# Biological Meaning of Nonlinear Mapping of Substitution Matrices in Metric Spaces

Juan Mendez and Javier Lorenzo

Institute of Intelligent Systems(SIANI)
Univ. Las Palmas de Gran Canaria
35017 Las Palmas, Canary Islands, Spain
Email: jmendez@iusiani.ulpgc.es

***Abstract*—Based on the Information Theory, a distance function between amino acids is induced to model the affinity relationship between them. Tables with mapping coordinates have been obtained by using nonlinear multidimensional scaling for different dimensions. These mapping coordinates are meaningless virtual data, but a high relationship with physical and chemical properties is found. The main conclusion is that the number of effective characteristics involved in substitution matrices is low. The hypothesis that hydrophobicity and secondary structure propensities are very important characteristics involved in substitution matrices is reinforced by the analysis of the results.**

***Index Terms*—Substitution Matrices; Molecular Evolution; Principal Component Analysis; Multidimensional Scaling; Information Theory**

## I. Introduction

The main assumption of the neutral theory of evolution[1] is that mutations at the molecular level are mainly neutral or weakly disadvantageous. Moreover, it asserts that substitution rates in low biologically relevant sites in proteins are greater than the active rates. The role of physical-chemical properties in evolutionary models has been emphasized by considering that local mutations tend to conserve important properties of the protein and that the knowledge implicitly contained in substitution matrices is related to the properties of protein domains rather the amino acids themselves[2]. This approach has been extended for the detection of protein domains in the space of property sequence instead of the primary sequence[3][4][5].

Most models of substitution matrices are computed as log-odds ratio of sequence probabilities. This is related to the concept of relative entropy or mutual information between two sequences. Intuitively, sequences with high homological relationship have high levels of common information coming from their ancestor, while distantly related sequences retain low levels of common information. Formally, the relationship between substitution matrices and information theory has been presented by Altschul[6] and is based on the statistical properties associated with general score matrices[7]. This theoretical approach provides a sound framework for substitution matrices that is used to propose an amino acid distance based on the evolutionary data.

Amino acids can be organized into physical-chemical groups with similar relationship between their members. The substitution matrices collected from sets of related sequences are statistical summaries of the evolution process[8]; they show that similar amino acids tend to replace each other more often than dissimilar ones in accordance with the rather neutral action of most mutations, as asserted by the neutral theory of evolution[1]. Yang [9] determines an amino acid substitution model from a codon substitution model at DNA level by using a Markov model. His conclusion is that a relationship exits between the acceptance rates of probabilities –from which the substitution matrices are computed – and the chemical properties, but this relationship may not be simple.

The conservational or homeostatic nature of physical-chemical characteristics in mutational process has also been considered [10] by focussing on the amino acid isoelectric point, which combines both electric and spatial information. Those authors study the effect of co-ordinated substitution (pairs of related amino acid substitutions) and the correlation with protein changes in physical-chemical properties. The study concludes that co-ordinated substitutions play a compensatory role. The relationship between mutation and place of amino acids in the protein was analyzed by Gromiha and Oobatake[11]. They studied the correlation between stability changes caused in buried amino acids and 48 physicochemical and conformal properties.

The study of the relationship between physical-chemical parameters in structure-dependent matrices[12] finds a high correlation between the matrices and some parameters, particularly for hydrophobicity and amino acid volume. Both of them seem to be important factors for the protein folding and in its functionality. Volume is important due to the physical constraints in the torsion angles of the peptide chain, while hydrophobicity is important in the interactions of residues with the aqueous environment. Koshi and Goldstein[12] use the correlation between the matrix $\Delta Q_{ab} = |Q(a) - Q(b)|$ related to a characteristic $Q$ of amino acids, such as the contained in the AAindex database[13], and the functional $\Phi_{ab}(M)$ where $M$ is the substitution matrix. The matrix $\Delta Q_{ab}$ is a distance in the characteristic domain, but they use the functional $\Phi_{ab}(M) = \ln(M_{ab}M_{ba})$. Since $\Delta Q_{aa} = \Delta Q_{bb} = 0$ and that in the general case, $\Phi_{aa} \neq \Phi_{bb}$, some perturbations could be introduced in the correlation values. A more homogeneous correlation must be between $\Delta Q$ and some functional verifying $\Phi_{aa}(M) = 0$ which is more similar to a distance. In this paper is suggested that some type of amino acid distance is a more correct measure to be correlated with $\Delta Q$, but it is proposed that this distance be obtained from substitution matrices due to their biological richness and the soundness of their theoretical background.

A set of orthogonal linear amino acid characteristics for clustering purposes has been obtained[14] from the eigenvectors of the PCA selection procedure from a wide characteristic set. They are significant among the characteristic set itself, but there are doubts about whether they are significant in the relationship with the substitution matrices. Also, an amino acid index was proposed to maximize its relationship to structural properties form a Machine Learning approach [15]

Venkatarajan and Brown [16] have represented the animo acids in a high dimensional space of 237 properties, and by using Principal Component Analysis(PCA) have used multidimensional scaling to generate linear mapping. Their conclusion is that 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, since the distances computed for pairs of amino acids in the five-dimensional property space are highly correlated with corresponding scores from similarity matrices

The converse approach to map the characteristic based multidimensional representation of amino acids is to map the evolutionary distance of amino acid. There are previous works in mapping substitutions matrices in virtual spaces from the Pattern Matching perspective in information retrieval[17]. An amino acid distance can be obtained from the Information Theory by mining the knowledge about the dimensionality of substitution matrices. The proposed distance is a *evolutionary distance* between amino acids; it is related to a biological environment, as general/local as the substitution matrix itself. The mapping of this distance in a multidimensional space provides knowledge about the intrinsic dimensionality of this information. However, the distance used by some autors[18] is heuristic. The main goal of this paper is to propose an well found distance based on the Information Theory and also an attempt to identify the meaning of some coordinates provided by the nonlinear mapping.

