Approaches to Measure Chemical Similarity – a Review

When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely advanced to the stage of science.

William Thomson, Lord Kelvin

Nina Nikolova and Joanna Jaworska*

Procter and Gamble, Eurocor, Central Product Safety, 100 Temselaan, B-1853 Strombeek-Bever, Belgium, Fax 322 5683098, Tel 322 456 2076, Tel 322 456 801, E-mail: jaworska.j@pg.com

Review Paper

Although the concept of similarity is a convenient for humans, a formal definition of similarity between chemical compounds is needed to enable automatic decision-making. The objective of similarity measures in toxicology and drug design is to allow assessment of chemical activities. The ideal similarity measure should be relevant to the activity of interest. The relevance could be established by exploiting the knowledge about fundamental chemical and biological processes responsible for the activity. Unfortunately, this knowledge is rarely available and therefore

* Corresponding author

Key words: similarity, neighborhood behavior, QSAR, applicability domain

Abbreviations: 2D – two dimensional, **3D** – three dimensional, **Ab initio calculations** – quantum chemical calculations using exact equations with no approximations which involve the whole electronic population of the molecule, **AiM** – The Theory of Atoms in Molecules, **AM1 calculations** – Austin Model 1 semi-empirical molecular orbital calculations, **BCP** – Bond Critical Point, **BCUT** – Burden Chemical Abstracts, **CoMMA** – Comparative Molecular Moment Analysis – a procedure that utilizes information from moment expansions of molecular mass and charge up through and inclusive of second order to perform molecular comparison, **CoMFA** – Comparative Molecular Field Analysis – 3D-QSAR method that uses statistical correlation techniques for the analysis of the quantitative relationship between the biological activity of a set of compounds with a specified alignment, and their three dimensional electronic and steric properties, **HOMO energy** – The Highest Occupied Molecular Orbital (HOMO) energy is obtained by molecular orbital calculations and relates to the ionization potential of a molecule and its reactivity as a nucleophile, **IUPAC** – The International Union of Pure and Applied Chemistry, **LUMO energy** – The Lowest Unoccupied Molecular Orbital (LUMO) energy is obtained from molecular orbital calculations and represents the electron affinity of a molecule or its reactivity as an electrophile, **MED-LA** – Molecular Electron Density Lego Assembler – utilizes the fact that electron densities decrease exponentially with distance from the nearest atomic nucleus and generates electron densities by an additive superposition of fuzzy density fragments, **MO calculations** – Molecular Orbital (MO) calculations are quantum chemical calculations based on the Schrödinger equation, which can be subdivided into semi-empirical and ab initio methods (see Ab initio calculations), **MOPAC** – Molecular Orbital Package, **MQSM** – Molecular Quantum Similarity Measure, **MSA** – Molecular Shape Analysis, **QSAR** – Quantitative Structure- Activity Relations, **SAR** – Structure-Activity Relations, **STERIMOL parameters** – defined by Verloop – set of substitutents for length and width, **TAE** – Transferable Atom Equivalent, **TAE/RECON** – An algorithm for rapid reconstruction of molecular charge densities and molecular electronic properties based on Atoms in Molecules theory, **WHIM descriptors** – Weighted Holistic Invariant Molecular descriptors

 DOI: 10.1002/qsar.200330831

QSAR & Combinatorial Science

different approximations have been developed based on similarity between structures or descriptor values. Various methods are reviewed, ranging from two-dimensional, three-dimensional and field approaches to recent methods based on "Atoms in Molecules" theory. All these methods attempt to describe chemical compounds by a set of numerical values and define some means for comparison between them. The review provides analysis of potential pitfalls of this methodology – loss of information in the representations of molecular structures – the relevance of a particular representation and chosen similarity measure to the activity. A brief review of known methods for descriptor selection is also provided. The popular "neighborhood behavior" principle is criticized, since proximity with respect to descriptors does not necessarily mean proximity with respect to activity. Structural similarity should also be used with care, as it does not always imply similar activity, as shown by examples. We remind that similarity measures and classification techniques based on distances rely on certain data distribution assumptions. If these assumptions are not satisfied for a given dataset, the results could be misleading. A discussion on similarity in descriptor space in the context of applicability domain assessment of QSAR models is also provided. Finally, it is shown that descriptor based similarity analysis is prone to errors if the relationship between the activity and the descriptors has not been previously established. A justification for the usage of a particular similarity measure should be provided for every specific activity by expert knowledge or derived by data modeling techniques.

1 Introduction

Quine and other philosophers of science argue that exploiting the similarity concept is a sign of immature science [1]. The notion of similarity is used mainly in early stages of the development of a particular science, and it may be quantified and explained accurately later as the theory of this science develops. For example, the periodic table was originally founded on similarity between elements and these "similarities" were later explained based on electrons and nucleus. Some philosophers believe that it is ill defined to say "A is similar to B" and it is only meaningful to say "A is similar to B with respect to C" [2]. This has important implications for toxicology – a chemical A cannot be similar to a chemical B in absolute terms but only with respect to some measurable key feature.

Regardless of its controversial status in philosophy, similarity is a widely used concept in toxicology. The objective of a similarity measure in toxicology and drug design is to allow assessment if chemicals have similar or dissimilar biochemical activity. When similarity is measured with respect to some feature, this feature has to be relevant to the activity of interest. The main applications are selection of compounds with similar activity to a given compound (similarity analysis), derivation of Structure-Activity Relations (SARs), and justification of read – across application. Similarity analysis is also used in a reverse way to select the most diverse subset (diversity selection) from a given set of compounds.

Similarity between chemical compounds is perceived often intuitively based on expert judgment. A chemist would describe "similar" compounds in terms of "approximately similar backbone and almost the same functional groups". A synthetic chemist may regard two molecules as similar when their topological descriptions of atoms and connecting bonds contain a sufficiently large number of common features [3] However, the use of computers when dealing with similarity-related problems requires unambiguous similarity criteria. Since one of the basic beliefs of chemistry is that similarity in structure implies similarity in activities or properties, the usual approach is to assess similarity by examining resemblance between molecular structures. Hence, the identification of the most informative representation of molecular structures is of great importance in similarity studies.

Similarity assessment, based on structure analogy, is very popular, but should be justified for every specific activity. There are numerous references, concerning the so called "similarity paradox" [4], where a small change in the chemical structure leads to a drastic change in the biochemical activity. The same is true for any other similarity assessment. For example, shape similarity is considered important because of hypothesized "lock and key" interaction with receptors. However, this may not always be the case. As an illustration, there are a number of such paradoxes in structure – odor relationship studies. Despite decades of investigations and many hypotheses (steric theory of odor, diffusion pore theory, infrared resonance, odotope theory), no one is yet able to give a reliable prediction of odor character, when faced with a novel molecule [5, 6]. Recent publications in the area started to explain odor character through vibrational spectra of molecules and electron tunneling, rather than shape and structure resemblance. When dealing with the similarity concept in chemistry, one should have in mind that the presence of reliable method for prediction verification is not by itself a justification for correctness. This is very well illustrated by the difficulties of structure – odor relationship studies to predict human perception of smells. In this case we have some natural ability to verify the validity of our models, which incites us to question their correctness.

The molecular structure is determined by three elements: constitution, configuration, and conformation [7]. Constitution means a certain manner and sequence of bonding of atoms and is expressed by topological descriptors, presence and absence of fragments, or other descriptors which

account for the two dimensional (2D) features of a molecule. Configuration is defined by a 3D (spatial) arrangement of atoms, which is in turn characterized by the valence angles of all atoms that are directly linked to at least two other atoms. Configuration is expressed by shape descriptors and any other approaches, accounting for the 3D arrangement of atoms. Finally, the conformations of a given molecule represent various thermodynamically stable spatial arrangements of its atoms. A vast number of methods of quantitative molecular structure description (topology, shape, physicochemical properties, quantum chemical descriptions, etc.) and comparison (similarity coefficients, etc.) have been proposed and applied to date.

Chemical compounds activity has been traditionally modeled using a variety of topological, physicochemical and electronic descriptors [8, 9]. This has provided the grounds for evaluating similarity between compounds by comparing numerical values of these descriptors. However, the most informative description of a molecule is its quantum mechanical wave function. In principle, it contains all the information about a given chemical. The structure diagram, 3D coordinates and some numerical descriptors can be regarded as direct manifestations of the underlying wave equations that describe the molecule. Recent research makes similarity assessment by electron density analysis much more attractive, because of the new developments in the algorithms and the increase in computer power.

This review summarizes structure-, descriptor- and field based approaches to similarity estimation. The intention is to highlight potential pitfalls of using similarity measures to decide on similarity between activities of chemical compounds, if the knowledge about the fundamental chemical and biological processes responsible for these activities is missing or ignored.

2 Similarity According to Constitution

Constitutional (topological) similarity assessment is based on representation of a chemical compound as a molecular graph. Numerous approaches of extracting information from molecular graphs have been proposed (*e.g.* sub-graph detection and comparison, calculation of graph invariants). All these approaches are limited to extraction and processing of 2D topological information only. Some attempts to involve 3D information into topological indices or fingerprints and overcome the insufficiency of the 2D topological information alone have been made more recently.

2.1 Classic Topological Descriptors

The simplest descriptors are counts of individual atoms, bonds, rings, pharmacophore points, degree of connectivity, etc. Two-dimensional fragment descriptors (*e.g.* atom-centered, bond-centered, ring-centered fragments) were studied in detail [10]. This led to the widespread application of augmented atom, atom sequence and ring fragments in systems and evaluating similarity by substructure searching techniques.

Besides the fragment approach, many graph theory based descriptors have been developed. Platt was the first to study paths in molecules as potential molecular descriptors for structure properties [11]. Some other methods have followed – the Hosoya's Z topological index [12], the Wiener number [13], the molecular connectivity indices as calculated by Randic and co-workers [14], the frequency of path lengths of varying size, the information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstic [15], the bonding connectivity indexes defined by Basak and co-workers [16] and Balaban's *J* indices [17]. Many different topological indices have been described in the QSAR literature, but most of them are highly correlated [10]. They were initially designed to account for branching, linearity, presence of cycles and other topological features. Topological indices have been used also for similarity evaluation [10, 16]. However, the characterization of a molecule by a single mathematical descriptor provides limited information. It is not surprising to find that several different structures have the same numerical value of a topological descriptor [18].

The most significant criticism on topological descriptors concerns their inability to express 3D molecular structure. The graph model of a chemical structure covers only its connectivity and cannot differentiate conformers. Recent attempts to rehabilitate topological indices consider the use of distance matrices for 3D structures, instead of adjacency matrices of molecular graphs [18]. The off-diagonal elements of the distance matrix represent the inter-atomic distances between the corresponding atoms. Different matrix invariants are proposed as 3D topological indices. Whereas these new descriptors indeed account for 3D structure, they can hardly bear the name "topological".

2.2 Molecular Fingerprints and Molecular Holograms

Fingerprints stand for the presence or absence of some properties (*e.g.* fragment substructures) within a molecule. Two-dimensional substructures are encoded by setting bits in a bit-string (or fingerprint). Molecules are estimated to be structurally similar if they have many such bits in common [10, 19]. Fingerprints may encode structural information (simple descriptors such as the numbers of atoms and bonds or the number of rotatable bonds) or distance information between pharmacophoric groups. When distance information is encoded, then fingerprints may account for conformational flexibility. Fingerprints are usually compared by the Tanimoto coefficient:

$$\tau = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}}, \quad (1)$$

where N_A is the number of features (bits) in the fingerprint A, N_B is the number of features in B, and $N_{A\&B}$ is the number of features common to A and B. The Soergel distance function $(1 - \tau)$ is often used to transform the Tanimoto association coefficient to a distance measure.

A molecular hologram is an array containing counts of molecular fragments instead of only ones or zeros in a fingerprint, denoting the presence or absence of some fragment. Since fragment counts depend only on molecular topology, they do not take into account conformational flexibility. However, holograms can account for chirality if chiral fragments are involved.

