Discovering the Intrinsic Dimensionality of BLOSUM Substitution Matrices Using Evolutionary MDS

Juan Méndez, Antonio Falcón, Mario Hernández, and Javier Lorenzo

Intelligent Systems Institute
Univ. Las Palmas de Gran Canaria, Spain

Abstract. The paper shows the application of the multidimensional scaling to discover the intrinsic dimensionality of the substitution matrices. These matrices are used in Bioinformatics to compare amino acids in the alignment procedures. However, the methodology can be used in other applications to discover the intrinsic dimensionality of a wide class of symmetrical matrices. The discovery of the intrinsic dimensionality of substitutions matrices is a data processing problem with applications in chemical evolution. The problem is related with the number of relevant physical, chemical and structural characteristic involved in these matrices. Many studies have dealt with the identification of relevant characteristic sets for these matrices, but few have concerned with establishing an upper bound of their cardinality. The methodology of multidimensional scaling is used to map the substitution matrix information in a virtual low dimensional space. The relationship between the quality of this process and the dimensionality of the mapping provides clues about the number of characteristics which better represents the matrix. To avoid the local minima problem, a genetic algorithm is used to minimize the objective function of the multidimensional scaling procedure. The main conclusion is that the number of effective characteristics involved in substitution matrices is small.

1 Introduction

The molecular evolution predicts that variations in spices are highly related to the physical-chemical factors involved in protein function and folding.The modelling of evolution at protein level is usually accomplished by using matrices of mutation probability among amino acids. These collect the co-occurrence probability of each amino acid pair in homologous sequences, which have some defined evolutionary distance or rate of conserved residues. Therefore, substitution rates at molecular level and physical-chemical properties seem to be of central interest in biological evolution. In special, the knowledge about the effects of amino acid properties in the substitution probability can be of great interest in order to understand the evolution mechanisms.

In practical computational tasks, substitution matrices are introduced to score the substitution of amino acid residues in sequence alignment procedures to reveal homologies. Different substitution matrices can be constructed according with the selection of the set of representative sequences in a biological framework. Eg. PAM matrices[1] are constructed from sequences with evolutionary relations, while BLOSUM matrices[2] are constructed from block sequences that have a similarity ratio.

In Data Mining, the problems concerning with dimensionality reduction or dimensionality discovery must deal with the concept of *intrinsic dimensionality* [3] of their data. In intrinsic dimensionality research, most approaches [4, 5] deal with high dimensional databases and try to reduce the data into a few dimensions by applying methods of multidimensional scaling while the precision in database continues to be high [6]. Intrinsic dimensionality of substitution matrices can be obtained by using two approaches: characteristic independent or dependent. Characteristics dependent analysis are weak techniques because they involve some problems. The main one is related with the choice of the property set. To avoid undesired exclusions, its cardinality must be high. There are some exhaustive compiled sets of amino acid characteristics which cardinality is about several hundreds; one of the best is the AAindex database[7]. The second problem is related to how obtain the intrinsic dimensionality from these massive sets. Feature selection procedures as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are the basic approaches. These procedures provide results as linear combinations of the characteristics contained in the original set. PCA is optimal in applications where the second order statistical parameters – as in gaussian case– define the probabilistic distribution. In the PCA procedure the results are eigenvalues and eigenvectors, that are relevant properties of the characteristics set. A set of orthogonal linear combination of amino acid characteristics for clustering has been obtained [8] from the eigenvectors of the PCA selection procedure from a wide characteristic set. They are significative among the characteristic set itself, but there are doubts about if they are significative in the relationship with the substitution matrices.

The main motivation of this paper is that the discovery of the intrinsic dimensionality of substitution matrices provides an upper bound about the cardinality of the set of relevant properties. Also, the main hypothesis is that the discovery can be directly obtained –independently of the set of properties– from the substitution data itself by using procedures of multidimensional scaling. This paper is mainly concerning with *how many* rather than *what* properties are important in the substitution matrices.

The following of this paper is organized in sections covering the methods, results and conclusion. In the methods section, the substitution matrices are coded based on derived distance matrices. Also, it includes the use of non-linear multidimensional scaling procedures to map the distance in a dimensional space. The result and conclusion sections argue about the intrinsic dimensionality of substitutions matrices.

