Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition

D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres

www.elsevier.com/locate/yjtbi

Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition

D.N. Georgiou, T.E. Karakasidis, J.J. Nieto, A. Torres

PII: S0022-5193(08)00580-8
DOI: doi:10.1016/j.jtbi.2008.11.003
Reference: YJTBI5356

To appear in: *Journal of Theoretical Biology*

Received date: 27 February 2008
Revised date: 14 October 2008
Accepted date: 1 November 2008

Cite this article as: D.N. Georgiou, T.E. Karakasidis, J.J. Nieto and A. Torres, Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition, *Journal of Theoretical Biology* (2008), doi:10.1016/j.jtbi.2008.11.003

Use of Fuzzy Clustering Technique and Matrices to Classify Amino Acids and Its Impact to Chou's Pseudo Amino Acid Composition

D.N. Georgiou[1], T.E. Karakasidis[2], J.J. Nieto[3], and A. Torres[4]

[1] *University of Patras, Department of Mathematics, 265 00 Patras, Greece.*
[2] *University of Thessaly, Department of Civil Engineering, 383 34 Volos, Greece.*
[3] *Departamento de Análisis Matemático, Facultad de Matemáticas, Universidad de Santiago de Compostela, 15782 Spain.*
[4] *Departamento de Psiquiatría, Radiología y Salud Pública, Facultad de Medicina, Universidad de Santiago de Compostela, 15782 Spain.*

Abstract

In this paper we present a study of classification of the twenty Amino Acids via a fuzzy clustering technique. In order to calculate distances among the various elements we employ two different distance functions: the Minkowski distance function and the NTV metric. In the clustering procedure we take into account several physical properties of the Amino acids. We examine the effect of the number and nature of properties taken into account to the clustering procedure as a function of the degree of similarity and the distance function used. It turns out that one should use the properties that determine in the more important way the behaviour of the amino acids and that the use of the appropriate metric can help in defining the separation into groups.

Key words: DNA, Amino acids, Protein, Fuzzy clustering.

Introduction

The genetic code is formed by strings of four letters (nucleic acids): A (adenosine), T (thymidine), C (cytidine), and G (guanosine). A string of three nucleic acids is an amino acid (or codon). Given that we have 4 letters with the possibility of being at 3 positions of the codon this results in 64 possible combinations and thus 64 possible amino acids (See the table, for example, in http://psyche.uthct.edu/shaun/SBlack/geneticd.html, Freeland and Hurst 1998). Three of these possible codons specify the termination of the polypeptide chain and thus they are called "stop codons". That leaves 61 codons to specify only 20 different amino acids (see Appendix A). The genetic code is degenerate in the sense that an amino acid can be represented by several triplets of nucleotides. For example, CAT is the amino acid Histidine (H), but CAC is also Histidine. Every triplet indicates a specific amino acid. A simple fuzzy cluster analysis of amino acids has been introduced by Mocz (Mocz 1995) to recognize secondary structure in proteins.

One can calculate the biological distance among the 20 amino acids according to their classification results. This is because ever since the concept of pseudo amino acid composition was proposed by Chou (Chou 2001), many efforts have been made trying to use various digital numbers to represent the 20 native amino acids in order to better reflect the sequence-order effects through the vehicle of pseudo amino acid composition (PseAA). In an earlier paper (Chou 2000), the physicochemical distance among the 20 amino acids (Schnieder and Wrede 1994) was adopted to define PseAA. Subsequently, some investigators used complexity measure factor (Xiao et al 2005), some used the values derived from the cellular automata (Xiao et al. 2005b, 2005c, 2006, 2006b), some used hydrophobic and/or hydrophilic values (Chou 2005, Chou et al 2005, Feng 2002, Wang et al 2006, Wang et al 2004, Gao et al 2005, Chen et al 2006, Mondal et al 2006), and some were through Fourier

transform (Guo et al 2006, Liu et al 2005). In view of this, the author's finding might have a series of impacts to the aforementioned work.

The pseudo amino acid composition was originally introduced to improve the prediction quality for protein subcellular localization and membrane protein type (Chou 2001), as well as for enzyme functional class (Chou 2005). The pseudo amino acid composition can be used to represent a protein sequence with a discrete model yet without completely losing its sequence-order information (Chou and Shen 2007a), and hence is particularly useful for analyzing a large amount of complicated protein sequences by means of the taxonomic approach. Actually, it has been widely used to study various protein attributes, such as protein structural class (Chen et al 2006a , Chen et al 2006b , Xiao et al. 2006a , Lin and Li 2007a , Ding 2007), protein subcellular localization (Chen and Shen 2008, Chou and Shen 2007a, Shen and Chou 2007a, Chou and Shen 2007b), protein subnuclear localization (Shen and Chou 2005, Mundra et al 2007) protein submitochondria localization (Du and Li 2006), protein oligomer type (Chou and Cai 2003), conotoxin superfamily classification (Mondal et al 2006,Lin and Li 2007b) membrane protein type (Liu et al 2005, Shen and Chou 2005, Wang et al 2006, Shen et al 2006, Chou and Shen 2007b0 apoptosis protein subcellular localization (Chen and Li 2007a, Chen and Li 2007b) enzyme functional classification (Chou 2005, Chou an Cai 2004, Zhou et al 2007, Shen and Chou 2007b) protein fold pattern (Shen and Chou 2006), and signal peptide [(Chou and Shen 2007c, Shen and Chou 2007 c).

Recent research works on the extension of these kind of parameters in the form of Markov Chain invariants of 2D graph or networks representation of aminoacid, DNA, and RNA sequences to codify psuedo-aminoacid and pseudo-nucleotide bases composition (Agüero-Chapín et al. 2008, González-Díaz et al 2007a,González-Díaz et al 2007b, Agüero-Chapin et al 2006). The reader can also consult some recent reviews which made a discussion of many of these previous results (González-Díaz et al 2008, González-Díaz et al 2007).

There are several methods that are used in order to extract characteristics of genomes and one of them is trying to find some common characteristics along its constituents. A method that can serve in this direction is the clustering procedure and more specifically fuzzy clustering. The fuzzy methods have the advantage to incorporate the uncertainties that exists for the data in the model (Torres and Nieto 2006). There are several methods of fuzzy clustering. Two of the most often employed methods are:

a) The fuzzy c-means algorithm (Bezdek 1981), which needs an a priori definition of the number of classes (called clusters) and its final result critically depends on this choice.

b) The fuzzy equivalence relation-based hierarchical clustering method (see, for example, Samaras et al. 2001, Klir and Yuan 1995) which avoids any a priori assumption on the number of classes. This is an immediate advantage whenever we want to extract unbiased results reflecting the structure of a given data set.

In the present work we employ the second method since we did not want to impose any a priori choice on the clustering of the amino acids.

When performing clustering, the elements that are to be classified are considered as points in a finite dimensional space where the axis correspond to the properties that we take into account in the clustering procedure. How similar are two elements is based on their "distance" in this space thus an important parameter is the metric used in order to calculate distances between the elements. There are several definition of metrics (Engelking 1977) and use of distances (Chou 1995, Chou and Zhang 1994, 1995), subcellular location (Chou and Elrod 1999, Chou 2000b), membrane protein type (Chou et al 2005, Chou and Elrod 1999b), enzyme family class (Chou and Elrod 2003, Chou and Cai 2004), GPCR type (Chou and Elrod 2002, Chou 2005b), protein-protein interaction (Chou and Cai 2006), metabolic pathways (Chou et al 2006), among many other protein attributes.

