amino acids, [50] which had the highest ability to discriminate in the Williams et al. study, also shows the highest correlation with component E1. We are currently in the process of determining the discriminatory power of our five components to find remote true homologues of proteins.

In a different approach to reduce the redundancy of existing property scales of amino acids, Tomii and Kanehisa [51] performed a cluster analysis of an exhaustive list of properties. Five out of six of their major clusters correspond to the five components in our studies. Although this work includes an exhaustive list of properties, its qualitative nature limits its application in sequence studies. In an earlier study Scheraga and coworkers [52] found ten factors by analyzing 188 physical–chemical properties in a combination of cluster analysis and multivariate factor analysis. In our work we show that five descriptors are sufficient to reproduce the original space, thus further reduction is possible, and the dominant factor 1 found in our studies correlates with hydrophobicity, whereas factor 1 in their study is related to α-helical propensity.

### Reduction of the amino acid alphabet

Our cluster analysis of the 20 amino acids based on the property distances can be used to define a reduced amino acid alphabet for protein folding studies. Wang and Wang [53] simplified the protein folding alphabet to five groups (CMFILVWY ATH GP SNQRK DE) by reducing the statistical contact potential of the MJ matrix. [39] Our clustering is similar, except that the amino acids G, P, A and C are in separate clusters up to a high distance cutoff in our method. This result is consistent with experiments that have shown that G and P are absolutely essential to preserve the 3D fold. [54, 55]

Our 5D descriptors measure differences in protein sequences by physical–chemical properties in a concise and quantitative way. Distances derived from these descriptors correlate well with similarity scores derived from substitution matrices. Future applications of the descriptors include property-based alignment procedures and fold recognition. Our descriptors should also be useful for finding sequence motifs based on conserved properties of protein families.

*Electronic supplementary material*. The list of 237 physical–chemical properties with references (properties.html) is available as electronic supplementary material.

**Acknowledgements** We thank Dr. C.H. Schein for critical reading of the manuscript and C.J. Orlea for help in editing. This work was funded by grants from the Department of Energy (DE-FG03-00ER63041), the Texas Higher Education Coordinating Board (ARP 004952-0084-1999) and the John Sealy Memorial Endowment Fund (2535-01).

## References

1. Abagyan RA, Batalov S (1997) J Mol Biol 273:355
2. Koshi JM, Goldstein RA (1997) Proteins: Struct Funct Genet 27:336
3. Buchler NE, Goldstein RA (1999) Proteins: Struct Funct Genet 34:113
4. Azarya-Sprinzak E, Naor D, Wolfson HJ, Nussinov R (1997) Protein Eng 10:1109
5. Grigoriev IV, Kim SH (1999) Proc Natl Acad Sci USA 96: 14318
6. Gribskov M, Veretnik S (1996) Methods Enzymol 266:198
7. Aravind L, Koonin EV (1999) J Mol Evol 48:291
8. Moult J (1999) Curr Opin Biotech 10:583
9. Koretke KK, Russell RB, Copley RR, Lupas AN (1999) Proteins: Struct Funct Genet 37:141
10. Karplus K, Sjolander K, Barrett C, Cline M, Haussler D, Hughey R, Holm L, Sander C (1997) Proteins: Struct Funct Genet Suppl 1:134
11. Karplus K, Barrett C, Hughey R (1998) Bioinformatics 14:846
12. Krogh A, Brown M, Mian I, Sjolander K, Haussler D (1994) J Mol Biol 235:1501
13. Levitt M, Gerstein M (1998) Proc Natl Acad Sci USA 95:5913
14. Huang ES, Subbiah S, Tsai J, Levitt M (1996) J Mol Biol 257:716
15. Jones DT, Bryson K, Tress ML, Hadley C (1999) Proteins: Struct Funct Genet Suppl 3:104
16. Rost B, Schneider R, Sander C (1997) J Mol Biol 270:471
17. Thiele R, Zimmer R, Lengauer T (1999) J Mol Biol 290:757
18. Sippl M, Lackner P, Domingues F, Koppensteiner W (1999) Proteins: Struct Funct Genet 37:226
19. Domingues FS, Lackner P, Andreeva A, Sippl M (2000) J Mol Biol 297:1003
20. Davison M (1983) Multidimensional scaling. In: Wiley series in probability and mathematical statistics. Wiley, New York
21. Trosset M (2000) Comput Optimization Appl 17:11
22. Tsogo L, Masson M, Bardot A (2001) IEEE Trans Syst Man Cybernetics Part A: Syst Humans 31:30
23. Havel TF, Kuntz IW, Crippen GM (1983) Bull Math Biol 45:665
24. Suyama M, Matsuo Y, Nishikawa K (1997) J Mol Evol 44:S163
25. Yona G, Levitt M (2000) Proc International Conference on Intelligent Systems for Molecular Biology 8:395
26. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, D.C. pp 345–352
27. Henikoff S, Henikoff J (1992) Proc Natl Acad Sci USA 89:10915
28. Naor D, Fischer D, Jernigan LR, Wolfson HJ, Nussinov R (1996) J Mol Biol 1996:924
29. Gonnet GH, Cohen MA, Benner SA (1992) Science 256:1443
30. Hänggi G, Braun W (1994) FEBS Lett 344: 147
31. Zhu H, Braun W (1999) Protein Sci 8:326

