MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **57** (2007) 51-67

ISSN 0340 - 6253

A new similarity/diversity measure for sequential data

R. Todeschini[a*], *D. Ballabio*[b], *V. Consonni*[a], *A. Mauri*[a]

[a] Milano Chemometrics and QSAR Research Group - Dept. of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1 – 20126 Milano (Italy)

[b] Dept. of Food Science and Technology, University of Milan, via Celoria, 2 – 20133 Milano (Italy)

Web site: www.disat.unimib.it/chm/

(Received June 23, 2006)

Abstract

The concept of similarity and its dual concept of diversity play a fundamental role in several QSAR strategies, chemometrics and library searching methods, virtual screening, as well as in relatively new fields such as genomics and proteomics.

In this paper, a new flexible similarity/diversity measure is proposed to deal with sequential data, both taking into account the differences in property values of the sequence elements and the ordering relationships among the sequence elements themselves.

Data such as DNA sequences, mass and NMR spectra, sequential molecular descriptors are all characterized by an ordering variable (the sequence) and by a property of the sequence elements.

Some examples on artificial DNA sequences, mass spectra, molecular descriptors and proteomic maps are given.

1. Introduction

The concept of similarity and its dual concept of diversity play a fundamental role in several QSAR strategies, chemometrics and library searching methods, virtual screening, as well as in relatively new fields such as genomics and proteomics. Several distance measures both for quantitative and binary

variables have been defined, such as, for example, Euclidean, Manhattan, Minkowski, Camberra distances for quantitative variables, and Hamming, Tanimoto, Jaccard distances for binary variables. Distances are the quantitative measure of diversity between a pair of objects, thus large distances indicate large diversity, i.e. small similarity, and small distances indicate small diversity, i.e. large similarity.

In this paper, a new similarity/diversity measure is proposed as a new approach to the analysis of sequential data, where useful information can be also obtained by the ordering relationships between the sequence elements.

The new proposed distance (weighted standardized Hasse distance) is evaluated between pairs of Hasse matrices derived from the classical partial ordering rules. It can be naturally standardized, thus allowing to interpret these distances as absolute values (e.g. percentage) and deriving simple similarity and correlation indices.

Simple examples of this methodology are given, showing its main characteristics and possible different applications.

2. Theory

The theory of the proposed approach to the similarity/diversity analysis of sequential data is presented introducing some partial ordering concepts, the Hasse matrix and the corresponding similarity/diversity measures. In the paragraph 2.4 a simple example of calculation is also given.

2.1 Partial Ordering (PO)

Partial Ordering is a ranking approach where the relationship of "incomparability" is added to the classical relationships of "greater than", "less or equal than", etc. [1-3].

Given a set Q of n elements, each described by a vector $\mathbf{x}$ of p variables (attributes), the two elements s and t belonging to Q are comparable if *for all* the variables x_j either $x_j(t) \geq x_j(s)$ or $x_j(s) \geq x_j(t)$. If $x_j(t) \geq x_j(s)$ for all x_j ($j = 1, ..., p$) then $t \geq s$, i.e. t covers s (or s is *covered* by t). The request "for all" is very important and is called the *generality principle*:

$$t \geq s \Leftrightarrow x_j(t) \geq x_j(s) \qquad \forall\, j \in [1,p] \qquad (1)$$

The ordering relationships between all the pairs of elements are collected into the Hasse matrix; for each pair of elements s and t the entry H_{st} of this matrix is:

$$H_{st} \begin{cases} +1 & \text{if } x_j(s) \geq x_j(t) \quad \forall\, j \in [1,p] \\ -1 & \text{if } x_j(s) < x_j(t) \quad \forall\, j \in [1,p] \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

If the entry s-t contains +1, the entry t-s contains –1; if the entry s-t contains 0, also the entry t-s contains 0. Then, the Hasse matrix is a square nxn matrix whose elements take only the values 0 and

±1; if pairs of equal elements are not present, it is also antisymmetric matrix. In fact, in presence of elements having the same variable values (for all the variables), in both the corresponding entries of the Hasse matrix (*s*-*t* and *t*-*s*), a value equal to 1 is stored.

It is interesting to observe that the Hasse matrix contains a holistic view of all the ordering relationships among the *n* elements belonging to the set Q. In other words, the Hasse matrix can be assumed as a fingerprint of the ordering relationships among the *n* elements.

In order to add more information to the Hasse matrix, the augmented Hasse matrix can be defined by adding to the main diagonal (zero in the original Hasse matrix) any property *P* of the elements. The property values of each set of *n* elements are scaled dividing each value by the maximum property value ($H_{ii} = P_i / P_{MAX}$).

