

Introduction to Similarity Searching in Chemistry*

Ongoing revision January 29, 2005

Valentin Monev

Institute of Organic Chemistry, Bulgarian Academy of Sciences, Sofia 1113, Bulgaria

Tel: (3592) 9606193. Fax: (3592) 8700225. Email: vmonev@orgchm.bas.bg

Match-Communications in Mathematical and in Computer Chemistry **51**: 7-38 (2004)

<http://ns.pmf.kg.ac.yu/match/>

Abstract

The similarity concept and its database implementation – similarity searching, are overviewed in the context of chemoinformatics.

Similarity is defined in terms of matches/overlap, dissimilarity in terms of mismatches/difference, for qualitative/quantitative characteristics. Similarity, dissimilarity and composite measures are constructed from similarity or/and dissimilarity components. Asymmetric measures are constructed by unequal weighting of dissimilarity components. Whole objects or local regions of them are compared, yielding global or local similarity. Asymmetric local similarity is obtained by treating the objects in the comparison unequally, e.g. by ignoring parts of them. Global characteristics provide overall descriptions of objects, local characteristics provide sufficient locational information for object alignment/superposition to be effected. Similar objects are likely to have similar properties – similar property principle.

In chemical similarity searching, molecules, fragments of molecules, reactions, mixtures, journal articles, etc. are selected as objects of interest. The selection of characteristics and their encoding is illustrated using the atom pair and topological torsion descriptors, as well as their variants of increased fuzziness. Similarity measure selection is still very much a matter of trial and error. Standard query object specification is made easier by using query by example, multiple searches using a single query yield a highly informative hyperlinked screen, and joint queries involve more than one object. Similarity scores illustrate results from similarity searches and measures of their effectiveness. Areas of application include direct and reverse property prediction, data mining, virtual screening, diversity analysis, pharmacophore searching, ligand docking, structure elucidation, pattern matching, and signature analysis.

* Dedicated to the memory of Professor Oscar E. Polansky.

Introduction

Similarity (fuzzy) searching is alternative and complementary (Willett 1998; Willett, Barnard and Downs 1998) to exact searching[†]. An illustration:

A similarity search for “Robert Anderson” at “123 Main Street” would produce not only any exact matches for these search terms but also “Bobby Andersen” at “122 Maine St.” and “Rob Enderson” at “321 Main Road”. Different variations on the name “Robert” and the street name “Main” are identified as being possible matches... even sound-alikes such as “Juan” and “John” (“Fraud investigator application”)

Similarity searching retrieves objects that are similar to a query, sorted in order of decreasing similarity. High-ranked objects are likely to have similar properties to the query (‘similar property principle’ (Johnson and Maggiora 1990), see below), and thus be of interest for property prediction. Pattern matching and signature analysis are names of similarity searching that originate from other application areas (“Queryplus”):

Industry	Sample Query	Typical Search Criteria
Drug Discovery	Find molecules that have similar properties to known entities	structure (2D and 3D), shape, reactivity, molecular weight
Commercial Images	Find images that look like a given query image	color, shape, texture
Satellite Images	Find specific features within an image or a set of images	shape, intensity, texture, hyper-spectral signatures
Manufacturing	Detect signatures that match known defects or deviate from specifications	electrical, vibration, x-ray, fluorescent signatures
Financial Services / Direct Mail	Find customers most similar to target	demographic data, risk factors, purchase behavior, usage patterns

In the next two sections, the basics of the similarity concept and its database implementation – similarity searching, are presented. Although the approach is quite general, most examples are taken from chemoinformatics.

[†] For a presentation of the full range of search methods, see (Kochnev, Monev and Bangov 2003) (textbook level) and the corresponding chapter in (Gasteiger 2003) (advanced level).

Similarity Basics

In general, *similarity* $S_{A,B}$ between two objects A and B is estimated by the number of matches or the overlap in the objects, with respect to one or more of their characteristics $\{X_{jA}\}, \{X_{jB}\}, j = 1, 2, \dots, n$. For identical objects, estimates of similarity $S_{A,B}$ take a *maximal* value. As a rule-of-thumb, in the mathematical expressions for calculating $S_{A,B}$ (*similarity measures*), the numerator contains component by component multiplication $X_{jA}X_{jB}$, or intersection \cap (set theory), or the logical operator AND.

Dissimilarity $D_{A,B}$ between two objects A and B is estimated by the number of mismatches or the difference between the objects, with respect to one or more of their characteristics $\{X_{jA}\}, \{X_{jB}\}, j = 1, 2, \dots, n$. For identical objects, the estimates of dissimilarity $D_{A,B}$ take a *minimal* value. Again, as a rule-of-thumb, in the mathematical expressions for calculating $D_{A,B}$ (*dissimilarity measures*), the numerator contains component by component subtraction $X_{jA} - X_{jB}$, or union \cup (set theory), or the logical operator XOR (exclusive OR).

Similarity is often used as a general term to encompass either similarity or dissimilarity, or both (see composite measures below and Table 1). The terms *resemblance*, *proximity* and *distance* are used in mathematics (Batagelj and Bren 1995) and statistical software packages (*Electronic Statistics Textbook* 2004), but have not gained wide acceptance in the chemical literature. Similarity and dissimilarity can in principle lead to different rankings (see example below).

The denominator, if present in a similarity measure, is just a normalizer (Gower 1985); it is the numerator that is indicative of whether similarity or dissimilarity is being estimated, or both. The characteristics chosen for the description/representation[‡] of the compared objects are interchangeably called descriptors, properties, features, attributes, qualities, representations, measurements, calculations, etc. In the formulations above, ‘matches’ and ‘mismatches’ refer to qualitative characteristics, e.g. binary ones (those which take one of two values: 1, or 0, present or absent, etc.), while the terms overlap and difference refer to quantitative characteristics, e.g. those whose values can be arranged in order of magnitude along a one-dimensional axis (Sneath and Sokal 1973) p. 148 (see also Measurement scales in (*Electronic Statistics Textbook*)).

Traditionally, similarity comparison is pairwise, although nothing in principle rules out estimates of similarity to be carried out for more than two objects at a time. Likewise, similarity estimates are traditionally

[‡] Let arrays $\{X_{jA}\}, \{X_{jB}\}, j = 1, 2, \dots, n$ represent objects A and B .

considered valid only if object alignment is carried out, so that ‘heads’ are not matched to ‘tails’. Thus, similarity estimates are *by default alignment-maximized*. However, estimates of similarity for partially aligned objects could be carried out, if such a need should arise. Indeed, maximal object alignment may be well-nigh impossible, as in protein sequence alignment. A third traditional assumption is that similarity is symmetric. This assumption is now abandoned, and asymmetric similarity (see below) has been coded into similarity search engines as a major option.

One of the first papers quantifying similarity in chemistry is that of Sneath in 1966 (Sneath 1966), also co-author of a definitive textbook on numerical taxonomy (Sneath and Sokal 1973).

Measures

Case of qualitative characteristics

Following Bradshaw (Bradshaw 2001), for two objects A and B , a is the number of features (characteristics) present in A and absent in B , b is the number of features absent in A and present in B , c is the number of features common to both objects and d is the number of features absent from both objects. Thus, c and d measure the (“present” and “absent”) matches, i.e. similarity; while a and b measure the mismatches, i.e. dissimilarity. The total number of features is $a + b + c + d = n$. The total number of bits set on A is $(a + c)$, and the total number of bits set on B is $(b + c)$. These totals form the basis of an alternative notation that uses a instead of $(a + c)$, and b instead of $(b + c)$ (Willett, Barnard and Downs 1998) p. 986. This notation, however, lumps together similarity and dissimilarity ‘components’ – a disadvantage when interpreting a similarity measure.

Constructing a similarity measure from the above ‘components’ is intuitive, e.g. all matches $(c + d)$ relative to all possibilities, i.e. matches plus mismatches $(c + d) + (a + b)$, yields $(c + d)/(a + b + c + d)$, called the simple matching coefficient[§] (Sneath and Sokal 1973) p.132, and equal weight is given to matches and mismatches. When absence of a feature in both objects is deemed to convey no information, then d should not occur in a similarity measure (Gower 1985). Omitting d from the above similarity measure, one obtains the Tanimoto (alias Jaccard, Gini??) similarity measure $c/(a + b + c)$. It is monotonic with that of Dice (alias Sorensen, Czekanowski) $c/0.5[(a + c) + (b + c)]$, which uses an ‘arithmetic mean’ normalizer, and gives double weight to the ‘present’ matches. Russell/Rao $c/(a + b + c + d)$ adds the matching absences to the normalizer in Tanimoto; Rogers/Tanimoto $(c + d)/(2a + 2b + c + d)$ gives double weight to mismatches, cosine (Sneath and Sokal 1973) p.172 (alias Ochiai) is $c/\sqrt{(a + c)(b + c)}$, and uses a ‘geomet-

[§] Normalization of similarity measures yields similarity indices or coefficients, see e.g. (Cioslowski 1998)

ric mean' normalizer; Baroni-Urbani/Buser is $(\sqrt{cd} + c)/(\sqrt{cd} + a + b + c)$; Kulczynski-2 is $\frac{1}{2}\left(\frac{c}{a+c} + \frac{c}{b+c}\right)$, etc.

To construct dissimilarity measures, one uses mismatches: $(a + b)$ is the Hamming (Manhattan, taxi-cab, city-block) distance (see below), $\sqrt{(a + b)}$ is Euclidean distance, $(a + b)/(a + b + c)$ is Soergel distance, complementary for binary characteristics to Tanimoto: $1 - [c/(a + b + c)]$. Both Hamming and Euclidean distances are not normalized, increasing with the number of characteristics used; to correct for this, mean Hamming is $(a + b)/(a + b + c + d)$, and this is identical to squared mean Euclidean distance. Pattern difference is $ab/(a + b + c + d)^2$, variance is $(a + b)/4(a + b + c + d)$, 'size' is $(a - b)^2/(a + b + c + d)^2$, 'shape' is $(a + b)/(a + b + c + d) - [(a - b)/(a + b + c + d)]^2$ (Sneath and Sokal 1973)p. 170 and (SPSS 2001).

Using just similarity or dissimilarity in a similarity measure may be misleading, as in the following case (James, Weininger and Delany 2000): "Consider the following two 1024-bit fingerprints. F1 and F2 each have 407 bits set, 402 of which are common to both. F3 and F4 each have 5 bits set, none of which are common. In both cases, the (squared mean) Euclidean distance between the fingerprints is 10/1024, or 0.0098, yet clearly the first pair of fingerprints are quite similar whereas the second pair have nothing in common". The similarity 'components' in this comparison are missing! Though rarely used in similarity searches so far, composite measures using both similarity and dissimilarity components exist: Hamann is $(c + d - a - b)/(a + b + c + d)$, Yule is $(cd - ab)/(cd + ab)$, Pearson is $(cd - ab)/\sqrt{(a + c)(b + c)(a + d)(b + d)}$, dispersion is $(cd - ab)/(a + b + c + d)^2$, McConnaughey is $(c^2 - ab)/(a + c)(b + c)$, Stiles is $\log_{10}\left(n(|cd - ab| - n/2)^2/(a + c)(b + c)(a + d)(b + d)\right)$ (Holliday, Hu and Willett 2002). Note that the two terms (in the numerator) should have opposite "signs" to avoid contradicting interpretation of the whole. A simple product of (1-Tanimoto) and squared Euclidean distance is used by Dixon (Dixon and Koehler 1999). The Grotch metric (Delaney, Hallowell and Warren 1985) is $(a + b) - \mu c$, where μ weights the relative contribution of the similarity component.