In the present paper we employ two metrics:

1) the Minkowski distance function employed in several clustering works (see, for example, Samaras et al. 2001, Karakasidis and Georgiou 2004), and

2) the NTV distance function introduced by Nieto et al. (Nieto et al. 2003, Dress and Lokot 2003) and employed in fuzzy properties of polynucleotides (Georgiou et al. preprint; Nieto et al 2006; Torres and Nieto 2003).

In the present paper we perform a clustering analysis of the twenty amino acids based on several physical properties: number of codons that code the protein, molecular weight, hydrophobicity, the number of atoms of different type and the corresponding number of protons as well as the number of total protons and we examine the influence of the properties on the classification procedure as well as the effect of the metric employed in the clustering procedure. There are many properties that can be employed in the clustering procedure. The reader can consult the AAindex database (Kawashima et al. 1999, Kawashima and Kanehisa 2000) and motivate selection of the considered properties. The obtained results may have potential for stimulating the development of predicting subcellular location of proteins and their other attributes, currently a very hot topic in bioinformatics and proteomics.

These clusters may help to explain the origin and emergence of the alphabet of amino acids encoded by the standard genetic code. Recently, Stepehn and Freeland (2008) have presented the first quantitative exploration of nature's "choices" set against various models for plausible alternatives with the help of computational chemistry. It is clear, that fuzzy technology, fuzzy clustering (Torres and Nieto 2006) and fuzzy cognitive maps (Stephen and Freeland 2008), will be useful in the protein content prediction methods and the prediction of protein structural classes (Zhang et al 2008).

The structure of the paper is as follows: First we present notions about fuzzy clustering using the fuzzy equivalence relation-based hierarchical clustering method. Then we present

clustering results as function of the degree of similarity and the number of physical properties taken into account in two cases : a) using the Minkowski distance function and b) using the NTV distance function. Finally we give the conclusions of the present work.

Fuzzy Clustering Preliminaries

In what follows by R we denote a *fuzzy relation* on a set X (see, for example, Bardossy and Duckstein 1995; Terano et al. 1992; Zimmermann 1991), that is, a fuzzy set in the direct product $X \times X = \{(x, y) : x, y \in X\}$ which is characterized by the membership function: $\mu_R : X \times X \to [0,1]$.

Also by $\Re$ we denote the set of all real numbers and by $\Re^+$ the set of all positive real numbers.

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a finite set. A fuzzy relation R in $X \times X$ can be expressed by a n x n matrix as following:

$$R = \begin{pmatrix} \mu_R(x_1, x_1) & \mu_R(x_1, x_2) & \cdots & \mu_R(x_1, x_n) \\ \mu_R(x_2, x_1) & \mu_R(x_2, x_2) & \cdots & \mu_R(x_2, x_n) \\ & & \ddots & \\ \mu_R(x_n, x_1) & \mu_R(x_n, x_2) & \cdots & \mu_R(x_n, x_n) \end{pmatrix}$$

A fuzzy relation R on X is:

(1) *reflexive* if $\mu_R(x, x) = 1$ for all $x \in X$,

(2) *symmetric* if $\mu_R(x, y) = \mu_R(y, x)$ for all $x, y \in X$, and

(3) *max-min transitive* if

$$\mu_R(x, z) \geq \sup\{ \min\{\mu_R(x, y), \mu_R(y, z)\} : y \in X \}$$

A fuzzy relation with the above properties is called *fuzzy similarity relation* or *fuzzy equivalence relation*.

A fuzzy relation on X that is reflexive and symmetric is usually called a *compatibility relation*.

The *max-min transitive closure* of a fuzzy relation R on X is defined as the smallest max-min fuzzy transitive relation containing R.

It is known that (see, for example, Hashimoto 1983) if R is a fuzzy compatibility relation on a finite set $X = \{x_1, \ldots, x_n\}$ then the max-min transitive closure R_T is the relation $R^{(n-1)} = R \circ \ldots \circ R$ i.e., relation R composed with itself *(n-1)* times.

We note that if R and S are two fuzzy relations on X the composition is characterized by the membership function:

$$\mu_{R \circ S}(x, z) \geq \sup\{ \min\{\mu_R(x, y), \mu_S(y, z)\} : y \in X \}$$

To illustrate the clustering method based on fuzzy equivalence relations (see, for example, Samaras et al. 2001, Klir and Yuan 1995), we consider a data set:

$X = \{x_1, \ldots, x_n\}$ where $x_i = \{x_{i1}, x_{i2} \ldots, x_{im}\} \in \Re^m$ and i=1,2,...,n.

Then we proceed with the following three steps:

1. We define a fuzzy relation R on X using the distance function of Minkowski, via the membership function:

$$\mu_R(x_i, x_k) = 1 - \frac{\left(\sum_{j=1}^{m} \left| x_{ij} - x_{kj} \right|^q \right)^{\frac{1}{q}}}{d}$$

for all $(x_i, x_k) \in X \times X$, where $q \in \Re^+$ and

$$d = \max \left\{ \left(\sum_{j=1}^{m} \left| x_{ij} - x_{kj} \right|^q \right)^{\frac{1}{q}} : i, k = 1,2, \ldots, n \right\}$$

Clearly R is a fuzzy compatibility relation but not necessarily a fuzzy equivalence relation (Klir and Yuan 1995).

Remark. Also we use (see the section entitled Fuzzy clustering using the NTV distance function) the distance function of NTV (Nieto et al. 2003; Dress et al. 2004; Dress and Lokot 2003) via the membership function:

$$\mu_R(x_i, x_k) = 1 - \frac{\sum_{j=1}^{m} |x_{ij} - x_{kj}|}{\sum_{j=1}^{m} \max\{x_{ij} - x_{kj}\}}$$

for all $(x_i, x_k) \in X \times X$ and where $x_i, x_k \neq (0,0,\ldots,0) \in \Re^m$.

2. We find the max-min transitive closure R_T

3. For every $a \in [0,1]$ called *the degree of similarity*, we define a new fuzzy matrix R_T^a as follows

$$\mu_{R_T^a}(x,y) = \begin{cases} 1 \ if \ \mu_{R_T}(x,y) \geq a \\ 0 \ if \ \mu_{R_T}(x,y) < a \end{cases}$$

The intervals of *a* that determine the partitions are derived from the values of the matrix R_T. In this way by the examination of the R_T matrix, we can determine the resulting partitions for all intervals of *a*-cuts.

Results

The clustering procedure is based on a number of physical properties of the amino acids that are presented in the Table1 of Appendix A. The columns c1 to c14 correspond to the properties of the amino acids. c1 contains the "Number of codons that code the protein", c2 the "Molecular weight", (the use of the molecular weight in the clustering procedure is motivated by (Homaeian et al. 2007, Kedarisetti et al 2006). Column c3 the "Hydrophobicity". There are different hydrophobic indices (Kurgan et al 2007, Wolfenden

2007). Here we use the normalized parameter for hydrophobicity of Table 2 in (Chechetkin 2003). It is of interest to note here work in (Kurgan et al. 2007) and applications of hydrophobic index to prediction of secondary structure content (Homaeian et al. 2007, Zhang et al. 2001, Lin and Pan 2001) and to prediction of structural classes (Kurgan and Chen 2007, Kedarisetti et al 2006).

Columns c4 to c8 the "Number of atomes" of various type (H, C, N, O, S) and c9 to c13 the corresponding "Number of protons" for each type of atom (H, C, N, O, S). Finally the "Total number of protons" of all the atoms appears in c14.

Of course there are several other properties that one could take into account in order to perform a clustering procedure however we should note that measures such as the aminoacid composition and hydrophobicity are employed in several methods of protein content prediction methods (Kurgan et al. 2007). Such properties also have been identified for their importance also by the work of Nakai et al. (1988).As is mentioned in (Kurgan et al 2007) and references there in hydrophobicity is not only one of the major structural forces but is also able to show periodicity of the secondary structure but can also show periodicity of the secondary structure.

