SIMILARITY MEASURES FOR SETS OF STRINGS AND APPLICATION IN CHEMICAL CLASSIFICATION

Borka JERMAN-BLAŽIČ
Institut Jožef Stefan, E. Kardelj University, Ljubljana, Yugoslavia

and

Milan RANDIĆ
Department of Computer Science, Drake University, Des Moines, Iowa 50311, USA

Abstract

This paper introduces new ideas for quantification of the similarity between chemical compounds. The method adopted makes use of similarity measures derived through comparison of two strings. The derived data on the similarity are then analyzed and applied in the identification of clusters in which the entities are more homogeneous and similar than those outside a cluster.

1. Introduction

The interest in "classification" is present in many scientific studies and arises in the context of many applications. In chemistry, classification is applied in a number of studies with the assumption that structurally similar compounds possess similar properties. Classification is an inherently multivariate problem. The high-dimensional nature of the classification provides an opportunity, but also involves difficulties, in the choice of the appropriate methodology. There are two broad categories of classification problems. In the first, the data of the entities and the group membership are known, while the unknown membership of other entities has to be determined through the analysis of the data. In the pattern recognition literature, this type of classification problem is referred to as "learning with a teacher". In statistical terminology, it falls under the heading of "discriminant analysis". In a second category of classification problems, the groups are themselves unknown and the primary purpose of the data analysis is to determine the groupings from the data so that entities within the same group are in some sense more similar or homogeneous than those which belong to different groups. This type of classification problem is referred to as "learning without a teacher". In statistical terminology, this falls under the heading of "cluster analysis".

Although discriminant analysis and cluster analysis are viewed as a dichotomy of the classification problems, in practice they are frequently combined. Classification problems are treated in a three-stage procedure: input, algorithm and output. All stages interact with each other, and the applied methods in any one stage play roles in the other two.

During the last two decades, stimulated by the easy access to numerical and graphic computing facilities, a number of new approaches and algorithms for discriminant analysis and cluster analysis were developed. As a rule, focus has been on the development of new algorithms.

The approach presented in this report introduces a quantification of the similarity of chemical compounds derived from comparison of two strings. The derived data about the similarity are then analyzed and applied in identification of clusters in which the present entities are more homogeneous and similar than those outside the cluster. The method is illustrated on a set of chemical compounds exhibiting biological activity.

2. Similarity measures derived from string comparison

Consider a pair of strings X and Y, made up of symbols from a finite alphabet A. Different quantifications [2] of the similarity and dissimilarity between them fall into two major groups: numerical and non-numerical. The *generalized Levenshtein distance* (GLD), the length of their *longest common subsequence*, and the length of their *shortest common supersequence* are numerical. The set of their common subsequences and the set of their common supersequences are non-numerical quantities. Both groups of measures make use of the elementary abstract measure of comparison, $D(X, Y)$ between X and Y, which is expressed in terms of a set of elementary measures $d(*.*)$. The abstract measure $D(X, Y)$ between $X, Y \in A^*$ is a map whose domain is $A^* \times A^*$. Formally:

$$D(X, Y) = I, \quad \text{the identity, if } X = Y.$$

$D(X, Y)$ is symmetric if and only if the function $d(*.*)$ inducing it obeys:

$$d(a, b) = d(b, a).$$

Here, we are interested in quantification of similarity and its application in chemical classification. Therefore, the numerical measure known under the name "generalized Levenshtein distance", which is a special of the abstract measure of comparison $D(X, Y)$, will be considered and applied.

2.1. THE SET OF ELEMENTARY MEASURES

Let A be a finite alphabet and A^* be the set of strings over A. A string $X \in A^*$ of the form $X = x_1, \ldots, x_N$, where each $x_i \in A$, is said to be of length $X = N$. $d(*.*)$ is a function whose arguments are a pair of symbols belonging to A^*.