Asymmetry in a similarity measure is the result of asymmetrical weighting of a *dissimilarity* component – multiplication is commutative by definition, difference is not. Weighting a and b unequally, one obtains *asymmetric similarity measures*, e.g. when $\alpha \neq \beta$ in the Tversky similarity measure $c/(\alpha a + \beta b + c)$, where α and β are user-defined constants (Bradshaw 1997; Willett, Barnard and Downs 1998). Tversky can be regarded as a generalization of the Tanimoto and Dice similarity measures; like them, it does not consider the "absence" matches d . A particular case is $c/(a + c)$, which measures the number of common features

Table 1. Types of similarity measures. For details see text.

Type	Name(s) of similarity measure	For qualitative characteristics	For quantitative characteristics	
			summation form	integration form
similarity measures	number of matches, overlap	c	$\sum_{j=1}^n X_{jA} X_{jB}$	$S_{A,B} = \iint \Gamma_{A}^*(\mathbf{r}, \mathbf{r}') \Omega(\mathbf{r}, \mathbf{r}') \Gamma_{B}(\mathbf{r}, \mathbf{r}') d\mathbf{r} d\mathbf{r}'$
	Tanimoto, Jaccard	$c/(a+b+c)$	$\frac{\sum_{j=1}^n X_{jA} X_{jB}}{\left(\sum_{j=1}^n X_{jA} X_{jA} + \sum_{j=1}^n X_{jB} X_{jB} - \sum_{j=1}^n X_{jA} X_{jB} \right)}$	(Good and Richards 1998)
	Dice, Sorensen, Czekanowski, Hodgkin-Richards	$c/0.5[(a+c)+(b+c)]$	$2 \sum_{j=1}^n X_{jA} X_{jB} / \left(\sum_{j=1}^n X_{jA} X_{jA} + \sum_{j=1}^n X_{jB} X_{jB} \right)$	$S_{A,B} / 0.5(S_{A,A} + S_{B,B})$
	Cosine, Ochiai, Carbo	$c/\sqrt{(a+c)(b+c)}$	$\sum_{j=1}^n X_{jA} X_{jB} / \sqrt{\sum_{j=1}^n X_{jA} X_{jA} \sum_{j=1}^n X_{jB} X_{jB}}$	$S_{A,B} / \sqrt{S_{A,A} S_{B,B}}$
dissimilarity measures	Hamming distance	$(a+b)$	$\sum_{j=1}^n X_{jA} - X_{jB} $	
	Euclidean distance	$\sqrt{(a+b)}$	$\sqrt{\sum_{j=1}^n (X_{jA} - X_{jB})^2}$	$D_{A,B} = \left(\iint \Gamma_{A}(\mathbf{r}, \mathbf{r}') - \Gamma_{B}(\mathbf{r}, \mathbf{r}') ^2 d\mathbf{r} d\mathbf{r}' \right)^{1/2}$
composite measures	Pearson	$\frac{(cd-ab)}{\sqrt{(a+c)(b+c)(a+d)(b+d)}}$	$\frac{\sum_{j=1}^n (X_{jA} - \bar{X}_{jA})(X_{jB} - \bar{X}_{jB})}{\sqrt{\sum_{j=1}^n (X_{jA} - \bar{X}_{jA})^2 \sum_{j=1}^n (X_{jB} - \bar{X}_{jB})^2}}$	
	Tversky contrast model	$\theta c - \alpha a - \beta b$		

relative to all the features present in \mathcal{A} , and gives zero weight to b^{**} . The composite measure Tversky contrast model $\theta c - \alpha a - \beta b$ (James, Weininger and Delany 2000) can be considered as allowing asymmetry in the Grotch metric. Taking the max or min is another way of weighting unequally a and b in a similarity measure, e.g. in the Simpson coefficient $c/\min[(a+c),(b+c)]$ (Bradshaw 2001), and in the Petke coefficient $c/\max[(a+c),(b+c)]$ (Petke 1993).

Case of quantitative characteristics

Overlap is usually expressed mathematically with component by component multiplication $X_{j\mathcal{A}}X_{j\mathcal{B}}$ followed by summation (integration). In the limit case of binary characteristics, $\sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{B}}$ is c ,

$\sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{A}}$ is $(a+c)$, $\sum_{j=1}^n X_{j\mathcal{B}}X_{j\mathcal{B}}$ is $(b+c)$, easily verified by substituting 0's and 1's in these expressions. d does not figure in them, and there seems to be no way to estimate 'empty' overlap within this mathematical apparatus. The summation form of some similarity measures follow. The Tanimoto coefficient

is $\sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{B}} / \left(\sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{A}} + \sum_{j=1}^n X_{j\mathcal{B}}X_{j\mathcal{B}} - \sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{B}} \right)$, Dice is

$2 \sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{B}} / \left(\sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{A}} + \sum_{j=1}^n X_{j\mathcal{B}}X_{j\mathcal{B}} \right)$, cosine is $\sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{B}} / \sqrt{\sum_{j=1}^n X_{j\mathcal{A}}X_{j\mathcal{A}} \sum_{j=1}^n X_{j\mathcal{B}}X_{j\mathcal{B}}}$, Pearson is

obtained from cosine by replacing $X_{j\mathcal{A}}$ with $X_{j\mathcal{A}} - \bar{X}_{j\mathcal{A}}$, and $X_{j\mathcal{B}}$ with $X_{j\mathcal{B}} - \bar{X}_{j\mathcal{B}}$, respectively, where

$\bar{X}_{j\mathcal{A}}$ is the mean $\frac{1}{n} \sum_{j=1}^n X_{j\mathcal{A}}$. Spearman is Pearson where the values of the characteristics have been replaced by their ranks (*Electronic Statistics Textbook*).

Replacing summation with integration, one obtains the integration form of the above-described similarity measures. Using different characteristics to describe the compared objects, one obtains 'different' similarity measures^{††}: the Carbo similarity measure (Carbo, Arnau and Leyda 1980) is $\mathcal{S}_{\mathcal{A},\mathcal{B}} = \iint \rho_{\mathcal{A}}(\mathbf{r})\Omega(\mathbf{r},\mathbf{r}')\rho_{\mathcal{B}}(\mathbf{r}')d\mathbf{r}d\mathbf{r}'$, where $\rho_{\mathcal{A}}(\mathbf{r})$ and $\rho_{\mathcal{B}}(\mathbf{r}')$ are the electron density functions (see section Characteristics) of quantum objects \mathcal{A} and \mathcal{B} , weighted by a positive definite operator $\Omega(\mathbf{r},\mathbf{r}')$, chosen e.g. as the Dirac function $\delta(\mathbf{r}-\mathbf{r}')$, the Coulomb operator $|\mathbf{r}-\mathbf{r}'|^{-1}$, etc. (Carbo-Dorca and Besalu 1998).

** For an interpretation in the case of structural features, see section Objects.

†† Some authors give a broader definition of the similarity measure concept by including the characteristics and weighting scheme in it, see e.g. (Willett, Barnard and Downs 1998) p. 985.

The resulting similarity measures are “overlap-like” $S_{A,B} = \int \rho_A(\mathbf{r})\rho_B(\mathbf{r})d\mathbf{r}$, “Coulomb-like”, etc. The Carbo similarity *coefficient* is obtained after geometric mean normalization $S_{A,B}/\sqrt{S_{A,A}S_{B,B}}$ (cosine), while the Hodgkin-Richards (Richards and Hodgkin 1988) similarity coefficient uses arithmetic mean normalization $S_{A,B}/0.5(S_{A,A} + S_{B,B})$ (Dice). The Cioslowski similarity measure NOEL (Cioslowski and Fleischmann 1991) $S_{A,B} = \iint \Gamma_A^*(\mathbf{r},\mathbf{r}')\Gamma_B(\mathbf{r},\mathbf{r}')d\mathbf{r}d\mathbf{r}'$ uses reduced first-order density matrices (one-matrices) (Szabo and Ostlund 1989; Committee 1995) rather than density functions to characterize A and B . No normalization is necessary, since NOEL has a direct interpretation – Number of Overlapping Electrons, at HF level of theory (Cioslowski and Surjan 1992). An atomic similarity index in integration form (Cioslowski and Nanayakkara 1993) is presented as an example of *local* similarity in section Objects.

Difference between two objects is usually expressed with component by component subtraction $X_{jA} - X_{jB}$, followed by summation (integration). Difference is asymmetric unless special care is taken, e.g. by taking its absolute value, its square, an arithmetic mean, a geometric mean, a harmonic mean, etc. A difference is called a *distance* in n-dimensional descriptor space (Sneath and Sokal 1973) p. 121 if it satisfies a number of mathematical conditions (Gower 1985; Batagelj and Bren 1995)(symmetry is one of them). These were previously considered all-important for similarity measures (see e.g. (Santini and Jain 1999), but now tend to be relaxed in favor of pragmatic considerations such as ease of computation and general usefulness.

In the limit case of binary characteristics, $\sum_{j=1}^n |X_{jA} - X_{jB}|$ yields $(a + b)$, easily verified by substituting 0's

and 1's in these expressions. The summation form of some dissimilarity measures follow: mean Hamming

distance is $\frac{1}{n} \sum_{j=1}^n |X_{jA} - X_{jB}|$, mean squared Euclidean distance is $\frac{1}{n} \sum_{j=1}^n (X_{jA} - X_{jB})^2$, power distance

(*Electronic Statistics Textbook*) is $\left(\sum_{j=1}^n |X_{jA} - X_{jB}|^p \right)^{1/r}$, where p and r are user-defined, divergence is

$\frac{1}{n} \frac{\sum_{j=1}^n (X_{jA} - X_{jB})^2}{\sum_{j=1}^n (X_{jA} + X_{jB})^2}$, Bray/Curtis is $\frac{\sum_{j=1}^n |X_{jA} - X_{jB}|}{\sum_{j=1}^n (X_{jA} + X_{jB})}$, Soergel distance is $\frac{\sum_{j=1}^n |X_{jA} - X_{jB}|}{\sum_{j=1}^n \max(X_{jA}, X_{jB})}$. Chebychev

distance is $\max_{j=1}^n |X_{jA} - X_{jB}|$ (*Electronic Statistics Textbook*). Mahalanobis distance ("Mahalanobis Metric") is

a generalization of standardized Euclidean distance, scaling the coordinate axes (characteristics) and correcting for correlation between them. Hausdorff distance ("Hausdorff distance") measures distance between two sets: maximum distance of a set to the nearest point in the other set. Levenshtein (edit) dis-

tance is the smallest number of insertions, deletions and substitutions required to change one string or tree into another ("Levenshtein distance"). Very many distances have been defined and used as dissimilarity measures, see e.g. Euclidean distance in quantum chemistry (Fratev, Polansky, Mehlhorn and Monev 1979), form and shape distance (Klein and Babic 1997).