We used several scenarios in order to perform the clustering procedure. These are summarized in the following cases:

Case 1 : All physical properties of Table in the Appendix B are employed (columns 1 to 14 of the Table are used).

Case 2 : All physical properties of Table1 in the Appendix A are employed except the total number of protons (columns 1 to 13 of the Table are used).

Case 3 : All physical properties of Table1 in the Appendix A are employed except the number of codons that code the protein (columns 2 to 14 of the Table are used).

Case 4 : All physical properties of Table1 in the Appendix A are employed except the number of codons that code the protein and the molecular weight (columns 3 to 14 of the Table are used).

Case 5 : All physical properties of Table1 in the Appendix A are employed except the number of codons that code the protein, the molecular weight and hydrophobicity (columns 4 to 14 of the Table are used).

First we present results concerning the clustering using the Minkowski distance function and then the results obtained using the NTV distance function

Fuzzy clustering using the Minkowski distance function.

We report in each case the values of *m* and *n* In all cases results are presented as function of the degree of similarity *a*. Amino acids that are in the same group appear within the same box.

Case 1 (n=20, m=14)

Results of the partitions as a function of the degree of similarity appear in Figure1. For *a*=0.60, 0.75 and 0.80 we obtain only one group of Amino Acids, which means that for these degree of similarity all amino acids appear similar to each other. For higher *a* values (*a*=0.83) only Trp appears different from all others and thus separated. We believe that this is due to its molecular weight and its total number of protons which are the largest among all amino acids. For *a*=0.85 further separation occurs. Here again we observe that the elements that are not part of the largest group of amino acids are the ones with the largest molecular weight as well as those with the largest number of total protons (although in fact these properties are related). Increasing the value of *a* (*a*=0.87) results in further separation. We can see again that the elements that are not part of the largest group are the ones with the largest molecular

weight and the largest number of protons. Phe and Tyr have similar atomic weights and the same number of codons while Arg, which has similar atomic weight but different number of codons, is separated. For a=0.90 we have more separation. We comment here that Asp and Asn have the same number of total protons. lle and Leu which are practically the same apart the number of codons so they are in the same group. Glu, His, Lys, Gln present quite similar molecular weights and number of codons and thus they are part of the same group. Finally for a=0.95 only lle and Leu, which are practically the same apart the number of codons, are in the same group. The rest of amino acids are separated apart. Such behaviour is expected since at such high values of a elements are expected to appear different from each other.

Case 2 (n=20, m=13)

Results are summarized in Figure 2. For a=0.60, 0.75 we obtain only one group of Amino Acids like in case 1 where we included the total number of protons in the clustering procedure. For a=0.80 only Trp appears separated from all other amino acids. We remind here that Trp corresponds to the largest molecular weight and number of total protons. Compared to case 1 we see that the separation in partitions starts at smaller values of the degree of similarity. Increasing a (a =0.83) we obtain further separation of Amino Acids. Again we observe, compared to case 1, that separation in partitions starts at smaller values of the degrees of similarity. For higher values (a=0.85, 0.87 and 0.90) we observe that increasing a results in increasing separation with slight differences compared to case 1 for the same a values. In fact separation starts at lower a values. Finally for a=0.95 we have similar behavior to case 1. In this case we can say that separation in the partitions starts at smaller values of similarity degree but we do not have significant differences in the tendency of group splitting with case 1 where we took into account the total number of protons.

Case 3 (n=20, *m*=13)

Results are summarized in Figure 3. For *a*=0.60, 0.75 and 0.80 we obtain a trivial partition of the elements of Amino Acids in one group. The behavior is similar to that of case 1 (it seems that the number of codons that code the protein does not result in any difference).

For *a*=0.83, 0.85 and 0.87 we observe similar behavior like in case 1 indicating that taking into account the number of codons that code the protein does not result in any difference. Only for *a*=0.90 we observe slight differences with case 1. For *a*=0.95 we obtain the same behaviour ike in the previous cases 1 and 2 as expected.

Concluding, we could say that if we omit the number of codons that code the protein in the clustering procedure we do not obtain any significant difference in the obtained partitions from the previous cases.

Case 4 (*n*=20, *m*=12)

Results are summarized in Figure 4. For *a*=0.60 and 0.75 we obtain only one group of Amino Acids: The behavior is similar to that of cases 1, 2 and 3 in the sense that for the lowest values amino acids appear to be similar. For higher *a* values (*a* =0.80 0.83, 0.85, 0.87) we have further separation and we observe differences from cases 1 and 2 and 3. It seems that when we neglect the molecular weight the separation in groups starts for smaller values of the similarity degree. For *a*=0.90 and 0.95 we obtain the same behaviour that we have seen for *a* =0.95 in all cases with only the lle Leu being in the same group and all other Amino Acids separated.

To conclude we could say that if we neglect the number of codons that code the protein and the molecular weight, we observe slight differences with the cases where these properties are taken into account. However the global tendency of separation into groups of Amino

Acids seems to be the same and the separation occurs at lower values of the degree of similarity.

Case 5 (n=20, m=11)

In fact in this case we take into account only "electronic properties" of the Amino Acids like the number of atoms and their protons. Results are summarized in Figure 5 where we can see that there is no difference with the previous case 4 where we took into account hydrophobicity. We can also say in comparison with case 4 that when we perform the clustering procedure using the number of different kind of atoms and their corresponding number of protons, adding hydrophobicity does not make any difference in the results.

Summarizing the obtained results it seems that when one has to perform a clustering procedure involving the amino acids the number of properties that must be taken into account may be limited to properties that are related to “electronic properties” like the number of the different types of atoms in the aminoacid and the corresponding protons and neglect other properties that are the simple outcome of the given configuration of atoms like the atomic weight and the total number of atoms. From a physical point of view it is these properties that determine the behaviour of the amicoacids.

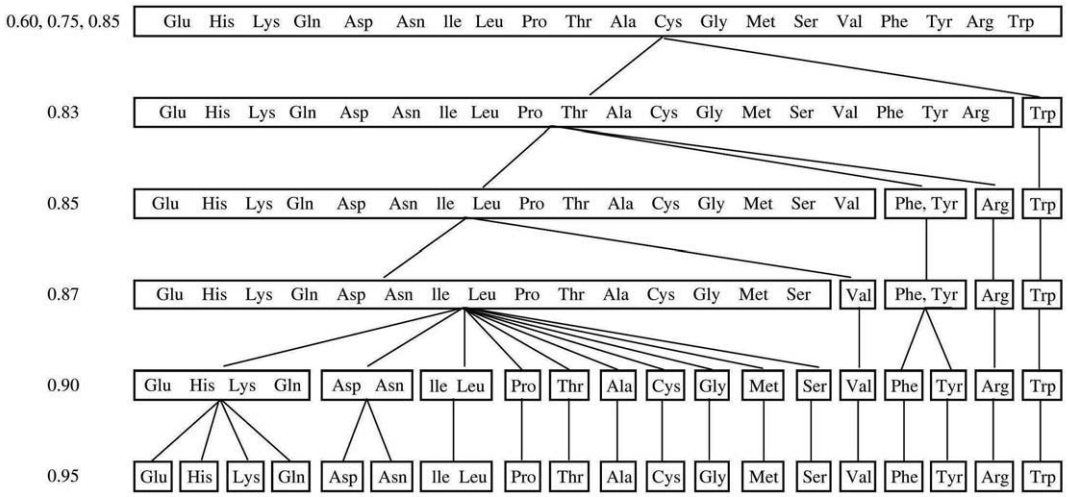

Figure1. Clustering results as a function of the degree of similarity for the case1 with the Minkowski distance function

