Bulgarian Chemical Communications, Volume 46, Special Issue A (pp. 209 – 214) 2014

Use of the descriptor fingerprints to clustering of chemical datasets

B. P. Stoyanov[1], P. R. Kostadinov[1], M. V. Kolev[1], Z. A. Mustafa[2], M. N. Moskovkina[3], R. S. Milina[2], I. P. Bangov[3]*

[1]*Department of Computer Informatics, Faculty of Mathematics and Informatics, Konstantin Preslavski University of Shumen, 115 Universitetska str., 9712 Shumen, Bulgaria*

[2]*Central Research Laboratory, Faculty of Natural Sciences, Prof. Assen Zlatarov University, 62 San Stefano str., 8001 Burgas, Bulgaria*

[3]*General Chemistry Chair, Faculty of Natural Sciences, Konstantin Preslavski University of Shumen, 115 Universitetska str., 9712 Shumen, Bulgaria*

Received April 28, 2014; Revised May 20, 2014

Dedicated to Acad. Dimiter Ivanov on the occasion of his 120th birth anniversary

A novel approach in the area of chemoinformatics, the use of descriptor fingerprints has been applied to the problem of clustering of chemical databases. This approach coupled with the clustering method of Butina (originally created for structural fingerprints) was tested with a set of 96 biodiesel fuels. The influence of the threshold values from 0.0 to 1.0 of the Tanimoto index on the clustering results was studied. The results show a good discrimination power of the method, biodiesels of the same oils fall in one the same clusters. It was also shown that similar cetane number (CN) values of biodiesels within the statistical accuracy fall in the same cluster.

Key words: clustering analysis, descriptor fingerprints, Tanimoto index, biodiesels

INTRODUCTION

There are many methods developed in the field of chemoinformatics, such as neuronal networks, partial least square (PLS), multivariable regression, discriminant analysis etc. They have their advantages and drawbacks. For a year we have developping a novel approach - *descriptor fingerprints* devised by one of the authors (IB) [1-3], toward the similarity of different chemical and/or non-chemical materials. Its potentials have been demonstrated with the creation of a server on the net for allergen and non allergen food proteins compared with servers based on the traditional chemoinformatics methods [2]. Hence, the purpose of this work is to study the influence of the *threshold value* on the results from similarity clustering with descriptor fingerprints.

The basic idea of descriptor fingerprints comes from the well known structural fingerprints [4-6]. Both structural and descriptor fingerprints are defined either as a character or binary string of *0*s and *1*s.

In the case of structural fingerprint an array of structural fragments is juxtaposed to the fingerprint array, each element of the former corresponding to the element of the latter being at the same location. The general scheme is shown on Fig. 1.

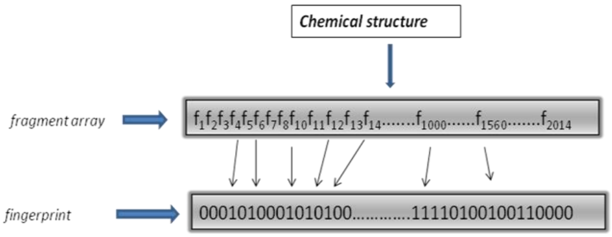
Chemical structure
fragment array
$f_1f_2f_3f_4f_5f_6f_7f_8f_{10}f_{11}f_{12}f_{13}f_{14}$.......$f_{1000}$......$f_{1560}$.......$f_{2014}$
fingerprint
0001010001010100.............11110100100110000

Fig. 1. Fingerprint generation of chemical structures.

The program automatically determines these minimum and maximum values for each descriptor from all the descriptors in the set. A precision (resolution) - *resValue* of each descriptor is user defined. Hence, for each real value descriptor its interval is divided into *N=(initValue-endValue)/ resValue* discrete sub-intervals. Some descriptor could be binary, e.g. the presence or absence of a property. They are encoded by one interval. Thus, the formation of the fingerprint from all its descriptors is carried out by concatenation of all their sub-intervals. In case a descriptor real value falls in an interval a 1 is put in the corresponding position of the fingerprint array, the other elements corresponding to the descriptor remaining 0s. The formation of a descriptor fingerprint is illustrated in Fig. 2.