2.2 Hasse similarity/diversity measures

Let be H^A and H^B two *n*x*n* Hasse matrices obtained by two different realizations of the variables defining *n* elements, i.e. representing two partial orderings *A* and *B*. The distance between the two partial orderings can be obtained by summing up the differences between the corresponding matrix elements. The distance between *A* and *B* can be considered as the contribution of two terms:

$$d_D(A,B) = \frac{\sum_{i=1}^{n} \left| H_{ii}^A - H_{ii}^B \right|}{n} \qquad d_H(A,B) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| H_{ij}^A - H_{ij}^B \right|}{n \cdot (n-1)/2} \tag{3}$$

where the first term d_D is the contribution to the distance due to the diagonal terms (the property values), while the second term d_H is the contribution to the distance due to the off-diagonal terms (the ranking relationships of the Hasse matrix). In both cases, the two distance terms *d* range from zero to one. This is obvious for the diagonal contribution using scaled values, but not for the off-diagonal contribution.

In case that only two variables are considered in building the Hasse matrix and that no discrepancy is observed between the ordering provided by the two variables, the corresponding Hasse matrix obtained contains only +1 and –1 values, meaning that a total ranking of the elements exists. If the Hasse matrix is obtained by using a second variable which provides an inverse ordering with respect to the first one, it will comprise only zero values, meaning that no ordering relationships exist among the elements based on these variables. Then, it is noticeable that the maximum theoretical distance between these two matrices is $n \times (n - 1)$.

From the two contributions, a weighted standardized Hasse distance (WSHD) can be defined as a trade-off between the ranking relationships and the property values. Therefore, the weighted standardized Hasse distance d_W can be defined as:

$$d_W(A,B) = (1-w) \cdot d_H(A,B) + w \cdot d_D(A,B) \qquad 0 \le d_W \le 1 \tag{4}$$

where w is a weighting term ranging between 0 and 1. Using a weight equal to zero, the distance is calculated taking into account only the ranking relationships, while a weight equal to one takes into account only the property values. In between, a weight equal to 0.5 takes equally into account both terms, resulting in a distance measure where both the ordering relationships among the elements and their property differences are equally considered.

Moreover, WSHD is a generalized Manhattan distance calculated on the corresponding pairs of elements of two Hasse matrices, thus preserving all the metric properties of the Manhattan distance.

This distance is straightforwardly interpretable as an absolute measure of distance (or as percentage $d\times100$), as an absolute measure of similarity after the transformation $s = 1 - d_W$ or as a correlation measure after the transformation:

$$r_W = (1 - d_W) \cdot 2 - 1 \qquad -1 \le r_W \le +1 \qquad (5)$$

The rank correlation r_H calculated for $w = 0$ (i.e. $d_W = d_H$) coincides with the Greiner-Kendall rank correlation index, defined as

$$\tau = \frac{4 \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^{+}}{n \cdot (n-1)} - 1 \qquad -1 \le \tau \le +1 \qquad (6)$$

where d^{+}_{ij} is defined as

$$d_{ij}^{+} = \begin{cases} 1 & \text{if } i < j \quad \text{and} \quad p_i < p_j \\ 0 & \text{otherwise} \end{cases}$$

and p are the ranks of the samples.

2.3 Hasse distance between Hasse matrices of different size

As explained above, Hasse matrices are square nxn antisymmetric matrices able to take into account the partial ordering of n elements. When two sets of different element size are considered, i.e. the two sets are constituted by n_1 and n_2 elements, respectively, with $n_1 > n_2$, two Hasse matrices H1 (n_1xn_1) and H2 (n_2xn_2) of different size have to be compared. In this case, the WSHD distance is not univocally defined and the algorithm has to be furtherly developed.

The distance between the two matrices can be calculated by overlapping n_1 - n_2 + 1 times the smaller matrix (n_2xn_2) to the bigger one (n_1xn_1, the reference matrix), starting from the top-left corner and shifting the smaller matrix diagonally until the bottom-right corner. Thus, only the elements of the bigger matrix corresponding to the size of the smaller matrix are considered in each calculation of the Hasse distance. Each distance between the pair of overlapped matrices is calculated as explained above and the smallest distance among the n_1 - n_2 + 1 distances is taken as the final distance. This procedure corresponds to search the subset of ordered elements of the bigger matrix which is more similar to the n_2 ordered elements of the smaller matrix.

2.4 Example of Hasse matrices

In order to better understand the theory presented in paragraphs 2.1 – 2.3, a simple example is given. The five sequential intensities (1, 2, ..., 5) of two samples A and B are given in Table 1.

The augmented Hasse matrices of the samples A and B, obtained by comparing the ordering and the corresponding property variables, are given in Table 2 and 3. The diagonal elements of the two Hasse matrices have been scaled with respect to the maximum property values (20 for the sample A, 18 for the sample B).

Table 1. 5-dimensional profiles of two artificial samples.

Ordering variable	*Property variable*	
ID	*sample A*	*sample B*
1	12	15
2	17	18
3	20	16
4	14	10
5	6	12

Table 2. Augmented Hasse matrix of sample A.

A	1	2	3	4	5
1	0.60	-1	-1	-1	0
2	+1	0.85	-1	0	0
3	+1	+1	1.00	0	0
4	+1	0	0	0.70	0
5	0	0	0	0	0.30

Table 3. Augmented Hasse matrix of sample B.