2.3 Distance Between Fingerprints

Fingerprints in similarity analysis are usually compared via the Tanimoto distance. However, Tanimoto distance could be used to compare any set of descriptors, which accounts for presence and absence of features.

Analysis of similarity assessment, based on the Tanimoto coefficient, has been done in [20, 21]. Tanimoto coefficients have been calculated between a query compound and all the compounds in the database for five different query compounds. The average similarity increases with more complex queries, or in other words, there are more compounds similar to the more complex query than to the simple one. The author notes that visual examination of the database does not agree with this conclusion, but this is exactly what has to be expected from Tanimoto's coefficient performance. Since it accounts only for presence or absence of fragments, there should be larger overlapping between complex query fragments and some target database than between a simple query and the same target.

The performance of similarity measures is often compared to the user intuitive expectations. Human reasoning is quite flexible and is able to notice different important features in various contexts. Conversely, a computer algorithm is designed to detect predetermined features which may be relevant in one context, but irrelevant in another. In the above example, the performance of the Tanimoto measure reflects its design. It should be applied when similarity could be measured only by the presence or absence of fragments.

2.4 BCUT (Molecular Eigenvalues)

BCUT indices are the eigenvalues of the modified connectivity matrices of the molecule. Reducing information from those matrices to eigenvalues effectively provides a one-dimensional measure, which is expected to reflect molecular structure.

BCUTS originally contained only 2D information in an adjacency matrix [22] but after some important contributions of Pearlman [23, 24] they evolved into a mix of all types of descriptors, depending on the values on the diagonal (*e.g.* charges) and the off-diagonal elements of the connection table (*e.g.* functions of inter-atomic distances). There are three classes of matrices: 2D or 3D connection tables, semi-empirical MO charges and AM1-derived atomic polarizabilities (requiring increasing amounts of CPU-time, respectively). The use of charges, polarizability and inter-atomic distances in modified connectivity matrices helps to take into account conformers. The largest and smallest eigenvalues are used most often. The smallest eigenvalue carries the least significant amount of information from the matrix, the largest – the most significant [25].

BCUT indices defined over a 2D connection table give results similar to Randic indices. For some selected matrices Burden suggests physical interpretations [26]. In general, the physical interpretation of BCUT indices for other matrices is difficult. Some parallels can be drawn between BCUT indices and graph spectrum of Laplacian matrices. The set of eigenvalues of a graph adjacency matrix is called graph spectrum. The adjacency matrix of a graph is a matrix with rows and columns labeled by graph vertices, with a 1 or 0 in position *(i, j)* according to whether nodes *i* and *j* are adjacent or not. The Laplacian matrix of a graph is again a matrix with rows and columns labeled by graph vertices, but with a (-1) in position *(i, j)* if *i* and *j* are adjacent and zero if they are not adjacent. Diagonal elements *(i, i)* hold the value equal to the degree of the node *i* (number of edges for the node). In general, there are many problems in physics and chemistry, where the Laplacian matrices of graphs and their spectra play a central role and they have a physical interpretation in various physical and chemical theories [27]. In similarity analysis specifically, BCUTs are very popular [23].

3 Similarity According to Configuration and Conformation

The ability to take into account the conformational flexibility of chemical compounds, the 3D arrangement of structural features and some 3D descriptors such as shape and volume may be of more importance than the topological information in some activity cases. In such cases the minimum energy conformer of a given compound may not be sufficient to model the activity and to allow successful similarity evaluation.

3.1 Distance-based and Angle-based Descriptors

The simplest of distance-based descriptors are distances between atoms or between functional groups. Individual distance descriptors comprise the inter-atomic distance between a pair of atoms. Angle-descriptors are based on generalized valence angles and torsion angles. Descriptors named "potential pharmacophore points" (PPP) are a generalized mix of distance and angle descriptors [28]. All the atoms of the molecule are analyzed to see whether they can be classified into one of the point types (H-bond donor,

H-bond acceptor, positively charged, negatively charged, hydrophobic). The atom pair distance (PPP-pair) and the three bonded atoms angle (PPP-triangle) form generalized distance distribution and valence angle descriptors respectively. The authors of this research claim that these triangle-based features provide a simple and effective mechanism for similarity searching based on size and shape. The frequency distance distribution [29] is another distance-based descriptor representing the whole molecule.

3.2 Three Dimensional and Field Similarity

Probably the most popular current method of analyzing 3D molecular similarity is the comparative molecular field analysis (CoMFA) technique [30]. The basic idea is to represent the 3D molecular field by a collection of sampled data points. Central to the method is the application of a 3D grid, which covers the structures of all the molecules in the data set. An attribute is assigned to each grid point and the set of attributes are collected into a vector to represent the 3D shape of the field. Thus, the shape information of the field is indirectly coded as attribute index numbers. Each attribute contains a scalar, which signifies the value of the field at the sampled grid point. The problem with this approach is that it is difficult to find comparable sampling points between molecules. To achieve this, the method requires a time consuming and error prone relative orientation of the molecules in the data set.

Distance between molecular fields

Molecular field similarity calculations are now being widely applied. One of the major techniques is the Carbo similarity index [31, 32]:

$$R_{AB}=\frac{\iiint P_A(x,\,y,\,z)P_B(x,\,y,\,z)\mathrm{dxdydz}}{\left(\iiint P_A^2(x,\,y,\,z)\mathrm{dxdydz}\right)^{1/2}\left(\iiint P_B^2(x,\,y,\,z)\mathrm{dxdydz}\right)^{1/2}} \tag{2}$$

Molecular similarity $\boldsymbol{R_{AB}}$ is determined from the structural properties P_A and P_B of the two compared molecules and the summations are over all components of the 3D grids (x, y, z) that surround these two. The numerator measures property overlap, while the denominator normalizes the similarity result. As originally applied by Carbo, quantum mechanically derived electron density is used as the structural property *P*. The technique has been extended to cover molecular electrostatic potentials and electric fields [3, 33, 34, 35, 36].

The same approach could be used in a shape similarity assessment [37]. An orthogonal grid is placed around the molecules and the structural property is evaluated at each grid point. For shape identification, every grid point is tested to see whether it falls inside the van der Waals surface of each molecule. The results are then applied in the following equation:

$$S_{AB}=\frac{B}{(T_A T_B)^2} \tag{3}$$

B is the number of grid points falling inside both molecules, while $\boldsymbol{T_A}$ and T_B are the total number of grid points falling inside each individual molecule. Grid-based shape and electrostatic potential similarity evaluations, while faster than the original quantum mechanically based calculations, are still time-consuming processes. Actually, shape calculations are slower than the electro-static potential ones, since very fine grids are required to obtain precise results (0.2-Å separation is generally used). Faster calculation procedure can be achieved through Gaussian approximation of electron densities [38].

3.3 Molecular Multi-pole Moments

A simple and fundamental characterization of molecular shape and charge is given by the moments of the shape and charge distributions. The lower order moments are clearly understood. Zero order moments of the mass and charge distributions are the total molecular mass and the net molecular charge. The second-order moments of the mass distribution are the moments of inertia. These moments, jointly with the principal inertial axes provide important information to be used in developing molecular descriptors. The first-order moment of the charge distribution is the dipole moment. For neutral molecules, it is invariant with respect to the location of the origin of the axes. This invariance is a consequence of the fact that the lowest order non-vanishing moment of the electrostatic multi-polar expansion does not depend upon the origin. The values of all higher order multi-polar moments depend upon the choice of origin of the multi-polar expansion.

Recently, a number of procedures have been proposed that eliminate the requirement of superposition between molecules [39, 40, 41, 42]. The alignment-free procedures are using relatively small set of descriptors that capture certain 3D molecular features. One of these procedures, named Comparative Molecular Moment Analysis (CoMMA) [39], uses the lower order moments of the molecular mass and charge distributions for comparison.

3.4 Shape

Besides topological approaches such as the molecular connectivity and kappa indices [43], a number of methods to quantify molecular shape have been proposed. Descriptors such as van der Waals volume and surface area can reflect the size of substituents, but they contain very little information about shape. The Taft steric parameter [44, 45] has found some applications, but its values cannot be determined for many substituents. In the STERIMOL [46]

parameter set, each substituent is represented by a length descriptor and four perpendicular width descriptors. While this approach more adequately describes the shape of a substituent, a much larger set of compounds is required to statistically accommodate this many descriptors. In addition, the STERIMOL parameters do not provide information on the orientations and distances between substituents of a molecule in space. Shape similarities and differences in Molecular Shape Analysis (MSA) [47] are described quantitatively in terms of common-overlap steric volume between pairs of molecules, representing atoms as spheres of standard van der Waals radii. Similar functional groups and fragments of the molecules are superimposed in order to maximize their shape similarities. This allows the use of one compound as a reference and the calculation of the MSA descriptor for each compound, based on the common overlap steric volume with this reference compound.

Shape comparison of dissimilar molecules can be performed by geometrically invariant molecular surface descriptors [48]. These quadratic shape descriptors are calculated by least squares fitting of a quadratic function to small sections of the molecular surface of a ligand. Invariant geometric properties of the approximated surface patch are then extracted and used to quantify the shape and to obtain a canonical orientation for this section of surface. The superimposition algorithm uses these geometric invariants to recognize similar regions of the surface shape on two molecules. The superimposing algorithm is insensitive to the connectivity and the relative sizes of the molecules being matched. The algorithm was applied to compare dissimilar ligands known to inhibit the same enzyme system. In all examined cases the algorithm generates superpositions that are in agreement with crystallographic results. The algorithm was also applied to align the two different proteins based on the shape of their active sites.

Duca and Hopfinger [49] developed a 4D-QSAR method to estimate molecular similarity as a function of conformation, alignment and atom type and applied it to study chiral and isosteric compounds, as well as for identification of common pharmacophores. The method allows molecular similarity to be measured with respect to the whole molecule as well as with respect to functional pieces of the molecule. Two types of similarity measures – relative and absolute – are distinguished. Relative similarity is defined as dependent upon an alignment constraint, while absolute similarity is alignment independent.

The first step in [49] is to estimate the conformational energy profile of the molecule. The second step is to construct the main distance-dependent matrix (MDDM), for each pair of interaction pharmacophore elements (IPEs) of the molecule. The IPEs are functional pieces of a molecule. Seven IPE types are proposed, which correspond to the major atom types composing any molecule (all atoms in the molecule, nonpolar atoms, polar atoms with positive charge, polar atoms with negative charge, hydrogen bond acceptor atoms, hydrogen bond donor atoms, aromatic atoms, and non-hydrogen atoms). A unique MDDM is constructed for each unique IPE pair for each molecule. Absolute molecular similarity MDDM matrix elements are defined as function of interatomic distances, associated with the atom pairs of two IPEs and the conformational energy profile. For relative molecular similarity estimation, a grid cell space is defined, alignment rule is selected and grid cell occupancy descriptors are calculated. MDDM elements are defined as a function of these descriptors. For both absolute and relative similarity, the set of normalized eigenvalues of MDDM matrix is defined to be the "essential molecular similarity measure", with respect to particular IPE types. Molecular dissimilarity between pair of compounds is then defined as sum of differences between normalized eigenvalues.

The WHIM descriptors (Weighted Holistic Invariant Molecular descriptors) [9, 50] are 3D-molecular descriptors, based on a consideration of the x, y, and z coordinates of a molecule and scaled with differing weighting schemes. The authors found these descriptors capable to model several physicochemical properties and biological activities for classes of heterogeneous compounds. However, WHIM indices are not able to describe the difference between linear and non-linear molecules [42].

4 Physicochemical Properties

The physicochemical properties are fundamental physical and chemical properties of the chemicals. They represent a macroscopic description of the substances [51]. Examples of global physicochemical properties are molecular weight, octanol-water partition coefficient (logP), total energy, heat of formation, ionization potential and molar refractivity [52]. Physicochemical properties are widely used in assessing similarity between chemicals [8, 9, 53, 54, 55]. Physicochemical property descriptors have been used for diversity profiling by several authors [56, 57, 58]. However, these descriptors are more frequently used in QSARs to establish relationship within con-generic compound series than in similarity assessment because the latter usually involves sets of diverse chemicals.