The elementary measure $d(a, b)$ can be interpreted as the measure associated with transforming "b" to "a", for $a, b \in A^*$ or, more explicitly, as a set of edit operations, i.e. $G_{X,Y}$. $G_{X,Y}$ is an exhaustive enumeration of the set of all the ways by which Y can

be edited to X using the edit operation of substitution, insertion and deletion without destroying the order of the occurrence of the symbols in X and Y.

The generalized Levenshtein distance (GLD) [3,4] between two strings X and Y is defined as the minimum of the sum of the elementary edit distances associated with edit operations required to transform Y to X. The elementary edit distances themselves are specified in terms of a map $d_1(*.*)$ from $A^* \times A$ to R, the set consisting of nonnegative real numbers and ∞.

The elementary edit distances obey the following conditions:

$$d(a, b) > 0, \qquad \text{for all} \quad a \neq b, \; a, b \in A;$$

$$d(a, b) = 0, \qquad \text{if} \quad a = b, \; a, b \in A;$$

$$d(a, b) \leq d(a, c) + d(c, b), \qquad \text{for all} \quad a, b, c \in A.$$

Subject to the above constraints, the GLD obeys the following for all $X, Y, Z \in A$:

$$\mathrm{GLD}(X, Y) > 0, \qquad \text{if} \quad X \neq Y;$$

$$\mathrm{GLD}(X, Y) = 0, \qquad \text{only if} \quad X = Y;$$

$$\mathrm{GLD}(X, Y) \leq \mathrm{GLD}(X, Z) + \mathrm{GLD}(Z, Y).$$

In general, a greater value of the GLD between two strings indicates a greater dissimilarity between them.

The edit operations are explained by the following example. Let us compare the strings X and Y:

string X $\quad a\ d\ b\ f\ c\ e$;
string Y $\quad b\ c\ f\ a\ e\ d$.

The operations required to transform X to Y are:

string X	$a\ d\ b\ f\ c\ e$
	delete a
$d\ b\ f\ c\ e$	
	delete d
$b\ f\ c\ e$	
	substitute a for c
$b\ f\ a\ e$	
	insert c
$b\ c\ f\ a\ e$	
	insert d
string Y	$b\ c\ f\ a\ e\ d$

After five edit operations, the string Y is obtained.

2.2 COMPUTING GLD

The computation of GLD in the case where the two strings have the same number of elements and the present elements are the same but differently ordered may be done by an algorithm which computes the necessary edit operations by calculation of the identities (matches) found in the traces of the strings X and Y. A trace between strings Y and X as illustrated in fig. 1 consists of the source string X above and the target sequence Y below, usually with lines from some elements in the source to some

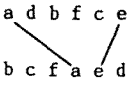

Fig. 1. A trace between two strings.

elements of the target. An element can have no more than one line, the lines must not cross each other, and the source elements with lines must have an ordered correspondence with the target elements. The lines provide a correspondence between the source and the target. The pair of elements connected by a line is an identity only when the elements are the same.

The number of identities in all possible traces derived for two strings is given by

$$E = \sum_{i=1}^{N} (c_i - 1),$$

where N is the number of string elements, and c_i is the number of identities in trace i on the right-hand side of element i. The method for counting identities in traces is explained elsewhere [8].

3. Classification of chemical compounds

Consider a set of N compounds which exhibit some similar physicochemical or other property as elements of the alphabet A. Let us order the elements of A in a string. The string is produced by comparing one element of A (for example "i") with the other elements present in A. The comparison may be done by any of the well-known methods for comparing two objects [6,7], i.e. Euclidian distance, Jaccard coefficient, etc. The ordering may be done according to decreasing value of the calculated distances (i.e. Euclidian or any other) of the elements of A to the element "i". If the procedure is repeated with the other elements, N strings will be obtained in which the elements of A are ordered according to one particular element belonging to A. As a result of the procedure, and $N \times N$ matrix is obtained in which the rows are strings of ordered elements with equal length and same content. If two such strings are compared, then they will differ at least in the placement of two elements, i.e. the first elements of the string according to which the other elements are ordered. If, for example, these two

elements are the most potent structures, exhibiting highest biological activity, then the two ranks will give a partial order of potency for the rest of the structures. Partial orders may be used for identification of a common strategic fragment responsible for the activity [9].