2 Methods

Substitutions matrices are computed as the log-odds between the relation probability q_{ab} between amino acids pairs in sets of proteins sequences and the independent probability $p_a p_b$

$$s(a,b) = \log \frac{q_{ab}}{p_a p_b} \tag{1}$$

These matrices are symmetrical $s(a,b) = s(b,a)$ and at usual mutation rates verifies $s(a,a) \geq s(a,b)$, but in general the diagonal terms are different: $s(a,a) \neq s(b,b)$. They are like similarity functions, but are not full similarity functions. Many heuristic distance expressions can be obtained from these matrices. The used in this paper is:

$$d(a,b) = s(a,a) + s(b,b) - 2s(a,b) \tag{2}$$

That has the properties of a distance matrix:

$$d(a,b) = d(b,a) \qquad d(a,b) \geq 0 \qquad d(a,a) = 0 \tag{3}$$

But it is no metric in the general case. The verification of additional properties required to be a metric, which are the if-only-if and triangular properties, depends on the $s(a,b)$ values. Eg. the if-only-if metric property, which requires the following property: $d(a,b) = 0 \leftrightarrow a = b$, is verified if the inequality $s(a,b) \leq s(a,a)$ can be transformed in the most restrictive condition $s(a,b) < s(a,a)$.

This distance is not a general purpose distance among amino acids. It is an evolutionary distance in a defined biological environment, as general or specific as the substitution matrix from which is obtained. A distance matrix among amino acids is a dimensional-less relationship, nothing relates it with a multidimensional coordinates system.

2.1 Multidimensional Scaling

Multidimensional scaling[9], or mapping, is the process of finding a configuration of points in a multidimensional space as lower dimensionality as possible, whose interpoint distances correspond, with the lowest error possible, to some previous existent similarities or dissimilarities data[10]. In some cases, the original data have a dimensional representation and mapping tries to find a good representation in a lower dimensional space. In this case, feature reduction procedures as PCA and ICA techniques are used. In other cases –as in this paper– the original distance does not imply a coordinate representation. It is a set of interrelations among concepts without any explicit dimensional counterpart.

Two distance types are involved in the mapping process. The first, $d(a,b)$, is the original coordinate-less measure. The second, $d_X(a,b)$, is the distance in a dimensional space where X denotes the coordinates system of amino acids. The problem to be solved is how to compute $d_X(a,b)$ as a good approximation of $d(a,b)$, which implies the computation of the vector set: $\mathbf{X}(a)$. The Sammon method [11] is a non-linear mapping procedure that provides a good ratio of result quality to computational complexity [12–14]. It maps a distance function to a reduced dimensionality space based on the minimization of an objective function by assigning trial coordinates to each amino acid.

The goal function is related to the relative error between the original dimensional-less distances, $d(a,b)$, and the dimensional ones, $d_X(a,b)$. Consequently, several solutions can be obtained if some local minima exist. The Sammon method requires the minimization of the goal function $G(X)$ which likes a relative error of the mapping process:

$$\min_X G(X) = \frac{\sum_a \sum_{b<a} \frac{[d_X(a,b) - d(a,b)]^2}{d(a,b)}}{\sum_a \sum_{b<a} d(a,b)} \tag{4}$$

The optimal solution $\mathbf{X}$ is not unique. There are some freedom degrees related to the geometrical transformations that preserve the distance d_X, eg. translations, rotations and sign inversion. The dimensional distance function is based on the L_2 norm:

$$d_X(a,b) = \left[\sum_{i=1}^{n} |X_i(a) - X_i(b)|^2\right]^{\frac{1}{2}} \tag{5}$$

To increase the efficiency of computational procedures, the vector $\mathbf{X}(a) \in \mathbb{R}^n$ provided by the optimization procedure is transformed into the $\mathbf{Y}(a)$ vector in the integer range $[0, 255]$ by using geometric transformations of translation and scaling. This discrete version can be coded by using integer arithmetic, more efficient that the float point one. The translation to coordinates origin does not modify the distances, but the scaling to fit the $[0, 255]$ range modifies the distance with a constant factor ρ related with the scaling factor. The relationship between the distances computed by mean of the two vector type is: $d_X(a,b) = \rho d_Y(a,b)$.

Some optimization methods, such as evolutionary and gradient based, can be used to achieve the minimization of the goal function. Gradient procedures have better convergence around a local minimum, while genetic procedures allow a better global optimization by considering several local minima. Many solutions are expected in the proposed problem covering a wide range of local minima due to non-linearity and geometrical transformations. In this paper, a genetic algorithm is used to obtain a solution which is afterward refined by applying a gradient procedure based on the Quasi-Newton algorithm. Genetic algorithms are good to explore the space domain, avoiding the local minima problem. However, in practice after a large number of iterations they are mainly working in the refinement of a local minimum, but the gradient procedures are more efficient for this task.