* To whom all correspondence should be sent:
E-mail: ivan.bangov@gmail.com

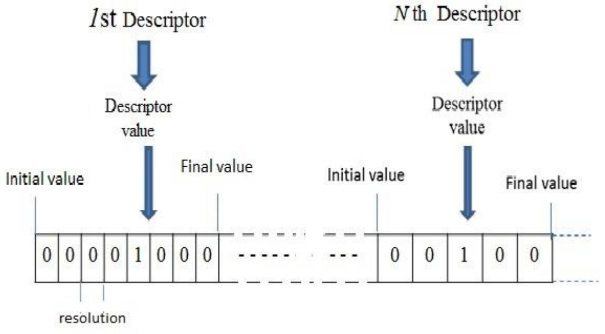

Fig. 2. Formation of a descriptor fingerprint.

The similarity search of two objects (chemical structures, materials, etc.) is carried out by comparing their fingerprints using any similarity measure [7,8]. In this work the Tanimoto measure has been employed to this end. It has the following form:

$$T=N_C/(N_A+N_B-N_C). \qquad (1)$$

Here, N_A is the number of elements having value 1 in the first fingerprint A, N_B is the number of elements having value 1 in the second fingerprint B, and N_C is the number of the common elements (being in the same position in both fingerprints) having value 1. This index has real values between 0.0 and 1.0. The higher is the value the more similar are the two objects.

CLUSTERING OF SETS OF OBJECTS

Clustering of set of objects is based on the cluster analysis. Here is a definition from Wikipedia [9]:

*“**Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.”*

There are several algorithms for clustering of different sets of objects. Some of them are the *connectivity based*, *distance measure based*, other are *centroid based* approaches. Here we use the clustering algorithm of Butina [10]. It consists of the following procedures:

- ◆First, the Tanimoto index is calculated pairwise for all the objects in the set within the given threshold value (the smallest Tanimoto value accepted for a pair).
- ◆Second, the objects are sorted according to the number of their neighbours, i.e. forming from most to less numerous pairs within the given threshold value, i.e., with Tanimoto value between the threshold value and 1.0.
- ◆Further, another pass of pairwise fingerprint Tanimoto calculation is carried out, starting from the first pair of the sorted set. All its neighbours form the first cluster and are subsequently flagged as “used”. Then, follows the same procedure with the second unused object, applied to the remaining unused fingerprints only, and the second cluster is formed, then the next clusters are formed in the same way. A last cluster is formed from the objects having 1 neighbour only.
- ◆An additional procedure of automatic perception of the threshold value (TV) was developed by us. An extra preliminary pairwise Tanimoto calculation of the data set with a subsequent ranking of the pairs was carried out. Further a procedure scan the names of the ranked pairs and finds out at what value the Tanimoto index produces false result. This value increase by 0.00001 is taken for a threshold value (TVauto).

APPLICATION, RESULTS AND DISCUSSION

We applied this clustering approach to a set of 96 biodiesel fuels of several vegetable oils: sunflower (SF), soya beans (SB), corn (CR), rape seed (RS), peanut (PN), palm (PM) and some mixed cases (MIX). They have been selected from different feed stocks (with different composition and properties respectively). In as much as the biodiesel fuel is a mixture of fatty acid methyl esters (FAME) its properties depend on the chemical structure of the individual FAME and their contents (FAME profile) [11-14]. Hence, the descriptor fingerprint method has been based on their gas chromatographic FAME profiles [15-19]. To the aim some available literature data on FAME profiles of biodiesels from several types of oils were used [11,14,20-22].

Our method of descriptor fingerprint clustering was applied to a set of different biodiesel fuels in order to derive a new alignment for them into some groups (clusters).

In this paper we study the influence of the threshold value (TV) on the clustering. The results are provided in Table 1.

One can see from Table 1, that the best clustering is with threshold value TV=0.7. Thus, it is seen that a perfect discrimination leading to separation of the different biodiesels, say, all sun-

Table 1. Results of the biodiesel oils clustering.