The remainder of this paper is organized in sections covering the methods, results and discussion. In the methods section, the concept of distance entropy is used to generate an amino acid distance based on the Information Theory. The results and discussion sections discuss about the meaning of some mapping coordinates. The virtual coordinates are meaning-less, but some information about their semantic can be discovered by obtaining correlations between individual coordinates with physical-chemical and structural properties.

## II. Methods

The methodology developed to study the intrinsic dimensionality of substitutions matrices is based on the use of an entropy distance among amino acids and its multidimensional scaling into a virtual space. Although the analytical expression of this distance is very simple and heuristically inferible, it can be obtained from the sound theoretical framework of information theory.

Protein sequences can be formally represented as random distributions of symbols contained in the set $\mathbf{A}$ of amino acid. Let $q_{ab}$ be the probability distribution on the $\mathbf{A}\times\mathbf{A}$ alphabet of the alignment of pairs of homologous sequences in a defined biological framework. Let $p_a$ be the probability distribution on the projection of the pairwise distribution: $\mathbf{A}\times\mathbf{A}\rightarrow\mathbf{A}$. This projection destroys the information about the probabilities of pair $(a, b)$, so that the probability $q_{ab}$ cannot be recovered from $p_a$. The choice of the set of homologous sequences used to compute the distribution $q_{ab}$ defines the biological context for the computation of alignments. Different biological environments are possible, as defined in the different BLOSUM or PAM matrices. The product value $p_a p_b \neq q_{ab}$ is the probability of a pair taken from two independent random sequences, while $q_{ab}$ is the probability among related sequences, thus it contains the information about these relations. The mutual information, $I(A, B)$, between two sequences $A$ and $B$ is defined as [19], [20]:

$$I(A, B) = H(A) + H(B) - H(A, B) \tag{1}$$

where $H(A, B)$ is the entropy of $\mathbf{A}\times\mathbf{A}$. Both, $H(A)$ and $H(B)$ are the entropy of random sequences with the same probability distribution, consequently they become identical to the entropy of $\mathbf{A}$. The mutual information is computed as:

$$I(A, B) = \sum_{ab} q_{ab} \log \frac{q_{ab}}{p_a p_b} \tag{2}$$

This expression also can be interpreted as the relative entropy – or Kullback-Leibler distance– between distributions $q_{ab}$ of related pairs and $p_a p_b$ of independent pairs. A goal of alignment statistics is to define useful tables that capture the biological significance of a set of related sequences. A substitution matrix $s(a, b)$ is introduced as the log-odds between the relationship probability $q_{ab}$ and the independent probability $p_a p_b$, so that the mutual information becomes the expected value of this matrix: $I(A, B) = E[s(a, b)] = \sum_{ab} q_{ab} s(a, b)$. The substitution matrix is the additional information needed to relate both probabilities. Thus, it can be interpreted as the information lost in the projection $\mathbf{A}\times\mathbf{A}\rightarrow\mathbf{A}$, such as[7]:

$$q_{ab} = p_a p_b e^{\lambda s(a,b)} \tag{3}$$

where $\lambda$ is introduced according to the base of the logarithm. Similarly to the the mutual information between sequence distributions, the distance entropy between two distributions is defined as[21]:

$$D(A, B) = H(A, B) - I(B, A) \tag{4}$$

which is related to the no common or uncorrelated properties of both distributions –denoting difference– while the mutual information is related to the common properties and denotes similarity. This distance entropy function is a metric; it verifies the basic properties of a distance: the symmetrical property: $D(A, B) = D(B, A)$, it has a null lower bound:

$D(A,B) \geq 0$ and also verifies that $D(A,A) = 0$. In addition, it is a metric because it verifies the triangular, $D(A,B)+D(B,C) \geq D(A,C)$, and the if-only-if properties, $D(A,B) = 0 \leftrightarrow A \equiv B$. The distance entropy can be computed from the distributions probability as a kind of relative entropy:

$$D(A,B) = \sum_{ab} q_{ab} \log \frac{p_a p_b}{q_{ab}^2} \tag{5}$$

This paper proposes that this distance can be considered as the expectation of the entropy distance between amino acids $D(a,b)$ such as: $D(A,B) = E[D(a,b)] = \sum_{ab} q_{ab} D(a,b)$. It can be expressed as:

$$D(a,b) = \frac{1}{2}[D(a,a) + D(b,b)] + d(a,b) \tag{6}$$

with the inclusion of an auxiliary distance $d(a,b)$ defined as: $d(a,b) = s(a,a) + s(b,b) - 2s(a,b)$, and where $D(a,a) = 2\log p_a/q_{aa}$. It is verified that $p_a \geq q_{aa}$, thus $D(a,a) \geq 0$. The substitution matrix $s(a,b)$ is not a "perfect" similarity value: $s(a,a) \neq s(b,b)$. Similarly, $D(a,b)$ is not a distance: $D(a,a) \neq D(b,b)$. However, the auxiliary distance $d(a,b)$ has the properties of a distance matrix:

$$d(a,b) = d(b,a) \qquad d(a,b) \geq 0 \qquad d(a,a) = 0 \tag{7}$$

but it is not a metric in the general case. The verification of the if-only-if and triangular properties depends on the $s(a,b)$ values. E.g. the if-only-if metric property which requires that: $d(a,b) = 0 \leftrightarrow a = b$ greatly depends on whether the inequality $s(a,b) \leq s(a,a)$ can be transformed to the most restrictive condition: $s(a,b) < s(a,a)$.