The relationship between the two elements according to which the others are ordered in the two strings can be expressed by a correlation index derived from the value E:

$$r_{i,j} = \frac{2E}{N(N-1)} - 1.$$

If r is near 1, the ordering of the chemical objects in string X is similar to the ordering of the chemical objects in string Y, whereas if r is near 0, the ordering is completely different.

The chemical compounds i and j are similar if they generate a similar ordering of the other elements of the set represented by the alphabet A. If both compounds generate similar rankings of the others and if both of them are close to each other in the other ranks, then they are similar. Similarity derived in this way (by considering the similarity to all compounds in the data set) ensures that many different structural features in the elements of the set of compounds are considered.

Classification of the compounds is done according to the values of $r_{i,j}$ obtained for each possible pair in the data set. The purpose of the classification is to identify groups of compounds with close values of r. The identification is done with a recursive procedure. The procedure is as follows:

In the first step, the data are presented in the form of an $N \times N$ matrix, i.e. the correlation matrix R is searched for the highest value of r. The compounds with the highest value of r are the kernels of the future clusters. A pair of compounds i, j which satisfies the prescribed value for r (for example, r greater or equal to 0.95) is chosen from the data. They are the first elements of the future cluster. If there are more elements which satisfy the same condition for r as the pair already chosen, then this cluster will have more than two kernel elements. The elements for the next kernel are generated in the same way. The process of kernel generation continues until no compounds with the prescribed value of r are left.

The second step completes the clustering. An element k is added to a particular kernel if the value of r for this element, i.e. $r_{k,n}$, is greater than or equal to a prescribed value ($r_{k,n}$ designates the correlation coefficient between this element and the elements present in the cluster). If two or more kernels satisfy this condition, then the element is added to the kernel with the higher value. This value is different from the threshold value for the kernel generation in the first step. The prescribed values of r in both steps may be changed during the clustering procedure according to the nature of the data classified. Sometimes it happens that the first prescribed threshold value of r is too high. This results in a small number of clusters. In this case the threshold value should be decreased. On the other hand, the kernel threshold value should not be too low because

in this case the similarity criterion is lost. The best and most valuable results are obtained when the threshold value varies between 0.8 and 0.95. The classification algorithm is presented in the appendix.

4. Classification of fluorophosphonic acid derivatives

As an example, we have selected a set of twenty-seven fluorophosphonic acid derivatives exhibiting inhibitor activity on the membrane enzyme of animal acetylcholinesterase [10,11]. The inhibitor activity of the compounds was measured as the bimolecular rate constant (of the purified acetylcholinesterase from the electrophorus electricus organ) and calculated according to Aldridge and Reiner [12]. The compounds, together with their bimolecular constants, abbreviated as in ref. [13], where interested readers can find their full chemical names and other details, are shown in table 1. All