5 Quantum-chemistry Approach

The quantum chemical approach to molecular similarity is an attempt to take advantage of the fundamental theory of matter. The quantum mechanical (QM) postulates assume that the wave function and the density function contain all the information of a system. The statement, applied to a chemical compound, means that all the information about any molecule could be extracted from the electron density. Bond creation and bond breaking in chemical reactions, as well as the shape changes in conformational processes, are expressed by changes in the electronic density of molecules.

The electronic density fully determines the nuclear distribution, hence the electronic density and its changes account for all the relevant chemical information about the molecule. In principle, quantum-chemical theory should be able to provide precise quantitative descriptions of molecular structures and their chemical properties.

All quantum mechanic methods (*ab initio* or semi-empirical) work by approximating a solution of the fundamental equation of quantum chemistry. The disadvantage of quantum mechanic methods has been that even an approximate solution of the Schrödinger equation can be extremely complex for all but the simplest systems. Long computational times have been required for meaningful calculations and hence the size of the system, which can be studied, has been limited. Modern developments in numerical algorithms and computer hardware have made QM calculations on practical systems much more feasible.

Quantum chemical methods allow derivation of molecular descriptors from the total molecular wave function and charge distribution (section 5.1). Other approaches include comparing electronic density between compounds (section 5.2) or analyzing topological features of the electron density (section 5.3).

5.1 Quantum Chemical Descriptors

Quantum chemical calculations are an attractive source of new molecular descriptors, which can, in principle, express all the electronic and geometric properties of molecules and their interactions. Quantum-chemical descriptors are able to characterize the reactivity, shape and binding properties of a complete molecule (*e.g.* HOMO and LUMO energies, total energy, heat of formation, ionization potential, number of filled orbitals, standard deviation of partial atomic charges and electron densities, dipole moment), as well as of molecular fragments and substituents (*e.g.* partial atomic charges, etc.). Consequently, the derived models will include information regarding the nature of the intermolecular forces involved in determining biological or other activity of the studied compounds [59].

In contrast to experimental measurements, no statistical errors exist in quantum-chemical calculations. There is an inherent error however, resulting from the assumptions made to facilitate the calculations. In most cases the direction but not the magnitude of the error is known. When using quantum chemistry-based descriptors with a series of related compounds, the computational error is considered to be approximately constant throughout the series. A basic weakness of quantum-chemical descriptors is the failure to directly address bulk effects, though this is also true for most available descriptors.

Quantum-chemical calculation is performed for a single structure at an energetic minimum corresponding to the hypothetical physical state of the gas at 0 K and infinitely low pressure. In addition, the zero-point vibrations of the molecule are neglected. Therefore, the quantum-chemical descriptors can't account in principle for entropy and temperature effects. If such effects dominate given property or process, quantum-chemical descriptors are not adequate for their representation and any correlation based on such descriptors, can be regarded as accidental. Most standard quantum-chemical program packages (*e.g.* AMPAC, MOPAC and Gaussian98) have an option to calculate the vibrational, rotational, and translational partition functions of the molecule for a specified temperature and their respective influences on the molecular enthalpy, entropy, and other thermodynamic functions. However, the thermodynamic functions provided by the quantum-chemical program packages mentioned above still refer to only a single conformation of a molecule only. A possible solution for flexible compounds is to average molecular descriptors over a set of conformers through arithmetic- or Boltzman average. Averaging will not work however in studies of biological activity in cases when only one conformation of the compound is active.

As most chemical reactions and all biochemical reactions refer to condensed (mostly liquid) media, it should be advantageous to use molecular descriptors calculated with some quantum-chemical scheme, which accounts for specific and non-specific bulk solvation effects. Specific effects on the molecular structure (primarily hydrogen bonding) can be taken into account by the super-molecule approach where the solute molecule is treated together with the specifically coordinated solvent molecules. A number of different calculation schemes are available for the description of the solvent bulk (reaction field) effects on the solute geometrical and electronic structure. In summary, it is clear that quantum chemical descriptors have considerable applicability potential in diverse areas of chemistry and biomedicine provided that their application is critically analyzed and justified for a given property or phenomenon.

5.2 Quantum Similarity Measure

The Carbo index is the most popular matching measure between the electron densities of two molecules (see Equation 2, section 3.2). Other indices, namely the Hodgkin-Richards index [60], the reactivity based similarity index [61], the overlap molecular quantum similarity measure (MQSM) and the Coulomb MQSM [31] have also been developed. The approach is attractive, due to the role of electron density as the ultimate information about a molecule and has received considerable attention in the literature. However, typical problems encountered are time-consuming computations, necessity to obtain the densities with reasonable quality and necessity to superimpose molecules in order to compare densities. The first problem is addressed by approximating electron density through fitting spherical Gaussian functions. As a solution for the last problem, a quantum self-similarity index was suggested [31].

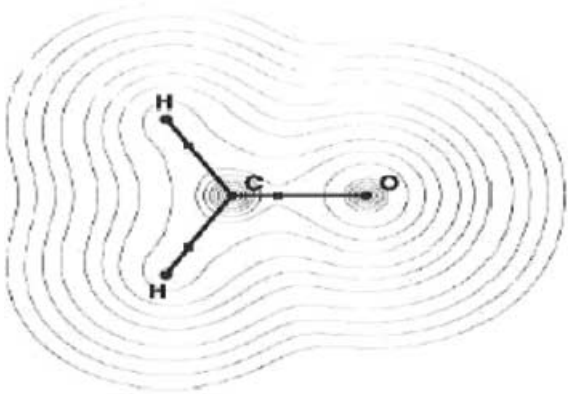


Figure 1. Contour lines of the electron density and bond critical points of the electron density of methanal.

5.3 The Theory of Atoms in Molecules

The theory of atoms in molecules was developed by Bader and co-workers [62, 63, 64]. It offers rigorous quantum chemical definition for atoms, bonds and functional groups providing a strong link between chemical intuition and the theory of quantum mechanics. This theory is based on the topology of the electron density distribution in molecules. The electron density in a molecule $\rho(r)$ has larger values around each nucleus (most of the density is concentrated around the nuclei). This can be depicted with contour lines of the electron density and/or with the gradient vector field of the electron density (Figure 1).

A gradient vector $\nabla\rho(r)$ points in the direction of increasing electron density. The gradient paths are all perpendicular to the contour lines that they cross. Some gradient paths terminate at the nuclei. Other gradient paths terminate at the critical points between the bonded atoms (bond critical points, BCP). Bonded atoms are characterized by bond paths between them that contain a BCP. Two gradient vectors that originate at the BCP and terminate at one of the nuclei define a bond path. Thus, the BCP represents a minimum of the electron density along the bond path, but it has a maximum in the electron density for a line perpendicular to the bond path. A BCP is located on an inter-atomic surface, which this theory uses to partition the molecule into its constituent atoms (or atomic basins). According to the AIM theory, the molecule can be uniquely partitioned into a set of bounded spatial regions. The form and properties of the groups defined by these regions truly recover the characteristics ascribed to the atoms and functional groups of chemistry.

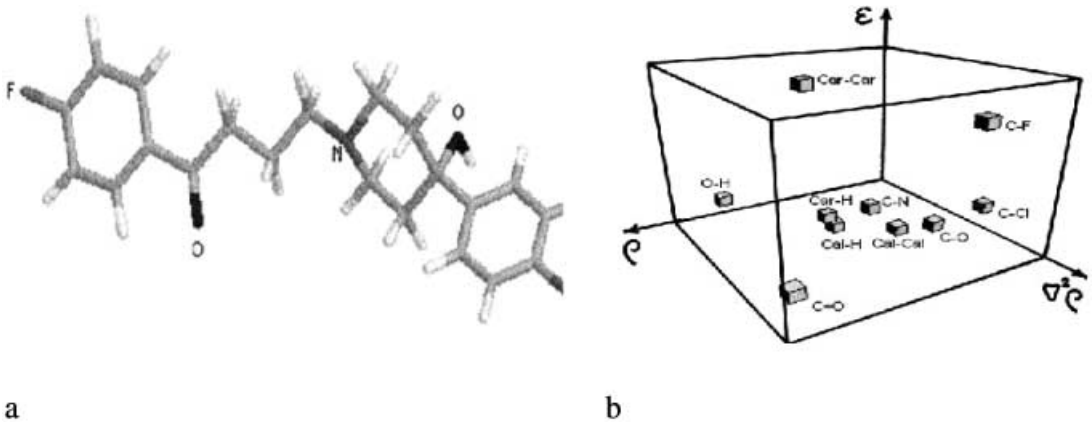
An interatomic surface (also referred to as a zero-flux surface) does not contain any gradient paths that terminate at nuclei, but instead its gradient paths terminate at a BCP. It has been demonstrated that several properties evaluated at the BCP summarize the characteristics of the corresponding bond. For example, the electron density at the BCP, denoted by $\rho_b(r)$, determines a bond order. The Laplacian of the electron density at the BCP, denoted by $\nabla^2\rho_b$, distinguishes two broad classes of bonds: if $\nabla^2\rho_b < 0$, the bond is a so-called *shared interaction*, but if $\nabla^2\rho_b > 0$, the bond is called a *closed-shell interaction*. Covalent bonds belong to the former class, and ionic bonds, hydrogen bonds, and van der Waals bonds belong to the latter. A third important quantity describing another facet of the electronic structure of a bond is the ellipticity at the BCP, denoted by ε_b. The ellipticity measures the susceptibility of ring bonds to rupture and provides a quantitative generalization of the $-\sigma-\pi$ character of a bond [54, 65]. BCP properties detect conjugation, subtle delocalization effects and hyperconjugation. They distinguish aromatic and anti-aromatic character and parallel bond order and prove that three-membered saturated hydrocarbon rings act like double bonds.

5.3.1 Quantum Similarity in BCP Space

The electron distribution, its Laplacian, and the ellipticity are in fact three components of a so-called chemical descriptor vector. Each vector describes a bond in a three-dimensional BCP space. Adding more components, such as the kinetic energy density, can increase the dimensionality of the BCP space. Thus, each molecule is represented by just a handful of numbers, being the components of the vectors describing its bonds. The basic working hypothesis is that – disregarding several technical issues – the molecule is completely and accurately described in a compact and abstract space called BCP space. As a result, similarity measures are reduced to *discrete* distance-like measures in BCP space without loosing their quantum mechanical basis.

Popelier illustrates the BCP space concept [54, 65, 66] with the representation of the drug haloperidol (Figure 2a). Haloperidol has 51 bonds, each of which is represented as one BCP in a 3D BCP space, spanned by following properties: ρ_b, $\nabla^{2\rho}{}_b$, ε_b. The complete representation shows that the BCPs cluster up in 10 well-resolved clusters. Figure 2b shows a representative BCP for each cluster. It has been observed that the C_{arom}-C_{arom} cluster is in fact split in two: the smaller sub cluster represents the two pairs of benzene carbon-carbon bonds adjacent to the C–F or C–Cl bond. These four bonds have a somewhat higher ellipticity than the other members of their cluster because halogens are π-donors. Moreover, such fine-tuning is correct even in predicting that fluorine is a stronger π-donor than chlorine, since fluorine causes the largest increase in ε_b. This example illustrates the power of the BCP space to describe the electronic structure of a molecule in a compact and reliable way.