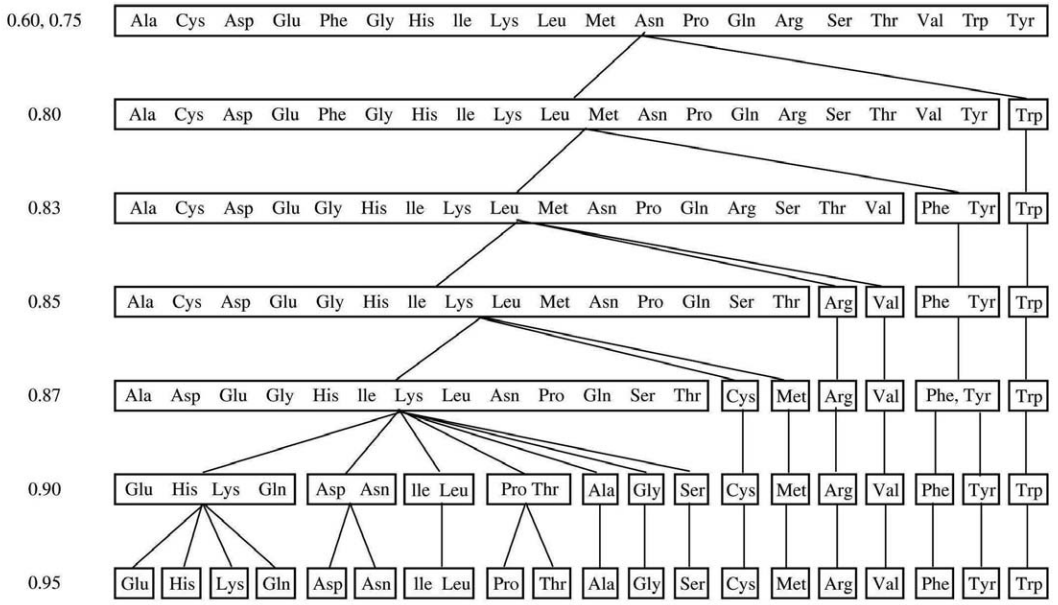

Figure2. Clustering results as a function of the degree of similarity for the case2 with the Minkowski distance function

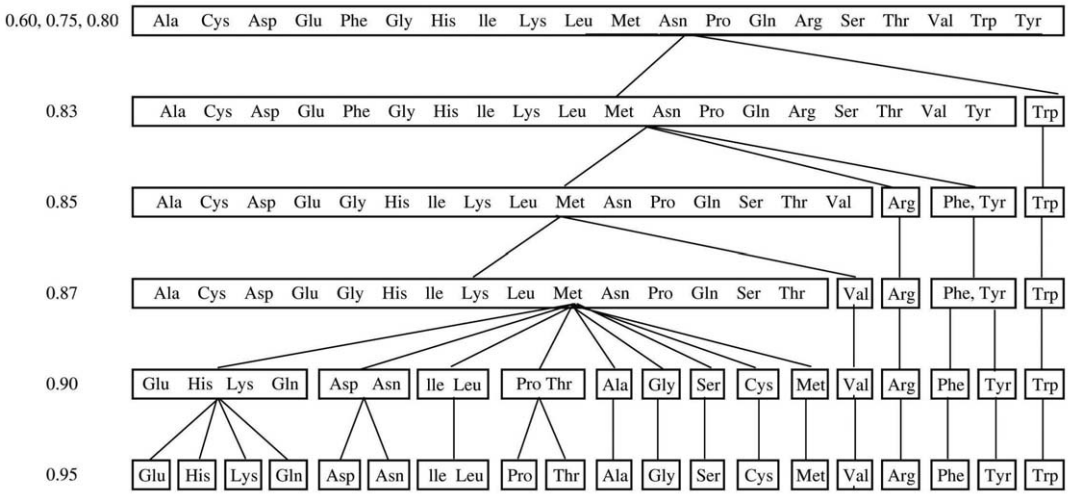

Figure3. Clustering results as a function of the degree of similarity for the case3 with the Minkowski distance function

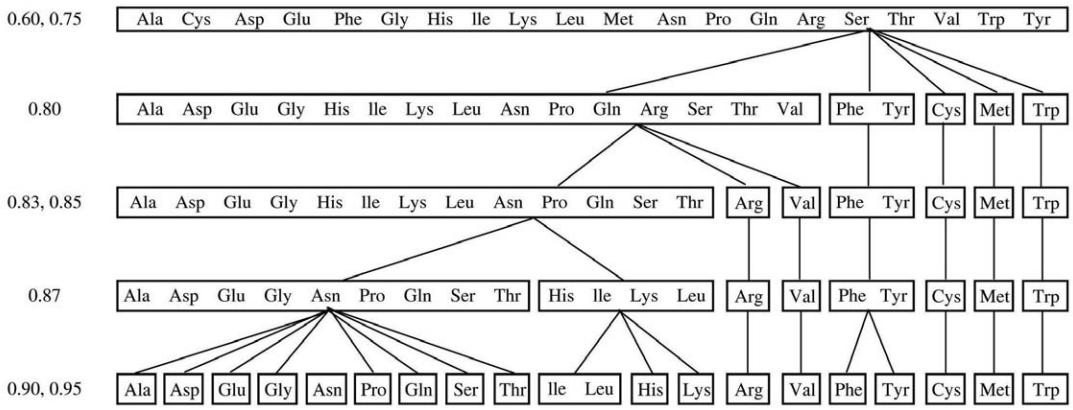

Figure4. Clustering results as a function of the degree of similarity for the case4 with the Minkowski distance function

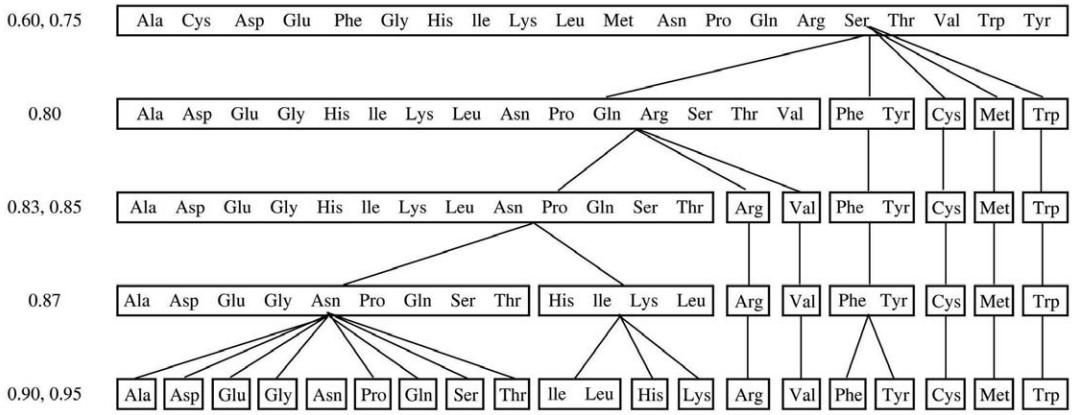

Figure5. Clustering results as a function of the degree of similarity for the case5 with the Minkowski distance function

Fuzzy clustering using the NTV distance function.

Case 1 (n=20, *m*=14)

Results are summarized in figure 6. For *a*=0.60, we have no discrimination for the elements of Amino Acids: We have similar behavior like in case 1 of the previous section. Increasing *a* (*a*=0.75) we obtain further separation. Compared to figure 1 we observe that we have separation of the amino acids and this is different than in case 1 of Minkowski distance function since the separated elements present molecular weights which are among the lowest ones and not among the highest ones like in case 1 of the previous section. For higher values of the degree of similarity (*a*=0.80, 0.83, 0.85, 0.87) we obtain more classes than in case 1 of Minkowski distance function. For *a*=0.90 and 0.95 we obtain the same partition as foir *a*=0.95 in the case 1 wit the Minkowski distance function.