Table 1

The studied set of compounds

No.	Label	R_1	R_2	Bimolecular constant $[mol^{-1}min^{-1}]$
1	MeSL	$-CH_3$	$-O-$	1.7×10^6
2	EtSL	$-C_2H_5$	$-O-$	1.6×10^3
3	PrSL	$-C_2H_7$	$-O-$	1.3×10^3
4	BuSl	$-C_4H_9$	$-O-$	6.0×10^4
5	HxSL	$-C_6H_{23}$	$-O-$	6.2×10^4
6	HpSL	$-C_7H_{15}$	$-O-$	1.3×10^5
7	OcSL	$-C_8H_{17}$	$-O-$	8.5×10^4
8	DoSL	$-C_{12}H_{25}$	$-O-$	8.7×10^4
9	MeSL3	$-CH_3-$	$-O-CH_2-CH_2-$	6.1×10^5
10	MeSL4	$-CH_3-$	$-O-CH_2-CH_2-O-$	2.1×10^6
11	MeSL5	$-CH_3-$	$-NH-CH_2-CH_2CH_2-O-$	3.6×10^6
12	MeSL6	$-CH_3-$	$=CH-CO-NH-CH_2-CH_2-O-$	2.0×10^2
13	MeSL7	$-CH_3-$	$-NH-CO-CH_2-S-CH_2-CH_2-O-$	1.9×10^7
14	MeSL8	$-CH_3-$	$-O-CH_2-O-NH-(CH_2)_3-O-$	9.0×10^5
15	MeSL10	$-CH_3-$	$-NH-CO-CH_2-O-(CH_2)_5-O-$	2.6×10^5
16	MeSL11	$-CH_3-$	$-O-(CH_2)_3-NH-CO-CH_2-S-(CH_2)_2-O-$	$10^3 - 10^4$
17	MeSL21	$-CH_3-$	$-CH_2-O-$	7.4×10^5
18	MeSL32	$-CH_3-$	$=CH-CH_2-O-$	6.6×10^5
19	MeSL53	$-CH_3-$	$-O-(CH_2)_3-O-$	3.0×10^6
20	MeSL64	$-CH_3-$	$-CH_2-O-(CH_2)_3-O-$	6.8×10^7
21	MeSL6A	$-CH_3-$	$=CH-CH_2-O-CH_2-CH_2-O-$	4.8×10^7
22	MeSL74	$-CH_3-$	$-O-(CH_2)_2-O-(CH_2)_2-O-$	3.6×10^7
23	MeSL85	$-CH_3-$	$-O-(CH_2)_3-O-(CH_2)_2-O-$	2.2×10^7
24	EtSL7a	$-C_2H_5-$	$-NH-CO-CH_2S-CH_2-CH_2-O-$	3.0×10^6
25	EtSL7E	$-C_2H_5-$	$-O-CO-CH_2-S-CH_2-O-$	2.5×10^7
26	EtSL10	$-C_2H_5-$	$-NH-CO-CH_2-O-(CH_2)_5-O-$	7.1×10^5
27	MeSL7E	$-CH_3-$	$-O-CO-CH_2-S-(CH_2)_2-O-$	2.9×10^7

compounds from table 1 display some inhibitory activity, which varies from molecule to molecule by several orders of magnitude.

As molecular descriptors, path counts with different lengths in the molecular graph were applied and the "gain of information" [8] was used as a measure for ordering the compounds in strings. The classification resulted in five clusters, depicted in fig. 2.

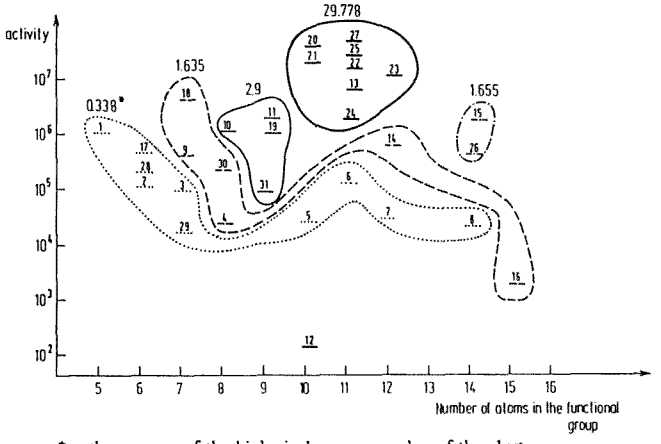

Fig. 2. The obtained clustering of the fluorophosphonic acid derivatives.