Once a compound is already represented in some descriptor space, a number of approaches to measure similarity are available. The simplest one is to calculate Euclidean distance between points in BCP. In this case, the distance between compounds A and B in BCP space can be defined as [65, 66]:



a b

Figure 2. 3D image of haloperidol (2a). and in the BCP space on (2b). Each molecular bond is represented as a point in a 3D space, spanned by the following properties: ρ_b, $\nabla^2\rho_b$, ε_b. Haloperidol has 51 bonds and results in 51 points in the BCP space.

$$d(A,B) = \sum_{i\in A}\sum_{j\in B} d_{ij},$$

where d_{ij} is the distance between BCP_i and BCP_j.

$$d_{ij} = \left[(\rho_{b,i} - \rho_{b,i})^2 + (\nabla^2\rho_{b,i} - \nabla^2\rho_{b,i})^2 + (\varepsilon_{b,i} - \varepsilon_{b,i})^2\right]^{1/2}$$

The BCP space concept was successfully used in QSAR studies [67, 68, 69, 70]. The method is able to suggest molecular fragments, responsible for activity, but it was applied so far only to sets of compounds with a common skeleton. The application of BCP space to a set of diverse chemicals is yet to be explored.

5.3.2 Transferable Atom Equivalent (TAE/RECON) Method

Another example of AiM approach application is the transferable atom equivalent (TAE/RECON) method [71, 72]. In the TAE/RECON method, atomic contributions are used to rapidly generate whole molecule electron-density-derived descriptors that approximate those available through *ab initio* calculations. A set of descriptors, suitable to use in QSAR and similarity analysis is generated.

Atoms in Molecules (AiM), is probably one of the most promising approaches to understand chemistry in terms of atoms and molecular fragments in the same way as an organic chemist is accustomed to think about them. In an increasing number of publications is indicated the applicability of AiM for quantitative structure activity relationships (QSAR), multi-pole moments (see 3.2), understanding of chemical reactivity, and its connection to X-ray crystallography [73]. BCP space provides efficient and informative description of the molecule. It becomes clear, that the potential of AIM has not been unleashed yet and new applications are to be expected.

5.4 Electron Density Theorems and Molecular Similarity

The local and global properties of molecular electron densities are interrelated by some of the fundamental theorems of molecular physics. These relations have significant consequences regarding the similarities between molecular properties and reactivities on various levels: in the comparisons of the roles of individual functional groups, and also – on the level of spatial requirements and global shapes of molecules.

The recently proven Holographic Electron Density Theorem [74] states that within any boundary-less molecular electron density cloud in a non-degenerate ground electronic state, any nonzero volume piece of the electron density cloud contains the complete information about the entire molecule. According to the classic and less powerful Hohenberg-Kohn Theorem, the entire electron density determines the energy of the molecule, whereas according to the Holographic Electron Density Theorem the entire electron density is not needed and any small nonzero volume piece of the electron density is already sufficient to determine the energy (and other properties) of the molecule. The theorem also gives grounds for treating host-guest interactions, macromolecular bonding and development of new computational tools for assessing molecular shape similarity [75, 76].

The Molecular Electron Density Lego Assembler (MED-LA) method [77, 78, 79] is a technique that can generate *ab initio* quality electron densities for large molecules, of a size that renders their description by conventional software for *ab initio* electron density computations unfeasible or impossible. The MED-LA approach is based on a simple electron density fragment additivity principle.

An electron density fragment data bank (storing only the fragment density matrices and basis set information) is generated, based on accurate, high quality *ab initio* quantum chemical calculations for small molecules, and the application of the electron density fragmentation principle. The

corresponding "fuzzy" density fragments also account for the inter-fragment interactions occurring within their molecular neighborhoods. These fuzzy fragment density matrices are combined according to an index assignment pattern based on the additivity principle. In this way, *ab initio* quality electron density matrices could be constructed for molecules of any size and for any nuclear arrangement. Experimental or theoretically determined nuclear coordinates or dynamic rearrangements, assumed to occur along reaction paths or in protein folding processes, could be used for this purpose.

This approach can be applied to the similarity assessment of biologically important macromolecules, through a combination of local and global macromolecular shape analysis [80]. The local features are described by the Shape Group Method [81, 78] and the associated topological shape matrices, reflecting the local curvature properties of the infinite set of iso-density contours of local functional groups within the macromolecule. The large-scale features are described by the functional group polyhedron approach. The local shape information is combined with orientation information as represented by the moment of inertia tensor information of local functional groups. The MED-LA method has been used for the study of correlations between local shape features and biochemical properties, including toxicities of various molecular series. Electron density fragmentation methods are applicable to the study of the combined effects of local and global shape features of molecular electron densities. In some instances the biochemical effect is a result of several factors, and in such cases no single shape feature can be expected to correlate well with experimental results. As an example, the MED-LA method has been used in a toxicological risk assessment of polyaromatic hydrocarbons (PAHs) to aquatic species *L.gibba*. The combination of a local (one-ring) and a global (complete molecule) similarity measures provided excellent correlation coefficient (0.96). This could be explained by the mechanistic hypothesis that among the many factors influencing overall toxicity, the global shape feature is probably relevant to the photosensitization step, involving the extended conjugated system of PAHs, whereas a local shape feature is important to photo modification [79].

6 Reactions

The importance of chemical similarity in reactions is the focus of a number of papers. Lawson [82] describes the concept of similarity between reactions and its use in information classification. Ponec [83] discusses the use of reaction similarity to rationalize various mechanistic features of pericyclic reactions. Sello [84] describes a program that predicts the likely products of specific reactions. Gasteiger [85, 86] defines similarity of chemical structures by generalized reaction types and by gross structural features. Two structures are considered similar if they can be converted by reactions belonging to the same predefined groups (for example oxidation or substitution reactions). The similarity of reactions is defined by physicochemical parameters, calculated for atoms and bonds at the reaction centre. These definitions correspond to structural transformations that can be made for an entire dataset of structures, prior to a search query, making possible an efficient and rapid search of databases of structures. Applications in the organization of databases of structures, reaction prediction, reaction planning, synthesis design, and in the automatic acquisition of knowledge about chemical reactions were proposed [85].

7 Distance Between Real Valued Descriptors

In this section similarity between compounds, expressed as a distance in real valued descriptor space, such as topological (section 2), physicochemical (section 4), or quantum chemical (section 5.1), is addressed. Comparing distances between points and a group of points, in order to decide how close a point to the group is, is an intuitive and widely used approach. Compounds are presented as points in n dimensional descriptor space and the distance between two points is taken as a similarity measure between two compounds. Euclidean distance is the most popular distance used. While this approach is apparently simple, its underlying assumptions are not always realized.

7.1 Classification Methods

Decision taking, based on distance comparison, is only a particular case in the broad domain of decision taking, based on classification. Different classification approaches have been developed over the years. All of them aim at deriving decision boundaries between groups in n dimensional space. Decision boundaries, derived by using Euclidean distances can only be linear, and thus are suitable only if the groups of points are linearly separable [87] (*e.g.*, active and inactive compounds in descriptor space can be separated by a line, a plane or a hyper plane). This boundary is guaranteed to be optimal (with minimum error of classification) only if the points in the groups have Gaussian distribution with uncorrelated descriptors with equal variance. The quantity

$$d^2 = (x - y)' C^{-1} (y - x),$$

is called the Mahalanobis distance between points x and y, where C is the covariance matrix. Mahalanobis distance accounts for different variance and correlation between descriptors and generates more flexible quadratic decision boundaries [87, 88]. However, they are again optimal only if the points have Gaussian distribution (in this case with unequal variance and specified covariance matrix).

Another popular classification technique are the artificial neural networks. Neural networks can generate complex

decision boundaries and therefore provide low error classification results when underlying data distributions differ from standard. However, as any complex model, neural network models are prone to over fitting. In addition, it has been proven that neural networks are equivalent to statistical classifiers, though they offer more effective computational algorithms [89].

The probabilistic approach is based on Bayes theorem and provides theoretically optimal decision rule [87, 88]. The Bayesian Decision Rule guarantees lowest classification error if the probability distributions of the chemical classes to be separated are known. This is rarely true and has led to two different approaches – parametric and non-parametric. The parametric approach assumes that probability distribution has a known shape (*e.g.* Gaussian) and estimates its parameters (*e.g.* mean and variance). In the non-parametric approach, the probability distribution is estimated from data.

One should be aware, that any distance scheme and classification technique could in principle provide low error classification results only if the underlying data distributions comply with the assumptions, hidden within the distance formulae and method basics.

8 Discussion

8.1 Similarity Approach and Mechanistic Understanding of Activity

A mechanistic model of a system is a representation of the physical or biological theory, governing this system, in contrast to an empirical (or statistical) model, which is determined by statistically fitting equations to data. Contemporary knowledge of mechanisms of activity is restricted to certain parts of the complex biochemical interactions (*e.g.* receptor based activity, cell membrane penetration, etc.) and for many endpoints is not yet available . This makes impossible to directly estimate activity purely based on chemical or biological theory. Similarity by activity is usually rephrased into similarity by structure, similarity by properties, similarity by descriptors or similarity by other molecular characteristics. This similarity does not always mean similarity in activity. For this purpose, it is necessary to refer to the basic tenets of QSAR modeling:

1) The properties of a chemical are implicit in its molecular structure;
2) Molecular structure can be measured and represented with a set of numbers (descriptors or other numerical representation);
3) Compounds with similar structure exhibit similar properties;
4) Compounds with dissimilar structure exhibit dissimilar properties.

The assumption that the biological property of a compound is implicit in its molecular structure effectively ignores its complex interaction with the environment. This complexity could be accounted through the presumed cause or mechanism of certain biological effect. In the case of drug discovery, the knowledge or hypothesis for receptor affinity is usually the most important, along with ADME processes (absorption, distribution, metabolism and excretion). This understanding resulted in a number of methods, ranging from simple heuristic rules (Lipinski rule of 5 [90]) to rule based systems, classification schemes, bioavailability models [91],docking techniques and descriptors reflecting the interaction with the target.

The second statement – "molecular structure can be measured and represented with a set of numbers" is the rationale behind all chemometric methods. However, it may be misleading if not used carefully. All information about the molecule is contained in its electron density. Any other representation of a molecule results into loss of information. The loss is especially obvious in representing molecules by a set of descriptors, because different molecules could have identical descriptors values. On the other hand, the loss of information from a specific representation may not be of importance for the studied endpoint. The knowledge of what causes the activity is the most reliable source of information when deciding which molecular characteristics are essential to the activity of interest. In the absence of such knowledge, diverse techniques from the areas of statistics, pattern recognition and data mining are available to select relevant descriptors.

Usually the modeler resorts to the similarity in structures with the hope that structurally similar compounds will also have the same mechanism of action [92]. This is a widely used approach, but such hope does not always come true. Several surprising structure-activity relationships demonstrate that chemically similar compounds may have significantly different biological actions and activities and different molecules can be very similar in their biological activities. Applying the results from one con-generic series to another one may lead to completely wrong conclusions [54, 93, 94, 95, 96]. As illustrated in [97], structurally similar compounds (eight compounds with the same connectivity and differing in only one or two substituents in this example) can have very different volume and surface potentials, hydrophobic and polar regions, hydrogen bond donor potentials, hydrogen bond acceptor potentials and molecular electrostatic potentials (Figure 3). This is also in contradiction with the long repeated "basics of QSAR", asserting that similar compounds have similar properties and dissimilar compounds have dissimilar properties.

The last two statements are sometimes cited as a logical consequence of the assumed existence of functional relationship between structure and activity. Considering the functional relationship between numerical descriptors and activity, it could be noticed, that this inference is generally wrong, except in special cases of relationships (*e.g.* linear).