The integration form of Euclidean distance as introduced in quantum chemistry by Carbo (Carbo and Domingo 1987) and Cioslowski (Cioslowski 1991) is e.g. $D_{A,B} = \left(\iiint |\Gamma_A(\mathbf{r}, \mathbf{r}') - \Gamma_B(\mathbf{r}, \mathbf{r}')|^2 d\mathbf{r} d\mathbf{r}' \right)^{1/2}$, and is calculated as $D_{A,B} = (S_{A,A} + S_{B,B} - 2S_{A,B})^{1/2}$, using the already defined quantities $S_{A,B}$ above.

Exercises

- Calculate the Tanimoto similarities and Euclidean distances for the two pairs F1/F2 and F3/F4 in the example above and compare them.
- Do the same for three objects represented by the following 10 bit binary strings. Interpret your results (Bradshaw 1997). Choose one object as a query, and rank the other two in order of decreasing similarity.

Object 1 1 0 0 0 1 1 0 0 1 0

Object 2 0 0 0 0 1 0 0 1 1 0

Object 3 0 0 0 0 0 0 0 1 0 0

- Verify that $\sum_{j=1}^n X_{jA} \text{ XOR } X_{jB}$ is $(a + b)$ in logical representation. Hint: The logical XOR (Exclusive OR) operation compares 2 bits and if exactly one of them is "1" (i.e., if they are different values), then the result is "1"; otherwise (if the bits are the same), the result is "0" ("Logical operations").
- Verify that $\sum_{j=1}^n X_{jA} \text{ AND } X_{jB}$ is c in logical representation. No hints!
- Verify that $\sum_{j=1}^n \min(X_{jA}, X_{jB})$ is c , where min is the standard minimum function.
- Verify that the binary form of Pearson, as given, is a limit case of the summation form. Hint: $n = a + b + c + d$.
- Verify that $\sum_{j=1}^n \max(X_{jA}, X_{jB}) = \sum_{j=1}^n X_{jA}, X_{jA} + \sum_{j=1}^n X_{jB}, X_{jB} - \sum_{j=1}^n X_{jA}, X_{jB}$ yields $(a + b + c)$ in the limit case of binary characteristics.

Objects

Any two objects of interest are legitimate candidates for quantification of their similarity with respect to a set of their characteristics. Objects of interest to a chemist include molecules, molecular substructures, reactions, mixtures, spectra, chromatograms, x-ray diffraction images, patents, journal articles, polymers, atoms, functional groups, complex chemical systems, molecular electrostatic fields, etc. An example from quantum chemistry (Cioslowski and Challacombe 1991) p. 82: Let A and B be two (not necessarily different) systems. A and B can be two different molecules, the same molecule described within two different quantum-chemical approximations, two different electronic states of the same molecule, or the same molecule with two different geometries.

Object alignment/superposition

If the matches/mismatches, overlap, or difference depend on the mutual ‘orientation’ of the compared objects A and B , an optimization procedure is required to locate an object alignment/superposition (Robinson, Lyne and Richards 1999) that maximizes similarity. Depending on the aim of the investigator, the characteristics used, the optimization procedure, etc., more than one alignment may be identified – indeed exploration of similarity space may be necessary to find a set of ‘best possible alignments’ (Mestres, Rohrer and Maggiora 1997). For objects that are too dissimilar, the assumption that the whole of one object can be aligned with the whole of another object becomes invalid – *local* alignment (Robinson, Lyne and Richards 2000) may be appropriate.

Global/local similarity

Depending on the interest of the investigator, *local regions* of objects can be compared, yielding local (or sub-) similarity (Willett, Barnard and Downs 1998) p. 984. Let the objects of interest A and B be specified as local regions of objects X and Y , respectively, and let the descriptors used[#] represent the complete objects. Local similarity can be obtained by counting or summation (Petke 1993; Mestres, Rohrer and Maggiora 1997; Amat, Besalu, Carbo-Dorca and Ponec 2001)/integration (Mezey 1999) in just the local regions of the complete objects. Again an example from quantum chemistry (Cioslowski and Nanayakkara 1993): Γ_X and Γ_Y are the reduced first-order density matrices (one-matrices) describing (molecules) X and Y , and we are interested in the similarity between (atoms) A and B . The diagonal elements of the continuous representation of the one-matrices are the electron densities $\Gamma_X(\mathbf{r}, \mathbf{r}) = \rho_X(\mathbf{r})$, $\Gamma_Y(\mathbf{r}, \mathbf{r}) = \rho_Y(\mathbf{r})$ of X and Y (Szabo and Ostlund 1989) p. 253. The similarity between the complete objects X and Y , as we have seen in section Measures, can be estimated by the ‘overlap’ in the electron densities or by the number of overlapping electrons. Integration of the molecular descriptors (molecular

[#] The descriptors used should be ‘local’, see next section.

electron densities) in just the sub-domains Ω_A and Ω_B (e.g. atomic basins), and in their common space $\Omega_{AB} = \Omega_A \cap \Omega_B$ (obtained by local alignment), allows quantification of the similarity between the local regions A and B e.g. atoms in molecules. The product of the number of electrons from each atom in the common space is a measure of the overlap of the electron densities of the two atoms. Each contribution is normalized by the total number of electrons in that atom (calculated as a local region of the respective molecule) $\left[\int_{\Omega_{AB}} \rho_X(\mathbf{r}) d\mathbf{r} / \int_{\Omega_A} \rho_X(\mathbf{r}) d\mathbf{r} \right] \left[\int_{\Omega_{AB}} \rho_Y(\mathbf{r}) d\mathbf{r} / \int_{\Omega_B} \rho_Y(\mathbf{r}) d\mathbf{r} \right]$ (Cioslowski and Nanayakkara 1993).

This symmetric index of atomic similarity depends not only on the spatial overlap of the two atoms (size and shape of the atomic basins), but also on how strongly they are electronically populated in their common space.

Global similarity is appropriate when the need is to identify complete objects similar to the query object. In global similarity, information about which local regions of the objects are similar is lost as a result of the match counting or summation (integration). Global similarity, however, may reflect the dominance of some local region.

Symmetric local similarity is appropriate when the need is to identify objects containing local regions similar to local regions in the query.

Another type of local similarity is obtained by using an *asymmetric* (directional) similarity measure, i.e. by treating the two objects in the similarity comparison on an unequal footing. Let the characteristics used be structural fragments, A be the query object, and B - a database object. The values of the similarity measure $c/(a+c)$ (see section Measures) do not depend on b - the fragments present in B and absent in A are ignored. Thus the highest ranking database objects B retrieved in a similarity search will be *superstructures* of A . Using $c/(b+c)$, the highest ranking database objects B retrieved in a similarity search will be *substructures* of A (Willett 1998) p. 2751. By ignoring fragments of B and of A , respectively, in the above two cases, similar local regions are identified.

Asymmetric local similarity is appropriate when the need is to identify i) objects containing local regions similar to the complete query or ii) complete objects similar to a local region of the query.

Exercises:

- Interpret the atomic similarity index in terms of $c/(a+c)$ and $c/(b+c)$, and the cosine index.
- Interpret intermediate values of α and β in the Tversky similarity measure when structural fragments are used.

Characteristics

Any set of characteristics can be used to describe/represent the compared objects. However, a similarity estimate is meaningless, unless the characteristics used contain in a direct or indirect way the information that is sought. Using different characteristics one obtains different estimates of similarity.

Global/local characteristics

Object characteristics can be roughly classified as global and local, with the latter providing sufficient locational information for object alignment/superposition to be effected (Downs and Willett 1996). Local similarity can only be estimated when such characteristics are used. Global characteristics are at the other extreme, providing overall descriptions of objects. Obviously, intermediate types of characteristics exist, and even ones with ‘variable resolution’ (Crippen 1999).

Examples of global characteristics are the atom pair (ap) (Carhart, Smith and Venkataraghavan 1985) and the topological torsion (tt) (Nilakantan, Bauman, Dixon and Venkataraghavan 1987). Atom pairs are defined as substructures of the form $AT_i - AT_j - (\text{distance})$, where (distance) is the distance in bonds along the shortest path between an atom of type AT_i and an atom of type AT_j (a slightly modernized definition is presented following (Hull, Fluder, Singh, Nachbar, Kearsley and Sheridan 2001)). Atom types encode the species of atom, the number of non-hydrogen atoms attached to it, and the number of incident π -bonds. For instance, ‘n21o1005’ is an atom pair of a nitrogen with 2 non-hydrogen neighbors and one π -bond, five bonds away from an oxygen with one neighbor and no π -bonds. Topological torsions are of the form $AT_i - AT_j - AT_k - AT_l$, where i, j, k, and l are consecutively bonded distinct atoms and the atom types are as described above. All of the ap’s and/or tt’s in a molecule are counted to form a frequency vector.

Figure 1 shows an example of a molecule parsed into atom pairs and topological torsions, and their fuzzier counterparts binding pairs (bp) and binding torsions (bt) (Kearsley, Sallamack, Fluder, Andose, Mosley and Sheridan 1996) (described in the next section).