The matrix $d(a,b)$, which is an evolutionary distance, has an advantage over other approaches in that it can be directly computed from the substitution matrix. However, *this is not a general purpose distance between amino acids but it is an evolutionary distance in a defined biological environment, as general or as specific as the substitution matrix from which it is obtained.*

This distance is a dimensionless concept. However, the possibility of a dimensional representation is considered. Moreover, it is used to make inferences about the intrinsic dimensionality, this possibility can be useful for many machine learning and data mining procedures that are mainly oriented to coordinate based representation. A distance matrix between amino acids is a dimensionless relationship, nothing relates it with a multidimensional coordinates system. By using nonlinear multidimensional scaling procedures, these dimensionless relationship were mapped into a virtual coordinate system[18]. The nonlinear mapping procedure[22] is based o the minimization of an error function between the amino acid distances $d(a,b)$ and their distances in a dimensional representation $\delta(a,b) = ||\mathbf{X}_a - \mathbf{X}_b||$, where $\mathbf{X}_a \in \mathbf{R}^n$ are the coordinates of an amino acid. This error function, $G(\mathbf{X}, n)$, depends on the coordinates, $\mathbf{X}$, and the dimensionality, $n$. An genetic algorithm generates the optimal solutions for each $n$ and the minimal value of $G^*(n)$ is shown in Figure 1. The coordinates $\mathbf{X} \in \mathbf{R}^n$ were modified to generate the $\mathbf{Y} \in \mathbf{Z}^n$ ones, such as $\delta(a,b) = \rho||\mathbf{Y}_a - \mathbf{Y}_{b||}$. This integer coordinates can be useful to implement advanced linear Pattern Matching procedures to provide fast sequence alignments. Table I contains these coordinates for BLOSUM 62 substitution matrix and different dimensions.

A quantitative analysis of the tentative semantics of mapping coordinates can be achieved by using the linear regression of some characteristic with the coordinate values. A linear model of the characteristic $Q$ in the $n$ dimensional space of $\mathbf{Y}$ coordinates can be obtained by the expression:

$$Q = \overline{Q} + \sum_{i=1}^{n} A_i (Y_i - \overline{Y}_i) \tag{8}$$

where $\overline{Q} = E[Q]$ is the mean of the characteristic, and $\overline{Y}_i = E[Y_i]$. The vector $\mathbf{A}$ can be computed as: $\mathbf{A} = \Sigma(\mathbf{Y},\mathbf{Y})^{-1}\Sigma(Q,\mathbf{Y})$, where $\Sigma(\mathbf{Y},\mathbf{Y})^{-1}$ is the inverse matrix of $\Sigma(\mathbf{Y},\mathbf{Y})$ the covariance of the coordinates defined as: $\Sigma(\mathbf{Y},\mathbf{Y})_{ij} = E[(Y_i - \overline{Y}_i)(Y_j - \overline{Y}_j)]$ and $\Sigma(Q,\mathbf{Y})$ is the covariance matrix between the characteristic $Q$ and the coordinates: $\Sigma(Q,\mathbf{Y})_i = E[(Q-\overline{Q})(Y_i - \overline{Y}_i)]$

The correlation coefficient of this multidimensional regression is computed as:

$$R_{Q,\mathbf{Y}} = \sqrt{\frac{\Sigma(Q,\mathbf{Y})^T \Sigma(\mathbf{Y},\mathbf{Y})^{-1} \Sigma(q,\mathbf{Y})}{\Sigma(Q,Q)}} \tag{9}$$

This coefficient is independent of rotations in the coordinate axes. The correlation coefficients between the characteristic and individual coordinates, which are dependent on rotations, are computed as:

$$R_{Q,Y_i} = \frac{\Sigma(Q,Y_i)}{\sqrt{\Sigma(Y_i,Y_i)\Sigma(Q,Q)}} \tag{10}$$

## III. Results

The methodology proposed in this paper can be applied to any substitution matrix. As an illustration, experimental results are generated for a test case: the BLOSUM 62 substitution matrix. Figure 1 shows the graphical representation of the optimal value $G^*$ of the goal function vs the dimensionality $n$. Fast convergence with monotonic decreasing in the goal function is obtained. As the dimensionality increases, a high decreasing of the marginal relevance is obtained. Therefore, after some small dimensionality values, little additional gain can be obtained with additional dimensions. This could be interpreted as most of the information contained in the substitution matrix is related with a few orthogonal –independent– factors.

The general formulation of the data mining and pattern analysis problem to be solved in order to discover the meaning of virtual coordinates is as follows: how to relate the virtual coordinates $X_i(a)$ or $Y_i(a)$ obtained from the mapping of the distance $d(a,b)$ to a set of characteristic $Q_j(a)$ with previous semantic, which are, in general, not orthogonal. This is an open problem, therefore, rather than provide a solution for this general problem, this paper analyses some clues about the

TABLE I
THE $\mathbf{Y}$ COORDINATES FROM 1 TO 5 DIMENSIONS FOR BLOSUM 62.

