Clustering of Red/White Wine and Allergen/Non-Allergen Data Sets by Using Descriptor Fingerprints

B P Stoyanov¹,², I E Dimitrov³, I A Doytchinova³ and I P Bangov¹

¹Konstantin Preslavsky University of Shumen, 115 Universitetska Str., Shumen, 9712, Bulgaria
²Chernorizets Hrabar Free University of Varna, Department of Computer Science, 84 Yanko Slavchev Str., Chayka Resort, 9007, Varna, Bulgaria
³Medical University of Sofia, 2 Dunav Str., 1000, Sofia, Bulgaria

e-mail: borislav.stoyanov@shu.bg

Abstract. An application of the descriptor fingerprints in relation to a clustering procedure of red and white wines and food allergen/non-allergens has been studied. The clustering method is based on a procedure using the Tanimoto similarity index. Different threshold values have been employed in order a best set of clusters to be obtained. It was shown that this procedure produces quite good discrimination between the red and white wines, but any correlation with the human sensation tastes (represented by a scale from 1 to 10) is more difficult to be obtained. The results from 4212 food allergens/non-allergens show a much better discrimination of those two groups starting from threshold value=0.5.

1. Introduction

Structural fingerprints have been widely used in structural similarity search procedures [1–4]. They represent the chemical structure as a unique array of fragments which is considered a fingerprint. Thus, each fingerprint, representing a structure is an array of 0s and 1s with elements one-to-one corresponding to the fragments in an array created in advance. The element 0 implies an absence of a given fragment from a fragment array, and an element of 1 implies a presence of such a fragment. These fingerprints we may call structure connectivity fingerprints. Obviously they have Boolean character, because of having any fragment 1 equals true, and not having it, 0 equals false.

Fingerprints have been exploited in other areas, such as 3D geometry of 3-and 4-point pharmacophores [5] and 3D fingerprints of potential pharmacophore point triangles [6], Computer-assisted structure elucidation (CASE) programs where fingerprint spectra (13C NMR, IR) [7] are employed in bit string chromosome representations [8], allergenicity prediction [9], and quantitative structure-activity relationship (QSAR) algorithms [10].

Further we have developed new type fingerprints, which are related with some properties of the studied objects [10–13]. Here instead of a fragment array we introduced an array of properties described by descriptors. Thus, each fingerprint, representing such an object is an array of 0s and 1s with elements one-to-one corresponding to such a descriptor array. We call these fingerprints, descriptor fingerprints.

2. Definition of Descriptor Fingerprints

We create descriptor fingerprints by a determination an interval and the precision of their possible values, for each descriptor an initial value (fromValue) and a final value (toValue) with accuracy step (resolution) resValue are assigned. The resValue value is assigned by the user. It determined the number
of sub-intervals that the interval of a descriptor has. The fromValue and toValue values were extracted by scanning the whole database and assigning the minimum value for each descriptor to its fromValue and the maximum value for each descriptor to its toValue. Hence, for each descriptor its interval is divided into \( N = \frac{(toValue - fromValue)}{resValue} \) sub-intervals (elements of the fingerprint array) and the program determines in which element the current descriptor value falls by putting an 1 in this element, the other elements remaining zeros. Thus, a descriptor fingerprint is formed for each one object by applying this procedure to all descriptors and concatenating them.

The fundamental problem here is, as in other QSAR algorithms, the selection of proper descriptors which describe adequately either the molecule species or the various products studied. They can be either real or discrete values. In the specific case of our study of wines and proteins we use continuous values, for the descriptors which are converted into discrete values by scaling and dividing into intervals the fingerprint array. If a descriptor value falls within a given interval, a number 1 is put in the corresponding fingerprint element; otherwise, the element takes 0.

Thus, the length of the fingerprint depends on the number of intervals and they depend on the step \( (resValue) \) we have chosen to divide the range between the fromValue and the toValue. Thus, the more intervals we have the longer is the fingerprint. The balance between the length and information content of the fingerprints must be found.

3. Cluster Generation by using Descriptor Fingerprints

We employ the method of Butina [14] for cluster generation in this paper. This method is based on the similarity between different objects. The similarity comparison between two objects (red and white wines and allergen and non-allergen proteins in our case) is carried out by using the Tanimoto index [15].