Cluster No	Vegetable oils in biodiesels	TV= 0.0	TV= 0.1	TV= 0.2	TV= 0.3	TV= 0.4	TV= 0.5	TV= 0.6	TV= 0.7	TV= 0.8	TV= 0.9	TV= 1.0	TVauto= 0.5263158
1	SF	33	33	32	30	28	25	21			5	3	13
	SB	19	19	19	18	12	1		11	7			
	CR	8	8	8	7	2							
	PN	6	5										
	RS	19	19	12									
	PM	6	6	5									
	Mix	5	5	1	1								
2	SF			1	3	2			11	6			
	SB				1	1	14	15			3	2	12
	CR					2							
	Mix			2									
3	SF						2	2	4	5	3	2	2
	SB					6	1						
	CR					2	2						
	RS			7	18								
	Mix			2									
4	SF						4	1	3	3			4
	PN			6									
	RS					14							
	PM				6								
	SB										2		
	CR							1					
5	SB						2			2	2		
	PN				4								
	RS					2							5
	SF							2					
	PM								4				
6	RS						6						
	PM					6							
	Mix				3								
	SF								3	3			
	SB							1			2		2
	CR							2					
7	CR					2							
	PN				2								
	PM						5						
	SF										2		
	SB								2	2			1
	RS							5					
8	SF					2		3					
	RS						3		3	2	2		
	SB												1
9	CR						3						
	PN					3							
	SF									2			3
	RS								3				
	PM							5					
10	RS						2	4	2	2			3
	Mix					2							
11	PN					2			2				
	RS						4	3					
	PM									2			
	CR												2

12	RS					3			2			
	SF							2				
	PN						3					
13	PN					3						
	RS						2					
	PM											3
14	PN					2						
	SF						2					2
15	SF					2						
	RS						2					
Last	SF				1		2	10	14	23	25	
	SB					1	3	6	8	10	17	
	CR			1		3	5	8	8	8	8	
	PN	1			1		3	4	6	6	6	
	RS			1	3		3	11	13	17	19	
	PM		1			1	1	2	4	6	6	
	Mix			1	3	5	5	5	5	5	5	

Table 2. Statistical data concerning cetane number of biodiesel fuels for the highest Tanimoto index values.

Cluster	N class	CN	Mean	Standard deviation	Confidence level (95%)
T=1					
1	14 SF	48.695	48.508	0.166	0.413
1	17 SF	48.453			
1	21 SF	48.377			
2	61 SB	49.260	49.020	0.340	3.056
2	67 SB	48.779			
3	6 SF	47.758	47.884	0.177	1.595
3	7 SF	48.009			
T=0.9					
1	14 SF	48.695	48.467	0.182	0.226
1	15 SF	48.584			
1	16 SF	48.224			
1	17 SF	48.453			
1	21 SF	45.377			
2	61 SB	49.260			
2	67 SB	48.779	48.774	0.489	1.214
2	76 SB	48.283			
3	5 SF	48.191			
3	6 SF	47.758	47.986	0.217	0.540
3	7 SF	48.009			

flower biodiesels in one cluster, all soy bean biodiesels in another etc., is not achieved. We have two clusters of sunflower biodiesels, two clusters of soy bean biodiesels. This situation can be explained with different fatty acid composition of vegetable oils.

The results from TV=0.6, TV=0.7, TV=0.8, and TV=0.9 providing best discrimination are visualized in Fig. 3, Fig.4, Fig. 5, and Fig. 6.

One can see from Table 1 and Fig. 3, Fig. 4, Fig. 5, and Fig. 6 that the number of biodiesels grouped in a separate cluster depends on the value of TV. At low TVs the number of objects in a cluster increases but the number of clusters decreases. Thus, at TV=0.0, and TV=0.1 we have only 1 cluster.

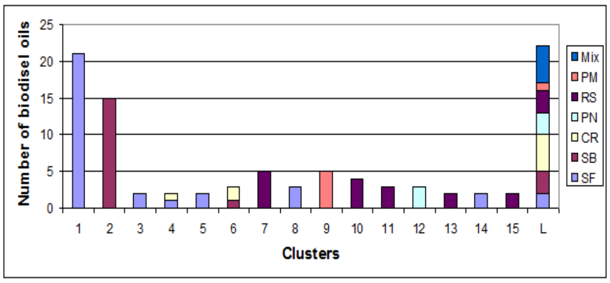

Fig. 3. Distribution of the different biodiesels at TV=0.6.

