MATH/CHEM/COMP 1988, Proceedings of an International Course and Conference on the Interfaces between Mathematics, Chemistry and Computer Science, Dubrovnik, Yugoslavia, 20–25 June 1988, A. Graovac (Ed.)
Studies in Physical and Theoretical Chemistry, Volume 63, pages 25–40

SIMILARITY MEASURES BETWEEN STRUCTURED OBJECTS [1]

Vladimir BATAGELJ

Department of Mathematics, University E.K. of Ljubljana, Jadranska 19, 61 111 Ljubljana, Yugoslavia

SUMMARY

The usual approach in chemistry to compare and/or distinguish compounds on the basis of their structure is by the use of topological indices; although the alternative approach - similarity measures should receive greater attention.

There are two main approaches to similarity measures between structured objects:

- indirect approach: to each object a vector with components measuring different properties of the object is assigned. The similarity between the objects is defined as a similarity between the corresponding vectors;

- direct approach: the similarity measure is based on the structural description of the objects. Often the similarity measures the "effort" needed to transform one object to the other.

In the paper different ways to define dissimilarities between structured objects and their possible applications are presented.

Already in 1839, the French chemist Gerhardt [37] maintained that "in chemistry, we determine the form of a compound not for its own sake, but to *compare it to others* and to *distinguish it from others.*"

How to operationalize this goal ? The usual approach in chemistry is by the use of topological (structural) indices [41]; although the alternative approach - similarity measures should receive greater attention.

Both approaches are closely interrelated. In the first part of the paper we present some general facts about indices and similarity measures. The rest of the paper is mainly devoted to different ways to define the dissimilarities between structured objects and their possible applications.

FIRST APPROACH: STRUCTURAL INDICES

From the measurement theory point of view [35,25,32,31] we can introduce structural indices in the following way:

Given a set of units (descriptions of objects) $\mathcal{E}$ and a structural property of units P (e.g. size, branching, cyclization, connectivity, symmetry, etc.) which is described by an intuitive or empirical relation $\mathbf{P}$:

$$XPY \equiv \text{unit } X \text{ is less-}P \text{ than unit } Y$$

we say that the mapping $i : \mathcal{E} \rightarrow R$ is an *index* measuring the property P if it satisfies the

[1] Supported in part by the Research Council of Slovenia, Yugoslavia.

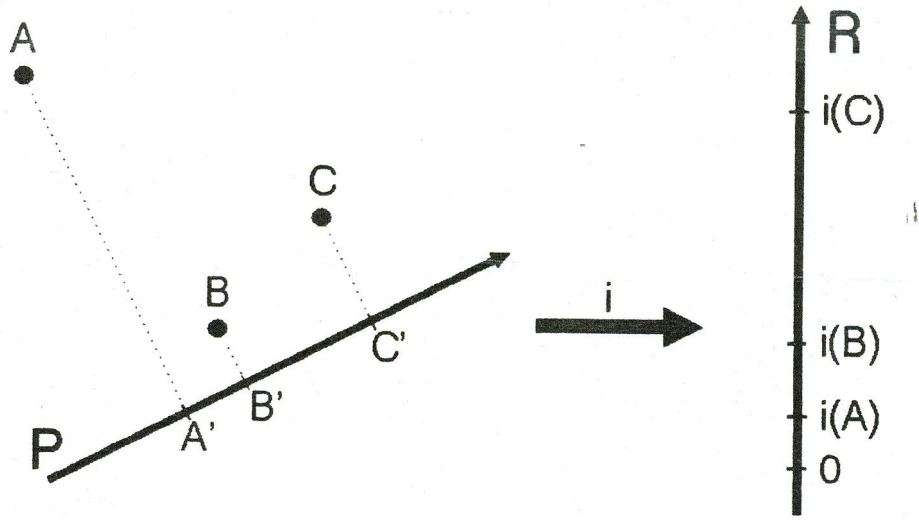

Figure 1: Structural index

condition:

$$X\mathbf{P}Y \Leftrightarrow i(X) < i(Y)$$

and possibly some other conditions reflecting transformations on units.

With some imagination we can visualize the situation as represented in Figure 1. Units are represented as points and the property P as a line in the plane. The relation $\mathbf{P}$ is determined by relative positions of projections of points on the line. Index i is a mapping of this line to the line of reals R which preserves the ordering of projections.

An important property of indices is the *unidimensionality* - each index induces a "dimension" in the "space" of units. This has some effects which should be considered in the use and interpretation of indices. For example, from Figure 1, we see that unit B is closer to A than to C *with respect to the property* P ; but in the "space" unit B is closer to C than to A.

With few exceptions [8] the inverse way to define an index is usually used. This leads to the *interpretation problem*: For a given index i what is the meaning of the corresponding property P ?

The ultimate goal of many inventors of indices is to produce an (efficiently computable) index i with the property [33]:

$$i(X) = i(Y) \Rightarrow X \approx Y$$

($\approx$ denotes isomorphism). It is evident that the complexity of such an index is not less than the complexity of the isomorphism problem for the corresponding class of (structured) units [7,14].

SECOND APPROACH: SIMILARITY MEASURES

Quantitatively we describe the similarity (association, resemblance) between units by a *similarity measure*:

$$r : (X, Y) \mapsto R$$

which assigns to each pair of units $X, Y \in \mathcal{E}$ a real number. Examples of similarity measures can be found in any book on data analysis and related topics [39,1,12,27,40,29,34,19,18].