Concluding we observe a different behavior than in case of Minkowski distance function since the separation in groups starts at smaller values of similarity degree and it follows a different tendency.

Case 2 (n=20, m=13)

Results are summarized in figure 7. We observe a different behavior from case 1 when we neglect the total number of protons. In fact for a=0.60 and 0.75 we obtain the trivial partition of the elements of Amino Acids. Increasing a results in further separation with differences compared to case1 of this section (NTV metric too). Only at a =0.95 we have the same results with the lle and Leu amino acids appearing in the same group and all other amino acids being separated.

It is of interest, compared to case 2 using the Minkoski distance function (Figure 2), that we have exactly the same behaviour.

Case 3 (n=20, m=13)

Results are summarized in figure 8. For a=0.60, we have the trivial partition of the elements of Amino Acids in only one group, like in all cases. For a=0.75 we see differences from cases 1 and 2. For higher values we have similar behaviour like in case 1. Globally speaking neglecting the number of codons that code the protein does not lead to significant differences.

Comparing to case 3 with the Minkowski distance the exact separation presents differences. However the global behavior is in qualitative agreement with the tendency observed in this case.

Case 4 (n=20, m=12)

Results for this case are summarized in Figure 9. For a=0.60 we obtain the following separation of Amino Acids. We have a different behavior than the one observed in case 3.

Increasing a (a=0.75) we obtain further separation with more groups than in case 3 and the behavior is different. The same behaviour occurs for a=0.80, 0.83, and 0.85 where we have a different behavior than the one observed in case 3. For a=0.95 we obtain the usual separation observed in all cases with the Minkowski or the NTV distance function.

In short, we can say that we have differences since separation starts at lower values of degree of similarity when neglecting molecular weight in the clustering procedure. Compared to the results of case 4 with the Minkowski distance function we observe that separation starts for lower values of the similarity degree.

Case 5 (n=20, m=11)

In fact in this case we take into account only "electronic properties" of the Amino acids like the number of atoms and their protons. Results are the same as in Figure 5 where we can see that there is no difference with the previous case where we took into account hydrophobicity.

Concluding we can say that when performing the clustering procedure using the number of different kind of atoms and their corresponding number of protons adding hydrophobicity in the procedure does not make any significant difference to the obtained partitions. This behavior is in qualitative agreement with that obtained in case 5 of the previous section (Minkowski distance function). Quantitatively we observe a difference since compared to the results based on Minkowski distance, separations in the NTV case starts at smaller values of degree of similarity.

Summarizing again the obtained results like in the case of the Minkowski metric we see that it is again the “electronic properties” that determine the basics of the partition since from a physical point of view it is these properties that determine the behaviour of the amicoacids. The fact that amino acids Ile and Leu are partioned always in the same group even is due to

the fact that they have nearly the same properties. This could indicate that further properties should be included in the analysis.

As far as the comparison of the two metrics used it turns out that in the case of the NTV metric although the results for high values of the similarity degree *a* are the same with that of the Minkowski metric, for lower a values the distinction of several groups starts earlier in the case of the NTV metric than in the case of Minkowski metric. This may be due to the fact that in the case of NTV metric the form of the membership function maximizes the effects of the difference between two amino acids. This is an important result that should be taken into account in clustering analysis of aminoacids but not only.

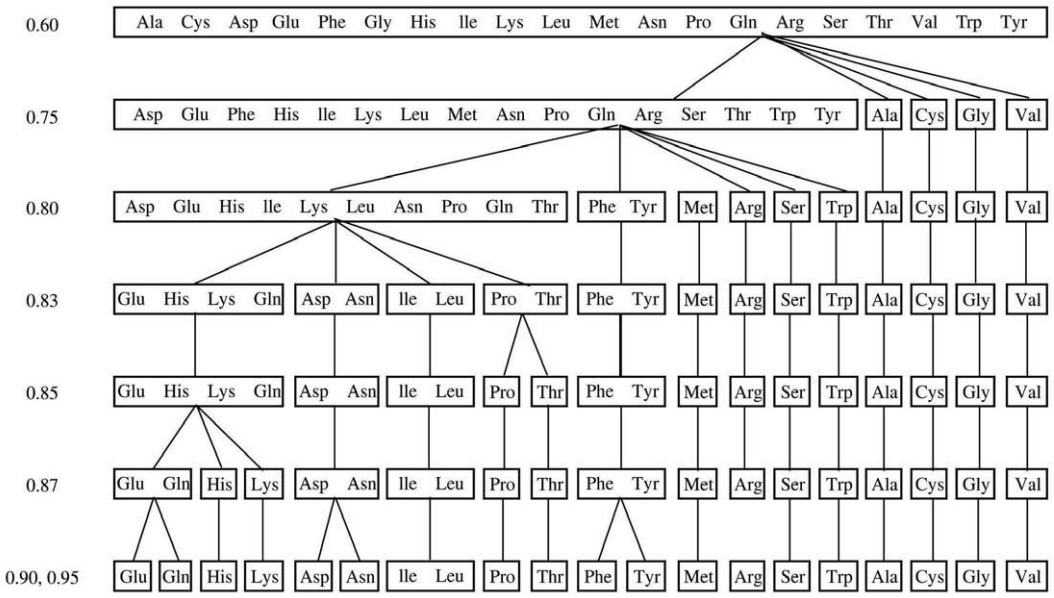

Figure 6. Clustering results as a function of the degree of similarity for the case1 with the NTV distance function

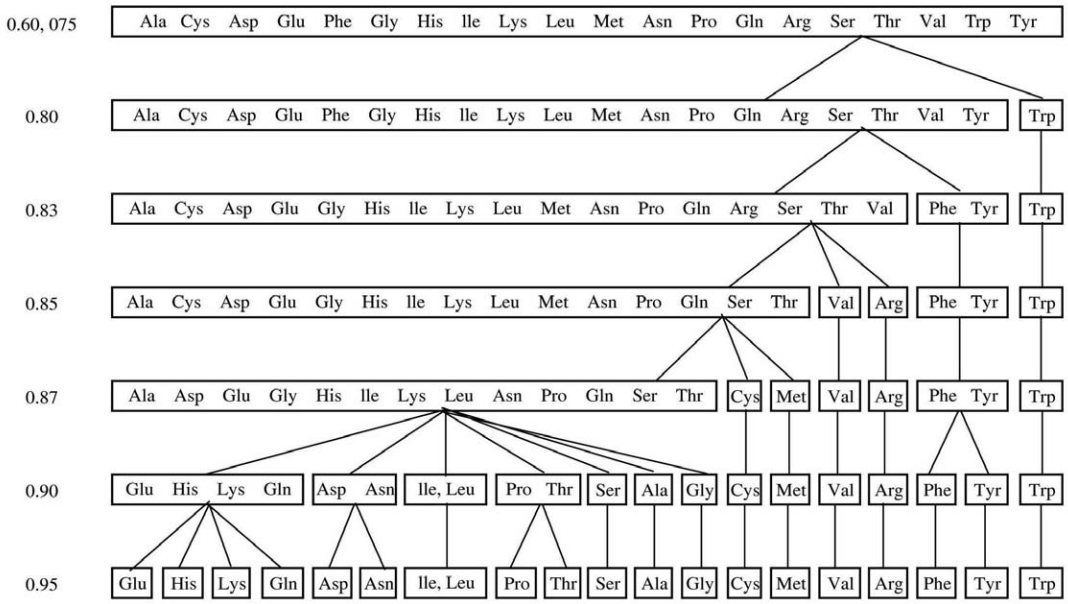

Figure 7. Clustering results as a function of the degree of similarity for the case2 with the NTV distance function