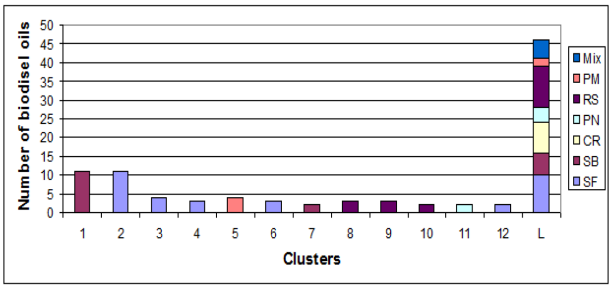

Fig. 4. Distribution of the different biodiesels at TV=0.7.

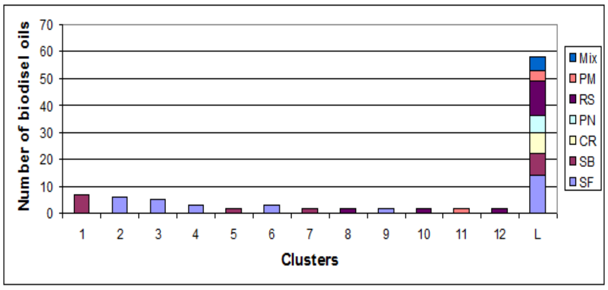

Fig. 5. Distribution of the different biodiesels at TV=0.8.

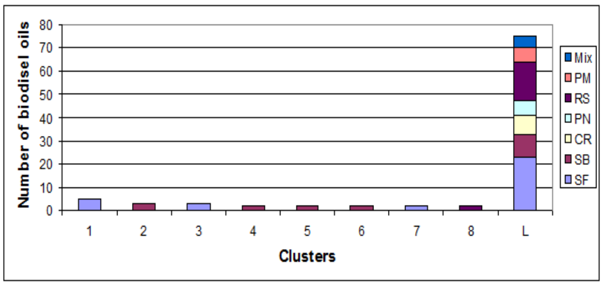

Fig. 6. Distribution of the different biodiesels at TV=0.9.

There are various parameters, used for characterization of different biodiesel grades. The determination of the cetane number (CN) is an important index from our clustering analyses. The CN values of the biodiesel fuels, grouped in clusters having T-values 1.0 and T=0.9 are shown in Table 2. Some statistical indicators such as mean and standard deviation (confidence level 95%) are presented also in Table 2.

One can see that the oils grouped in a given cluster are of the same type on the one hand and on the other hand they produce similar CN values.

The lowering of Tanimoto index value extends the frontiers of the cluster; this tendency is characterized by the statistical indicator of the Standard Deviation value, presented in Table 2.

CONCLUSION

A group of 96 biodiesel oils of different origin was used for creation of descriptor fingerprints based on their gas chromatographic FAME profiles. The influence of the threshold value on the clustering results was studied. As a result of the clustering analysis based on descriptor fingerprints biodiesel oils of the same type were grouped into distinct clusters. It was observed that in general, biodiesels having similar values of some important parameters (properties) fall in the same clusters. It was shown the CN values are correlating with the Tanimoto index are within the statistical accuracy. The obtained results allow us to carry out in-depth studies on the ability of the method of descriptor fingerprints for prediction of various analytical properties.