Thus, for two fingerprints \( a \) and \( b \) the Tanimoto index has the following form: \( T(a,b) = \frac{Nc}{Na + Nb - Nc} \). Here \( Nc \) is the number of the common 1s (being in the same location of both fingerprint arrays), \( Na \) and \( Nb \) are the total numbers of 1s in the fingerprints of objects \( a \) and \( b \), respectively. In as much as in our case of descriptor fingerprints there one 1 in any sub-interval of each descriptor interval then the numbers of 1s for both fingerprints is always equal to the number of descriptors. Hence, our Tanimoto expression will take the form of: \( T(a,b) = \frac{Nc}{2Nd - Nc} \), where \( Nd \) is the number of the descriptors.

The Tanimoto (T) index takes real values between 0.0 and 1.0. The larger is the value the more similar are the two structures. Thus, two structures having Tanimoto index \( T = 0.98 \) are considered much more similar than two structures of Tanimoto index \( T = 0.56 \).

Consider a database of objects (chemical structures, chemical products, proteins, fuels, etc.) with known properties such as biological activity, exploitation properties, etc., characterized by their descriptors fingerprints. If a new product is present, its descriptors are determined and its fingerprint generated. Then a similarity search can be carried out on the database comparing this fingerprint with all database fingerprints by calculating the Tanimoto index for fingerprints. Thus, the Tanimoto value will show how close the new product to any of the database objects is.

Butina clustering procedure consists in the following steps:

Step 1: The fingerprints of the database are compared within the database (without comparing any fingerprint with itself) by calculating the Tanimoto indecies for each pair of fingerprints. There is an input threshold Tanimoto value, and only pairs of values higher than this value are selected. To each fingerprint the fingerprints that form pair of Tanimoto value higher the threshold value are considered its neighbors.

Step 2: The fingerprints are sorted according to their number of the neighbors.

Step 3: A new scan is carried out, as the fingerprints with their neighbors form separate clusters. If a fingerprint is selected in a cluster it is omitted for a further use within the clustering process.
4. Results and Discussion

The main purpose of this work is to investigate the different factors that influence the clustering process of our descriptor fingerprints. In this study we have applied the clustering process to two sets, one of white and red wines and one of allergen/non-allergen food proteins.

There are two factors which can influence the clustering process. The first one is the resolution (the number of sub-intervals corresponding to a descriptor), the second is the threshold value.

4.1. Clustering of Red and White Wines

A large dataset was formed of a group of 1599 red wines (Vinho Verde samples from Northern regions of Portugal) and a group of 3898 white wines, each one of 11 analytic laboratory benchmarks. Laboratory analytical test values (11 descriptors) have been taken from literature [16] and used for descriptor fingerprints creation. Descriptor fingerprints were generated on the basis of physicochemical laboratory data routinely used for wine characterization such as fixed and volatile acidity; residual sugar, total and free sulfur dioxide, citric acid, chlorides, sulfates, density, pH and alcohol content. The FPCHEM software developed by us has been used to this end. The sensory tests are usually carried out by human senses such as flavor and taste and they require extremely experienced persons. The relationships between the human sensor tests and physics-chemical analyses are too much complex and still poorly understood [17]. Thus wine classification becomes a serious problem. Our investigation aimed to check and expose the possibility of using a descriptor fingerprints to rank and distinguish different classes of wines based on laboratory test data (classified by the sensory testers from 1 to 10). This work aims at a study of the influence of the different factors on the clustering of our database of red and white wines based on the descriptor fingerprint created from objective analytic tests that are available at the certification step and studying their correlations with the human sensitive tests.

We have the following results which characterize the efficiency of our clustering: number of clusters, differentiation between red and white wines (number of red/white bad hits), and the correctness of the prediction of the results obtained by the humans-tests (number of test bad hits) are the values which are to demonstrate the possibility of the method of descriptor fingerprints to this end. Here, the dependence of those values on resolution and the threshold value is has been investigated, and the results are provided in Table 1.

One can see that with the increase of the threshold value the number of the clusters increases and the number of bad hits both red/white and the relation to the human tests bad hits decreases. This increase of the number of the clusters and decrease of both the red/white wine bad hits and tester bad hits is sharper with the increase of both the threshold value and the resolution. Thus in resolution equal to 30 we have very good results for the red/white recognition at threshold values starting from 0.6. However, such a relation to the human based tests appears to be not so good. It is obvious that the human senses do not feel the whole chemical factors within the wine flavor and taste.