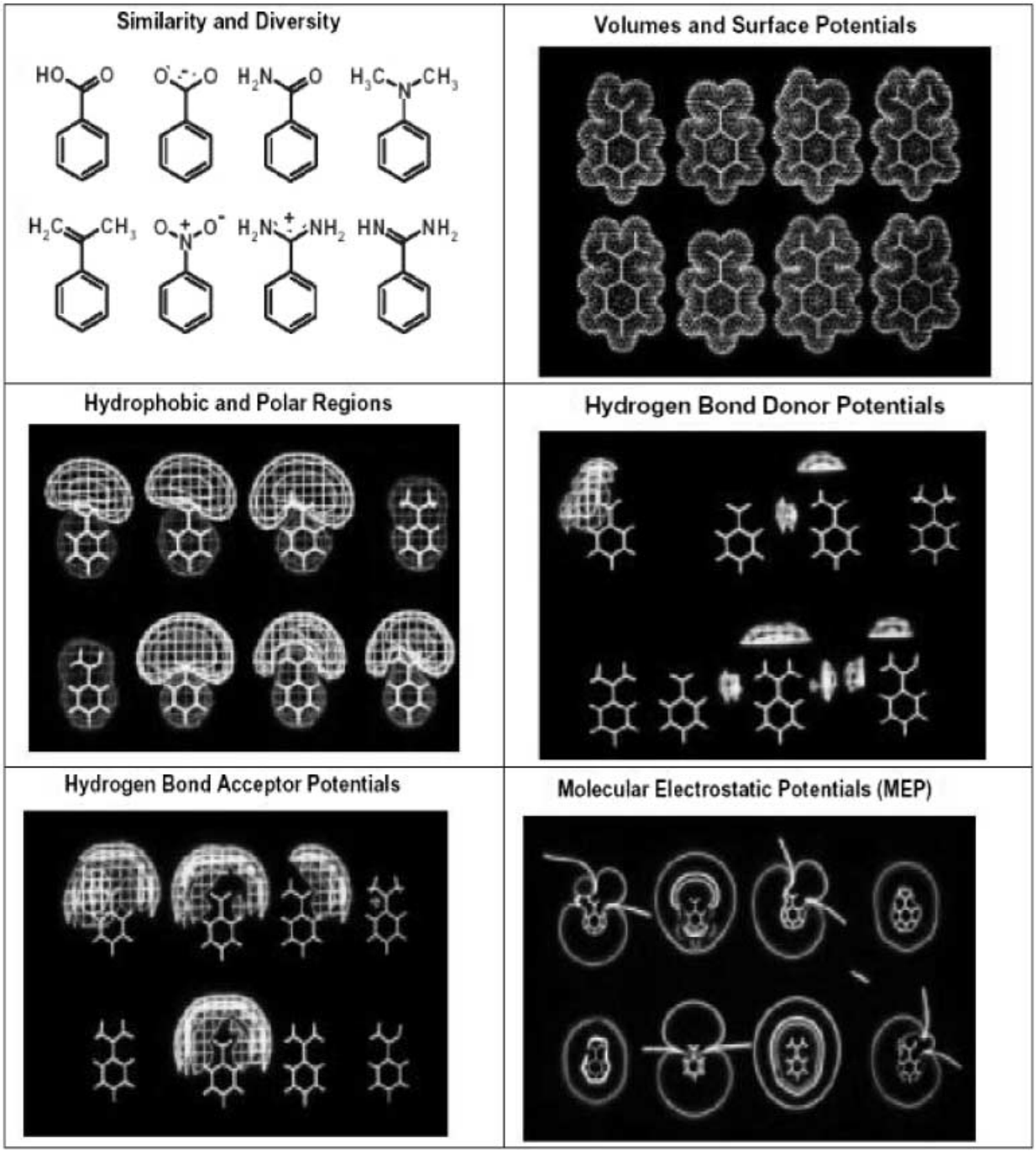


Figure 3. Examples from [117] with permission from the author. Structurally similar compounds can have very different volume and surface potentials, hydrophobic and polar regions, hydrogen bond donor potentials, hydrogen bond acceptor potentials and molecular electrostatic potentials.

Furthermore, QSAR is usually in search for continuous and smooth functional relationship between numerical descriptors and a property. Within restricted descriptor space, covered by small con-generic series of chemicals this is a reasonable assumption. It is also rational within data set, where activity is known to be elicited by a common mechanism. However, within data set, consisting of broad chemical classes, and with activity possibly caused by different and unknown mechanisms, the functional relationship is unlikely to be the same. An approach for avoiding non-homogeneous (different relationships in different regions) and rigid (discontinuous) descriptor space is to restrain analysis to data sets, consisting of chemicals with common and relevant mechanism, with the hope the same

relationship holds for these chemicals.. The other approach is to use data analysis techniques (for example clustering) in order to split up the descriptor space to regions with hopefully smooth descriptor/activity relationship.

8.2 Similarity Approach to Applicability Domain Selection

A very important application of similarity analysis is in QSAR applicability domain determination [98]. The usual QSAR practice is to work with small series of compounds when deriving a model of a property. While this has its historical grounds in modeling properties of con-generic series of compounds, it is hard to determine whether a certain model will be appropriate for predicting a property of a new compound. Having in mind the great diversity of chemical compounds, it is obvious that a model, obtained by analysis of a small data set could not stand for the global relationship between compounds and the endpoint. It rather lights up the tiny island populated with compounds of the data set at hand. Therefore, it is of great importance to identify these islands in chemical space.

The applicability domain of multiple QSARs for a certain common property should be searched within a descriptor space, comprising at least all the descriptors, involved in these QSARs. Any other concerned descriptors should be justified to be relevant to the endpoint. This is easy to demonstrate (Figure 4).

Let us have two models of the same endpoint, each of them over different descriptor: $y = f_1(x_1)$ and $y = f_2(x_2)$. A nonlinear relationship $f_1(x_1)$ is derived from *data set 1*, represented by rectangles; and a linear relationship $f_2(x_2)$ is derived from *data set 2*, represented by triangles. All points from *data set 2* belong to the interval $[x_2' \; x_2'']$.

Let's suppose that a true relationship f_3 exists and is in the form of the 2D figure shown above. In fact, the relationships f_1 and f_2 are just projections of the true $f_3(x_1, x_2)$ relationship. The actual descriptor space is (x_1, x_2), where the clusters are defined by the compounds in the corresponding training sets (triangles and rectangles in the example). For a new compound cluster membership has to be assessed.

Let us consider the point Q in a 2D space (x_1, x_2). Its projection on one-dimensional space (x_2) is point Q′, which belongs to the interval $[x_2', x_2'']$ and is considered within *data set 2*. However, its true endpoint value (y) is very different from the one, predicted by the model $y = f_2(x_2)$. The source of the error is that the model $y = f_2(x_2)$ ignores an important descriptor x_1, but this cannot be realized just by analyzing *data set 2*.

On the other side, if the model $y = f_2(x_2)$ generalizes well (imagine that the linear relationship holds outside of x_2 ranges), prediction results could be correct. The point P' is not in the interval $[x_2', x_2'']$, but the predicted value (y) is correct. Again, this cannot be justified only by analyzing *data set 2*.

However, these conclusions are based on the the assumption that the true relationship is known. If f_3 is not the true relationship, but rather a projection of another high-dimensional relationship $f_n(x_1, x_2, .., x_n)$, then looking only in 2D space (x_1, x_2) would not be sufficient to assess the model prediction quality.

Therefore, there is no way to justify whether the predictions for points P and Q will be wrong or correct, only by data set examination. The presence of a point within or outside the part of descriptor space, covered by a data set, is to be used only as a warning for model applicability, but not as a final decision on prediction quality. In reality, the true relationship is not known and is the objective of modeling. However, mechanistic understanding of the property modeled, or at least a hypothesized mechanism, could help manage insufficient data.

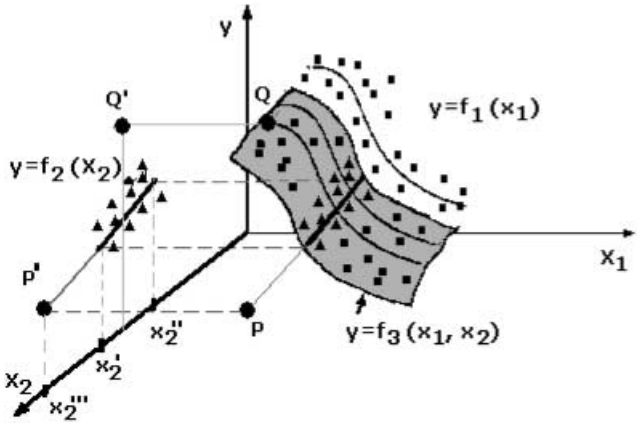


Figure 4. QSAR applicability domain within a hypothetical descriptor space and 2 relationships using 2 different and 1 common descriptor. The applicability domain of multiple QSAR models for a certain common property should be searched within a descriptor space, comprised at least of all descriptors involved in these QSARs. Any other descriptors involved should be justified to be relevant to the endpoint. The presence of a point within (Q) or outside (P) the region of the descriptor space, covered by the data set, is to be used only as a warning for model applicability, but not as a final decision on prediction quality.

8.3 The Need of Information Preserving and Relevant Descriptions

Finding an optimal set of descriptors from a large set of available descriptors is a problem, which occurs not only in QSAR and similarity analysis, but also in many contexts (general modeling, machine learning, pattern recognition and data mining). The main issues in developing descriptor selection techniques are: choosing a small descriptor set in order to reduce the cost and running time of the model and achieving of an acceptably low error rate. This has led to the development of a variety of techniques for selection of optimal subsets from larger sets of possible descriptors.

Available methods for descriptor selection are still not well exploited in QSAR and similarity studies. Besides random selection [20, 99], the most popular approach used is

the Principal Component Analysis method, which selects descriptors with largest variance as the most important ones. However, most variable descriptors are not necessarily most important. The only case when the descriptors with largest variance have the largest contribution to the activity is when there exists a monotonous relationship between the descriptors and the activity and (*i.e.* the function has no maximums or minimums). Moreover, there are a number of examples in the literature, illustrating that most variable descriptors are not always able to provide the best classification results [87, 88]. We will provide below a brief review of descriptor selection methods.

The descriptor selection techniques fall into two main categories. In the first approach domain knowledge is used for the reduction of the descriptor set. The second approach is used when domain knowledge is unavailable or expensive. In this case heuristic algorithms are applied to select a subset of the available descriptors.

Applying domain knowledge in the case of similarity analysis stands for exploiting biological and chemical knowledge about the activity. The knowledge could provide indications which specific characteristic of a molecule are essential and should be used in order to decide on similarity between compounds.

Available methods allow representation of a single molecule by hundreds of descriptors. Most successful applications use in parallel different descriptor types (topological, physicochemical, electronic, fragment based) [28, 100, 101, 102, 103]. If a descriptor is successful in the prediction of a certain property, this is because it covers the important for the considered property characteristic of the chemical compound. Descriptors that fail do not reflect these important critical features of the compound. Typically, only few descriptors (or combination of descriptors) encode features of the molecule, related to some activity. Most of the variables contain no or little information about the activity of interest. It has been known for a long time among the pattern recognition community, that in order to solve some classification problem successfully, the extracted descriptors should satisfy the following conditions:

1) The descriptors should be information preserving or allowing controlled loss of information;
2) The resulting multidimensional points should cluster tightly within a class and be far apart for different classes.

In the context of molecular compounds this means that descriptors should allow the reconstruction of all the information about a compound. This is almost never the case for QSAR descriptors. As discussed earlier, the most information rich description of a molecule is the electron density. The electron density knowledge allows the calculation of the molecular topology, atom positions and bond length, as well as 3D properties. By contrast, the knowledge of a set of descriptor values does not allow reconstruction of all the information for a given compound.