REFERENCES

1. I. Bangov, L. Naneva, M. N. Moskovkina, I. Dimitrov, I. Doytchinova, *Annual of Konstantin Preslavski University of Shumen, Natural Sciences, Chemistry*, **XXII B1**, 28 (2013).
2. I. Dimitrov, L. Naneva, I. Doytchinova, I. Bangov, *Bioinformatics*, **30**, 846 (2014).
3. I. Dimitrov, N. Kochev, M. Moskovkina, L. Naneva, V. Paskaleva, R. Milina, Z. Mustafa, I. Doychinova, I. Bangov, *Acta Sci. Natur.*, (2013) submitted.
4. J. M. Barnard, *in Handbook of Chemoinformatics*, **1**, 27 (2003).
5. J. Tomczak, *in Handbook of Chemoinformatics*, **2**, 392 (2003).
6. T. Engel, *in Chemoinformatics. A textbook*, 15 (2003).
7. P. Willett, *in Handbook of Chemoinformatics*, **2**, 904 (2003).
8. N. Kochev, V. Monev, I. Bangov, *in Chemoinformatics. A textbook*, 291 (2003).
9. http://en.wikipedia.org/wiki/Cluster_analysis.
10. D. Butina, *J. Chem. Inf. Comput. Sci.*, **39**, 747 (1999).
11. S. K. Hoekman, A. Broch, C. Robbins, E. Ceniceros, M. Natarajan, *Renew. Sust. Energ. Rev.*, **16**, 143 (2012).
12. Y. C. Sharma, B. Singh, S. N. Upadhyay, *Fuel*, **87**, 2355 (2008).
13. A. E. Atabania, A. S. Silitonga, I. A. Badruddina, T. M. I. Mahlia, H. H. Masjuki, S. Mekhilef, *Renew. Sust. Energ. Rev.*, **16**, 2070 (2012).
14. M. J. Ramos, C. M. Fernandez, A. Casas, L. Rodriguez, A. Perez, *Bioresource Technol.*, **100**, 261 (2009).
15. C. P. Prados, D. R. Rezende, L. R. Batista, M. I. R. Alves, N. R. A. Filho, *Fuel*, **96**, 476 (2011).
16. E. D. Dodds, M. R. McCoy, L. D. Rea, M. J. Kennish, *Lipids*, **40**, 419 (2005).
17. L. Mondello, A. Casilli, P. Q. Tranchida, R. Costa, B. Chiofalo, P. Dugo, *J. Chromatogr. A*, **1035**, 237 (2004).
18. E. Bravy, G. Perrety, L. Montanary, *J. Chromatogr. A*, **1134**, 210 (2006).

19. R. Milina, Z. Mustafa, *Petroleum & Coal*, **55**, 12 (2013).
20. B. R. Moser, S. F. Vaughn, *Biomass Bioenerg.*, **37**, 31 (2012).
21. D. Brodnjak-Voncina, Z. C. Kodba, M. Novich, *Chemometr. Intell. Lab.*, **75**, 31 (2005).
22. E. G. Giakoumis, *Renew. Energ.*, **50**, 858 (2013).

ИЗПОЛЗВАНЕ НА ДЕСКРИПТОРНИТЕ ОТПЕЧАТЪЦИ НА ПРЪСТИТЕ ЗА КЛЪСТЕРИРАНЕ НА ХИМИЧНИ МНОЖЕСТВА ОТ ДАННИ

Б. П. Стоянов[1], П. Р. Костадинов[1], М. В. Колев[1], З. А. Мустафа[2], М. Н. Московкина[3], Р. С. Милина[2], И. П. Бангов[3]*

[1]Катедра по компютърна информатика, Факултет по математика и информатика, Шуменски университет "Константин Преславски", ул. Университетска 115, 9712 Шумен, България
[2]Централна изследователска лаборатория, Факултет по природни науки, Университет "Проф. Асен Златаров", ул. Сан Стефано 62., 8001 Бургас, България
[3]Катедра Обща химия, Факултет по природни науки, Шуменски университет "Константин Преславски", ул. Университетска 115, 9712 Шумен, България

Постъпила на 28 април 2014 г.; Коригирана на 20 май 2014 г.

(Резюме)

Един нов подход в областта на хемоинформатиката - използването на дескрипторните отпеачатъци на пръстите е приложен към проблема за клъстериране на бази от химически данни. Този подход, придружен с метода на Бутина (първоначално създаден за структурни дескриптори) беше тестват с група от 96 биодизелни горива. Влиянието на праговата стойност от 0.0 до 1.0 на индекса на Танимото върху резултатите от клъстерирането беше изследвано. Резултатите показаха една добра разделителна способност на метода, като биодизели с еднакви масла попадат в едни и същи клъстери. Беше показано също, че подобни стойности на цетановото число на биодизелите попадат в едни и същи клъстери в рамките на статистическата точност.