B	1	2	3	4	5
1	0.83	-1	-1	0	0
2	+1	1.00	0	0	0
3	+1	0	0.89	0	0
4	0	0	0	0.56	-1
5	0	0	0	+1	0.67

Table 4. Matrix of the difference between the matrices A and B.

|A-B|	1	2	3	4	5
1	0.23	0	0	1	0
2	0	0.15	1	0	0
3	0	1	0.11	0	0
4	1	0	0	0.14	1
5	0	0	0	1	0.37

The differences between the two matrices are collected in Table 4.

The sums of the diagonal and off-diagonal terms (on the half matrix) are 1.00 and 3.00, respectively, and the distances:

$$d_D = \frac{1.00}{5} = 0.20 \qquad d_H = \frac{3}{5 \cdot (5-1)/2} = 0.30$$

Let now suppose that the sample B is represented only by the first four signals. In this case, the fifth row and the fifth column of Table 3 are not present. The distance between the samples A and B is calculated as the minimum distance between the two distances from the 4x4 B Hasse matrix overlapped 1) to the first 4 rows/columns of the A matrix and 2) to the last 4 rows/columns of the A matrix. In this example, the two Hasse distances are both 4/12, i.e. $d_H = 0.34$.

3. Applications of Hasse distance to sequential data

Data including an ordering variable can be considered as sequential data. These can be characterized by an ordering variable (sequential integer numbers, variable X1) and a property variable (real numbers, variable X2).

Examples of sequential data are mass spectrometry signals, which are ordered by increasing masses, the intensity of signals being the property variable and their position in the spectrum the ordering variable; IR/UV signals, the signal intensity being the property variable and the wave length the ordering variable; 1D – NMR spectra, the signal intensity being the property variable and the chemical shifts the ordering variable. In general, all the spectra achieved along time are intrinsically ordered and can be analysed as sequential data. Analogously, data based on natural sequences can be also considered as sequential data. In effect, a sequence of integer numbers representing the positions of the elements into the sequence is the ordering variable, while any property characterizing the elements of the sequence is the property variable. A word can be thought of a sequence of characters whose position in the sentence is the ordering variable, while the position in the alphabet is the property variable. In the case of DNA sequences, which are sequences of the four nucleic acids, the molecular weight can be chosen as the property characterizing the elements of the sequence, i.e. the nucleic acids. For proteins, any physico-chemical property of the 20 aminoacids of protein sequences can be used as the property variable, while the most relevant protein abundances can be used in the case of proteomic maps.

Table 5. Examples of sequential data for applying the Hasse distance.

Sequential data	*Ordering variable*	*Property variable*
DNA sequences	1, ... , sequence length	A, C, G, T property
NMR spectra	1, ... , 1500 (from spectra resolution)	signal intensity

Mass spectra	1, ... , 250 (from spectra resolution)	signal intensity
Molecular descriptors	1, ..., sequence length	descriptor value
Proteomic maps	number of considered proteins, ..., 1	protein abundance

This kind of data can be easily characterized by Hasse matrices and their similarity/diversity assessed by the previously defined Hasse distance. In this case, the maximum information about the sequence is obtained by using only two variables, i.e. the ordering variable (X1) and the property variable (X2). In fact, in this case, the incomparabilities between two samples *s* and *t* can be due to only one condition, i.e. when the two variables X1 and X2 show an opposite rank:

$$X1(s) > X1(t) \text{ and } X2(s) < X2(t) \quad \text{or} \quad X1(s) < X1(t) \text{ and } X2(s) > X2(t)$$

For example, if three variables are taken into account, the incomparabilities between two samples can be obtained by opposite ranks of X1-X2 or X1-X3 or X2-X3, with a loss of information. In fact, in this case, the presence of zero values in the Hasse matrix cannot univocally related to a specific relationship. Examples of sequential data are collected in Table 5 and briefly discussed.

All the calculations have been performed by a MATLAB module [4] produced by the Authors.

3.1 DNA sequences

DNA sequences are sequences of four nucleic acid bases (adenine, thymine, guanine, cytosine) and can be denoted by the letters A, T, G, C, respectively. Even when sequences are not too long, the searching for their similarity/diversity is not usually easy as shown by several sequence comparisons considered in literature papers.

Comparisons among DNA sequences are performed using as the ordering variable the integer numbers correponding to the element positions in the sequence and as the second variable some properties of the nucleic bases such as the molecular weight (Table 6).

Table 6. Different representations of the DNA sequences. *MW* is the molecular weight.

Label	*ID*	*MW*	*Scaled ID*	*Scaled MW*
C	1	111.1	0.25	0.735
T	2	126.0	0.50	0.834
A	3	135.13	0.75	0.894
G	4	151.13	1.00	1.000

In order to illustrate the characteristics of the Hasse matrix and the corresponding Hasse diagram, a random 20-length sequence S1 constituted by 4 different elements has been arbitrarily defined:

ATGGTGCACCTGACTCCTGA

The two variables used for building the Hasse matrices are shown in bold characters in Table 7.