Atom pairs are descriptors which can be applied straightforwardly, no need for descriptor selection and/or reduction, no adjustable parameters, no ad hoc assumptions about what substructures are important. Long-range relationships between atoms are captured, and the descriptors are generalizable to 3D and properties (see next section). They are easily perceived, compact in representation, usable for large numbers of objects and for large objects (in chemistry). As defined, atom pairs do not encode stereochemical or conformational information. They can conveniently be computed from a constitutional representation of chemical structure e.g. a connection table. 3-methyl-1*H*-pyridin-4-one shown in Figure 1 has 8 non-hydrogen atoms and thus 28 ($n(n-1)/2$ for $n=8$) atom pairs, 23 of which are distinct. An individual ap does not convey much structural information, but the set of ap’s of a molecule captures fairly

well structural information. The ap's are general enough that a significant number may be found in common among diverse structures yet specific enough that in the aggregate they can discriminate even closely related topological isomers from one another. For example, among about 170000 structures in a public database of structures, fewer than 340 structures share identical sets of atom pairs with other, topologically distinct structures. Most of these correspondences are related to a pair of isomers (1,4- and 1,5-disubstituted naphthalene) which cannot be distinguished by their set of ap's (Carhart, Smith and Venkataraghavan 1985).

unique ap	frequency	unique tt	frequency	unique bp	frequency			
1	c31c2102	3	1	o11c31c31c21	1	1	4603	3
2	c31c2101	2	2	o11c31c31c10	1	2	6601	3
3	o11c2103	2	3	c31c31c21c21	1	3	6602	3
4	n20c2101	2	4	c21c31c31c21	1	4	6603	3
5	c31c2103	1	5	c21c31c31c10	1	5	6702	3
6	c21c2102	1	6	n20c21c31c31	1	6	3601	2
7	c21c2101	1	7	o11c31c21c21	1	7	3602	2
8	c21c2103	1	8	c31c21n20c21	1	8	4602	2
9	c31c3101	1	9	n20c21c21c31	1	9	6701	2
10	n20c3103	1	10	n20c21c31c10	1	10	3404	1
11	n20c3102	1	11	c21n20c21c21	1	11	3603	1
12	o11c3102	1	11 total		12	3703	1	
13	o11n2004	1	-----		13	4701	1	
14	n20c2102	1	-----		14	6604	1	
15	o11c3101	1	-----		28 total			
16	o11c2102	1	-----		-----			
17	c31c1001	1	-----		unique bt	frequency		
18	c21c1004	1	-----		1	4766	3	
19	c31c1002	1	-----		2	6676	3	
20	o11c1003	1	-----		3	3667	2	
21	c21c1003	1	-----		4	6366	2	
22	c21c1002	1	-----		5	3666	1	
23	n20c1003	1	-----		11 total			
28 total		-----		-----				

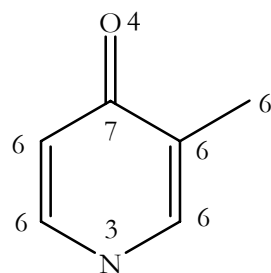


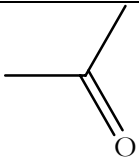
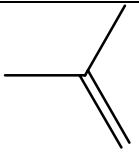
Figure 1. Sample molecule parsed into descriptors: atom pairs, topological torsions, binding pairs, binding torsions. Numbering at atoms indicates binding property atom type.

Computing the set of atom pairs from the connection table of a structure requires an algorithm for finding the length of the shortest path between each pair of non-hydrogen atoms in the structure. The properties of atom type are easily extracted from the topological description of the structure. The number of incident

π -bonds is 1 for aromatic compounds. Tautomer bonds are reduced to a specific pattern of double and single bonds, with common cases such as carboxylic acids, amides, and ureas handled in a self-consistent manner. Dative bonds, such as N-O in an N-oxide, are treated as double bonds. Further details of the encoding of the atom pair descriptor can be found in the original paper (Carhart, Smith and Venkataraghavan 1985).

Exercise

- Enumerate the unique ap's contained in the molecules acetone and isobutylene, and estimate their similarity using the Dice index. ap 1 occurs once in both molecules, and ap 4 occurs twice in both molecules, i.e. $c = (1+2) = 3$. There are 6 ap's in total in acetone ($a + c$) = 6, and there are 6 ap's in total in isobutylene ($b+c$) = 6. The similarity score is $3/0.5(6 + 6) = 0.5$.

unique ap's			descriptor average representing e.g. a joint query, a centroid, or a mixture of the two molecules at left (see text in next section)	
1	c10c1002	1	1	1
2	c10o1102	2	0	1
3	c10c1102	0	2	1
4	c31c1001	2	2	2
5	c31o1101	1	0	0.5
6	c31c1101	0	1	0.5

The topological torsion is another descriptor developed at Lederle laboratories. The rationale for the tt descriptor is that the torsion angle is the minimal structural unit in terms of which the conformation of a molecule can be completely described. The 3D structure of a molecule can be completely built by using a series of torsions as basic building blocks. tt is the topological analogue of the torsion.

Encoding the tt's in a molecule is straightforward: starting from a connectivity table, the algorithm finds all the possible tt's by looping first over all the atoms and then over three successive levels of branching. Checks are made that the atoms in the tt quartet are distinct and that the same tt is not counted twice in opposite directions.

Unlike the atom pair, tt is not a long-range descriptor. A small change in one part of a large molecule does not affect the tt's in a distant part of the molecule. In contrast, a change of a single atom in a molecule

alters all atom pairs involving that atom. Thus the tt complements rather than replaces the atom pair descriptor.

The performance of the atom pair and topological torsion descriptors in similarity searching is compared in section Similarity scores.

Examples of *local* descriptors are quantum-mechanical reduced first-order density matrices (one-matrices) and electron density functions, as we have already seen. The complete information for an N-electron system is contained in the density matrix operator $\Gamma^{(N)}$, whose coordinate representation is $\Gamma^{(N)}(\mathbf{r}_1, \dots, \mathbf{r}_N; \mathbf{r}'_1, \dots, \mathbf{r}'_N) = \Psi(\mathbf{r}_1, \dots, \mathbf{r}_N) \Psi^*(\mathbf{r}'_1, \dots, \mathbf{r}'_N)$, where Ψ is the normalized N-electron state of the system in coordinate representation, and \mathbf{r}_i are the space-spin coordinates of the i-th electron. Integrating over the coordinates of $N - 1$ electrons, one obtains the one-particle reduced density matrix operator Γ , whose coordinate representation is $\Gamma(\mathbf{r}_1, \mathbf{r}'_1) = N \int d\mathbf{r}_2 \cdots d\mathbf{r}_N \Psi(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \Psi^*(\mathbf{r}'_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$. The diagonal elements $\Gamma(\mathbf{r}, \mathbf{r}) = \rho(\mathbf{r})$ are called electron density functions because they give the probability of finding an electron at space-spin coordinate \mathbf{r} . The momentum-space representation of the diagonal elements of Γ has also been used as a local descriptor (Cooper and Allan 1989). Two-particle reduced density matrices (obtained after integration over the coordinates of $N - 2$ electrons), and their diagonal elements, are descriptors with even higher information content (Ponec and Strnad 1990).

As with all local descriptors, a major stumbling block for their widespread use is the high computational cost of object alignment. Workarounds include the use of lower-resolution local descriptors for the initial stages of object alignment.

Similar property principle

Similar objects are likely to have similar properties (Johnson and Maggiora 1990). At the basis of this assumption is the so-called Principle of Continuity “Changes in nature are gradual”, which can be traced to Aristotle’s “Natura non facit saltus”. A more practical formulation is “Systems which differ little in their mathematical properties will differ little in their physical, chemical and biological properties” (Trinajstić, Klein and Randić 1986). The similar property principle is inapplicable for cases where abrupt changes occur e.g. singularities and bifurcation phenomena.

The expectation that similar objects will frequently show similar properties has a statistical interpretation, e.g. in medicinal chemistry, the density of actives in a set of molecules that are structurally similar to an active lead will exceed the density of actives in a set of randomly chosen molecules (Horvath and Jeandéans 2000).

Similarity Searching

Similarity searching is the database implementation of the similarity concept. One of the first reviews on chemical similarity searching is in the mid 1980-s (Willett 1987); comprehensive reviews include (Downs and Willett 1996; Willett 1998; Willett, Barnard and Downs 1998).

A do-it-yourself (DIY) approach to similarity searching would involve selection of the objects to use and assembling, or acquiring access to, an appropriate database; selection of the characteristics to describe the objects, the way to encode the characteristics and processing of the encoded data to allow searching; selection of the similarity measures for use and encoding them; developing an algorithm for searching and building/deploying a search engine; building a graphical user interface including possibly Internet-based access; testing of the similarity search implementation by specifying query objects, similarity measures and similarity cutoff values as well as possibly a subset of characteristics to use, and interpreting the validity of the results; using the similarity search application in areas such as property prediction, pattern matching and signature analysis. Some of these steps are overviewed next, in the context of chemoinformatics.

Object selection

The most common objects of interest to a chemist are *molecules*. One source of drug-like compounds is the MDL Drug Data Report (MDDR) (MDDR 2002) a licensed database compiled from the patent literature and conference proceedings. Contains about 115 000 compounds, coverage starting from 1988. A small percentage of molecules in the MDDR are very large (e.g. peptides) and some are very small. If one wants to consider drug-sized molecules, only those molecules within the range of 7-50 non-hydrogen atoms could be retained. Molecules in the MDDR are assigned a “therapeutic category” by the vendor. Some therapeutic categories (e.g. “antihypertensive”) contain molecules that work by different mechanisms. There are 647 therapeutic categories. A molecule may be in more than one therapeutic category, and some therapeutic categories are nearly synonymous (Sheridan 2002). MDL Drug Data Report-3D is also available. A comparison of eight large chemical databases is given in (Voigt, Bienfait, Wang and Nicklaus 2001). Considerations for compound acquisition can be found in (Rhodes, Willett, Dunbar and Humblet 2000).

A mind-opening example of a *fragment*-based search space is given in (Rarey and Stahl 2001): “the search space could either be an explicitly enumerated compound database, the closed form of a combinatorial library, or a more generally defined “chemistry space”. Under chemistry space, we understand a large set of diverse fragments together with generic definitions of how the fragments can be combined to molecules... a chemistry space created by shredding the World Drug Index (World Drug Index 2001) into small fragments. The space contains about 17000 fragments which can be connected to each other via 12 different link types. The space is theoretically infinite. Limited to reasonably sized molecules (consisting of

less than 6 fragments), it still contains about 2.15^{18} molecules. Depending on the size of the query molecule, this space can be searched in between 2 and 20 minutes on a single CPU workstation.” Searching this virtual search space can be done using the feature tree descriptor which represents a molecule by its fragments (Rarey and Dixon 1998).

Reactions can be considered as composite systems containing reacting and product molecules, as well as reaction sites. The familiar atom pair can be used as a descriptor (Grethe and Moock 1990).

Journal articles can be described by the citations contained in them. The assumption is that articles whose reference lists include some of the same sources have a subject relationship, regardless of whether their titles, abstracts, or keywords contain the same terms. The number of matching citations can be used as a similarity measure. One can specify which citations in the query article to use as descriptors (those that seem relevant to the aims of the investigator); if not, by default, all the citations are used in the similarity search. Articles that share the largest number of citations with the query article are listed first (ISI). Similarity searching via a citation index can be combined with a conventional reaction database, to obtain a *reaction similarity and retrieval* tool. Reactions are especially suitable for citation-based similarity searching because citations are “based on all important features of the reaction, not just the molecular structures or bonds broken and made in the reaction” (Garfield 2002).

Mixtures containing up to several thousand distinct chemical entities are often synthesized and tested in mix-and-split combinatorial chemistry. The descriptor representation of a mixture may be approximated as the descriptor average of its individual component molecules, using e.g. atom pair and topological torsion descriptors (Sheridan 2000) (see example for acetone and isobutylene).

Exercise

- Define a search space for materials suitable for nonlinear optics, a possible project in materials science. Bear in mind that materials with appropriate *bulk* properties are sought. Send your suggestions to the author, who is currently investigating the viability of such a project ☺.