Since in reality profound knowledge of mechanisms of activity is rarely fully available, a method or a descriptor, proposed for similarity analysis, should be assessed on its ability to discriminate between compounds with different biological activities. Various measures of descriptor importance (sometimes named "interestingness") and descriptor selection techniques have been developed in machine learning, pattern recognition and data mining. An extensive work on descriptor selection for classification and a number of references could be found in [104, 105]. In most cases it is usual to find redundant or ineffective descriptors, if they exist, and try to eliminate them. The process is known in the literature as "feature selection" or "dimensionality reduction". Feature selection methods search for the best subset of features through the competing candidate subsets, according to an evaluation function. Searching for the best feature subset is an exhaustive procedure even for a medium-sized feature set. Over the past few years, a number of search algorithms have been designed to prevent an exhaustive search of subsets and reduce computational complexity. A typical feature selection method involves the following tasks:

- Generation of a next candidate subset;
- Evaluation of the generated subset;
- A stopping criterion;
- Validation of the selected subset;

A number of generation (or search) and evaluation algorithms exist. The most widely known generation procedures include the Branch & Bound algorithm [106], RELIEF and RELIEF-F [107,108], and the wrapper method [109]. Branch & Bound performs a complete search through the feature subsets, but avoids an exhaustive search by exploiting the principle that a subset of features should not be better than any larger set that contains the subset. The rest of the above mentioned methods use a heuristic search where generation of subsets is incremental (increasing or decreasing). The Sequential Forward Selection (SFS) method starts from an empty set and in each iteration generates subsets by adding a descriptor, selected by an evaluation function, while the Sequential Backward Selection (SBS) works backwards, *i.e.* it starts from the complete feature set and in each iteration generates a subset by discarding a feature selected by an evaluation function. The evaluation function can use different measures in order to determine whether a feature or a subset of features will be selected or rejected.

Feature selection methods are grouped into two categories: *filter* methods, which are independent of the modeling algorithm and *wrapper* methods, which use the modeling algorithm as the evaluation function [105] (*i.e.* for each descriptor subset a model is build and prediction accuracy is evaluated). Different types of evaluation functions use one of the following measures [104]: variability, distance, uncertainty, dependence, consistency and classifier error rate, with the latter being used by wrapper methods.

Other well-known measures of descriptor importance, which have been extensively used in several areas of physical, social, management and computer sciences are Shannon Entropy, Gini Index and Kullback Mutual Entropy, to name a few. As an example, an analysis of sixteen importance measures is presented in [110].

Shannon entropy measures the average information content of a variable X. It can be interpreted as the average amount of surprise one receives upon learning the value of X or the amount of uncertainty that exists as to the value of X. Shannon Entropy is defined as:

$$SE = -\sum_i p_i \log_2 p_i,$$

where the distributions of different descriptor values are represented as histograms with the same number of bins $p_i = c_i / \sum_i c_i$, and c_i is the count in histogram bin i. Narrower distributions of descriptor values result in lower entropy than broader distributions. Uniformly distributed discrete variables have the highest entropy value (*i.e.* low information content – no surprise upon learning the value of X). Information theory considers as more promising the low-entropy descriptors, because of their high information content. High-entropy descriptors could be of interest in diversity analysis, which aims at obtaining uniformly populated regions in descriptor space. Recently Lin has proposed a diversity metric, based on low-entropy descriptors [111]. This diversity metric has been criticized by Agrafiotis [112] because it tends to over-sample remote areas of the feature space and produces unbalanced designs. It is correctly noticed that an increase in the diversity results in an *increase* in entropy, not a *decrease*. However, Agrafiotis makes the one-sided conclusion that the notion of information as defined in Information Theory is inappropriate in diversity analysis. As we have already noted, for the molecular diversity task high entropy should be looked for.

A uniformly populated descriptor region still does not imply a uniformly populated activity area. The information theory provides methods for assessing mutual information between random variables. It may be more efficient to select descriptors with maximum mutual information for a given activity or use other criteria, besides maximum variance. The information theory has successful applications in image retrieval where again similarity between pictures isn't a well-defined concept.

8.4 "Neighborhood Principle" Assumptions

The efforts to computerize similarity assessment resulted in a number of different methods. One of the most popular among them is the search for compounds with similar activity in the descriptor space. This approach presupposes the existence of a set of descriptors, such that molecules in the same local region ("neighborhood") of this descriptor space tend to have similar values of a desired property [113]. This is assumed to be the fundamental axiom of molecular similarity in descriptor space and is often called the "neighborhood principle" or "neighborhood behavior axiom".

The similarity according to the neighborhood axiom is defined with respect to a molecular property of interest, which leads to multiple definitions of similarity, one for each property. As a result, it allows "similarity" to be defined in an objective way, well suited for computer analysis.

The axiom allows taking the decision that two chemicals have close values of certain property if they have close descriptor values. There are numerous methods for exploiting this idea and statements that it is supported by the experience of synthetic chemists, but some publications claiming the opposite also exist [114]. A formal analysis whether (and when) this assumption holds is necessary before its application.

The presence (or absence) of neighborhood behavior with respect to certain descriptors and properties may be revealed by examination of the plot of differences in descriptor values vs. differences in biological activities [115]. Differences between descriptor values for a single descriptor are plotted on horizontal axis, while differences between property values are plotted on vertical axis. If a good neighborhood behavior holds, then the upper left triangle region would be empty, because there will be no large changes of the property due to small descriptor changes.

To illustrate when good neighborhood behavior holds, plots on Figure 5 are generated using artificial data. Figure 5a presents the differences' plot of the data set, having the linear relationship plus random noise ($y = ax + b + \varepsilon$). The random noise is added to all data sets on Figure 5, to reflect the fact that real data sets rarely exhibit exact relationship, but rather they are obscured by noise due to different reasons. Apparently, in Figure 5a, local neighborhood could be revealed, in which small changes in descriptor (horizontal axis) will lead to small changes in property (vertical axis). The linear relationship ensures that anywhere within the descriptor space, this neighborhood region will cover the same amount of the space. If the linear relationship is steep, then the neighborhood will enclose smaller, but constant volume anywhere inside the descriptor space.

On the contrary, on Figure 5b – f such a neighborhood could not be revealed. Figure 5b – e presents a differences' plot, where underlying relationships are exponential, logarithmic, 1/x and parabolic, respectively. A neighborhood in which small changes in descriptor values give rise to small differences in property, could be found for the logarithmic and parabolic relationships, however the volume (or number of points) enclosed will differ widely. Hence, it will not be possible to specify a single threshold and decide on property similarity, based only on proximity between descriptor values.

Similarity searching in descriptor space could be deceiving. Unless linear relationship holds between descriptors

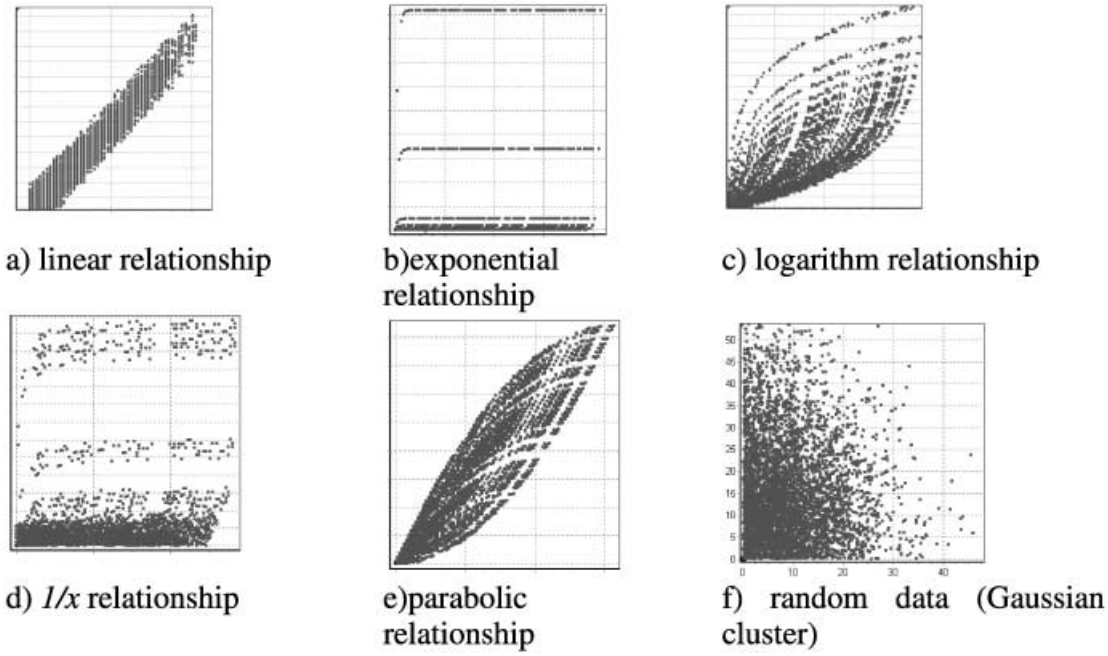


Figure 5. Illustration of the pitfalls of a "neighborhood behavior". The differences between the descriptor values are plotted on the horizontal axis, while the differences between the property values are plotted on the vertical axis. If a good neighborhood behavior holds, then the upper left triangle region would be empty, because there will be no large changes of the property due to small descriptor changes. It can be noted that this is true for linear relationships between a descriptor and an endpoint (a), but not necessarily for other types of relationships (b – f).

and activity, discovering proximity with respect to descriptors does not necessarily imply proximity with respect to activity. However, the linear relationship is only a special case, given the complexity of biochemical interactions, and its frequent use is not always justified.

When non-linear (and/or non-monotonous) relationships are considered, "the neighborhood behavior" is not a necessary condition [113, 116] for similarity in a biochemical activity. This can be illustrated with a simple graph on Figure 6a. A small variation between x_1 and x_2 leads to a small variation of the endpoint (activity) dA, but rather large variation between x_2 and x_3 leads to the same amount of small variation dA in the activity. The same concept is illustrated on Figure 6b with a real data set (the well-known training set used to derive bio-concentration/bio-accumulation factor from octanol/water partition coefficient [117]). It could also be noticed, that small variation of a descriptor (dX on Figure 6b) can lead to very different activity values.

In similarity analysis this will prevent discovering compounds with similar activity, because of the large difference in descriptors. This could also hamper diversity analysis, where one is trying to find the most diverse points in the descriptor space. The basic assumption is that a large difference in descriptor values indicates large differences in activity. Once again, this is true and is the sufficient condition when the underlying relationship between activity and descriptors is linear. However, it is not sufficient, if the function is monotonous (only increasing or only decreasing), but not linear (*e.g.* exponential). It even could be wrong if the underlying relationship has maxima or minima, as illustrated in Figure 6a, or if it is not continuous. To summarize, the neighborhood principle is not applicable to every data set, endpoint and descriptor. While being a useful concept, the modeler should be extremely careful when using methods, exploiting the principle and its validity should be justified for every specific data set.

Similarity analysis in descriptor space could become more complicated when the underlying relationship is discontinuous, or if there are different relationships in different areas of the same descriptor space (non-homogeneous space), which is often the case. In any case, if the underlying relationship is known, we could map distances between points in descriptor space to distances between properties. However, similarity analysis usually is performed exactly because the underlying relationship is not known. In this common case, differences of the property values are not proportional to differences in descriptors (except if there is linear relationship between property and descriptors) and similarity decisions are prone to errors. In diversity analysis one should take into account that well spread compounds in descriptor space do not always mean that the compounds will be well spread in respect to activity and that large distances in descriptor space do not always mean large distances in activity.