In Figure 1 the Hasse diagram of this sequence is represented. As it can be easily noted, the information contained in the diagram not only considers the absolute sequence of the elements, but also four linear extensions are highlighted, one for each different element (A, C, G, T). For example, the sequence of the element A is characterized by the path 1-8-13-20, while for the element C the path is 7-9-10-14-16-17. The links between pairs of nodes represent ordering relationships between the elements, while elements on the same horizontal level are incomparable elements (not linked among them).

Table 7. The variables selected in this work for building the Hasse matrices (columns 1 and 4 in bold characters).

ID	*Base*	*MW*	***Scaled ID***
1	A	135.13	**0.75**
2	T	126.0	**0.50**
3	G	151.13	**1.00**
4	G	151.13	**1.00**
5	T	126.0	**0.50**
....	...	....	**....**
....	...	....	**....**
18	T	111.1	**0.50**
19	G	135.13	**1.00**
20	A	151.13	**0.75**

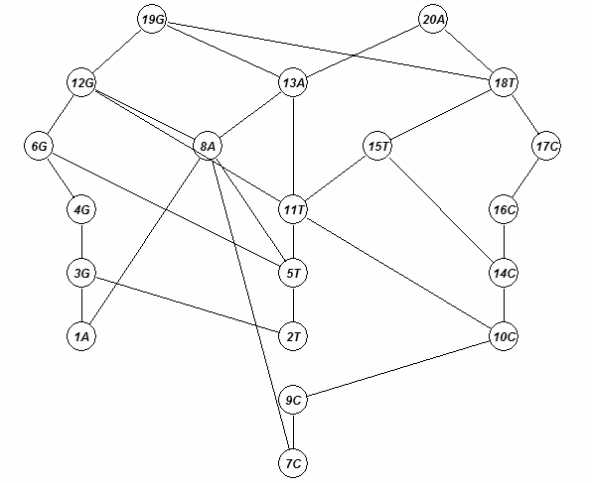

Figure 1. The Hasse diagram of the artifical sequence ATGGTGCACCTGACTCCTGA.

A simple example of WSHD calculation showing its sensitivity to small changes is discussed by substituting the fifth element T of the sequence S1 by C (S2), A (S3) and G (S4), respectively.
The calculated distances using the three weights 0, 0.5 and 1 are shown in Table 8.
As expected the distances between S1 – S2 and S1 – S3, calculated taking into account only the property values, are equal (both differences between the pair of elements T-C and T-A are 0.25); nevertheless, the corresponding distances calculated by using the ordered property are different: S1 – S2 is equal to 3.684 and S1 – S3 is equal to 2.105. For all the weigths, the most dissimilar is S1 - S4, being the substitution more influent in the global ordering of the sequence. The presence of four A and six C characters in the original sequence S1 justifies the greater difference when T is substituted by C.

Table 8. The distances (as percentages) of the three modified sequences with respect to the original sequence S1, calculated by using three different weights.

w	*S2 - C*	*S3 - A*	*S4 - G*
0	3.684	2.105	4.737
0.5	2.467	1.678	3.618
1	1.250	1.250	2.500

3.2 NMR and mass spectra

In general, experimental spectra constitute a tipical case where an ordering variable (time, masses, chemical shifts, vawe lengths, etc.) and a property variable (signal intensities) can be naturally associated.
In the case of mass and NMR spectra, the ordering variable is a sequence of integer numbers from one to the maximum number of numerically different signals (given by the spectra resolution). For example, in NMR spectroscopy, assuming the chemical shifts take values from 0.01 to 15.00, with a resolution of 0.01, the total number of resolved signals is 1500. The signals with intesities greater than zero are registered and embedded into the 1500 signals; all the other signals have intensities equal to zero.
The two variables used for building the Hasse matrices are shown in bold characters in Table 9, for 24 NMR signals whose intensities are greater than zero. These signals are successively embedded into an 1500-array. Then the distances are calculated from pairs of Hasse matrices of size 1500 x 1500.
In an analogous way, the ordering variable for mass spectra is constituted by integer numbers ranging from 1 to the spectral resolution (assuming a mass resolution of 0.1 in the range 0 – 25, 250 signals are obtained). The sensitivity of WSHD is here evaluated on a real mass spectrum of pentobarbital (SP1).
The original mass spectra (SP1) is shown in Figure 2, where the differences of SP2, SP3 and SP4 are also highlighted.
The data are collected in Table 10, where only the signals different from zero are reported.

Table 9. Example of 24 NMR signals with intensities greater than zero. In bold characters the ordering variable (*ID*) and the property variable (*Height*) selected for building the Hasse matrices are shown.