4.2. Allergenicity Clustering of Food Proteins by Descriptor Fingerprints

In the present study we use 4212 proteins both allergens and not allergens characterized by 200 E-descriptors. A more detailed description on the generation of these descriptors is reported in the paper [9]. In short, the E-descriptors for the 20 naturally occurring amino acids, defined by Braun and Venkatarajan [18], were derived by principal component analysis of a data matrix consisting of 237 physicochemical properties. The first principal component (E1) reflects the hydrophobicity of amino acids; the second (E2) – their size; the third (E3) – their helix-forming propensity; the forth (E4) correlates with the relative abundance of amino acids; and the fifth (E5) is dominated by the β-strand forming propensity. In the present study the five E-descriptors were used to describe the protein sequences.

To make the length of the proteins used in the present study uniform, an auto-cross covariance (ACC) transformation was used [19]. Auto-covariance $A_{jj}(l)$ and cross-covariance $C_{jk}(l)$ were calculated according to the following equations:

$$A_{jj}(l)=\sum_{i=1}^{n-l}E_{i,j}E_{i+l,j}$$

$$C_{jk}(l)=\sum_{i=1}^{n-l}E_{i,j}E_{i+l,k}$$

(1)
\[ C_{jk}(l) = \sum_{n-l}^{n-1} \frac{E_{j,i} \times E_{k,i+l}}{n-l} \] 

(2)

Table 1. Results of red/white wines clustering.

| Threshold value | Number of clusters | Number of taste bad hits | Number of red/white wine bad hits |
|-----------------|--------------------|--------------------------|-----------------------------------|
| Resolution 10   |                    |                          |                                   |
| 0.1             | 7                  | 3407                     | 1437                              |
| 0.2             | 13                 | 3202                     | 1420                              |
| 0.3             | 85                 | 3125                     | 951                               |
| 0.4             | 188                | 3125                     | 673                               |
| 0.5             | 447                | 2853                     | 246                               |
| 0.6             | 835                | 2219                     | 78                                |
| 0.7             | 1199               | 1205                     | 19                                |
| 0.8             | 1199               | 1205                     | 19                                |
| 0.9             | 1028               | 346                      | 4                                 |
| Resolution 20   |                    |                          |                                   |
| 0.1             | 24                 | 3275                     | 1354                              |
| 0.2             | 70                 | 3342                     | 1174                              |
| 0.3             | 495                | 2790                     | 331                               |
| 0.4             | 960                | 2145                     | 96                                |
| 0.5             | 1258               | 1095                     | 16                                |
| 0.6             | 1100               | 405                      | 3                                 |
| 0.7             | 920                | 135                      | 3                                 |
| 0.8             | 920                | 135                      | 3                                 |
| 0.9             | 834                | 30                       | 3                                 |
| Resolution 30   |                    |                          |                                   |
| 0.1             | 52                 | 3214                     | 1296                              |
| 0.2             | 70                 | 3342                     | 1174                              |
| 0.3             | 495                | 2790                     | 331                               |
| 0.4             | 960                | 2145                     | 96                                |
| 0.5             | 1258               | 1095                     | 16                                |
| 0.6             | 1100               | 405                      | 3                                 |
| 0.7             | 920                | 135                      | 3                                 |
| 0.8             | 920                | 135                      | 3                                 |
| 0.9             | 834                | 30                       | 3                                 |

Indices \( j \) and \( k \) refers to the \( E \)-descriptors \((j = 1-5, k = 1-5, j \neq k)\), \( n \) is the number of amino acids in a sequence, index \( i \) points the amino acid position \((i = 1, 2, \ldots, n)\) and \( l \) is the lag \((l = 1, 2, \ldots, L)\). Short lags \((L = 8)\) were chosen, as only the influence of close amino acid proximity was investigated. The subsets of antigens and non-antigens were transformed into matrices with 200 variables \((5^2 \times 8)\) each.

The results of allergen/non-allergen activity cluster analysis (number of clusters and activity bad hits) as a function of the threshold value and the resolution are presented in Table 2. Here we can see that again as in the case of red/white wines we have an increase of the number of clusters and a decrease of the bad hits with the increase of the threshold value and of the resolution. It appears that a good database can be created with threshold value 0.5—0.6 and resolution 30.
Table 2. Results of allergen/non-allergen activity clustering.