Similarity assessment in the descriptor space will work if the true relationship between descriptors and activity is

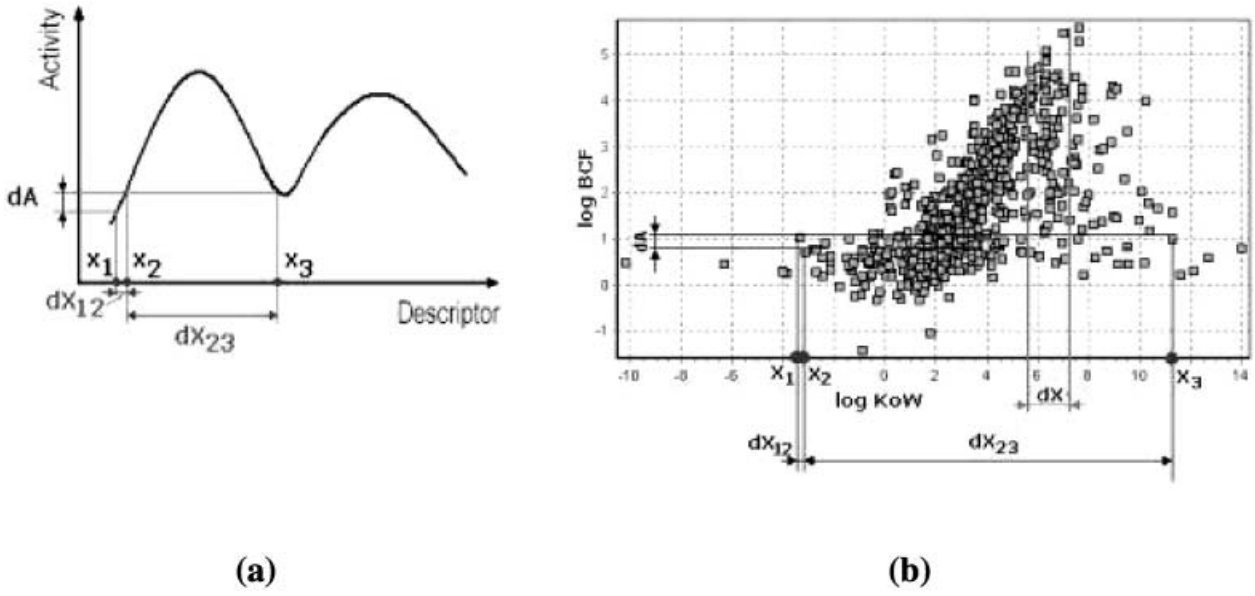


Figure 6. The "neighborhood behavior" of a descriptor in case of nonmonotonous relationship. A small variation between x_1 and x_2 leads to a small variation of the endpoint (activity) dA, but rather large variation between x_2 and x_3 leads to the same amount of small variation dA in the activity (6a). The same concept is illustrated on (6b) with a real data set for bio-concentration factor and $\log K_{ow}$).

linear, or if applied (intentionally or by chance) on a restricted subspace where the relationship could be approximated with a linear one. The hard question is how to reveal these "restricted" spaces. Clustering in descriptor space is an option, but it will not take into account property values. Probably the best way is to find "which combinations of descriptors are most successful in grouping of active compounds together and away from inactive compounds", if other chemical or biological considerations are not available.

The conclusion is that deciding on similarity in respect to activity, based on similarity between descriptor values may be misleading. Here are some ways to avoid mistakes: 1) have a known (or at least hypothesized) relationship between activity and a set of descriptors 2) determine a relationship by using the training set of compounds and some learning algorithms. The relationship does not need to be specified as mathematical function, it may take any form, resulting from utilized classification or data mining algorithm.

The suggestion that a similarity measure should be determined (fitted) using the training set of compounds appears in some publications [24, 118]. Such approach to the measure construction can be considered as a universal one and may be automated in contrast to usually used approaches in guessing the measure for a particular task. In fact, this approach could be employed not only for similarity assessment in descriptor space, but in any other, including structural and field similarity. The basic idea is to exploit learning algorithms to find the common characteristics between compounds with similar activity.

8.5 Search for the "best" Measure

The abundance of available similarity measures for molecular similarity has led to comparative studies in which researchers try to identify a single, "best" measure, using some quantitative performance criterion. Such comparisons, are limited because they assume, usually implicitly, that there is some specific type of structural feature (similarity coefficient, weighting scheme or whatever it is that is being investigated) that is uniquely well suited for describing the type(s) of biological activity that are being sought for in a similarity search. The assumption cannot be expected to be generally valid, due to the complexity of biological activity. A possible solution is to combine results from different measures. Another solution is to use some method that takes into account all the information about the molecule and in this way partially to overcome the need to understand the mechanism of activity and the necessity to identify the relevant descriptor from a chemical point of view. It should be noted that in this case the interactions with the target are still not addressed. Presently the AIM approach (section 5.3) seems to provide the most rigorous approach in this respect, but the theory is relatively new and novel applications are to be expected.

9 Conclusions

Similarity is often a very convenient concept for humans, but a formal definition of similarity is necessary to enable automatic decision-making. The similarity measure should

be relevant to the activity of interest. The relevance could be established by exploiting the knowledge about fundamental chemical and biological processes, responsible for these activities. Since such knowledge is rarely available, various approximations have been developed, based on similarity between descriptor values or structure analogy. This approach has to be used with caution. First, description of chemical compounds should not lose relevant information. Secondly, similar activity need not imply vicinity in the chosen descriptor space, in particular if the functional relationship between property and descriptors is not monotonous and continuous. Therefore, to develop a SAR or explore similarity one needs to determine a *quantitative* relationship for the activity within the training set. The relationship does not need to be specified as a mathematical function, it may take any form, resulting from the classification or data mining algorithm being used. In other words, when a mechanistic understanding is missing, the discovery of the correct similarity measure between molecules is equivalent to the development of a QSAR.

References

[1] W. V. Quine, Natural kinds. In Ontological relativity and other essays, Columbia University Press, New York, NY, **1977.**
[2] N. Goodman (Ed.), Seven structures on similarity. Problems and Projects, 437 – 447. Bobbs-Merril, New York, **1972.**
[3] V. J. Gillet, D. J. Wild, P. Willett, J. Bradshaw, Similarity and Dissimilarity Methods for Processing Chemical Structure databases, *Comput. J.* **1998**, *41*, No.8
[4] J. Bajorath, Virtual screening in drug discovery: Methods, expectations and reality, Current Drug Discovery, http://www.current-drugs.com/CDD/CDD/CDDPDF/issue2 – 03/BAJORATH.pdf (March 2002)
[5] Trends in Fragrance Research: About Structure-Odour Relationships, The BASICS archives, http://www.xs4all.nl/~bacis/bnb01081.html
[6] L. Turin, Y. Fumiko, Structure-odor relations: a modern perspective, http://www.physiol.ucl.ac.uk/research/turin l/review final.pdf
[7] A. McNaught, A. Wilkinson (Eds.), IUPAC Compendium of Chemical Terminology. The Gold Book, Second Edition, Blackwell Science **1997.**
[8] H. Kubinyi, QSAR: Hansch Analysis and Related Approaches, VCH, Weinheim, **1993.**
[9] H. Kubinyi (Ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM Science Publishers B. V., Leiden, **1993.**
[10] P. Willett, Chemoinformatics – similarity and diversity in chemical libraries, *Analytical Biotechnology* **2000**, *11*, 85 – 88.
[11] M. Randic, On Characterization of Chemical Structure, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672 – 672.
[12] H. Hosoya, M. Gotoh, M. Murakami, S. Ikeda, Topological Index and Thermodynamic Properties. 5. How Can We Explain the Topological Dependency of Thermodynamic Properties of Alkanes with the Topology of Graphs? *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 192 – 196.
[13] H. Wiener, Structural determination of Paraffin Boiling Points, *J. Am. Chem. Soc.* **1947**, *69*, 17 – 20.
[14] M. Randic, Characterization of Molecular Branching, *J. Am. Chem. Soc.* **1975**, *97*, 6609 – 6615.
[15] D. Bonchev, N. Trinajstic, Information Theory, Distance Matrix, and Molecular Branching, *J. Chem. Phys.* **1977**, *67*, 4517 – 4533.
[16] S. Basak, V. Magnuson, Determining structural similarity of chemicals using graph-theoretic indices, *Discrete Appl. Math*. **1988**, *19*, 17 – 44.
[17] A. Balaban, Topological indices based on topological distances in molecular graphs, *Pure Appl. Chem.* **1983**, *55*, 199 – 206.
[18] M. Randic, On Characterization of Chemical Structure, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672 – 672.
[19] P. Willett, J. Barnard, G. Downs, Chemical Similarity Searching, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983 – 996.
[20] D. M. Bayada, H. Hamersma, V. J. van Geerestein, Molecular Diversity and Representativity in Chemical Databases, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1 – 10.
[21] D. R. Flower, On the Properties of Bit String-Based Measures of Chemical Similarity, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379 – 386.
[22] F. R. Burden, Molecular identification number for substructure searches, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225 – 227.
[23] R. S. Pearlman, Novel Software Tools for addressing Chemical Diversity http://www.netsci.org/Science/Combichem/feature08.html
[24] R. S. Pearlman, K. M. Smith, Metric Validation And The Receptor-Relevant Subspace Concept, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28 – 35.
[25] J. M. Blaney, E. J. Martin, Computational approaches for combinatorial library design and molecular diversity analysis, *Curr. Opin. Chem. Biol.* **1997**, *1*, 54 – 59.
[26] F. R. Burden, D. A. Winkler, New QSAR Methods Applied to Structure-Activity Mapping and Combinatorial Chemistry, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 236 – 242.
[27] B. Mohar, Laplace eigenvalues of graph – a survey. *Discrete Math.* **1992**, *109*, 171 – 183.
[28] W. Fisanick, K. Cross, A. Rusinko, Similarity Searching on CAS Registry Substances 1. Global Molecular Property and Generic Atom Triangle Geometric Searching, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664 – 674.
[29] P. Willett, Searching for pharmacophoric patterns in databases of three-dimensional chemical structures, *J. Mol. Recognit.* **1995**, *8*, 290 – 303.
[30] R. D. Cramer III, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959 – 5967.
[31] R. Carbo-Dorca, D. Robert, L. Amat, X. Girones, E. Besalu, University of Girona, Spain, Molecular Quantum Similarity in Qsar and Drug Design Coulson's Challenge Series, Lect. Notes Chem., Vol. 73
[32] R. Carbo, L. Leyda, M. Arnau, How Similar is a Molecule to Another? An Electron Density Measure of Similarity Between two Molecular Structures, *Int. J. Quantum Chem.* **1980**, *17*, 1185 – 1189.
[33] E. E. Hodgkin, W. G. Richards, Molecular Similarity Based on Electrostatic Potential and Electric Field, *Int. J . Quantum Chem.* **1987**, *14*, 105 – 110.
[34] E. E. Hodgkin, W. G. Richards, A Semi-Empirical Method for Calculating Molecular Similarity, *Chem. Commun.* **1986**, *19*, 1342 – 1344.
[35] M. Manaut, F. Sanz, J. Jose, M. Milesi, Automatic Search for Maximum Similarity between Molecular Electrostatic Po-