Signal	*ppm*	***ID***	***Height***
1	0.01	**1**	**0.1159**
2	1.17	**117**	**0.1278**
3	1.18	**118**	**0.1247**
4	2.17	**217**	**1**
5	2.42	**242**	**0.0314**
6	2.43	**243**	**0.0255**
....	....	....	....
....	....	....	....
....	....	....	....
21	11.74	**1174**	**0.0709**
22	11.88	**1188**	**0.0703**
23	13.79	**1379**	**0.0629**
24	13.97	**1397**	**0.0668**

Table 10. Signal intensities for the pentobarbital (SP1) and the three modified simulated spectra (SP2 – SP4). In bold characaters the modified intensities.

Mass	*SP1*	*SP2*	*SP3*	*SP4*	*Mass*	*SP1*	*SP2*	*SP3*	*SP4*	*Mass*	*SP1*	*SP2*	*SP3*	*SP4*
36	3	3	3	3	75	3	3	3	3	139	6	6	6	6
37	3	3	3	3	77	10	10	10	10	140	20	20	20	20
38	12	12	12	12	78	5	5	5	5	141	826	826	826	**1000**
39	139	139	139	139	79	12	12	12	12	142	71	71	71	71
40	38	38	38	38	80	22	22	22	22	143	11	11	11	11
41	364	364	364	364	81	18	18	18	18	144	1	1	1	1
42	83	83	83	83	82	13	13	13	13	151	1	1	1	1
43	378	378	378	378	83	30	30	**0**	30	152	1	1	1	1
44	84	84	84	84	84	13	13	13	13	153	6	6	6	6
45	6	6	6	6	85	38	38	38	38	154	0	0	0	0
50	5	5	5	5	86	8	8	8	8	155	133	133	133	133
51	11	11	11	11	87	9	9	9	9	156	1000	1000	1000	**826**
52	14	14	14	14	88	1	1	1	1	157	253	253	253	253
53	64	64	64	64	91	5	5	5	5	158	29	29	29	29
54	31	31	31	31	92	3	3	3	3	159	3	3	3	3
55	162	162	162	162	93	4	4	4	4	165	1	1	1	1
56	30	30	30	30	94	21	21	21	21	166	1	1	1	1
57	17	17	17	17	95	17	17	17	17	167	2	2	2	2
58	5	5	5	5	96	20	20	20	20	168	3	3	3	3
59	1	1	1	1	97	55	55	55	55	169	5	5	5	5
60	5	5	5	5	98	127	127	127	127	179	1	1	1	1
61	2	2	2	2	99	9	9	9	9	181	4	4	4	4
62	1	1	1	1	100	3	3	3	3	183	7	7	7	7
63	3	3	3	3	101	3	3	3	3	185	2	2	2	2

64	1	1	1	1	102	0	0	0	0	191	1	1	1	1
65	13	13	13	13	103	1	1	1	1	193	1	1	1	1
66	13	13	13	13	105	2	2	2	2	195	2	2	2	2
67	58	58	58	58	106	3	3	3	3	197	65	65	65	65
68	33	33	33	33	107	2	2	2	2	198	10	10	10	10
69	120	120	120	120	133	3	**2**	3	3	199	2	2	2	2
70	67	67	67	67	134	2	**3**	2	2	204	1	1	1	1
71	89	89	89	89	135	2	**3**	2	2	207	6	6	6	6
72	5	5	5	5	136	2	**1**	2	2	208	2	2	2	2
73	9	9	9	9	137	6	6	6	6	209	1	1	1	1
74	8	8	8	8	138	7	7	7	7	227	6	6	6	6
										228	1	1	1	1

A total of 250 signals are considered and three spectra are arbitrarily obtained by performing small modifications of the signal intensities of the first spectrum. In particular, for SP2 only some small signals have been modified (signals 133 – 135), for SP3 one signal has been modified from 30 to 0 (signal 83), for SP4 the two greatest signals have been inverted (signals 141 and 156).

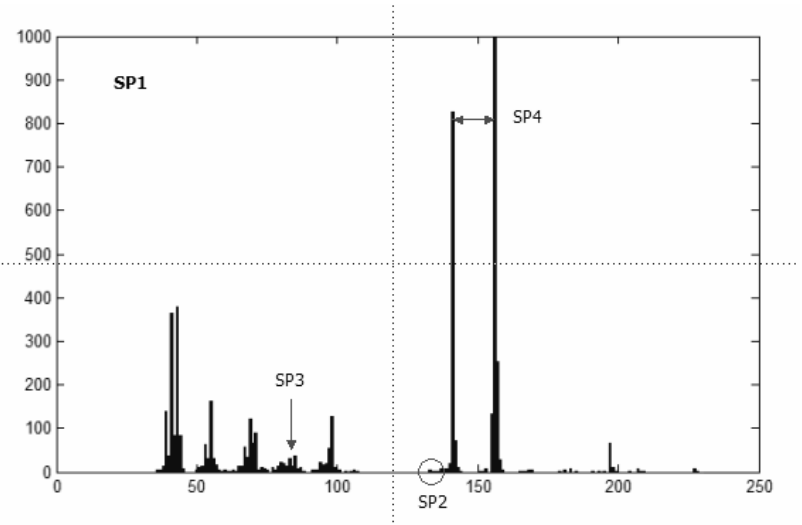

Figure 2. Mass spectrum of SP1, together with the modifications performed for spectra SP2 – SP4 (see also Table 10).