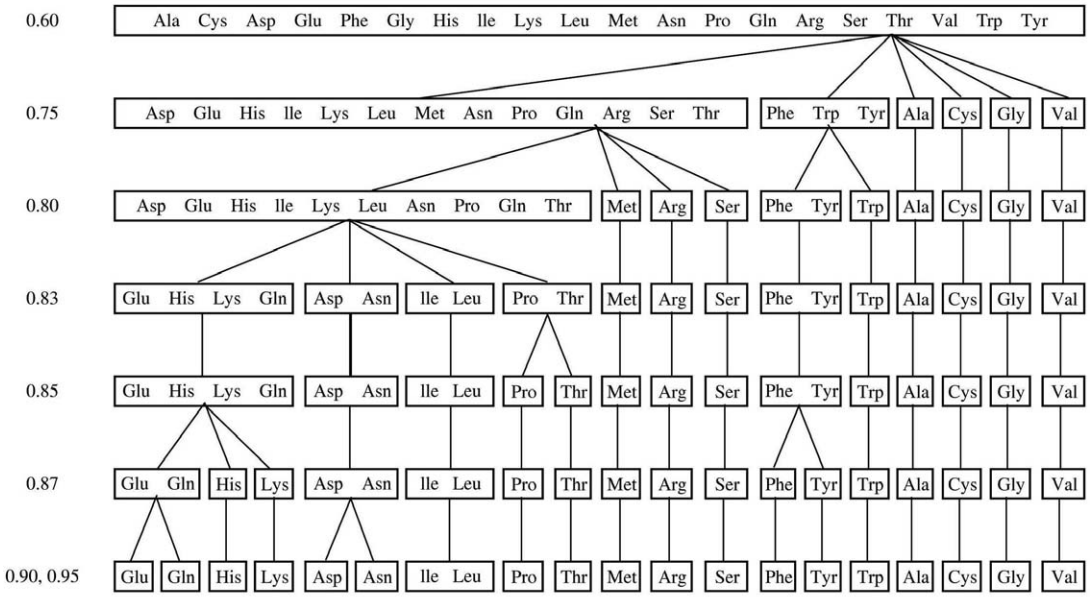

Figure 8. Clustering results as a function of the degree of similarity for the case3 with the NTV distance function

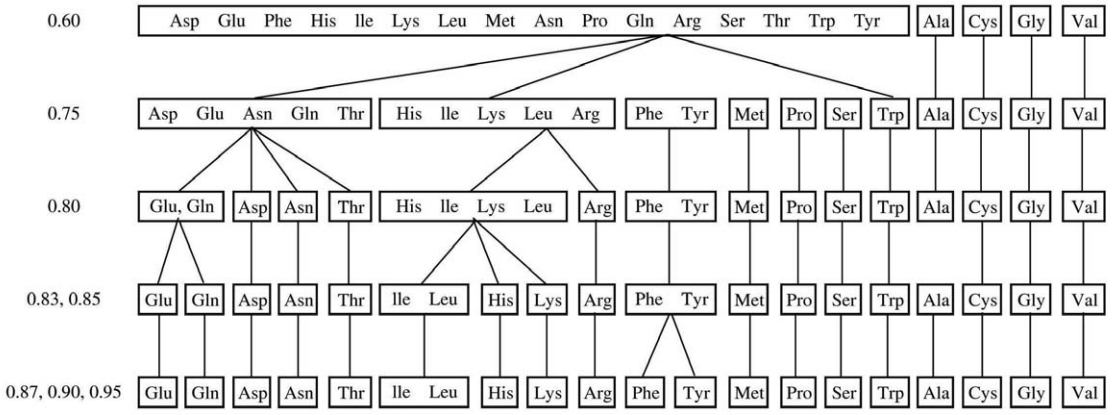

Figure 9. Clustering results as a function of the degree of similarity for the case4 with the NTV distance function

Conclusions

In the present work we perform a clustering analysis of the twenty Amino Acids using the fuzzy equivalence relation-based hierarchical clustering method. We examined the influence of using different distance function definitions by applying the Minkowski distance function and the NTV function. The effect of the number and nature of properties of the amino acids taken into account in the clustering procedure is also examined.

It turns out that in the case of NTV definition separation in partitions starts for smaller values of the degree of similarity than in the case of the Minkowski distance definition. This is due to the fact that due to its definition maximizes the possible differences on properties of the studied aminoacids. It seems that the main role in the partitioning comes from the number of atoms and their contribution to the number of protons. Adding hydrophobicity does not change results. We observe slight differences when we include the molecular weight and no difference when we add the number of codons that code the protein. This behavior is probably due to the fact that all additional properties (apart the number of different kind of atoms and their corresponding protons) are related to the number of atoms and the corresponding number of protons. Thus we could think that the number of atoms and the corresponding number of protons may be considered as the basic properties for comparing amino acids.

These indications show could be very useful in the taxonomy of larger sequences occurring in biological system since they indicate that one can use a minimum of information to perform the clustering, especially when one has to deal with a large amount of data and on the other hand they indicate that the proper choice of metric can lead to more clear separations of the data. It would be of interest in a future work to take into account more properties or examine the effect of use of a hydrophobicity scale instead of hydrophobicity index since recent work (Kurgan et al 2007) has shown that secondary structure content along

the protein sequence is characterized by about 2.5 times stronger relation with the two proposed hydrophobicity scales when compared with the currently used raw index values.

Acknowledgements

J.J. Nieto and A. Torres partially supported by Ministerio de Educacion y Ciencia and FEDER, projects MTM 2004-06652 and MTM2007-61724, and by Xunta de Galicia and FEDER, project PGIDIT02PXIC507002PN.

References

Agüero-Chapin G., González-Díaz H., Molina R., Varona-Santos J., Uriarte E., González-Díaz Y., 2006, Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L. FEBS Lett., Feb 6;580(3):723-30.

Agüero-Chapín G., Gonzalez-Díaz H., Riva G. D., Rodríguez E., Sanchez-Rodríguez A., Podda G., Vazquez-Padrón R. I., 2008, MMM-QSAR Recognition of Ribonucleases without Alignment: Comparison with an HMM Model and Isolation from Schizosaccharomyces pombe, Prediction, and Experimental Assay of a New Sequence. J Chem Inf Model., Feb 25;48(2):434-448.

Bardossy A. and Duckstein L., 1995, Fuzzy Rule-Based Modeling with Applications to Geophysical, Biological and Engineering Systems, CRC Press, Boca Raton.

Bezdek J.C., 1981, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Chechetkin V. R., 2003, Block structure and stability of the genetic code. J. Theoretical Biology 222, 177-188

Chen C., Zhou X., Tian Y., Zou X., Cai P., 2006, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Anal Biochem 357, 116-121.

Chen, C., Tian, Y. X., Zou, X. Y., Cai, P. X. and Mo, J. Y., 2006b, Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol, 243, 444-448.

Chen, C., Zhou, X., Tian, Y., Zou, X. and Cai, P., 2006a, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem, 357, 116-121.

Chen, Y. L. and Li, Q. Z., 2007a, Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. Journal of Theoretical Biology, 248, 377-381.

Chen, Y. L. and Li, Q. Z., 2007b, Prediction of the subcellular location of apoptosis proteins. Journal of Theoretical Biology, 245, 775-783.

Chou K. C., 1995, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins: Structure, Function & Genetics 21, 319-344.

Chou K. C., 2000, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochemical & Biophysical Research Communications 278, 477-483.

Chou K. C., 2000b, Review: Prediction of protein structural classes and subcellular locations, Current Protein and Peptide Science 1, 171-208.