Descriptor selection and encoding

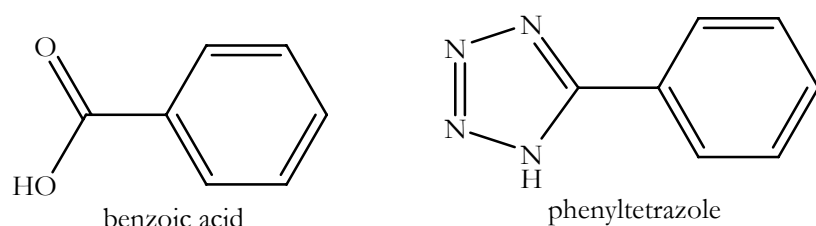
The atom pair and topological torsion descriptors are selected for illustrative purposes in the similarity searching context. A descriptor generator (Kearsley, Sallamack, Fluder, Andose, Mosley and Sheridan 1996) is used to generate ap and tt descriptors from the connection table of each molecule in the chemical database (e.g. MDDR). A first pass through the database is performed to create a catalog of unique descriptors ($\sim 10\,200$ unique ap's and $\sim 5\,900$ unique tt's) and another catalog of each molecule name. Then, a second pass creates a list of the frequency of each descriptor found in each molecule (Hull, Singh, Nachbar, Sheridan, Kearsley and Fluder 2001).

A limitation of the atom pair descriptor is its topological nature – the 3D shape of a structure is not captured. A 3D analogue of ap - the *atom pair geometric ag* - is defined by changing the definition of (distance) from through-bond (topological) to through-space (geometric) distance (Sheridan, Miller, Underwood and Kearsley 1996). Since geometric distance is continuous, it is partitioned into discreet “bins”. Distance is divided into 30 bins starting at 1Å and ending at 75.3. The interval between the first and second bin starts at 0.5 and thereafter the interval increases. The centers of the bins are <1.0, 1.5, 2.1, 2.7, 3.3, 4.1, 4.9, etc. Distances are “fuzzy” in that a particular pair of atoms contributes to two adjacent bins according to how close it is to the bin centers. This is done so that, for instance, a distance of 4.7 would be perceived as similar to 5.0, even though the distances might fall in different bins. For example, a nitrogen with 3 non-hydrogen neighbors and zero π -bonds and a carbonyl oxygen at a geometric distance of 4.7 generate 0.25 of ag n30o1106 (bin 6) and 0.75 of ag n30o1107 (bin 7). The number of times a given descriptor occurs in a molecule, i.e. its count, may not be an integer, but the sum of the counts over all descriptors is still $n(n-1)/2$. The ag’s for the conformations in a 3D database are generated just like the ap’s for the chemical structure diagrams in a 2D database.

Exercise:

- Obtain the ag’s for the molecule in Figure 1. Hint: Copy/paste “3-methyl-1H-pyridin-4-one” into a 2D software package (e.g. Chemdraw (Chemoffice Ultra 2003)); copy/paste the obtained chemical structure diagram into a 3D software package (e.g. Chem3D (Chemoffice Ultra 2003)) to obtain a 3D conformation; select an atom and point to a second atom to obtain the interatomic distance.

A limitation of the ap, tt and ag descriptors is the specificity of the atom typing, e.g. benzoic acid and phenyltetrazole would not be perceived as very similar, even though carboxylates and tetrazoles are both anions at physiological pH.



A fuzzier atom type participating in these descriptors has been defined that is pharmacologically relevant – physiochemical type at near-neutral pH (Kearsley, Sallamack, Fluder, Andose, Mosley and Sheridan 1996), one of seven binding property classes: 1 = cation; 2 = anion; 3 = neutral hydrogen-bond donor; 4 = neutral H-bond acceptor; 5 = polar atom (atoms which are both donors and acceptors, e.g. hydroxy

oxygen or either donor or acceptor via tautomerization, e.g. the nitrogens of imidazole); 6 = hydrophobe; 7 = other (nonpolar atoms in a polar environment or polar atoms that cannot accept or donate H-bonds). Figure 1 shows an example of a typed molecule, parsed into binding pair (bp) and binding torsion (bt) descriptors (bg requires a 3D conformation). An automated method to assign these types has been described.

The physiochemical atom type, however, is too fuzzy an atomic descriptor for the purpose of identifying common substructures^{§§} (symmetric local similarity based on atoms and topological distances as descriptors). In this case, atom type has been defined as a string containing chemical element (all halogens equivalenced to the “element” Hal), number of incident π -bonds, and physiochemical type (Sheridan and Miller 1998).

Two other atomic properties have been used in the definition of atom type, thereby increasing its fuzziness relative to that in the ap and tt descriptors – atomic log P contribution (yielding hydrophobic pairs hp’s and torsions ht’s) and partial atomic charges (charge pairs cp’s and torsions ct’s) (Kearsley, Sallamack, Fluder, Andose, Mosley and Sheridan 1996). Both properties take continuous values, so a set of 7 equidistant and overlapping bins is used to represent ranges from -0.50 or below to 0.50 or above. Each particular value is binned fuzzily^{***} by getting assigned to two adjacent bins, as a result of the 50% overlap of the bins.

Increasing the fuzziness of object description reduces the number of descriptors used and broadens the scope of a similarity search. At the same time, increasing fuzziness may reduce the discriminatory power of descriptors to unacceptable levels. Thus, it is desirable to be able to control the degree of fuzziness of descriptors.

Several approaches to the estimation of descriptor fuzziness have been proposed (Kearsley, Sallamack, Fluder, Andose, Mosley and Sheridan 1996). For a single object, the ratio $R = \text{total number of descriptors} / \text{number of unique descriptors}$ can be regarded as a measure of fuzziness of a descriptor, e.g. $R = 28/23 = 1.22$ for the ap descriptor in the molecule in Figure 1. For a population of objects, a median of R can be taken over a large sample of the objects. A second approach is to estimate how similar any two objects are likely to be – the median similarity for a large sample of pairs of objects can be regarded as an estimate of the fuzziness of the used descriptor. A standard object/sample of objects should be used, for comparisons to be valid.

^{§§} According to this definition of (2D) common substructures (Sheridan and Miller 1998), corresponding atoms in \mathcal{A} and \mathcal{B} that are local regions of X and Y , respectively, must have the same atom type, and the topological distances between the atoms in \mathcal{A} must be the same as the distances between the corresponding atoms in \mathcal{B} . Only non-hydrogen atoms are considered. Substructures (cliques) \mathcal{A} and \mathcal{B} may be discontinuous.

^{***} For another example of fuzzy binning see (Chen, Rusinko, Tropsha and Young 1999).

For 35635 molecules from the World Drug Index (World Drug Index 2001), a median of R is ap 2.17, bp 4.68, hp 5.43, cp 6.20, tt 1.68, bt 3.75, ht 4.26, ct 4.56. The median pairwise similarity for 10000 randomly selected pairs of compounds from the same source is ap 0.15, bp 0.36, hp 0.39, cp 0.43, tt 0.04, bt 0.26, ht 0.31, ct 0.35. Both approaches indicate that the fuzziness of the newly defined descriptors is indeed increased, and in the following order original \ll binding property $<$ hydrophobic $<$ charge. Torsions are always less fuzzy than the corresponding pairs $tt < ap$, $bt < bp$, $ht < hp$, $ct < cp$. The least fuzzy descriptor is tt.

A general approach to increasing the fuzziness of descriptors has recently been proposed (Hull, Singh, Nachbar, Sheridan, Kearsley and Fluder 2001). Latent semantic structure indexing (LaSSI) employs the singular value decomposition (SVD) technique from linear algebra to obtain a reduced number of correlated descriptors. By varying the number of singular values (the choice of k), the user can control the level of fuzziness of a similarity search: larger values of k produce better approximations of the original descriptor space than smaller values.

Similarity measure selection

In general, different similarity measures yield different rankings, except when they are monotonic. Improved results are obtained by using ‘data fusion’ methods to combine the rankings resulting from different coefficients (Gillet, Holliday, Hu, Khatib and Willett 2002). According to Skvortsova (Skvortsova, Baskin, Stankevich, Palyulin and Zefirov 1998), for each study a similarity measure must be fitted, using a training set of objects. Composite measures may need to be optimized, as the Grotch metric and Tversky contrast model.

Empirically, the Dice coefficient has worked better than cosine similarity in retrieving actives (Hull, Fluder, Singh, Nachbar, Kearsley and Sheridan 2001) and is the standard choice for use with the ap and tt descriptors. For any given probe, Dice and Tanimoto give identical ranks – they are monotonic, so Dice, the classical choice of Carhart, is kept.

For the geometric atom pair descriptors, the summation form of the Dice coefficient is used, rather than the binary form.

Although the ap and tt descriptors are also used as original descriptors in LaSSI, the cosine similarity index is employed in this technique because “the coefficients on the new descriptors are floating point numbers and no longer represent frequencies”(Singh, Sheridan, Fluder and Hull 2001). There is no data in the literature for the use of Dice with LaSSI.

Asymmetric similarity measures allow fuzzy super- and sub-structure searching. A superstructure search is defined as: look for structures containing given query. A substructure search is defined as: look for structures embedded with given query (James, Weininger and Delany 2000). In both cases asymmetric *local*

similarity is estimated. A fuzzy superstructure search implementation that uses as a similarity measure the ratio of the number of bonds in the maximum common structure to the number of bonds in the query $c/(a + c)$, is given in (Hagadone 1992).

Query object specification

The user either enters, or copies, a query object at search time, using the graphical user interface. An object may be selected from existing objects in the database, or a previously found object may be specified by referencing a file by pathname or URL.

Query by Example (QBE) ("Query by example - the Viper Project") is a method of query creation that allows the user to search for objects based on an example in the form of selected objects, or a list of names of objects. The QBE parser performs an analysis and formulates a query to submit to the search engine. QBE can be thought of as a "fill-in-the blanks" method of query creation. It is easier to learn than formal query languages, such as the standard Structured Query Language (SQL), while still enabling powerful searches. Results from a QBE may be more variable than those from a formal query entry.

Multiple searches can be carried out using a single query. The results are presented in a single highly informative screen containing hyperlinks. Specifying a single journal article as a query object, for example, one gets similar articles retrieved and ranked according to full-text word- and citation-based measures (citation and word matches), with the option to see just the first 3 highest-ranking ones or all; the citations themselves are also similar objects (subject matter match), and are presented sorted by decreasing similarity; documents on the same web site are given (site match); the home pages of the authors are given (containing author-matched objects); articles which cite the query article are given (subject matter match); context of the citations to the query article, expandable to several sentences for each citing article (subject matter match); objects in the same category, as classified by the web-crawler (category match); comments and corrections by online users (subject matter match); even the articles which have been viewed by the user who read the query article online (online user match) – see a sample web page (Santini and Jain 1999).