The distances calculated by using the weights 0, 0.25, 0.50, 0.75 and 1 are collected in Table 11.

Table 11. The distances (as percentages) between SP1-SP4 spectra calculated with different weights.

w	d(1-2)	d(1-3)	d(1-4)	d(2-3)	d(2-4)	d(3-4)
0.00	0.216	0.630	0.004	0.847	0.220	0.634
0.25	0.163	0.476	0.041	0.639	0.203	0.517
0.50	0.109	0.322	0.078	0.431	0.187	0.399
0.75	0.055	0.167	0.114	0.223	0.170	0.282
1.00	0.002	0.013	0.151	0.015	0.153	0.164

As can be observed, when only the signal intensities are considered ($w = 1$), the two most similar spectra are SP1 and SP2, while the most dissimilar are SP3 - SP4 (0.164), SP2 - SP4 (0.153) and SP1 - SP4 (0.151). However, when only the signal ranking is considered ($w = 0$), the two most similar spectra are SP1 - SP4 (0.004), while the most dissimilar are SP2 - SP3 (0.847), SP3 - SP4 (0.634) and SP1 - SP3 (0.630). It is interesting to observe the opposite behaviour of the distances between SP1 and SP4, where the intensities of two highest signals have been exchanged. In this case, the contribution of the intensity differences is maximal (case $w = 1$), while in the Hasse diagram only the two cells corresponding to the two highest signals take opposite values, being the ranking of all the other signals not influenced (they remain smaller with respect to them as in the case SP1). The strong sensitivity of the Hasse distance ($w = 0$) in changes of small/medium signals is highlighted by the distances between SP1 - SP2 (0.216) and SP1 - SP3 (0.630).

3.3 Molecular descriptors

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, health research and quality control, being obtained when molecules, thought as real objects, are transformed into a molecular representation allowing some mathematical treatment. Many molecular descriptors have been proposed until now derived from different theories and approaches [5-6]. The information content of a molecular descriptor depends on the kind of molecular representation that is used and on the defined algorithm for its calculation. There are simple molecular descriptors derived by counting some atom-types or structural fragments in the molecule, other derived from algorithms applied to a topological representation (molecular graph) and usually called topological or 2D-descriptors, and there are molecular descriptors derived from a geometrical representation that are called geometrical or 3D-descriptors.

In chemistry, molecular descriptors are the basic elements used by all the methods for assessing molecular similarities.

In order to apply the proposed approach to the similarity/diversity in QSAR/QSPR problems, a set of convenient molecular descriptors has to be found. Several ordered descriptors are defined in literature, such as autocorrelation descriptors of different lags, connectivity indices of different orders, radial

distribution function (RDF) descriptors, etc. However, not all the ordered descriptors can be properly used in this approach. In fact, when the ranking of the descriptors values largely depends from the descriptor definition and less from the molecular structure, the descriptors cannot be used being equal all the Hasse matrices, i.e. the similarity/diversity measure does not depend from the descriptor ranking.

For example, the values of connectivity indices χ_0, χ_1, ..., χ_5 calculated for different molecules largely differ among them, but their ranking is the same in almost all the cases, i.e. they decrease from χ_0 to χ_5. Then, the information related to the ranking is lost and the differences in similarity/diversity arise only from the differences in descriptor values. A set of ordered descriptors showing a ranking independence is the set of RDF descriptors [7]. They are defined as:

$$RDF_{R,w} = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i \cdot w_j \cdot e^{-\beta\left(R-r_{ij}\right)^2}$$

where A is the number of atoms, w the atomic property, and r_{ij} the geometric distance between the i-th and j-th atoms. The parameter f is a scaling factor (equal to 1), β a smoothing parameter (assumed equal to 100) and R represent the radius related to the spherical volume (range of 1 – 15), with a step assumed equal to 0.5 Å. Five different properties have been used (Table 12) and thirty RDF descriptors for each property have been calculated using DRAGON software [8], giving a total of 150 descriptors for molecule. An example of RDF spectrum is shown in Figure 3 for cyclohexane.