Chou K. C., 2001, Prediction of protein cellular attributes using pseudo amino acid composition, PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60) 43, 246-255.

Chou K. C., 2005, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21, 10-19.

Chou K. C., 2005b, Prediction of G-protein-coupled receptor classes, Journal of Proteome Research 4, 1413-1418.

Chou K. C., Cai Y. D., 2004, Predicting enzyme family class in a hybridization space, Protein Science 13, 2857-2863.

Chou K. C., Cai Y. D., 2005, Prediction of membrane protein types by incorporating amphipathic effects, Journal of Chemical Information and Modeling 45, 407-413.

Chou K. C., Cai Y. D., 2006, Predicting protein-protein interactions from sequences in a hybridization space, Journal of Proteome Research 5, 316-322.

Chou K. C., Cai Y. D., Zhong W. Z., 2006b, Predicting networking couples for metabolic pathways of Arabidopsis, EXCLI Journal 5, 55-65.

Chou K. C., D. W. Elrod, 2002, Bioinformatical analysis of G-protein-coupled receptors, Journal of Proteome Research 1 (2002) 429-433.

Chou K. C., Elrod D. W., 1999, Protein subcellular location prediction, Protein Engineering 12, 107-118.

Chou K. C., Elrod D. W., 1999b, Prediction of membrane protein types and subcellular locations, PROTEINS: Structure, Function, and Genetics 34, 137-153.

Chou K. C., Elrod D. W., 2003, Prediction of enzyme family classes, Journal of Proteome Research 2 (2003) 183-190.

Chou K. C., Zhang C. T., 1994, Predicting protein folding types by distance functions that make allowances for amino acid interactions, Journal of Biological Chemistry 269, 22014-22020.

Chou K. C., Zhang C. T., 1995, Review: Prediction of protein structural classes, Critical Reviews in Biochemistry and Molecular Biology 30, 275-349.

Chou, K. C. and Cai, Y.D., 2003, Predicting protein quaternary structure by pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics, 53, 282-289.

Chou, K. C. and Cai, Y.D., 2004, Predicting enzyme family class in a hybridization space. Protein Science, 13, 2857-2863.

Chou, K. C. and Shen, H. B., 2007a, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research, 6, 1728-1734.

Chou, K. C. and Shen, H. B., 2007a, Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry, 370, 1-16.

Chou, K. C. and Shen, H. B., 2007b, Large-scale plant protein subcellular location prediction. Journal of Cellular Biochemistry, 100, 665-678.

Chou, K. C. and Shen, H. B., 2007b, MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Comm, 360, 339-345.

Chou, K. C. and Shen, H. B., 2007c, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm, 357, 633-640.

Chou, K. C. and Shen, H. B., 2008, Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nature Protocols, 3, 153-162.

Chou, K. C., 2001 Prediction of protein cellular attributes using pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics (Erratum: ibid, 2001, Vol44, 60), 43, 246-255.

Chou, K. C., 2005, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics, 21, 10-19.

Ding, Y. S., Zhang, T. L. and Chou, K. C., 2007, Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein & Peptide Letters, 14, 811-815.

Dress A. and Lokot T., 2003, A simple proof of the triangle inequality for the NTV metric, Applied Mathematics Letters, 16, 809-813.

Dress A., Lokot T., and Pustylnikov L.D., 2004, A new scale-invariant Geometry of L1 space, Applied Mathematics Letters, 17, 815-820.

Du, P. and Li, Y., 2006, Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics, 7, 518.

Engelking R., General Topology, Warszawa 1977.

Feng Z. P., 2002, An overview on predicting the subcellular location of a protein, In Silico Biol 2, 291-303.

Freeland S.J. and Hurst L.D., 1998, The Genetic Code is one in a million, Journal of Molecular Evolution, 47, 238-248.

Gao Y., Shao S. H., Xiao X., Ding Y. S., Huang Y. S., Huang Z. D., Chou K. C., 2005, Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter, Amino Acids 28, 373-376.

Georgiou D.N., Karakasidis T.E., Nieto J.J., and Torres A., A study of genetic sequences using Metric Spaces and Fuzzy Sets, Preprint.

González-Díaz H., Agüero-Chapin G., Varona J., Molina R., Delogu G., Santana L.,Uriarte E., Podda G., 2007b, 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J Comput Chem., Apr 30;28(6):1049-56.

González-Díaz H., González-Díaz Y., Santana L., Ubeira F. M., Uriarte E., 2008, Proteomics, networks and connectivity indices. Proteomics., Feb;8(4):750-78.

González-Díaz H., Pérez-Castillo Y., Podda G. 2007a, Uriarte E.Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices.J Comput Chem., Sep;28(12):1990-5.

González-Díaz H., Vilar S., Santana L., Uriarte E. 2007 Medicinal chemistry and bioinformatics-current trends in drugs discovery with networks topological indices. Curr Top Med Chem., 7(10):1015-29.

Guo Y. Z., Li M., Lu M., Wen Z., Wang K., Li G., Wu J., 2006, Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform, Amino Acids 30, 397-402.

Hashimoto H., 1983, Szpilrajn's theorem on fuzzy orderings, Fuzzy Sets and Systems, 10, 101-108.

Homaeian L., Kurgan L. A., Cios K. J., Ruan J, Chen K., 2007, Prediction of Protein Secondary Structure Content for the Twilight Zone Sequences. Proteins, 69(3):486-498

Karakasidis T. E. and Georgiou D. N., 2004, Partitioning elements of the periodic table via fuzzy clustering technique, Soft Computing (Springer-Verlag), 8, 231-236.

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

Kawashima, S., Ogata, H., and Kanehisa, 1999, M.; AAindex: amino acid index database. Nucleic Acids Res., 27, 368-369

Kedarisetti K, Kurgan L, Dick S, 2006, Classifier Ensembles for Protein Structural Class Prediction with Varying Homology. Biochemical and Biophysical Research Communications, 348(3):981-988

Klir G.J. and Yuan B., 1995, Fuzzy Sets and Fuzzy Logic (Theory and Applications), Prentice Hall PRT New Jersey.

Kurgan L. A., Stach W., Ruan J., 2007, Novel scales based on hydrophobicity indices for secondary protein structure. J. Theoretical Biology 248, 354-366

Kurgan L., Chen K., 2007, Prediction of Protein Structural Class for the Twilight Zone Sequences. Biochemical and Biophysical Research Communications, 357(2):453-460

Lin Z, Pan X. 2001, Accurate prediction of protein secondary structural content. J Protein Chem., 20:217-220.

Lin, H. and Li, Q. Z., 2007a, Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components. Journal of Computational Chemistry, 28, 1463-1466.

Lin, H. and Li, Q. Z., 2007b, Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun, 354, 548-551.

Liu H., Wang M., Chou K. C., 2005, Low-frequency Fourier spectrum for predicting membrane protein types, Biochem Biophys Res Commun 336, 737-739.

Liu, H., Wang, M. and Chou, K. C., 2005, Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun, 336, 737-739.

Mocz G., 1995, Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins, Protein Science 4, 1178-1187.

Mondal S., Bhavna R., Mohan Babu R., Ramakumar S., 2006, Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification, J Theor. Biol. 243, 252-260.

Mondal, S., Bhavna, R., Mohan Babu, R. and Ramakumar, S., 2006, Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol, 243, 252-260.

Mundra, P., Kumar, M., Kumar, K.K., Jayaraman, V.K. and Kulkarni, B.D., 2007, Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. Pattern Recognition Letters, 28, 1610-1615.

Nakai, K., Kidera, A. and Kanehisa, M., 1988, Cluster-Analysis Of Amino-Acid Indexes For Prediction Of Protein-Structure And Function, Protein eng., 2, 93-100

Nieto J. J. and Torres A., 2003, Midpoints for fuzzy sets and their application in medicine, Artificial Inteligence in Medicine 17, 81-101.