Two or more objects may be specified as a *joint* query. Joint queries are represented by descriptor averages – the sum of the frequencies of the descriptors of the members of the joint query divided by the number of members (Singh, Sheridan, Fluder and Hull 2001) (see example for acetone and isobutylene). *Relevance feedback*^{†††} may be used to select the members of the joint query. For instance, let a single query search yield 11 active compounds in the top-ranking 300 compounds; three structurally diverse representative compounds from these 11 can be used to construct a joint query.

^{†††} Positive feedback involves summing descriptors found in “successfully” retrieved objects with those in the original query ("Query by example - the Viper Project"; Hull, Singh, Nachbar, Sheridan, Kearsley and Fluder 2001). Negative feedback involves subtracting descriptors found in “unsuccessfully” retrieved objects from those in the original query. With neutral feedback no action is taken.

Similarity scores

The atom pair (ap) and topological torsion (tt) descriptors and their fuzzy analogues bp, bt, hp, ht, cp, ct and the 3D extensions ag, bg are again selected for illustrative purposes (Carhart, Smith and Venkataraghavan 1985; Nilakantan, Bauman, Dixon and Venkataraghavan 1987; Kearsley, Sallamack, Fluder, Andose, Mosley and Sheridan 1996; Sheridan, Miller, Underwood and Kearsley 1996; Hull, Fluder, Singh, Nachbar, Kearsley and Sheridan 2001; Hull, Singh, Nachbar, Sheridan, Kearsley and Fluder 2001; Sheridan, Singh, Fluder and Kearsley 2001; Singh, Sheridan, Fluder and Hull 2001). The Dice similarity index is used in standard similarity searching implementations – TOPOSIM (T) using topological descriptors and GEOSIM (G) using 3D descriptors; cosine is used in LaSSi (L). The tuning of the degree of fuzziness in L is beyond the scope of this overview, and is taken to be optimal (roughly $k = 300$). To evaluate the performance of the descriptors one needs a database of compounds for which the biological activities are known e.g. MDDR. Queries are selected that are typical of a drug-like molecule and from therapeutic categories that i) contain a sufficient number of actives (e.g. > 50) for reasonable statistics, ii) have several chemical classes present in them, and iii) are fairly specific, so that most of the molecules probably work by the same mechanism.

The connection table of the query object (similarity probe) is processed to obtain the set of atom pairs, and then the database file is scanned to evaluate the similarity between the query and each of the database structures. The maximum number of structures that the program will select is specified, as well as the minimum similarity score that a database compound must show to be selected. Within these limits, the program selects from the database the structures that are most similar (highest similarity value) to the query and creates an output file of compound numbers and similarity values, sorted by decreasing similarity, for the selected compounds.

As an illustration Carhart gives a similarity probe for diazepam, a widely prescribed drug with sedative and anticonvulsant properties. The 100 most similar compounds give similarity values with diazepam ranging from 1.00 (diazepam itself) to 0.660. Of these, 78 are easily recognized as analogues of the probe compound. They range from structures which differ from the query by 1 atom through structures which possess a few different atoms and bonds to more heavily altered analogues. Interspersed with these are 22 other structures that bear resemblance to diazepam but that are not usually thought of as benzodiazepine analogues. An unusual number of these show psychotropic activity or analogous activities in animal models. These results are consistent with the expectation that similar structures will frequently show similar properties (similar property principle).

A similarity probe with nicotine as query was applied to a database using atom pair as well as topological torsion descriptors. Interpretation of the similarity scores of the first 1000 compounds selected by the two methods indicates that i) tt similarity drops off more steeply – a similarity score of 0.65 obtained by the atom pair method is roughly equivalent to a score of 0.50 obtained by the tt method; ii) the first few com-

pounds are the same in both sets, but beyond this there are differences. For example, a derivative of nicotine is not selected by the ap method, because the additional atoms contribute a large number of ap's. Another compound whose cyclic analogue is closely related to nicotine is also not selected by the ap method; iii) the tt's perform better than ap's at low levels of screening.

Similarity searches using the ap and tt descriptors are typically able to discover active compounds which are in different chemical classes than the probe (Kearsley, Sallamack, Fluder, Andose, Mosley and Sheridan 1996).

Once all similarity scores are calculated, they are sorted from high to low score. Ranks are then assigned: the molecule with the highest score is rank 1, the next highest rank 2, etc. When many types of descriptors are used in an investigation, only the ranks of the compounds are used, because the distribution of absolute scores varies from one descriptor to another. More than one descriptor can be used as a basis for ranking of compounds. A *combination score* is defined as the mean of the similarities for two single descriptors (e.g. ap_{tt} similarity = 0.5 ap similarity + 0.5 tt similarity). A *minimum rank score* e.g. mr(ap,tt) for each compound in a database is defined as its rank in the ap list or its rank in the tt list, whichever is lower. The compounds are then sorted by the new score, and new ranks are assigned, the smallest score being rank 1, the next rank 2, etc. This is analogous to taking the union of the top-scoring compounds from the separate ap and tt lists.

A *measure of effectiveness* – initial enhancement (IE) – of a similarity search is defined as the ratio of the number of actives for a particular therapeutic category retrieved in the top-scoring 300 compounds (actives@300) to the number of actives that are expected by pure chance $IE = \frac{\text{actives@300}}{\text{nactives} \times 300/N}$, where nactives

is the total number of actives belonging to the corresponding therapeutic category and N is the number of compounds in the database. Since IE is based on the similar property principle, it is a measure of how effective a similarity search is for activity prediction. For cases where the number of actives is small, so that moving compounds across the arbitrary boundary of 300 makes a large difference in IE, a robust initial enhancement (RIE) has recently been proposed that decreases the weight of an active with increasing of its rank (Sheridan, Singh, Fluder and Kearsley 2001).

Measures of diversity of the active compounds retrieved in a similarity search are i) mean similarity to the centroid, and ii) number of structural classes. For each therapeutic category, the retrieved actives are used to construct the centroid as their descriptor average (see example for two molecules acetone and isobutylene). The mean similarity to this virtual compound (using the least fuzzy descriptor tt and Dice similarity coefficient) is a measure of the diversity of the set of retrieved actives. Cluster analysis is used to enumerate the number of structural classes within the set of retrieved actives, clustering compounds together when similarity between any two of them is greater than or equal to 0.65.

Correlation of ranks between methods (or descriptors) is a measure of whether two methods (or descriptors) select different actives for a given probe. A simple scatterplot for the ranks of the compounds obtained via the two methods that shows little or no correlation would be indicative that the two methods rank the compounds very differently, i.e. they select different actives as the top-scoring compounds.

Ten activities (narcotic, antihistamine, tranquilizer, dopaminergic, ace-inhibitor, estrogen, sympathomimetic, parasympatho-mimetic, serotonergic, gabaminergic) are studied using the ap, bp, hp, cp, tt, bt, ht, ct descriptors. The effectiveness of the similarity searches (mean initial enhancements) are ap 26.6, bp 25.0, hp 18.0, cp 23.7, tt 28.3, bt 22.3, ht 19.1, ct 22.8. All descriptors give IE's $\gg 1$, indicating general usefulness. The original ap and tt descriptors do better on the average than their fuzzy analogues. Hydrophobic descriptors are consistently the worst performers. Average values can be misleading, however, and detailed analysis (data not shown) reveals that for different activities different descriptors may perform optimally.

Of the 28 possible pairs of combination and of minimum rank scores five are studied as representative – four combine the scores of a specific descriptor ap or tt with that of a corresponding fuzzy descriptor bp or cp, and the fifth combines the scores of the original ap and tt. The mean IE's for the combination (minimum rank) scores are apbp 29.4 (28.6), apcp 27.9 (28.1), ttbt 32.6 (28.0), ttct 30.7 (28.1), aptt 29.8 (28.9). Both scores do better on the average than the component scores, e.g. ttbt 32.6 is much better than tt 28.3 or bt 22.3. Combination scoring gives somewhat higher IE's than minimum rank scoring; hence it is preferred for future use. Of the five combinations, ttbt is the clear winner.

Correlation of ranks between pairs of descriptors varies from 0.10 to 0.98 with no apparent pattern. Any two descriptors may rank the same set of actives very differently – the ranks of actives are very sensitive to atom type definition. For practical similarity searches, where one takes a relatively small number of top-scoring compounds, different descriptors will seem to select different subsets of actives. Even though some are with lesser enhancements, all eight descriptors are kept for future use for generating as diverse a set of actives as possible.

Eight therapeutic categories are studied using 2D (ap, bp) and 3D (ag, bg) descriptors and their combination scores. The mean IE's obtained are ap 17.8, bp 20.4, apbp 22.5, ag 18.5, bg 19.7, agbg 21.5. The combination scores do better in all cases, hence they are preferred. agbg has an IE lower than apbp – 3D descriptor similarity searches G (using ag and bg) do *not* find more actives than topological searches T (using ap and bp).

Correlation of ranks between 2D and 3D descriptors (apbp and agbg) indicates that generally the ranks cluster along the diagonal (coefficient 0.99), i.e. the overall rankings by T and G are similar. However, there are compounds that fall far from the diagonal – compounds where similar chemical groups are held

via a different topological connectivity. To avoid missing such compounds it is useful to keep the 3D option for similarity searching.

The standard implementation T and variable fuzziness implementation L are evaluated using the same original descriptors ap and tt. Mean initial enhancements are ap T 33, ap L 42, tt T 37, tt L 40, aptt T 39, aptt L 45, hence the combination score is preferred. Initial enhancement using aptt score for single probes varies from 7 to 83, mean 45 for L, from 6 to 109, mean 39 for T. Hence L is as good as, or better than T for selecting active compounds from a large database of drug-like molecules (Hull, Fluder, Singh, Nachbar, Kearsley and Sheridan 2001).

An example of a scatterplot between the ranks of actives retrieved by L and T (figure not shown) indicates that there is little to no correlation for any of the probes – the actives are scattered and do not fall near the diagonal. This lack of correlation of ranks between L and T indicates that the two implementations select different sets of actives (Hull, Singh, Nachbar, Sheridan, Kearsley and Fluder 2001).

IE using aptt for *joint* probes varies from 19 to 113, mean 71 for L, from 37 to 113, mean 69 for T. Hence the use of joint probes significantly enhances the rate of retrieval of active compounds compared with the single molecule probe.

The average for five therapeutic classes of the mean similarity of the retrieved actives to the centroid is 0.51 for L, ranging from 0.10 to 0.70; for T it is 0.52, ranging from 0.23 to 0.74. L retrieves 6.6 structural classes on average, T – 5.5. Thus the diversity of the retrieved actives is somewhat greater for L than for T. This is not surprising since L retrieves diverse chemical structures through fuzzier descriptors whereas T retrieves only compounds that share descriptors with the probe. L (fuzzy descriptors) could be used initially to identify most diverse structural classes. Subsequent similarity searches with less fuzzy descriptors (T) could retrieve in depth the actives corresponding to these classes.