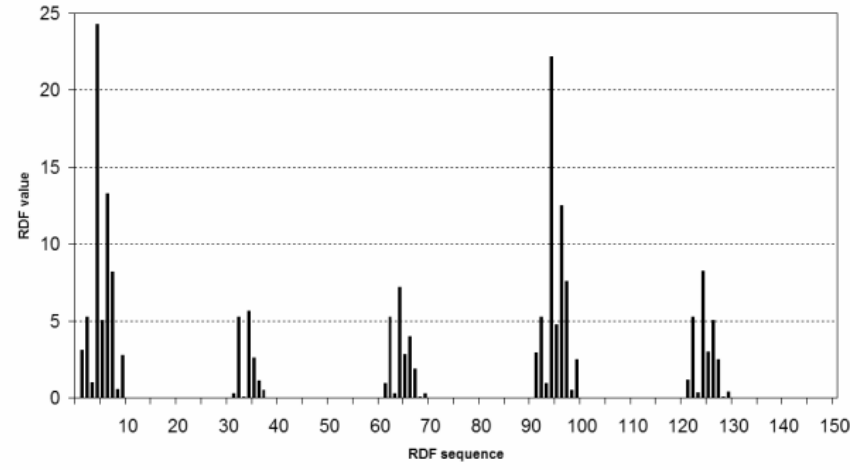

Figure 3. RDF spectrum of cyclohexane. The five blocks of signals correspond to the five different properties (Table 12).

Table 12. Atomic properties (weights) used for the calculation of the RDF descriptors.

Weight	*Description*
u	no weight (all weights equal to 1)
m	atomic mass
v	van der Waals volume
e	Sanderson electronegativity
p	atomic polarizability

In order to check the similarity/diversity measures based on the RDF descriptors, a small set of 10 molecules has been used and the corresponding distances calculated with $w = 0$ (upper matrix) and $w = 0.5$ (lower matrix) are shown in Table 13. Being the set of 150 RDF descriptors built considering all the weights of Table 12, the molecule representation contains several different chemical information (geometric, mass-related, electronic, etc.). The use of selected subsets can obviously highlight different levels of chemical similarity.

As it can be noted, only considering the distances calculated by the Hasse matrices ($w = 0$), the most similar pairs of molecules are cyclohexane – benzene, benzene – toluene, Br-benzene – I-benzene, F-benzene – Cl-benzene; moreover, the molecules less dissimilar from anthracene are n-hexane and naphtalene. By considering the weighted Hasse distances ($w = 0.5$), similar cosiderations can be done. The most dissimilar pair is I-benzene – anthracene.

Table 13. WSHD distances (as percentages) for $w = 0$ (upper matrix) and $w = 0.5$ (lower matrix). In bold characters are highlighted the most similar molecule pairs.

ID	*molecule*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	n-hexane	0	18.398	18.685	14.058	17.575	16.206	16.617	17.888	11.579	14.130
2	cyclohexane	*11.791*	0	**1.718**	8.510	**3.436**	5.486	4.412	5.754	11.579	27.302
3	benzene	*12.077*	***2.180***	0	7.204	**2.291**	4.788	4.823	6.309	10.416	26.067
4	toluene	*9.685*	*6.280*	*4.512*	0	5.736	5.047	6.497	7.427	4.358	21.280
5	F-benzene	*11.533*	***3.270***	***1.687***	***3.666***	0	**3.302**	6.738	8.260	9.092	25.011
6	Cl-benzene	*11.659*	*5.540*	*4.429*	*4.478*	***3.571***	0	8.591	9.951	8.421	24.931
7	Br-benzene	*11.818*	*4.205*	*4.345*	*5.599*	*5.514*	*7.293*	0	**2.720**	9.065	25.074
8	I-benzene	*13.256*	*6.034*	*6.152*	*6.758*	*7.266*	*8.789*	*3.803*	0	10.353	26.488
9	naphtalene	*8.215*	*7.833*	*6.516*	***3.168***	*5.752*	*6.572*	*7.086*	*8.428*	0	17.888
10	anthracene	*10.358*	*17.619*	*16.278*	*13.383*	*15.643*	*16.581*	*16.507*	*18.245*	*11.003*	0

3.4 Proteomic maps

The evaluation of complex therapeutic and toxic behaviour of chemicals from their effects on simpler biological systems such as cells is among the most interesting trends in drug discovery, environmental

safety studies, molecular pharmacology and hazard assessment. This scientific cross-breeding is going under the name of "toxicogenomics".

In these last years, special attention has been payed to the cellular proteome which characterizes the different abundance of thousands of proteins belonging to the same cell.

A typical proteomic map is a planar map constituted by two axes representing charge (x-axis) and mass (y-axis) where even 2000 cell's proteins may appear as separated spots accordingly to their charge vs mass values; the spot diameters are related to the abundance of the proteins.

Toxicological studies on proteomic maps consist in perturbing the control cell with a chemical and evaluate the resulting differences in the abundance of protein expressions with respect to the control cell. A very interesting mathematical challenge is to understand the complexity of cellular events, and then describe and characterize changes in proteomic maps.

The application of the proposed approach to the proteomic maps requires the selection of a control map. The protein abundances of the control map are ordered from the most to the less abundant, among the considered expressed proteins. The ordering variable is the set of integers from n (the number of considered proteins) to 1: then, the Hasse matrix of the control map represents a total order. The Hasse matrices of the other proteomic maps are build using the ordering variable defined for the control map and the corresponding abundances.