32. Zhu H, Schein CH, Braun W (2000) Bioinformatics 16:950
33. Press WH, Teukolsky SA, Vettering WT, Flannery BP (1999) Numerical recipes in C. Cambridge University Press, New York
34. Felsenstein J (1985) Evolution 39:783
35. Fitch WM, Margolish E (1967) Science 15:299
36. Bairoch A, Apweiler R (1999) Nucleic Acids Res 27:49
37. Kawashima S, Ogata H, Kanehisa M (1999) Nucleic Acids Res 27:368
38. Grantham R (1974) Science 185:862
39. Miyazawa S, Jernigan RL (1985) Macromolecules 1985:534
40. Rao JKM (1987) Int J Pept Protein Res 29:276
41. Johnson MS, Overington JP (1993) J Mol Biol 233:716
42. Prlic A, Domingues FS, Sippl MJ (2000) Protein Eng 13:545
43. Brenner SE, Levitt M (2000) Protein Sci 9:197
44. Brenner SE, Chothia C, Hubbard TJ (1998) Proc Natl Acad Sci USA 95:6073
45. Kim SH (2000) Curr Opin Struct Biol 10:380
46. Wood RD, Mitchell M, Sgouros J, Lindahl T (2001) Science 291:1284
47. Rohde K, Bork P (1993) CABIOS 9:183
48. Otsuka J, Kikuchi N, Kojima S (1999) Biochim Biophys Acta 1434:221
49. Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, Takayama S, Brown CJ, Dunker AK (2001) Pac Symp Biocomput 89
50. Nishikawa K, Ooi T (1986) J Biochem 100:1043
51. Tomii K, Kanehisa M (1996) Protein Eng 9:27
52. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA (1985) J Protein Chem 4:23
53. Wang J, Wang W (1999) Nat Struct Biol 6:1033
54. Schafmeister CE, LaPorte SL, Miercke LJW, Stroud RM (1997) Nat Struct Biol 4:1039
55. Riddle DS, Santiago JV, Bray HST, Doshi N, Grantcharova VP, Yi Q, Baker D (1997) Nat Struct Biol 4:805
56. Meirovitch H, Rackovsky S, Scheraga HA (1980) Macromolecules 13:1398
57. Parker JMR, Guo D, Hodges RS (1986) Biochemistry 25:5425
58. Cid H, Bunster M, Canales M, Gazitua F (1992) Protein Eng 5:373
59. Pliska V, Schmidt M, Fauchere JL (1981) J Chromatogr 216:79
60. Biou V, Gibrat JF, Levin JM, Robson B, Garnier J (1988) Protein Eng 2:185
61. Manavalan P, Ponnuswamy PK (1978) Nature 275:673
62. Oobatake M, Ooi T (1977) J Theor Biol 67:567
63. Fauchere JL, Pliska V (1983) Eur J Med Chem 18:369
64. Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V (1988) Int J Pept Protein Res 32:269
65. Levitt M (1976) J Mol Biol 104:59
66. Chothia C (1976) J Mol Biol 105:1
67. Hutchens JO (1970) Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds. In: Sober HA (ed) Handbook of biochemistry, 2nd edn. Chemical Rubber Co, Cleveland, Ohio, pp B60–B61
68. Fasman GD (1976) Handbook of biochemistry and molecular biology. CRC Press, Cleveland, Ohio
69. Sneath PHA (1966) J Theor Biol 12:157
70. Chothia C (1975) Nature 254:304
71. Crawford JL, Lipscomb WN, Schellman CG (1973) Proc Natl Acad Sci USA 70:538
72. Beghin F, Dirkx J (1975) Arch Int Physiol Biochim 83:167
73. Robson B, Suzuki E (1976) J Mol Biol 107:327
74. Chou PY, Fasman GD (1978) Annu Rev Biochem 47:251
75. Kanehisa MI, Tsong TY (1980) Biopolymers 19:1617
76. Palau J, Argos P, Puigdomenech P (1981) Int J Pept Protein Res 19:394
77. Isogai Y, Nemethy G, Rackovsky S, Leach SJ, Scheraga HA (1980) Biopolymers 19:1183
78. Maxfield FR, Scheraga HA (1976) Biochemistry 15:5138
79. Cohn EJ, Edsall JT (1943) Proteins, amino acids, peptides. Reinhold, New York
80. Nakashima H, Nishikawa K, Ooi T (1990) Proteins: Struct Funct Genet 8:173
81. McCaldon P, Argos P (1988) Proteins 4:99
82. Bull HB, Breese K (1974) Arch Biochem Biophys 161:665
83. Jones DT, Taylor WR, Thornton JM (1992) CABIOS 8:275
84. Jukes TH, Holmquist R, Moise H (1975) Science 189:50
85. Burgess AW, Ponnuswamy PK, Sheraga HA (1974) Isr J Chem 12:239
86. Munoz V, Serrano L (1994) Proteins 20:301
87. Geisow MJ, Roberts RDB (1980) Int J Biol Macromol 2:387
88. Browne CA, Bennet HPJ, Solomon S (1982) Anal Biochem 124:201
89. Klein P, Kanehisa M, DeLisi C (1984) Biochim Biophys Acta 787:221
90. Tanaka S, Scheraga HA (1977) Macromolecules 10:9
91. Benner SA, Cohen MA, Gonnet GH (1994) Protein Eng 7:1323
92. Levin JM, Robson B, Garnier J (1986) FEBS Lett 205:303
93. Luthy R, McLachlan AD, Eisenberg D (1991) Proteins 10:229