Nieto J. J., Torres A., and Vazquez-Trasande M. M., 2003, A metric space to study differences between polynucleotides, Applied Mathematics Letters, 16, 1289-1294.

Nieto J. J., Torres A., Georgiou D. N. and Karakasidis T. E., 2006, Fuzzy Polynucleotide spaces and Metrics, Bulletin of Mathematical Biology, 68, 703-725

Samaras P., Kungolos A., Karakasidis T., Georgiou D., Perakis K., 2001, Statistical Evaluation of PCDD/F Emission Data During Solid Waste Combustion by Fuzzy Clustering Techniques, Journal of Environmental Science and Health, Marcel Dekker, Inc.(part A), 36, 153-161.

Schneider G., Wrede P., 1994, The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site, Biophysical Journal 66, 335-344.

Shen, H. B. and Chou, K. C., 2005, Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochemical & Biophysical Research Communications, 334, 288-292.

Shen, H. B. and Chou, K. C., 2006, Ensemble classifier for protein fold pattern recognition. Bioinformatics, 22, 1717-1722.

Shen, H. B. and Chou, K. C., 2007a, Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun, 355, 1006-1011.

Shen, H. B. and Chou, K. C., 2007b, EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm, 364, 53-59.

Shen, H. B. and Chou, K. C., 2007c, Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Comm, 363, 297-303.

Shen, H. B., Yang, J. and Chou, K. C., 2006, Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. Journal of Theoretical Biology, 240, 9-13.

Shen, H. B. and Chou, K. C., 2005, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Comm, 337, 752-756.

Stephen Y. L., Freeland J., 2008, A quantitative investigation of the chemical space surrounding amino acid alphabet formation. J. Theoretical Biology 250, 349-361

Terano T., Asai K., and Sugeno M., 1992, Fuzzy Systems Theory and its Applications, Academic Press, Harcount Brace Jovanovich Publishers, San Diego, California.

Torres A. and Nieto J.J., 2003, The fuzzy polynucleotide space:basic properties, Bioinformatics, 19, 587-592.

Torres A., and Nieto J. J., 2006, Fuzzy logic in medicine and bioinformatics, Journal of Biomedicine and Biotechnology, Article ID 91908.

Wang M., Yang J., Liu G. P., Xu Z. J., Chou K. C., 2004, Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition, Protein Engineering, Design, and Selection 17, 509-516.

Wang S. Q., Yang J., Chou K. C., 2006, Using stacked generalization to predict membrane protein types based on pseudo amino acid composition, Journal of Theoretical Biology 242, 941-946.

Wang, S. Q., Yang, J. and Chou, K. C., 2006, Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. Journal of Theoretical Biology, 242, 941-946.

Wolfenden R., 2007, Experimental measures of amino acid hydrophobicity and the topology of transmembrane and globular proteins. J. Cell Biology 177, i10-i10

Xiao X., Shao S. H., Ding Y. S., Huang Z. D., Chou K. C., 2006b, Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location, Amino Acids 30, 49-54.

Xiao X., Shao S. H., Huang Z. D., Chou K. C., 2006, Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, Journal of Computational Chemistry 27, 478-482.

Xiao X., Shao S., Ding Y., Huang Z., Chen X., Chou K. C., 2005b, Using cellular automata to generate Image representation for biological sequences, Amino Acids 28, 29-35.

Xiao X., Shao S., Ding Y., Huang Z., Chen X., Chou K. C., 2005c, An Application of Gene Comparative Image for Predicting the Effect on Replication Ratio by HBV Virus Gene Missense Mutation, Journal of Theoretical Biology 235, 555-565.

Xiao X., Shao S., Ding Y., Huang Z., Huang Y., Chou K. C., 2005, Using complexity measure factor to predict protein subcellular location, Amino Acids 28, 57-61.

Xiao, X., Shao, S.H., Huang, Z.D. and Chou, K.C., 2006a, Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. Journal of Computational Chemistry, 27, 478-482.

Zhang T. L., Ding Y. S., Chou K. C., 2008, Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern. J. Theoretical Biology 250, 186-193

Zhang Z. D., Sun Z. R., Zhang C. T. 2001, A new approach to predict the helix/strand content of globular proteins. J Theor Biol, 208:65-78.

Zhou, X. B., Chen, C., Li, Z. C. and Zou, X. Y., 2007, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology, 248, 546-551.

Zimmermann H. J., 1991, Fuzzy Theory and its Applications, Kluwer Academic Publishers, New York.

APPENDIX A

TABLE 1. The twenty amino acids and their properties presented in columns c1 to c14.

				c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14
							Number of atoms					Number of protons					
Amino Acid	3-Letter Code		Reverse codon table	Number of Codons that code the protein	Molecular Weight	Hydrophobicity	H	C	N	O	S	H	C	N	O	S	Total of protons
1	Ala	Alanine	(GCU, GCC, GCA, GCG)	4	89.09	0.616	3	1	0	0	0	3	6	0	0	0	9
2	Cys	Cysteine	(UGU, UGC)	2	121.16	0.68	3	1	0	0	1	3	6	0	0	16	25
3	Asp	Aspartic acid	(GAU, GAC)	2	133.1	0.028	3	2	0	2	0	3	12	0	16	0	31
4	Glu	Glutamic acid	(GAA, GAG)	2	147.13	0.043	5	3	0	2	0	5	18	0	16	0	39
5	Phe	Phenylalanine	(UUU, UUC)	2	165.19	1	7	7	0	0	0	7	42	0	0	0	49
6	Gly	Glycine	(GGU, GGC, GGA, GGG)	4	75.07	0.501	1	0	0	0	0	1	0	0	0	0	1
7	His	Histidine	(CAU, CAC)	2	155.16	0.165	5	4	2	0	0	5	24	14	0	0	43
8	Ile	Isoleucine	(AUU, AUC, AUA)	3	131.18	0.943	9	4	0	0	0	9	24	0	0	0	33
9	Lys	Lysine	(AAA, AAG)	2	146.19	0.283	10	4	1	0	0	10	24	7	0	0	41
10	Leu	Leucine	(UUA, UUG, CUU, CUC, CUA, CUG)	6	131.18	0.943	9	4	0	0	0	9	24	0	0	0	33
11	Met	Methionine	(AUG)	1	149.21	0.738	7	3	0	0	1	7	18	0	0	16	41
12	Asn	Asparagine	(AAU,AAC)	2	132.12	0.236	4	2	1	1	0	4	12	7	8	0	31
13	Pro	Proline	(CCU, CCC, CCA, CCG)	4	115.13	0.711	6	3	0	0	0	6	18	0	0	0	24
14	Gln	Glutamine	(CAA, CAG)	2	146.15	0.251	6	3	1	1	0	6	18	7	8	0	39
15	Arg	Arginine	(CGU, CGC, CGA, CGG, AGA, AGG)	6	174.2	0	10	4	3	0	0	10	24	21	0	0	55
16	Ser	Serine	(UCU, UCC, UCA, UCG, AGU, AGC)	6	105.09	0.359	3	1	0	1	0	3	6	0	8	0	17
17	Thr	Threonine	(ACU, ACC, ACA, ACG)	4	119.12	0.45	5	2	0	1	0	5	12	0	8	0	25
18	Val	Valine	(GUU, GUC, GUA, GUG)	4	117.15	0.825	7	3	0	0	0	0	7	18	0	0	25
19	Trp	Tryptophane	(UGG)	1	204.23	0.878	8	9	1	0	0	8	54	7	0	0	69
20	Tyr	Tyrosisne	(UAU, UAC)	2	181.19	0.88	7	7	0	1	0	7	42	0	8	0	57