| n | 1 | 2 | | 3 | | | 4 | | | | 5 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aa | $Y_1$ | $Y_1$ | $Y_2$ | $Y_1$ | $Y_2$ | $Y_3$ | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $Y_5$ |
| A | 140 | 111 | 84 | 148 | 142 | 97 | 102 | 105 | 63 | 116 | 102 | 50 | 67 | 207 | 45 |
| R | 180 | 69 | 175 | 21 | 48 | 115 | 72 | 2 | 162 | 96 | 200 | 45 | 15 | 202 | 156 |
| N | 203 | 3 | 134 | 74 | 166 | 199 | 42 | 95 | 174 | 47 | 235 | 40 | 72 | 136 | 64 |
| D | 216 | 0 | 80 | 45 | 212 | 139 | 64 | 114 | 100 | 0 | 184 | 147 | 92 | 97 | 43 |
| C | 31 | 220 | 37 | 229 | 160 | 0 | 95 | 44 | 26 | 251 | 39 | 34 | 255 | 225 | 54 |
| Q | 164 | 65 | 131 | 52 | 93 | 154 | 134 | 44 | 149 | 51 | 178 | 127 | 85 | 181 | 158 |
| E | 190 | 31 | 112 | 24 | 138 | 144 | 104 | 56 | 123 | 22 | 190 | 136 | 62 | 138 | 123 |
| G | 228 | 31 | 34 | 166 | 209 | 170 | 0 | 160 | 92 | 99 | 127 | 8 | 7 | 113 | 0 |
| H | 239 | 35 | 219 | 38 | 61 | 239 | 115 | 103 | 252 | 63 | 187 | 26 | 47 | 47 | 181 |
| I | 97 | 173 | 99 | 174 | 57 | 68 | 182 | 102 | 95 | 155 | 40 | 20 | 86 | 242 | 126 |
| L | 89 | 171 | 125 | 148 | 30 | 59 | 170 | 76 | 116 | 175 | 41 | 70 | 92 | 241 | 144 |
| K | 173 | 53 | 153 | 24 | 92 | 90 | 89 | 0 | 120 | 73 | 173 | 106 | 0 | 209 | 117 |
| M | 115 | 142 | 137 | 108 | 34 | 62 | 162 | 44 | 142 | 157 | 87 | 65 | 88 | 241 | 180 |
| F | 59 | 177 | 181 | 176 | 1 | 148 | 150 | 169 | 158 | 187 | 12 | 24 | 65 | 127 | 184 |
| P | 255 | 87 | 0 | 0 | 156 | 18 | 129 | 73 | 0 | 31 | 86 | 199 | 65 | 205 | 16 |
| S | 154 | 74 | 99 | 103 | 149 | 119 | 59 | 88 | 103 | 97 | 147 | 66 | 94 | 164 | 59 |
| T | 131 | 107 | 57 | 103 | 139 | 47 | 63 | 57 | 83 | 146 | 171 | 39 | 137 | 228 | 86 |
| W | 0 | 200 | 255 | 238 | 22 | 255 | 38 | 128 | 241 | 255 | 0 | 86 | 170 | 0 | 180 |
| Y | 71 | 133 | 213 | 127 | 0 | 193 | 139 | 165 | 207 | 140 | 78 | 0 | 103 | 101 | 215 |
| V | 105 | 158 | 90 | 166 | 78 | 65 | 178 | 96 | 89 | 133 | 62 | 10 | 84 | 248 | 115 |
| $\rho$ | 0.144 | 0.113 | | 0.088 | | | 0.089 | | | | 0.075 | | | | |

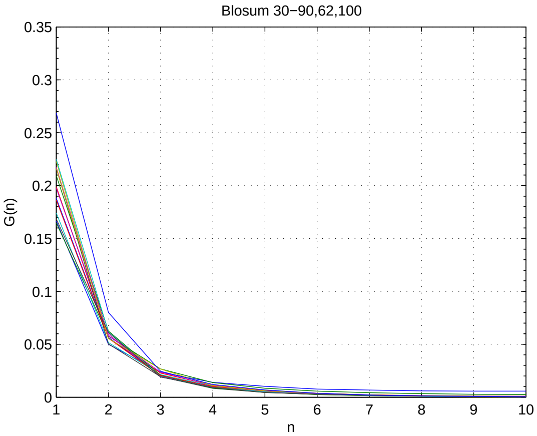

Fig. 1. Optimal goal value $G^*(n)$ for BLOSUM matrices from 30 to 90, 62 and 100.

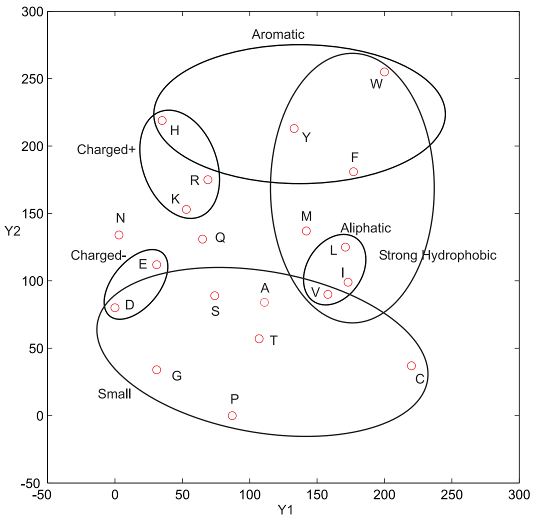

Fig. 2. Spatial representation for $n = 2$ of the mapping for BLOSUM 62 of some amino acid and their groups.

relationship with previously established relevant characteristics such as hydrophobicity, volume, solvent accessibility and secondary structure propensity.

Firstly, as an illustrative contribution on the discovery of some semantic in the virtual coordinates, a high spatial organization of the amino acid groups can be found in the results provided. Amino acids can be grouped according to their physical-chemical properties. Some groups –aliphatic or aromatic– are related to the chemical structure; others –tiny or small– are grouped by molecular size; the polar and charged groups are related to the electric activity, and hydrophobic group is related to their affinity with water. It has been shown [23] that the amino acid groups have a high level of spatial organization when they are mapped in a two dimensional space obtained from the reduction of the whole AAindex database [13] to two index according to their correlations. On that line, but from a different approach, Figure 2 shows the two dimensional mapping of the amino acids using the $\mathbf{Y}$ coordinates obtained from Table I.