| Threshold value | Resolution 10 | Resolution 20 | Resolution 30 |
|-----------------|---------------|---------------|---------------|
|                 | Number of clusters | Activity bad hits | Number of clusters | Activity bad hits | Number of clusters | Activity bad hits |
| 0.1             | 4              | 2197          | 22             | 2168           | 35             | 2112           |
| 0.2             | 23             | 2166          | 60             | 2009           | 195            | 1543           |
| 0.3             | 33             | 2097          | 199            | 1453           | 403            | 724            |
| 0.4             | 65             | 1865          | 391            | 800            | 527            | 85             |
| 0.5             | 156            | 1600          | 527            | 213            | 536            | 1              |
| 0.6             | 326            | 1035          | 546            | 11             | 479            | 0              |
| 0.7             | 515            | 252           | 501            | 0              | 434            | 0              |
| 0.8             | 533            | 2             | 422            | 0              | 324            | 0              |
| 0.9             | 412            | 0             | 326            | 0              | 135            | 0              |

5. Conclusions
These preliminary results show that the descriptor fingerprints proposed by us could be used both for clustering on the basis of Tanimoto similarity index hence, for a QSAR prediction of unknown materials. Please note that in the case of red and white wines, we have a comparison between some chemical values and the tasting carried out by the human sensations. We have comparatively good results with the red and white wine discrimination (number of red/white wine bad hits=3, at threshold value=0.6), but to obtain the precise tasting results (number of taste bad hits=30, just at threshold value=0.9) is more difficult. Obviously the tasters do not fill all the chemicals constituting the wine’s flavor. Much better results are observed with the discrimination between food allergens and non-allergens. By using a threshold value of 0.5 we have activity bad hits =1, and above activity bad hits=0.

Acknowledgements
The paper was partially supported by the National Scientific Program “Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)”, financed by the Ministry of Education and Science, Bulgaria, for Borislav Stoyanov and Ivan Bangov.

References
[1] Barnard J M 2008 Representation of Molecular Structures-Overview (Handbook of Chemoinformatics) ed J Gasteiger (Weinheim, Germany: Wiley-VCH Verlag GmbH) chapter 3 pp 27–50
[2] Engel T 2003 Representation of Chemical Compounds Chemoinformatics (Chemoinformatics: A Textbook) ed J Gasteiger and T Engel (Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA) chapter 2 pp 15–168
[3] Kochev N, Monev V and Bangov I 2003 Searching Chemical Structures Chemoinformatics (Chemoinformatics: A Textbook) ed J Gasteiger and T Engel (Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA) chapter 6 pp 291–318
[4] Willett P 2008 Similarity Searching in Chemical Structure Databases (Handbook of Chemoinformatics) ed J Gasteiger (Weinheim, Germany: Wiley-VCH Verlag GmbH) chapter 33 pp 904–912
[5] Mason J S, Morize I, Menard P R, Cheney D L, Hulme C and Labaudiniere R F 1999 J. Med. Chem. 42 3251–3264
[6] Brown R D and Martin Y C 1996 J. Chem. Inf. Comput. Sci. 36 572–584
[7] Von Homeyer A 2008 Evolutionary Algorithms and Their Applications (Chemistry Handbook of Chemoinformatics) ed J Gasteiger (Weinheim, Germany: Wiley-VCH Verlag GmbH) pp 1239–1280
[8] Cho S J and Hermesmeier M A 2002 J. Chem. Inf. Comput. Sci. 42 927–936
[9] Dimitrov I, Naneva L, Doychinova I and Bangov I 2014 Bioinformatics 30 846–851
[10] Dimitrov I, Kochev N, Moskovkina M, Naneva L, Paskaleva V, Doychinova I, Milina R, Mustafa Z and Bangov I 2014 Acta Sci. Nat. 1 45—50
[11] Mustava Z, Moskovkina M, Milina R and Bangov I 2017 Int. J. Sci. Eng. Res. 8 1720—1738
[12] Mustava Z, Moskovkina M, Milina R and Bangov I 2015 Acta Sci. Nat. 1 164—170
[13] Bangov I, Moskovkina M, Mustava Z and Milina R 2015 Acta Sci. Nat. 2 43—46
[14] Butina D 1999 J. Chem. Inf. Comput. Sci. 39 747–750
[15] Tanimoto T T 1958 An Elementary Mathematical Theory of Classification and Prediction (New York, International Business Machines Corporation) p 10
[16] Cortez P, Cerdeira A, Almeida F, Matos T and Reis J 2009 Decis. Support Syst. 47 547–553
[17] Ebeler S E 1999 Linking Flavor Chemistry to Sensory Analysis of Wine (Flavor Chemistry) ed R Teranishi, E L Wick and I Hornstein (Boston, MA: Springer US) chapter 35 pp 409–421
[18] Braun W and Venkataraman M S 2001 J. Mol. Model. 7 445–453
[19] Nystrom A, Andersson P M and Lundstedt T 2000 Quant. Struct.-Act. Relat. 19 264–269