Application areas

Similarity searching allows ranking and retrieval of objects in databases according to their similarity to a query. Similarity searching is *fuzzy* (“fuzzy-match searching” (Fisanick, Lipkus and Rusinko 1994), “fuzzy version of exact searching” (Willett, Barnard and Downs 1998)) in that it retrieves not only the query object (as the most similar one), but also (other) similar^{##} objects. For example, *pattern matching* finds patterns that may be missed because the exact patterns specified may be slightly off; *signature analysis* allows identification of objects that have signatures which are never identical and may be missed by an exact search. Similarity searching is *fault-tolerant* in that it reduces the effect of query specification errors. A similarity

^{##} It is not clear whether all instances of fuzziness (Wang 1996; FT&T 2003) can be reduced to using similarity, see e.g. (Bilgic and Turksen 1999). Fuzzification can be achieved using measurement theory (see e.g. fuzzy binning), fuzzy descriptors (see above), fuzzy similarity measures (e.g. fuzzy Hamming distance (“Fuzzy Hamming distance")), etc.

search 'reduces' to an exact search if the set of characteristics used are identical to those used for an exact search.

Similarity searching is *content-based* because it uses characteristics of the objects as search criteria rather than descriptions about the objects such as keywords (meta-data). Content-based ranking and retrieval is independent of nomenclature, location, etc.

Basing on the similar property principle, objects with similar properties may be searched for – *property prediction*. Given an object with desired properties, the *direct* approach is to look for (e.g. structurally) similar objects, and expect that among the high-ranking retrieved objects some will exhibit the desired properties or even improve on them. In the *reverse* approach the properties of the query are 'predicted': if significant similarity is found between a novel query object and objects in a database these similarities may provide information about the properties (e.g. structure and function) of the new object. One may suspect a relationship between two or more objects that one would like to explore. Similarity searches may uncover previously unnoticed relationships between the objects. Objects may be globally similar, or have similar regions. Two (or more) objects can be compared to identify regions of similarity. Significant similarity may suggest an 'evolutionary' relationship between the similar objects.

In chemical similarity searching, *browsing* of ranked similar objects may be used for evaluation of the uniqueness of proposed or newly synthesized compounds, finding starting materials or intermediates in *synthesis design*, handling of chemical *reactions* and *mixtures* – finding the right chemicals for one's needs, even if one does not know exactly what he is looking for ☺.

'Direct' property prediction is a standard technique in drug discovery. Given a compound with an interesting biological activity, compounds that are similar to it in structure are likely to have a similar activity. In practice, an investigator provides a chemical structure as a probe, searches over a database of sample-available compounds, and finds those that are most similar, which are then submitted for testing (Hull, Fluder, Singh, Nachbar, Kearsley and Sheridan 2001).

Automated, miniaturized, and parallelized synthesis and testing (combinatorial chemistry/high-throughput screening) are accelerating the development of a complex of methods for *data mining* (Gillet, Willett and Bradshaw 1998; Chen, Rusinko, Tropsha and Young 1999; *Electronic Statistics Textbook*) and computer screening (*virtual screening*) ("Virtual screening") of object libraries. Classes of objects are recognized (*cluster analysis* (Downs and Willett 1996; *Electronic Statistics Textbook*)) basing on estimation of distances in descriptor space (dissimilarities). In object selection, as diverse classes as possible are chosen so that all the different types of properties (e.g. bioactivities) within a larger collection are sampled using as few objects as possible (*diversity analysis*) (Gillet, Willett, Bradshaw and Green 1999). Key chemical features and the spatial relationships among them that are considered to be responsible for a desired biological activity may be identified (*pharmacophore recognition*) using local similarity e.g. via common substructures in sets of active

molecules (Sheridan and Miller 1998); *pharmacophore searching* in 3D databases may be carried out using a pharmacophore as the query (Chen, Rusinko, Tropsha and Young 1999). Shape similarity of ligands to a receptor site (*ligand docking*) may be used for finding structures that fit into proteins. *Molecular superposition* may use similarity optimizer techniques.

'Reverse' property prediction is used with chromatography application databases that contain separations, including method details and assigned chemical structures for each chromatogram. Retrieving compounds present in the database that are similar to the query allows the retrieval of suitable separation conditions for use with the query (*method selection*). (This is analogous to a doctor doing a similarity search on digital x-ray images: by retrieving similar x-rays present in the database, he can find other patients whose condition is similar to his patient, and thus learn from their treatment and experience).

Similarity searching may be used as a step in *structure elucidation*, e.g. i) similarity search the available (NMR) spectra database using the experimental spectrum as a query (*pattern matching*); ii) do an exact search for the (sub)structures corresponding to the high ranking spectra retrieved in i); iii) use these (sub)structures and additional information (e.g. IR, MS spectra) to determine common fragments which have a high probability of being contained in the structure of the unknown compound; iv) validate the results by using spectrum prediction software for the identified fragments (Williams 2000). In stage i), not only the similarity ranking is important – usually the most similar spectrum is the target one, but also the similarity values are important - a value below a certain threshold indicates a probably invalid solution.

Signature analysis also uses elements of reverse 'property' prediction. For synthetic drugs, practical experience has shown that the impurity profiles of the products from a given illicit laboratory are characteristic. Provided that there is no change in the method or the conditions of drug synthesis, variations in the impurity content of drugs synthesized at different times by the same chemist in a clandestine laboratory are believed to be relatively small. Consequently, based on their chemical characteristics, samples of seized drugs can be classified into groups identified by their chemical impurity profiles, and a given sample or group of samples may be associated with an individual chemist or laboratory operating clandestinely. It is thus possible to link together illicit drug consignments from the same source and to build up a database of related drug samples over a period of time.

Similarly, starting materials used in illicit drug manufacture may also contain certain characteristic impurities. The impurity content and the type of impurities may vary depending on the nature of the starting material, on whether a precursor chemical was diverted from legitimate sources or was itself manufactured clandestinely. The identification of characteristic impurities (or impurity patterns) in precursors may therefore help to link them to a commercial or clandestine source ("Signature analysis").

Conclusion

Similarity (fuzzy) searching is a powerful alternative to exact searching that is being added to all contemporary database implementations.

References

- Amat, L., E. Besalu, R. Carbo-Dorca and R. Ponec (2001). "Identification of active molecular sites using quantum self-similarity measures." *J. Chem. Inf. Comput. Sci.* **41**(4): 978-991.
- Batagelj, V. and M. Bren (1995). "Comparing Resemblance Measures." *Journal of Classification* **12**(1): 73-90, <http://vlado.fmf.uni-lj.si/pub/Cluster/compare.pdf>.
- Bilgic, T. and I. B. Turksen (1999). Measurement of Membership Functions: Theoretical and Empirical Work. *Handbook of Fuzzy Sets and Systems*. D. Dubois and H. Prade, Kluwer. **1**: 195-232, www.ie.boun.edu.tr/~taner/publications/papers/membership.pdf.
- Bradshaw, J. (1997). "Introduction to Tversky similarity measure." www.daylight.com/meetings/mug97/Bradshaw/MUG97/tv_tversky.html.
- Bradshaw, J. (2001). "YAMS - Yet Another Measure of Similarity." *Euromug01, Cambridge UK*, www.daylight.com/meetings/emug01/Bradshaw/Similarity/YAMS.html.
- Carbo, R., M. Arnau and L. Leyda (1980). "How similar is a molecule to another? An electron density measure of similarity between two molecular structures." *Int. J. Quant. Chem.* **17**: 1185-1189.
- Carbo, R. and L. Domingo (1987). "LCAO-MO similarity measures and taxonomy." *Int. J. Quant. Chem.* **32**: 517-545.
- Carbo-Dorca, R. and E. Besalu (1998). "A general survey of molecular quantum similarity." *J. Mol. Struct. - Theochem* **451**(1-2): 11-23.
- Carhart, R. E., D. H. Smith and R. Venkataraghavan (1985). "Atom pairs as molecular features in structure-activity studies: definition and applications." *J. Chem. Inf. Comput. Sci.* **25**: 64-73.
- Chemoffice Ultra, 2003, www.cambridgesoft.com.
- Chen, X., A. Rusinko, A. Tropsha and S. S. Young (1999). "Automated pharmacophore identification for large chemical data sets." *J. Chem. Inf. Comput. Sci.* **39**(5): 887-896.
- Cioslowski, J. (1991). "Quantifying the Hammond Postulate - Intramolecular Proton- Transfer in Substituted Hydrogen Catecholate Anions." *J. Am. Chem. Soc.* **113**(18): 6756-6760.
- Cioslowski, J. (1998). Electronic wavefunctions analysis. *Encyclopedia of Computational Chemistry*. P. R. Schleyer. New York, John Wiley: 892-905.
- Cioslowski, J. and M. Challacombe (1991). "Maximum similarity orbitals for analysis of the electronic excited states." *Int. J. Quant. Chem.: Quant. Chem. Symp.* **25**: 81-93.
- Cioslowski, J. and E. D. Fleischmann (1991). "Assessing Molecular Similarity from Results of Abinitio Electronic-Structure Calculations." *J. Am. Chem. Soc.* **113**(1): 64-67.
- Cioslowski, J. and A. Nanayakkara (1993). "Similarity of Atoms in Molecules." *J. Am. Chem. Soc.* **115**(24): 11213-11215.
- Cioslowski, J. and P. R. Surjan (1992). "An Observable-Based Interpretation of Electronic Wave-Functions - Application to Hypervalent Molecules." *Theochem-Journal of Molecular Structure* **87**: 9-33.
- Committee, N. R. C. (1995). The N- and V-Representability Problems. *Mathematical Challenges from Theoretical/Computational Chemistry*, National Academy Press, www.nap.edu/readingroom/books/mctcc/chap4.3.html.
- Cooper, D. L. and N. L. Allan (1989). "A novel-approach to molecular similarity." *J. Comp.-Aid. Mol. Des.* **3**(3): 253-259.
- Crippen, G. M. (1999). "VRI: 3D QSAR at variable resolution." *J. Comp. Chem.* **20**(14): 1577-1585.
- Delaney, M. F., J. R. Hallowell and F. V. Warren (1985). "Optimization of a similarity metric for library searching of highly compressed vapor-phase infrared spectra." *J. Chem. Inf. Comput. Sci.* **25**: 27-30.