The small group –with the exception of N amino acid– is mapped at the bottom of the map. It forms a regular convex region at low $Y_2$ coordinate and extends along the whole range of $Y_1$ one. A subsequent quantitative analysis based on correlations will show that the $Y_2$ coordinate has a high correlation level with physical amino acid dimensions; relationship that is qualitatively suggested by Figure 2. The aromatic group –VIL– comprises a well defined cluster included within the strong hydrophobic group –WFYMVIL– located at higher values of the $Y_1$ coordinate. The opposite groups –charged– have the lowest values in this coordinate. Also as in amino acid size in relation to $Y_2$, it is shown that the $Y_1$ coordinate has a high correlation with hydrophobicity.

The hydrophobic –or its opposite hydrophilic– property of amino acids is fundamental in the dynamics and structure of proteins[24]. Due to that the biological matter is basically an aqueous solution, the water affinity is essential in the relation of a protein with its environment. The mutations with significant changes in the water affinity have a high probability of generating disfunctions, and consequently they have a low survival probability, therefore, there are lost in the evolution process. There are many hydrophobicity scales for amino acid residues in proteins [25], [26], [27], [28]. Every scale is obtained from different experimental measurement criteria, but all are qualitatively related to the hydrophobicity property.

Some characteristics are used for a tentative analysis to discover the meaning of some virtual coordinates. Hydrophobic scales,such as Levitt, $H_l$, Kyte-Doolittle, $H_{kd}$, and Karplus, $H_k$ are considered. Size related characteristics, such as amino acid volume [29], $V$, and active surface [30], $S$, are used, as well as secondary structure properties, such as the classical Chou-Fasman parameters for $\alpha$-helix, $\beta$-sheet and turn propensity[31], $A_{cf}$,$B_{cf}$,$T_{cf}$. One advantage of the Chou-Fasman propensity scales is that they are averaged over all possible environments[32]. Additional scales[33], [34], [35], $A_b$,$B_{kb}$,$T_m$, contained in the AAindex database are also included. Three scales,$L_1, L_2, L_3$, of solvent accessibility derived from Bordo-Argos[36] are included. To avoid problems related to the different scale magnitudes, a normalization is performed as: $Q'_i = (Q_i - \overline{Q_i})/\sigma_i$, where $\overline{Q_i}$ is the mean value and $\sigma_i$ is the standard deviation of $Q_i$.

Ideally, a good set must contain low correlation values for characteristics of different meaning and high values for those with the same semantic. In practice, this is far from being achieved. Table II shows the correlation coefficient between these characteristics.

Table III shows the correlation coefficient $R_{q,\mathbf{Y}}$ among characteristics and mapping coordinates for a dimensionality of from 1 to 3. Also, the correlations with individual coordinates $R_{q,Y_i}$ are included for $n = 2$. For a dimensionality of from 1 to 3, the more significant characteristics are emphasized. The procedure to interpret this Table is as follows: for each dimensionality $n$, the $n$ characteristics with highest correlation values but with lowest correlation values between themselves must be selected. This can be a difficult task depending on the correlation values between characteristics.

For $n = 1$ the Monne *et al.* turn propensity (MTP) is the most significant, but is little different from the Chou-Fasma sheet propensity. For $n = 2$ volume has the highest correlation value(0.894). The next characteristic is surface(0.866), but this characteristic is strongly correlated (0.953) with volume, both are size related characteristics. Instead of surface the next selected characteristic is KDH which is independent(0.047) of volume. KDH and volume are the most correlated with mapping coordinates with the lowest inter-correlation. This means that they can be identified with the two virtual coordinates for this mapping. As qualitatively suggested previously in the graphical representation of two dimensional mapping, there is a high correlation value(0.786) between KDH and the $Y_1$ coordinate, while the volume is better correlated (0.817) with the $Y_2$ coordinate. This is in line with the relevance of both KDH and volume in substitution matrices.

In general, coordinate rotations can be used to maximize the correlation coefficient between any coordinate of the rotated space and any desired characteristic. This enables coordinate axes with well defined semantic meaning to be obtained. However, the complexity can explode at higher dimensions because many simultaneous –or sequenced– rotations are required for different alignments of each coordinate. Also in high dimensions, complex rotation matrices are required.

For $n = 3$ the $L_1$ solvent accessibility scale of Bordo-Argos is the most relevant. The next characteristic in order of relevance is the scale $L_3$ also of the solvent accessibility kind, but both are strongly correlated(-0.945); therefore, next one must be selected. In this case, it is volume and the following one is KDH. Therefore, solvent accessibility, volume and KDH seem to be the three main characteristics but, unfortunately KHD and $L_1$ scale are also strongly correlated (0.881).

These illustrates the difficulty of characteristic oriented analysis. A PCA procedure is used to facilitate this task. This provides a set of lineal combination of characteristics that are orthogonal. It requires the eigenvalues $\lambda$ to be obtained in the equation: $(A - \lambda I)X = 0$, where the matrix $A$ is the characteristic covariance. The eigenvectors $V$ are also obtained.

Table IV shows the correlation values for every dimensionality with the eigenvectors. The procedure to analyze the results is that described previously. In this case, the problem related to the correlations between characteristics is avoided because they are orthogonal. The procedure becomes simple because the eigenvector $V_1$ is the most relevant in all cases. It is followed by the $V_2$, and then $V_6$ and finally is $V_{10}$ up to the $n = 4$ dimensionality. The analysis for higher dimensionalities is irrelevant, as concluded from the inference of the intrinsic dimensionality.