The space allocation of the points representing a particular compound in fig. 1 is a function of the activity, i.e. the bimolecular constant and the number of atoms in the attached functional group to the piperidine ring (this parameter being chosen because an assumption based on experimental findings exists about the correlation between the activity and the geometrical arrangement of the functional group). The measured biochemical properties, i.e. the bimolecular constants of the compounds in the identified cluster, were quite homogeneous, therefore an average value of the activity was calculated. It may be noticed that the values of biological activity vary from cluster to cluster by an order of magnitude of two. Careful observation of the represented clusters confirms the previous assumption that the geometry and the spatial arrangement of the attached chains in the functional group largely determine the biochemical properties of the molecules. The structural properties of the compounds and the way in which they were classified suggest an explanation of their biochemical behaviour. The highest activity was observed within a cluster containing R_2 with 6 to 8 atoms and short R_1. The other compounds with 11, 12 or more atoms as well as with 4 to 5 atoms in R_2 and short R_1 displayed weak inhibitory activity. These facts may be explained with the optimal geometry of the extended conformation of an R_2 chain composed of 6–8 atoms and its best fit to the active site pocket.

The distance between the active serine hydroxyl of m-AChE and the anionic site on the pocket wall is approximately the same as the length of an extended R_2 chain with the optimal number of atoms. The small differences in activity between the compounds of this group are due to the differences in the chemical content of the chain, i.e. some of the peptide bonds are present and some of them contain carbonyl bonds. The intensive drop in activity observed within the compounds with larger R_2 chains may be explained in terms of the increasing volume of the molecules and the larger mechanical strain on the active center walls, which force the molecules to unfavourable conformations.

5. Conclusions

The method presented makes use of the similarity postulate and introduces a new method for the classification of chemical objects. The approach involves ordering of the strucures and classification of the structures by GLD. The results imply a decision of which two among the studied chemical objects, i.e. structures, are the most similar. This information is used as a basis for identification and generation of clusters in which all elements show considerable similarity. The method is suitable for QSAR studies and property prediction.

Appendix

The algorithm:

```
generation of kernels
  m := number of objects in the data set;
  n := 0;
  repeat
    choose a pair of compounds [i, j]: r(i, j) ≥ kmax;
    n := n + 1;
    cluster n := [i, j];
      repeat
        for k := 1 to m do
        if r(k, c) ≥ kmax for an element c being in cluster n then
          cluster n := cluster [n + k];
      until (no such k)
  until (no such i, j);
joining other elements to the obtained kernel
  for i := 1 to m do
    choose all clusters that r(i, c) = cmax for all
    elements c in the cluster and then join the element i
    to the cluster having the greatest average correlation.
```

References

[1] B. Jerman-Blažič, I. Fabič-Petrač and M. Randić, Chem. Int. Lab. Syst. 6(1989)49.
[2] R.L. Kashyap and B.J. Ommen, Int. J. Comp. Math. 13(1983)95.
[3] A. Levenshtein, Sov. Phys. Dokl. 10(1966)707.
[4] P.H. Seller, J. Comb. Theory 16(1974)253.
[5] J.D. Morrisett, C.A. Broomfield and B.E. Hackley, J. Biol. Chem. 224(1969)5758.
[6] V. Batagelj, Preprint Series Dept. Math., University E. Kardelj 26(1988)252.
[7] M. Johnson, J. Math. Chem. 3(1989)117.
[8] B. Jerman-Blažič I. Fabič-Petrač and M. Randić, J. Comp. Chem. 7(1986)176.
[9] M. Randić, B. Jerman-Blažič, D.H. Rouvray, P. Seybold and S. Grossman, Int. J. Quant. Chem.: Quant. Biol. Symp. 14(1987)245.
[10] A. Štalc, A.O. Župančič, M. Šentjurč, M. Schara and S. Pečar, Period. Biol. 82(1980)369.
[11] A. Štalc, A. Šentjurč and S. Pečar, Pharmacol. 16(1980)247.
[12] W.N. Aldreidge and E. Reiner, *Enzyme Inhibitors as Substrates* (North-Holland, Amsterdam, 1972), p. 38.
[13] A. Štalc, M. Šentjurč, M. Schara, S. Pečar and A.O. Župančič, Period. Biol. 84(1982)91.