3 Results

The proposed methodology can be applied to any symmetrical matrix which can generate a distance matrix. The substitution matrices used in Bioinformatics for alignments procedures verify this property. The most used substitution matrices are the BLOSUM family. Figure 1 shows the graphical representation of the optimal value $G(n)$ of the goal function in Equation (4) vs the dimensionality n of the mapping space. A fast convergence toward null values is obtained when the dimensionality increases, which implies a high decreasing in the marginal relevance of additional dimensions. Therefore, after small dimensionality values of three or four, few 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.

Table 1 shows coordinates form 1 to 5 dimensionality. These dimensions are the most relevant. Remark that due to the random nature of genetic algorithms, two different runs of the code can provide different solutions in the $\mathbf{X}$ vector, but similar –no too much different– values in G.

The coordinates generated by the mapping process are virtual meaning-less data. An arbitrary set of rotations, translations and sign inversions can be involved in the

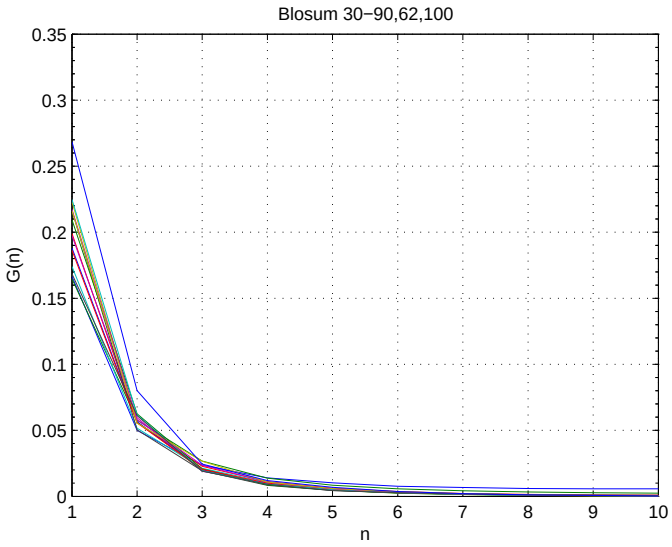

Fig. 1. Optimal goal value $G(n)$ vs the dimensionality n of the mapping space for BLOSUM matrices from 30 to 90 every 5, and also for 62 and 100 values. The result suggests that after dimensionality three or four few gain is obtained by introducing additional dimensions

optimization process because these transformations are distance invariant. No control exists on the spatial organization of the solutions, and it is a fact that random conditions are involved in their generation. Therefore, a first appreciation is that no relationship exists among the mapping coordinates and any variable with meaning at biological, physical or chemical level. However, without a full refusing of these appreciations, some kind of semantic organization can be found in the mapped space.

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 $d(a, b)$ with a set of characteristic $Q_j(a)$ with previous semantic, which in general are no orthogonal. This is an open problem which requires future studies.

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 provided results. Amino acids can be grouped according with their physical-chemical properties. Some groups –aliphatic or aromatic– are related to the chemical structure; others –tiny or small– are grouped with the molecular size; the polar and charged groups are related to the electric activity, and hydrophobic group is related with their affinity with water. It has been shown [15] 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 to two index according with their correlations. In this 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 1.

Table 1. The $\mathbf{Y}$ coordinates from 1 to 5 dimensionality of BLOSUM 62 matrix.

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
ρ	0.144	0.113		0.088			0.089				0.075				

The small group –except N amino acid– is mapped at the bottom of the map. It forms a region at low Y_2 coordinate and extends along the whole range of Y_1 one. The aromatic group –VIL– conforms a cluster included into the strong hydrophobic group –WFYMVIL– located at higher values of the Y_1 coordinate. The opposite groups –charged– have the lower values in this coordinate. The hydrophobic –or its opposite hydrophilic– property of amino acids is fundamental in the dynamics and structure of proteins[16]. 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 significative changes in the water affinity have a high probability of generate disfunctions, and consequently they have a low survival probability, therefore, there are lost in the evolution process.