Two conclusions can be obtained. The first is that the most important eigenvectors are $V_1$ and $V_2$ with high correlation values in all cases. The $V_6$ eigenvector is relevant, but not excessively. Also, the $V_{10}$ eigenvector is weak, its correlation value(0.531) is low. The second conclusion is that the eigenvalues are not important at all, but the decreasing order in eigenvalues is related to the decreasing order in relevance among eigenvectors and the substitution matrix. To obtain useful results, it is necessary to interpret the eigenvectors. Table V shows the weights that define how to compute each of the main eigenvectors from the normalized characteristics. A simple analysis reveals that the main weight in the $V_1$ vector is the Levitt scale, but due to the negative sign (-0.729) the eigenvector is proportional to hydrophobicity rather than to the hydrophilicity of the Levitt scale. The highest weight in the $V_2$ vector are related to turn propensity while for $V_6$ are in sheet propensity. Finally, one vector $V_{10}$ is low relevant but is mainly related to hydrophobicity –hydrophilicity to be

TABLE II
CORRELATION COEFFICIENT BETWEEN CHARACTERISTICS.

| | $H_{kd}$ | $H_k$ | $V$ | $S$ | $A_{cf}$ | $A_b$ | $B_{cf}$ | $B_{kb}$ | $T_{cf}$ | $T_m$ | $L_1$ | $L_2$ | $L_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $H_l$ | -0.679 | -0.577 | -0.325 | -0.127 | 0.130 | 0.083 | -0.701 | 0.257 | 0.481 | 0.685 | -0.725 | -0.333 | 0.748 |
| $H_{kd}$ | 1.000 | 0.511 | 0.047 | -0.218 | 0.159 | 0.160 | 0.666 | -0.305 | -0.642 | -0.842 | 0.881 | 0.421 | -0.915 |
| $H_k$ | | 1.000 | 0.726 | 0.528 | 0.353 | 0.043 | 0.584 | -0.463 | -0.703 | -0.465 | 0.416 | 0.192 | -0.430 |
| $V$ | | | 1.000 | 0.953 | 0.346 | 0.302 | 0.519 | -0.670 | -0.588 | -0.177 | 0.092 | 0.017 | -0.087 |
| $S$ | | | | 1.000 | 0.319 | 0.256 | 0.319 | -0.615 | -0.419 | 0.054 | -0.148 | -0.074 | 0.155 |
| $A_{cf}$ | | | | | 1.000 | 0.526 | -0.006 | -0.398 | -0.659 | -0.172 | -0.010 | 0.229 | -0.068 |
| $A_b$ | | | | | | 1.000 | 0.386 | -0.513 | -0.539 | -0.366 | 0.165 | 0.182 | -0.206 |
| $B_{cf}$ | | | | | | | 1.000 | -0.611 | -0.683 | -0.697 | 0.683 | 0.215 | -0.672 |
| $B_{kb}$ | | | | | | | | 1.000 | 0.679 | 0.448 | -0.216 | -0.110 | 0.226 |
| $T_{cf}$ | | | | | | | | | 1.000 | 0.633 | -0.575 | -0.328 | 0.614 |
| $T_m$ | | | | | | | | | | 1.000 | -0.733 | -0.447 | 0.793 |
| $L_1$ | | | | | | | | | | | 1.000 | 0.198 | -0.945 |
| $L_2$ | | | | | | | | | | | | 1.000 | -0.509 |

TABLE III
MULTIDIMENSIONAL REGRESSION OF THE CHARACTERISTIC SET VS THE DIMENSIONALITY.

| n | $H_l$ | $H_{kd}$ | $H_k$ | $V$ | $S$ | $A_{cf}$ | $A_b$ | $B_{cf}$ | $B_{kb}$ | $T_{cf}$ | $T_m$ | $L_1$ | $L_2$ | $L_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.624 | 0.606 | 0.489 | 0.549 | 0.382 | 0.189 | 0.447 | 0.761 | 0.618 | 0.683 | $\boxed{0.787}$ | 0.675 | 0.181 | 0.654 |
| 2 | 0.750 | $\boxed{0.846}$ | 0.713 | $\boxed{0.894}$ | 0.866 | 0.318 | 0.509 | 0.800 | 0.682 | 0.783 | 0.812 | 0.814 | 0.300 | 0.815 |
| $Y_1$ | -0.745 | 0.786 | 0.637 | 0.455 | 0.242 | 0.125 | 0.163 | 0.767 | -0.523 | -0.724 | -0.808 | 0.797 | 0.298 | -0.800 |
| $Y_2$ | -0.166 | -0.221 | 0.390 | 0.817 | 0.853 | 0.305 | 0.498 | 0.313 | -0.495 | -0.379 | -0.012 | -0.073 | 0.010 | 0.061 |
| 3 | 0.777 | $\boxed{0.901}$ | 0.827 | $\boxed{0.906}$ | 0.868 | 0.396 | 0.482 | 0.864 | 0.684 | 0.828 | 0.879 | $\boxed{0.925}$ | 0.404 | 0.912 |

precise– and solvent accessibility.

Another significant result obtained from Table III is the low correlation values for the two $\alpha$-helix propensity scales. At high dimensionality the value increases; this means that it is well correlated with some coordinate axes, but its marginal relevance at high dimensionality is low according to what is asserted above.

## IV. DISCUSSIONS

The characteristic oriented analysis shows that hydrophobicity and volume are very relevant. The characteristic pair of volume and Kyte-Doolittle hydrophobicity seems to be a statistically independent set with the highest possibilities of being the main important characteristic. The structural propensity factors are also important but they are high correlated with hydrophobicity. An orthogonal representation based on PCA was used in characteristic dependence analysis. The result shows that hydrophobicity is the most important, followed by turn and sheer propensity. However, neither volume nor solvent accessibility seem to be directly relevant. Turn regions are mainly in the surface of proteins where they are highly exposed to the solvent[37]. The evolution process of insertion and deletion of amino acids is stronger in turn regions than in helix and sheet ones, so they are more flexible in conformational changes but are rich in hydrophilic residues[38]. Therefore, *turn propensity, hidrophiliciy and solvent accessibility are different experimental measures that are related to the dynamic of the protein-water and protein-protein interaction, which globally seems to be the main factor*.