- Dixon, S. L. and R. T. Koehler (1999). "The hidden component of size in two-dimensional fragment descriptors: Side effects on sampling in bioactive libraries." *J. Med. Chem.* **42**(15): 2887-2900.
- Downs, G. M. and P. Willett (1996). Similarity searching in databases of chemical structures. *Reviews in Computational Chemistry*. K. B. Lipkowitz and D. B. Boyd. New York, VCH Publishers, Inc. **7**: 1-66.
- Electronic Statistics Textbook* (2004). StatSoft, Tulsa, OK, www.statsoft.com/textbook/stathome.html.
- Fisanick, W., A. H. Lipkus and A. Rusinko (1994). "Similarity Searching on Cas Registry Substances .2. 2d Structural Similarity." *J. Chem. Inf. Comput. Sci.* **34**(1): 130-140.
- Fratev, F., O. E. Polansky, A. Mehlhorn and V. Monev (1979). "Application of distance and similarity measures - comparison of molecular electronic structures in arbitrary electronic states." *J. Mol. Struct.* **56**(2): 245-253.
- "Fraud investigator application." www.infoglide.com.
- FT&T (2003). *9th International Conference on Fuzzy Theory and Technology*, www.ee.duke.edu/JCIS/FTT.html.
- "Fuzzy Hamming distance." <http://link.springer.de/link/service/series/0558/bibs/2089/20890086.htm>.
- Garfield, E. (2002). "Reaction Similarity and Retrieval." www.isinet.com/essays/chemicalliterature/15.html.
- Gasteiger, J., Ed. (2003). *Handbook of Chemoinformatics: From Data to Knowledge*. Weinheim, Wiley-VCH.
- Gillet, V. J., J. D. Holliday, C. Y. Hu, I. Khatib and P. Willett (2002). "Combining different types of information in similarity searching and library design." www.ibcp.fr/GGMM/Nimes/C3.html.
- Gillet, V. J., P. Willett and J. Bradshaw (1998). "Identification of biological activity profiles using substructural analysis and genetic algorithms." *J. Chem. Inf. Comput. Sci.* **38**(2): 165-179.
- Gillet, V. J., P. Willett, J. Bradshaw and D. V. S. Green (1999). "Selecting combinatorial libraries to optimize diversity and physical properties." *J. Chem. Inf. Comput. Sci.* **39**(1): 169-177.
- Good, A. C. and W. G. Richards (1998). "Explicit calculation of 3D molecular similarity." *Perspectives in Drug Discovery and Design* **9-11**: 321-338.
- Gower, J. C. (1985). Measures of similarity, dissimilarity, and distance. *Encyclopedia of statistical sciences*. S. Kotz. New York, Wiley: 397-405.
- Grethe, G. and T. E. Moock (1990). "Similarity Searching in Reacs - a New Tool for the Synthetic Chemist." *J. Chem. Inf. Comput. Sci.* **30**(4): 511-520.
- Hagadone, T. R. (1992). "Molecular substructure similarity searching - efficient retrieval in 2-dimensional structure databases." *J. Chem. Inf. Comput. Sci.* **32**(5): 515-521.
- "Hausdorff distance." www-cgrl.cs.mcgill.ca/~godfried/teaching/cg-projects/98/normand/main.html.
- Holliday, J. D., C. Y. Hu and P. Willett (2002). "Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit-strings." *Comb. Chem. High Through. Scr.* **5**(2): 155-166, www.daylight.com/meetings/emug01/Bradshaw/Similarity/Hu.htm.
- Horvath, D. and C. Jeandenans (2000). "Molecular similarity and virtual screening. In silico methods to retrieve active analogs in the context of discovering therapeutic compounds." *Actual Chim.*(9): 64-67.
- Hull, R. D., E. M. Fluder, S. B. Singh, R. B. Nachbar, S. K. Kearsley and R. P. Sheridan (2001). "Chemical similarity searches using latent semantic structural indexing (LaSSI) and comparison to TO-POSIM." *J. Med. Chem.* **44**(8): 1185-1191.
- Hull, R. D., S. B. Singh, R. B. Nachbar, R. P. Sheridan, S. K. Kearsley and E. M. Fluder (2001). "Latent semantic structure indexing (LaSSI) for defining chemical similarity." *J. Med. Chem.* **44**(8): 1177-1184.
- ISI "Science Citation Index Expanded." www.isinet.com/isi/products/citation/scie/index.html.
- James, C. A., D. Weininger and J. Delany (2000). Fingerprints - Screening and Similarity. *Daylight Theory Manual*. Irvine, CA and Santa Fe, NM, Daylight Chemical Information Systems, Inc., www.daylight.com/dayhtml/doc/theory/theory.finger.html.
- Johnson, M. A. and G. M. Maggiora (1990). *Concepts and Applications of Molecular Similarity*. New York, Wiley.
- Kearsley, S. K., S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley and R. P. Sheridan (1996). "Chemical similarity using physicochemical property descriptors." *J. Chem. Inf. Comput. Sci.* **36**(1): 118-127.
- Klein, D. J. and D. Babic (1997). "Partial orderings in chemistry." *J. Chem. Inf. Comput. Sci.* **37**(4): 656-671.

- Kochev, N., V. Monev and I. Bangov (2003). Searching chemical structures. *Chemoinformatics: A textbook*. J. Gasteiger and T. Engel. Weinheim, Wiley-VCH: 291-318.
- "Levenshtein distance." www.merriampark.com/ld.htm.
- "Logical operations." www.ralphb.net/IPSubnet/logical.html.
- "Mahalanobis Metric." www.galactic.com/Algorithms/discrim_mahaldist.htm.
- MDDR (2002). Molecular Design Drug Data Report. San Leandro, CA, Molecular Design Ltd., www.mdli.com/products/knowledge.html.
- Mestres, J., D. C. Rohrer and G. M. Maggiora (1997). "MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches." *J. Comp. Chem.* **18**(7): 934-954.
- Mezey, P. G. (1999). "The holographic electron density theorem and quantum similarity measures." *Mol. Phys.* **96**(2): 169-178.
- Nilakantan, R., N. Bauman, J. S. Dixon and R. Venkataraghavan (1987). "Topological torsion - a new molecular descriptor for SAR applications. Comparison with other descriptors." *J. Chem. Inf. Comput. Sci.* **27**(2): 82-85.
- Petke, J. D. (1993). "Cumulative and discrete similarity analysis of electrostatic potentials and fields." *J. Comp. Chem.* **14**(8): 928-933.
- Ponec, R. and M. Strnad (1990). "A novel-approach to the characterization of molecular similarity - the 2nd-order similarity index." *Coll. Czech. Chem. Comm.* **55**(4): 896-902.
- "Query by example - the Viper Project." <http://viper.unige.ch/research/index.html>.
- "Queryplus." www.queryplus.com.
- Rarey, M. and J. S. Dixon (1998). "Feature trees: A new molecular similarity measure based on tree matching." *J. Comp.-Aid. Mol. Des.* **12**(5): 471-490.
- Rarey, M. and M. Stahl (2001). "Similarity searching in large combinatorial chemistry spaces." *J. Comp.-Aid. Mol. Des.* **15**(6): 497-520.
- Rhodes, N., P. Willett, J. B. Dunbar and C. Humblet (2000). "Bit-string methods for selective compound acquisition." *J. Chem. Inf. Comput. Sci.* **40**(2): 210-214.
- Richards, W. G. and E. E. Hodgkin (1988). "Molecular similarity." *Chem. Brit.* **24**(11): 1141.
- Robinson, D. D., P. D. Lyne and W. G. Richards (1999). "Alignment of 3D-structures by the method of 2D-projections." *J. Chem. Inf. Comput. Sci.* **39**(3): 594-600.
- Robinson, D. D., P. D. Lyne and W. G. Richards (2000). "Partial molecular alignment via local structure analysis." *J. Chem. Inf. Comput. Sci.* **40**(2): 503-512.
- Santini, S. and R. Jain (1999). "Similarity Measures." *IEEE Trans. Pattern Anal. Mach. Intell.* **21**(9): 871-883, <http://citeseer.nj.nec.com/santini99similarity.html>.
- Sheridan, R. P. (2000). "The centroid approximation for mixtures: Calculating similarity and deriving structure-activity relationships." *J. Chem. Inf. Comput. Sci.* **40**(6): 1456-1469.
- Sheridan, R. P. (2002). "The most common chemical replacements in drug-like compounds." *J. Chem. Inf. Comput. Sci.* **42**(1): 103-108.
- Sheridan, R. P. and M. D. Miller (1998). "A method for visualizing recurrent topological substructures in sets of active molecules." *J. Chem. Inf. Comput. Sci.* **38**(5): 915-924.
- Sheridan, R. P., M. D. Miller, D. J. Underwood and S. K. Kearsley (1996). "Chemical similarity using geometric atom pair descriptors." *J. Chem. Inf. Comput. Sci.* **36**(1): 128-136.
- Sheridan, R. P., S. B. Singh, E. M. Fluder and S. K. Kearsley (2001). "Protocols for bridging the peptide to nonpeptide gap in topological similarity searches." *J. Chem. Inf. Comput. Sci.* **41**(5): 1395-1406.
- "Signature analysis." www.undcp.org/bulletin/bulletin_1999-01-01_1_page008.html.
- Singh, S. B., R. P. Sheridan, E. M. Fluder and R. D. Hull (2001). "Mining the chemical quarry with joint chemical probes: An application of latent semantic structure indexing (LaSSI) and TOPOSIM (dice) to chemical database mining." *J. Med. Chem.* **44**(10): 1564-1575.
- Skvortsova, M. I., Baskin, II, I. V. Stankevich, V. A. Palyulin and N. S. Zefirov (1998). "Molecular similarity. I. Analytical description of the set of graph similarity measures." *J. Chem. Inf. Comput. Sci.* **38**(5): 785-790.
- Sneath, P. H. A. (1966). "Relations between chemical structure and biological activity in peptides." *J. Theoret. Biol.* **12**: 157-195.
- Sneath, P. H. A. and R. R. Sokal (1973). *Numerical taxonomy*. San Francisco, W. H. Freeman and Co.
- SPSS® Reference manual, 2001, www.spss.com.

- Szabo, A. and N. Ostlund (1989). *Modern quantum chemistry*. New York, McGraw-Hill.
- Trinajstić, N., D. J. Klein and M. Randić (1986). "On some solved and unsolved problems in chemical graph theory." *Int. J. Quant. Chem.: Quant. Biol. Symp.* **20**: 699-742.
- "Virtual screening." http://discuss.foresight.org/critmail/sci_nano/5011.html.
- Voigt, J. H., B. Bienfait, S. M. Wang and M. C. Nicklaus (2001). "Comparison of the NCI open database with seven large chemical structural databases." *J. Chem. Inf. Comput. Sci.* **41**(3): 702-712.
- Wang, P. (1996). "The interpretation of fuzziness." *IEEE Trans. Syst. Man Cybern. Part B-Cybern.* **26**(2): 321-326, www.cogsci.indiana.edu/pub/wang.fuzziness.ps.
- Willett, P. (1987). *Similarity and clustering in chemical information systems*. New York, Research Studies Press Ltd, John Wiley & Sons.
- Willett, P. (1998). Structural similarity measures for database searching. *Encyclopedia of Computational Chemistry*. P. R. Schleyer, N. L. Allinger, T. Clark et al. Chichester, John Wiley: 2748-2756.
- Willett, P., J. M. Barnard and G. M. Downs (1998). "Chemical similarity searching." *J. Chem. Inf. Comput. Sci.* **38**(6): 983-996.
- Williams, A. (2000). "Recent advances in NMR prediction and automated structure elucidation software." *Current Opinion in Drug Discovery & Development* **3**: 298-305.
- World Drug Index, 2001, www.derwent.co.uk.