4 Conclusion

The aim to be exhaustive in the discovery of relevant characteristics in the substitution matrices increases the cardinality of the characteristic set. Finding the most relevant characteristics is more critical as the number of properties is increased, thus the definition of an upper limit is a good choice to avoid the computational explosion. This assertion expresses qualitatively the advantage of determining the intrinsic dimensionality of substitution matrices from themselves, instead of estimates it from a big and

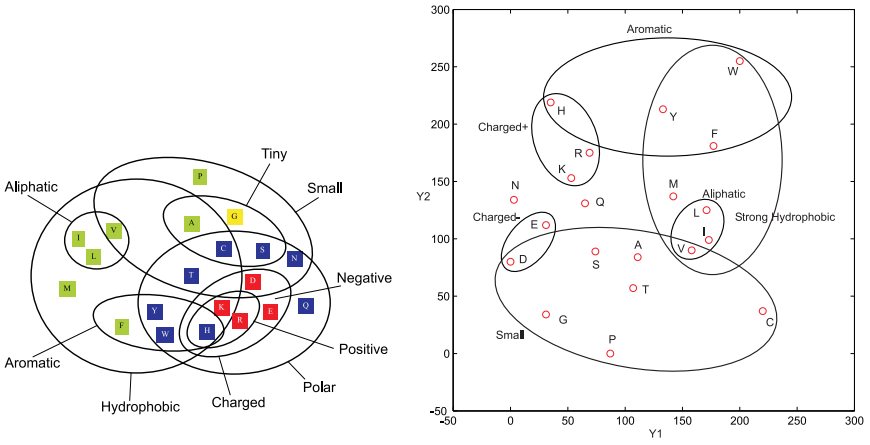

Fig. 2. At left the biochemical groups of amino acids. At right a representation for the $n = 2$ mapping of BLOSUM 62 showing the spatial organization of some amino acid groups. Relevant groups are spatially organized as clusters. Small group tends to be in low Y_2 values while hydrophobicity tends to high Y_1 values.

unclosed set of characteristic. This paper hypothesizes that the computation of the intrinsic dimensionality – the *how many characteristics* problem– is better achieved as a characteristic independent procedure. However, this does not exclude that the *what characteristic* analysis is necessary.

A lot of factors could be implied in the substitutions matrices, but the main result of this paper suggested that no too much must be considered. About three or four factors are important for the BLOSUM test case. Multidimensional mapping of substitution matrices could be considered an useful methodology to know about how many independent factors are involved in these matrices. The minimizing of a goal function related with the relative error of mapping has been used. The fast decreasing of goal function at small dimensionality suggests a small number of independent characteristics, and also that additional factors have a small contribution in the substitution matrix.

References

1. Dayhoff, M., Schwartz, R., Orcutt, B.: Atlas of Protein Sequence and Structure. Volume 5. Nat. Biomed. Res. Found. (1978)
2. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. **89** (1992) 10915–10919
3. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Morgan Kaufmann (1990)
4. Chakrabarti, K., Mehrotra, S.: Local dimensionality reduction: A new approach to indexing high dimensional spaces. In: The VLDB Journal. (2000) 89–100
5. Kanth, K.V.R., Agrawal, D., Abbadi, A.E., Singh, A.: Dimensionality reduction for similarity searching in dynamic databases. Computer Vision and Image Understanding: CVIU **75**(1–2) (1999) 59–72

6. Aggarwal, C.C.: On the effects of dimensionality reduction on high dimensional similarity search. In: Symposium on Principles of Database Systems. (2001)
7. Kawashima, S., Ogata, H., Kanehisa, M.: Aaindex: amino acid index database. Nucleic Acids Res. **27** (1999) 368–369
8. Venkatarajan, M.S., Braun, W.: New quantitative descriptors of amino acids based on multidimensional scaling of a large number of pysical-chemical properties. J. Mol. Model **7** (2001) 445–453
9. Cox, T., Cox, M.A.: Multidimensional Scaling. Chapman and Hall (1994)
10. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley and Sons (2001)
11. Sammon, J.: A nonlinear mapping for data structure analysis. IEEE Trans. Computers **18** (1969) 401–409
12. Li, S., de Vel, O., Coomans, D.: Comparative performance analysis of non-linear dimensionality reduction methods. Technical report, James Cook Univ. (1995)
13. Backer, S.D., Naud, A., Scheunders, P.: Nonlinear dimensionality reduction techniques for unsupervised feature extraction. Pattern Recognition Letters **19** (1998) 711–720
14. Scheunders, P., Backer, S.D., Naud, A.: Non-linear mapping for feature extraction. Lecture notes in computer science **1451** (1998) 823–830
15. Hagerty, C., Kulikowski, C., Muchnik, I., Kim, S.: Two indeces can approximate 402 amino acid properties. In: Proc. IEEE Int. Symp. Intelligent Control, Intelligent Systems and Semiotics. (1999) 365–369
16. Gerstein, M., Levitt, M.: Simulating water and the molecules of life. Scientific American (1998) 100–105