The Koshi and Goldstein assertion that $\alpha$-helical propensity is poorly conserved during evolution, and also that mutations correlate better with $\beta$-sheet than with $\alpha$-helical[41] is confirmed. It is not an specially important factor in modeling the effects of biological evolution at molecular level. It shows high levels of correlation at high dimension but the correlation are poor at low dimensions.

## V. CONCLUSIONS

An amino acid distance based on the evolutionary data contained in the substitution matrix is useful to achieve a characteristic independent procedure to discover its intrinsic dimensionality. The Information Theory provides a sound theoretical background to induce these amino acid distances. The entropy distance defined as a relative entropy has been used to generate a simple but theoretically well-founded distance, which can be obtained directly from the substitution matrix values. In addition, a tentative analysis based on multidimensional linear regression model of the contribution of physical, chemical and structural characteristics has been used to assign possible meaning to the virtual coordinates. The analysis was focused on some previously referenced relevant characteristics such as hydrophobicity, amino acid size, secondary structure propensities and solvent accessibility.

## REFERENCES

[1] M. Kimura, *The neutral theory of molecular evolution*. Cambridge University Press, 1983.

[2] J. M. Koshi, D. P. Mindell, and R. A. Goldstein, “Beyond mutation matrices: physical-chemistry based evolutionary models,” in *Second Int. Conf. on Comp. Mol. Biol.*, 1998, pp. 140–145. [Online]. Available: citeseer.nj.nec.com/314531.html

[3] I. Ladunga, “Physean: Physical sequence analysis for the identification of protein domains on the basis of physical ans chemical properties of amino acids,” *Bioinformatics*, vol. 15, no. 12, pp. 1028–1038, 1999.

[4] I. V. Grigoriev and S.-H. Kim, “Detection of protein fold similarity based on correlation of amino acid properties,” *Proc. Natl. Acad. Sci.*, vol. 96, no. 25, pp. 14 318–14 323, 1999.

[5] G. Mocz, “Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins,” *Protein Science*, vol. 4, pp. 1178–1187, 1995.

TABLE IV
CORRELATION COEFFICIENT FOR PCA

| $n$ | $V_{14}$ | $V_{13}$ | $V_{12}$ | $V_{11}$ | $V_{10}$ | $V_9$ | $V_8$ | $V_7$ | $V_6$ | $V_5$ | $V_4$ | $V_3$ | $V_2$ | $V_1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.039 | 0.113 | 0.031 | 0.087 | 0.221 | 0.208 | 0.236 | 0.053 | 0.027 | 0.070 | 0.270 | 0.067 | 0.068 | $\boxed{0.828}$ |
| 2 | 0.124 | 0.193 | 0.020 | 0.076 | 0.341 | 0.160 | 0.205 | 0.115 | 0.393 | 0.249 | 0.067 | 0.215 | $\boxed{0.766}$ | $\boxed{0.911}$ |
| 3 | 0.107 | 0.179 | 0.191 | 0.049 | 0.329 | 0.193 | 0.487 | 0.139 | $\boxed{0.653}$ | 0.218 | 0.404 | 0.238 | $\boxed{0.814}$ | $\boxed{0.956}$ |
| 4 | 0.292 | 0.247 | 0.160 | 0.525 | $\boxed{0.531}$ | 0.398 | 0.448 | 0.246 | $\boxed{0.603}$ | 0.360 | 0.309 | 0.291 | $\boxed{0.853}$ | $\boxed{0.967}$ |
| $\lambda$ | 0.000 | 0.003 | 0.015 | 0.041 | 0.054 | 0.140 | 0.214 | 0.237 | 0.371 | 0.724 | 0.943 | 1.569 | 3.083 | 6.605 |

TABLE V
WEIGHT OF THE MORE RELEVANT EIGENVECTORS

| | $H_l$ | $H_{kd}$ | $H_k$ | $V$ | $S$ | $A_{cf}$ | $A_b$ | $B_{cf}$ | $B_{kb}$ | $T_{cf}$ | $T_m$ | $L_1$ | $L_2$ | $L_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $V_1$ | $\boxed{-0.729}$ | 0.014 | -0.216 | 0.085 | -0.026 | -0.245 | -0.006 | 0.306 | -0.268 | 0.005 | -0.024 | 0.042 | 0.288 | 0.322 |
| $V_2$ | -0.244 | 0.012 | 0.033 | -0.030 | 0.001 | 0.022 | -0.166 | -0.130 | 0.041 | $\boxed{0.659}$ | $\boxed{0.571}$ | -0.294 | -0.128 | -0.165 |
| $V_6$ | 0.000 | -0.024 | -0.117 | 0.155 | -0.301 | -0.065 | -0.058 | $\boxed{0.415}$ | $\boxed{0.658}$ | -0.167 | 0.269 | 0.045 | -0.304 | 0.250 |
| $V_{10}$ | -0.000 | $\boxed{-0.641}$ | -0.119 | 0.177 | -0.252 | 0.212 | 0.171 | -0.143 | 0.200 | 0.153 | 0.051 | 0.251 | $\boxed{0.491}$ | -0.122 |

[6] S. Altschul, "Amino acid substitution matrices from an information theoretic perspective," *J. Mol. Biol.*, vol. 219, pp. 555–565, 1991.

[7] S. Karlin and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," *Proc. Natl. Acad. Sci.*, vol. 87, pp. 2264–2268, March 1990.

[8] M. Betts and R. Russell, *Bioinformatics for Geneticists*. John Wiley and Sons, 2003, ch. Amino acid properties and consequences of subsitutions.

[9] Z. Yang, "Relating physicochemical properties of amino acids to variable nucleotide substitution patterns among sites," in *Pacific Symposium on Computational Biology*, 2000, pp. 81–92.