Table 14. The coordinates and the abundances of 20 proteins of the proteomic maps. The third entry (12,10 – spot 3) is taken as reference for artificial modifications, subtracting 7 in each step.

n	*Rank ID*	*x*	*y*	M_0	....	M_4
1	20	21	23	144.4	....	144.4
2	19	28	9	143.6		143.6
3	***18***	***12***	***10***	***136.7***		***108.7***
4	17	22	9	125.3		125.3
5	16	27	12	118.6		118.6
6	15	15	8	114.9		114.9
7	14	13	14	112.3		112.3
8	13	29	8	108.9		108.9
9	12	14	11	98.2		98.2
10	11	26	13	94.1		94.1
11	10	25	4	93.6		93.6
12	9	16	6	90.0		90.0
13	8	30	8	86.7		86.7
14	7	21	8	84.8		84.8
15	6	6	7	82.5		82.5
16	5	29	19	82.0		82.0
17	4	20	9	80.0		80.0
18	3	28	8	79.8		79.8
19	2	23	10	72.8		72.8
20	1	11	9	72.2	....	72.2

In order to exemplify the proposed approach, a calculation has been performed producing artificially some proteomic maps which differ systematically from the control (Table 14).

Following the strategy proposed by Randic et al. [9], the abundance of the third spot (map M_0) corresponding to the coordinates (12, 10) is modified iteratively subtracting 7 to the initial abundance (136.7), obtaining other 4 different proteomic maps (M_1 – M_4). The four abundance values of these modified proteomic maps are 129.7, 122.7, 115.7, and 108.7.

The Hasse distances between the control map and the four modified maps are reported in Table 15, for the weights 0, 0.5, 1. All the distances are presented as percentages, i.e. multiplied by 100.

When only the structure of the Hasse matrix is considered ($w = 0$), the distance between the control and the first modified map is equal to zero, because the new value of the spot 129.7 does not modify the ranking of the abundances; in the three other cases, the ranking is modified in increasing manner (1, 2 and 5 positions, respectively) and the distances reflect these modified rankings.

Table 15. The distance (as percentages) of the four artificially modified proteomic maps from the reference one (spot 3), for three different weights.

w	1	2	3	4
0	0.000	0.526	1.053	2.632
0.5	0.121	0.506	0.890	1.801
1	0.242	0.485	0.727	0.970

In the case of $w = 0.5$, the distances are calculated taking into account both the Hasse matrix off-diagonal terms and the diagonal terms. In this case, also the first modified map shows a distance greater than zero from the control. In the last case ($w = 1$), only the diagonal contributions of the Hasse matrix are considered, and the distances from the control differ in an uniform way (0.243).

Thus, it can be observed that when the off-diagonal terms are taken into account ($w = 0$ and $w = 0.5$), the distances from the control increase not-linearly, due to the relevant role of the global ordering relationships of the abundances. Moreover, the Hasse distances appear sensitive to small changes in abundances, but they also show a robustness with respect to changes which does not change the ranking of the abundances.

4. Conclusions

The proposed similarity/diversity measure appears as a new approach to sequential data, where useful information can be also obtained by the ordering relationships between the sequence elements. In particular, the weighted Hasse distance shows some advantages: a) the Hasse matrices and the corresponding distances are calculated by a straightforward algorithm; b) the distance is naturally

standardized, allowing a natural interpretation of the obtained values; c) the distances are able to take into account the whole structure of the ranking relationships of the sequences; d) the distances can be obtained by a flexible strategy (the weights) depending on the specific similarity/diversity study; e) a simple rank correlation measure is derived, also taking into account incomparabilities among sequence elements.

Specific studies are in progress about the characterization of DNA sequences, proteomic maps and molecular similarity.

References

[1] Brüggemann, R. and Bartel, H.-G. (1999). A theoretical Concept to Rank Environmentally Significant Chemicals. *J. Chem. Inf. Comput. Sci*., **39**, 211-217.

[2] Pavan, M. and Todeschini, R. (2004). New indices for analysing partial ranking diagrams. *Anal. Chim. Acta*., **515**, 167-181.

[3] Brüggemann, R, Franck, H, Kerber, A. (2004). Proceedings of the Conference "Partial Orders in Environmental Sciences and Chemistry". *MATCH Commun. Math. Comput. Chem*., **54**, 485-689.

[4] MATLAB (ver. 6.5); The MathWorks Inc., Natick (MA), USA.

[5] Todeschini, R. and Consonni, V. (2000). *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, Germany, 668 pp.

[6] Devillers, J. and Balaban, A. (2000). *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon & Breach, Amsterdam, The Netherlands, 812 pp.

[7] Hemmer, M.C., Steinhauer, V. and Gasteiger, J. (1999). Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra. *Vibrat. Spect*., **19**, 151-164.

[8] Talete srl, DRAGON for Windows (Software for Molecular Descriptor Calculations) - Ver. 5.4 – 2006 – http://www.talete.mi.it

[9] M. Randić, J. Zupan and M. Nović (2001). On 3-D Graphical Representation of Proteomics Maps and Their Numerical Characterization. *J. Chem. Inf. Comput. Sci*., **41**, 1339-1344.