[10] D. Afonnikov, D. Oshchepkov, and N. Kolchanov, "Detection of conserved physico-chemical characteristics of proteins by ananlyzing clusters of position with co-ordinated substitutions," *Bioinformatics*, vol. 17, no. 11, pp. 1035–1046, 2001.

[11] M. M. Groniha and M. Oobatake, "Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutattions," *Protein Engineering*, vol. 12, no. 7, pp. 549–555, 1999.

[12] J. M. Koshi and R. A. Goldstein, "Mutation matrices and physical-chemical properties: Correlation and implications," *PROTEINS: Structure, Function and genetics*, vol. 27, pp. 336–344, 1997.

[13] S. Kawashima, H. Ogata, and M. Kanehisa, "Aaindex: amino acid index database," *Nucleic Acids Res.*, vol. 27, pp. 368–369, 1999.

[14] M. S. Venkatarajan and W. Braun, "New quantitative descriptors of amino acids based on multidimensional scaling of a large number of pysical-chemical properties," *J. Mol. Model*, vol. 7, pp. 445–453, 2001.

[15] R. H. Leary, J. B. Rosen, and P. Jambeckz, "An optimal structure-discriminative amino acid index for protein fold recognition," *Biophysical Journal*, vol. 86, pp. 411–419, 2004.

[16] M. S. Venkatarajan and W. Braun, "New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physicalchemical properties," *J Mol Model*, vol. 7, pp. 445–453, 2001.

[17] J. Mendéz, A. Falcón, and J. Lorenzo, "A procedure for biological sensitive pattern maching in protein sequences," *Lecture Notes in Computer Science*, vol. 2652, pp. 547–555, 2003.

[18] J. Mendez, A. Falcon, M. Hernandez, and J. Lorenzo, "Discovering the intrinsic dimensionality of BLOSUM substitution matrices using evolutionary MDS," in *Innovations in Hybrid Intelligent Systems HAIS 2007*, ser. Advances in Soft Computing, A. Abraham, E. Corchado, and J. M. Corchado, Eds., vol. 44. Springer, 2007, pp. 369–376.

[19] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[20] P. Baldi and S. Brunak, *Bioinformatics, The Machine Learning Approach*. MIT Press, 2001.

[21] D. J. MacKay, *Information Theory, Inference and Learning Algoritms*. Cambridge Univ. Press, 2003.

[22] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. John Wiley and Sons, 2001.

[23] C. Hagerty, C. Kulikowski, I. Muchnik, and S. Kim, "Two indeces can approximate 402 amino acid properties," in *Proc. IEEE Int. Symp. Intelligent Control, Intelligent Systems and Semiotics*, 1999, pp. 365–369.

[24] M. Gerstein and M. Levitt, "Simulating water and the molecules of life," *Scientific American*, pp. 100–105, Nov. 1998.

[25] J. Cornette, K. Cease, H. Margalit, J. Spouge, J. Berzofsy, and C. DeLisi, "Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins," *J. Mol. Biol.*, vol. 195, pp. 659–685, 1987.

[26] M. Levitt, "A simplified representation of protein conformations for rapid simulation of protein folding," *J. Mol. Biol.*, vol. 104, pp. 59–107, 1976.

[27] J. Kyte and R. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, p. 105132, 1982.

[28] P. Karplus, "Hydrophobicity regained," *Protein Sci*, vol. 6, no. 6, pp. 1302–1307, 1997.

[29] A. Zamyatim, "Protein volume in solution," *Prog. Biophys. Mol. Biol.*, vol. 24, pp. 107–123, 1972.

[30] C. Chothia, "The nature of the accessible and buried surfaces in proteins," *J. Mol. Biol.*, vol. 105, pp. 1–14, 1975.

[31] P. Y. Chou and G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence," *Adv. Enzymol.*, vol. 47, pp. 45–148, 1978.

[32] T. S. Niwa and A. Ogino, "Multiple regression analysis of beta-sheet propensity of amino acids," *Jour. Mol. Structure(Teochem)*, vol. 419, pp. 155–160, 1997.

[33] M. Blaber, X. J. Zhang, and B. W. Matthews, "Structural basis of amino acid alpha helix propensity," *Science*, vol. 260, pp. 1637–1640, 1993.

[34] C. A. Kim and J. M. Berg, "Thermodynamic beta-sheet properties measured using a zinc-finger host peptide," *Nature*, vol. 362, pp. 267–270, 1993.

[35] M. Monne, M. Hermansson, and G. Heijne, "A turn propensity scale for transmembrane helices," *J. Mol. Biol.*, vol. 288, pp. 141–145, 1999.

[36] D. Bordo and P. Argos, "Suggestions for safe residue sunstitutions in site-directed mutagenesis," *J Mol Biol*, vol. 217, no. 4, pp. 721–729, 1991.

[37] C. Branden and J. Tooze, *Introduction to Protein Structure*. Garland Pub. Inc., 1999.

[38] A. M. Lesk, *Introduction to Protein Architecture*. Oxford University Press, 2001.

[39] C. D. Bustamante, J. P. Townsend, and D. L. Hartl, "Solvent accessibility and puriying selection within proteins of *Escherichia coli* and *Salmonella enterica*," *Mol. Biol. Evol.*, vol. 17, no. 2, pp. 301–308, 2000.

[40] N. Goldman, J. L. Thorne, and D. T. Jones, "Assessing the impact of secondary structure and solvent accessibility on protein evolution," *Genetics*, vol. 149, pp. 445–458, 1998.

[41] M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, and A. Sarai, "Relationship between amino acid properties and protein stabiity: buried mutations," *J Protein Chem.*, vol. 18, no. 5, pp. 565–578, 1999.