tential Distributions, *J. Comput.-Aided Mol. Design* **1991**, *5*, 1–380.
[36] J. Cioslowski, E. D. Fleischmann, Assessing Molecular Similarity from Results of ab Initio Electronic Structure Calculations, *J. Am. Chem. Soc.* **1991**, *113*, 64–67.
[37] A. M. Meyer, W. G. Richards, Similarity of Molecular Shape, *J. Comput.-Aided Mol. Design* **1991**, *5*, 426–439.
[38] A. C. Good, W. G. Richards, Rapid Evaluation of Shape Similarity Using Gaussian Functions, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112–116.
[39] B. D. Silverman, D. E. Platt, Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition, *J. Med. Chem.* **1996**, *39*, 2129–2140.
[40] R. Bursi, T. Dao, T. van Wijk, M. de Gooyer, E. Kellenbach, P. Verwer, Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861–867.
[41] D. B. Turner, P. Willett, A. M. Ferguson, T. W. Heritage, Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies: 2. Model validation using a benchmark steroid dataset, *J. Comput.-Aided Mol. Design* **1999**, *13*, 271–296.
[42] H. Patel, M. T. D. Cronin, A Novel Index for the Description of Molecular Linearity, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1228–1236.
[43] L. B. Kier, A Shape Index from Molecular Graphs, *Quant. Struct.-Act. Relat.* **1985**, *4*, 109–116.
[44] R. W. Taft, Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters, *J. Am. Chem. Soc.* **1952**, *74*, 3120–3128.
[45] R. W. Taft, The General Nature of the Proportionality of Polar Effects of Substituent Groups in Organic Chemistry, *J. Am. Chem. Soc.* **1953**, *75*, 4231–4238.
[46] A. Verloop, The STERIMOL Approach to Drug Design, Marcel Dekker, New York, **1987.**
[47] D. E. Walters, A. J. Hopfinger, Case studies of the application of molecular shape analysis to elucidate drug action, *J. Mol. Struct.: THEOCHEM*, **1986**, *134*, 317–323.
[48] B. B. Goldman, W. T. Wipke, Quadratic Shape Descriptors. 1. Rapid Superposition of Dissimilar Molecules Using Geometrically Invariant Surface Descriptors, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 644–658.
[49] J. S. Duca, A. J. Hopfinger, Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367–1387.
[50] R. Todeschini, P. Gramatica, 3D-Modelling and prediction by WHIM descriptors. Part 5. Theory Development and Chemical Meaning of WHIM descriptors, *Quant. Struct.-Act. Relat.* **1997**, *16*, 113–119.
[51] D. R. Lide, CRC Handbook of Chemistry and Physics, 83rd Edition, National Institute of Standards & Technology , USA, CRC Press.
[52] H. Kubinyi, QSAR: Hansch Analysis and Related Approaches, in: *Methods and Principles in Medicinal Chemistry, Vol.1*, R. Manhold, P. Krogsgaard-Larsen, H. Timmermann (Eds.), VCH, Weinheim, **1993**, pp. 21–36.
[53] M. A. Johnson, G. M. Maggiora (Eds.), Concepts and Applications of Molecular Similarity, Wiley, New York, **1990.**
[54] P. M. Dean (Ed.), Molecular Similarity in Drug Design, Chapman & Hall, New York, **1995.**
[55] K. Sen (Ed.), Molecular Similarity I and II, *Topics Curr. Chem.* **1995**, 173–174
[56] J. M. Blaney, E. J. Martin, Computational approaches for combinatorial library design and molecular diversity analysis, *Curr. Opin. Chem. Biol.* **1997**, *1*, 54–59.
[57] P. Willett, Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, **1987.**
[58] R. D. Brown, Y. C. Martin, Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection, *J. Chem. Inf. Comput. Sci.* **1996**, *36,* 572–584.
[59] M. Karelson, V. S. Lobanov, A. R. Katritzky, Quantum-Chemical Descriptors in QSAR/QSPR Studies, *Chem. Rev.* **1996**, *96*, 1027–1044.
[60] P. E. Bowen-Jenkins, W. G. Richards, Quantitative Measures of Similarity between Pharmacologically Active Compounds, *Int. J. Quantum Chem.* **1986**, *30*, 763–768.
[61] G. Boon, W. Langenaeker, F. De Proft, H. De Winter, J. P. Tollenaere, P. Geerlings, Systematic Study of the Quality of Various Quantum Similarity Descriptors. Use of the Autocorrelation Function and Principal Component Analysis, *J. Phys. Chem. A* **2001**, *105*, 8805–8814.
[62] R. F. W. Bader, Atoms in Molecules: A Quantum Theory, Clarendon Press, **1990.**
[63] R. F. W. Bader, S. G. Anderson, A. J. Duke, Quantum Topology of Molecular Charge Distributions. 1, *J. Am. Chem. Soc.* **1979**, *101*, 1389–1395.
[64] P. L. A. Popelier, Atoms in Molecules: an introduction, H. Pearson, (Ed.), London, **2000.**
[65] P. L. A. Popelier, Quantum Molecular Similarity. 1. BCP Space, *J. Phys. Chem. A* **1999**, *103*, 2883–2890.
[66] S. E. O'Brien, P. L. A. Popelier, Quantum molecular similarity. Part 2: The relation between properties in BCP space and bond length, *Can. J. Chem.* **1999**, *77*, 28–36.
[67] S. E. O'Brien and P. L. A. Popelier, Quantum Molecular Similarity. 3. QTMS Descriptors, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 764–775.
[68] P. L. A. Popelier, P. J. Smith, Quantum Topological Atoms, in *Chemical Modelling: Applications and Theory*, Vol. 2, A. Hinchliffe (Ed.), Royal Society of Chemistry Specialist, Periodical Report, **2002**, pp. 391–448.
[69] S. E. O'Brien, P. L. A. Popelier , Quantum Molecular Similarity. Part 4: Anti-Tumour Activity of Phenylbutenones, *Perkin Trans. II* **2002**, 478–483.
[70] P. L. A. Popelier, U. A. Chaudry, P. J. Smith, Quantum Topological Molecular Similarity. Part 5: Further Development with an Application to the toxicity of Polychlorinated dibenzo-p-dioxins (PCDDs), *Perkin Trans.II* **2002**, 1231–1237.
[71] C. B. Mazza, N. Sukumar, C. M. Breneman, S. M. Cramer, Prediction of Protein Retention in Ion-Exchange Systems Using Molecular Descriptors Obtained from Crystal Structure, *Anal. Chem.* **2001**, *73*, 5457–5461.
[72] http://www.chem.rpi.edu/chemweb/recondoc/WinRecon.html
[73] A. Hinchliffe (Ed.), Chemical Modelling: Applications and Theory, Vol. 1, Royal Society of Chemistry, Cambridge, **2000.**
[74] P. G. Mezey, Theorems on Molecular Shape-Similarity Descriptors: External T-Plasters and Interior T-Aggregates, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1076–1081.
[75] P. G. Mezey, Shape Analysis, in *Encyclopedia of Computational Chemistry*, Vol. 4, P.v.R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. R. Kollman, H. F. Schaefer III, P. R. Schreiner (Eds.), John Wiley & Sons, Chichester, UK, **1999**, pp. 2582–2589.

[76] P. G. Mezey, Shape in Chemistry: An Introduction to Molecular Shape and Topology, VCH Publishers, New York, **1993.**
[77] P. G. Mezey, The Shape of Molecular Charge Distributions: Group Theory without Symmetry, *J. Comput. Chem.* **1987**, *8*, 462 – 469.
[78] P. D. Walker, G. A. Arteca, P. G. Mezey, A Complete Shape Group Characterization for Molecular Charge Densities Represented by Gaussian-Type Functions, *J. Compur. Chem.* **1990**, *12*, 220 – 230.
[79] P. G. Mezey, Z. Zimpel, P. Warburton, P. D. Walker, D. G. Irvine, D. G. Dixon, B. Greenberg, High-Resolution Shape-Fragment MEDLA Database for Toxicological Shape Analysis of PAHs, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 602 – 611.
[80] P. G. Mezey, Local and Global Similarities of Molecules: Electron Density Theorems, Computational Aspects, and Applications, European Congress on Computational Methods in Applied Sciences and Engineering, ECCOMAS 2000, Barcelona (11 – 14 September **2000**).
[81] G. A. Arteca, V. B. Jammal, P. G. Mezey, Shape Group Studies of Molecular Similarity and Regioselectivity in Chemical Reactions, *J. Comput. Chem.* **1988**, *9*, 608 – 619.
[82] A. Lawson, Organic reaction similarity in information processing, *J. Chem. Inf. Comput. Sci.* **1992**, 32, 675 – 679.
[83] R. Ponec, M. Strnad, Similarity ideas in the theory of pericyclic reactivity, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 693 – 699.
[84] G. Sello, Reaction prediction: the suggestions of the Beppe program, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 713 – 717.
[85] J. Gasteiger, W. D. Ihlenfeldt, R. Fick, J. R. Rose, Similarity concepts for the planning of organic reactions and syntheses, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 700 – 712.
[86] Y. C. Martin, Diverse Viewpoints on Computational Aspects of Molecular Diversity, *J. Comb. Chem.* **2001**, *3*, 231 – 250.
[87] R. O. Duda, P. E. Hart, Pattern Classification and Scene Analysis, John Wiley, New York, **1973.**
[88] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, **1990.**
[89] S. Haykin, Neural networks. A comprehensive foundation, Macmillan/IEEE Press, **1994.**
[90] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Adv. Drug Deliv. Rev.* **1997**, *23*, 3 – 25.
[91] D. K. Agrafiotis, V. Lobanov, F. Salemme, Combinatorial informatics in the post-genomic era, Nature Rev. **2002**, www.nature.com/reviews/drugdisc
[92] M. D. Barratt, J. V. Castell, M. Chamberlain, R. D. Combes, J. C. Dearden, J. H. Fentem, I. Gerner, A. Giuliani, T. J. B. Gray, D. J. Livingstone, W. McLean Provan, F. J. J. A. L. Rutten, H. J. M. Verhaar, P. Zbinden, The Integrated Use of Alternative Approaches for Predicting Toxic Hazard The Report and Recommendations of ECVAM Workshop 8, http://altweb.jhsph.edu/publications/ECVAM/ecvam08.htm
[93] A. Burger, Isosterism and bioisosterism in drug design, *Prog. Drug. Res.* **1991**, *37*, 287 – 371.
[94] G. A. Patani, E. J. LaVoie, Bioisosterism: A rational approach in drug design, *Chem. Rev.* **1996**, *96*, 3147 – 3176.
[95] H. Kubinyi, Similarity and Dissimilarity – A Medicinal Chemist's View, in *3D QSAR in Drug Design. Volume II. Ligand-Protein Interactions and Molecular Similarity*, H. Kubinyi, G. Folkers, Y. C. Martin (Eds.), Kluwer/ESCOM, Dordrecht, **1998**, pp. 225 – 252; also published in: *Persp. Drug Design Discov.* **1998**, 9/10/11, 225 – 252.
[96] H. Kubinyi, Chemical Similarity and Biological Activity, 3rd Workshop on Chemical Structure and Biological Activity: Perspectives on QSAR 2001 (November 8 – 10, **2001**) Sao Paolo, Brazil, http://arara.iq.usp.br/l6.htm
[97] H. Kybinyi, Chemical Similarity and Biological activity. Hugo Kubinyi Lectures, http://home.t-online.de/home/kubinyi/dd-06.pdf
[98] ICCA Workshop "(Q)SARS For Human Health And The Environment: Workshop on Regulatory Acceptance, Setubal, Portugal, March 4 – 6, **2002.**
[99] T. Potter, H. Matter, Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases, *J. Med. Chem.* **1998**, *41,* 478 – 488.
[100] R. A. Lewis, J. S. Mason, I. M. McLay, Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599 – 614.
[101] R. D. Brown, Y. C. Martin, The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1 – 9.
[102] G. M. Downs, P. Willett, W. Fisanick, Similarity Searching and Clustering of Chemical Structure Databases using Molecular Property Data, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094 – 1102.
[103] S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, R. P. Sheridan, Chemical Similarity Using Physiochemical Property Descriptors, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118 – 127.
[104] M. Dash, H. Liu, Feature Selection for Classification, *Intell. Data Anal.* **1997**, *1*, 131 – 156.
[105] M. Hall, Correlation-based feature selection of discrete and numeric class machine learning, in *Proceedings of the International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, CA, **2000**, pp. 359 – 366.
[106] P. M. Narendra, K. Fukunaga, A branch and bound algorithm for feature selection. *IEEE T. Comp.* **1977**, *C-29*, 917 – 922.
[107] K. Kira, L. A. Rendell, The feature selection problem: Traditional methods and a new algorithm, in *Proceedings of Ninth National Conference on Artificial Intelligence*, **1992**, pp. 129 – 134.
[108] I. Kononenko, Estimating attributes: Analysis and extension of RELIEF, in *Proceedings of European Conference on Machine Learning*, Morgan Kaufmann, **1994**, pp. 171 – 182.
[109] R. Kohavi, D. Sommerfield, Feature subset selection using the wrapper method: Overfitting and dynamic search space topology, in *Proceedings of First International Conference on Knowledge Discovery and Data Mining*, Morgan Kaufmann, **1995**, pp. 192 – 197.
[110] R. J. Hilderman, H. J. Hamilton, Heuristic measures of interestingness, in J. Zytkov, J. Rauch (Eds.), *Proceedings of the 3rd European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'99)*, **1999**, pp. 232 – 241.
[111] S. K. Lin, Molecular diversity assessment: logarithmic relations of information and species diversity and logarithmic relations of entropy and indistinguishability after rejection of Gibbs paradox of entropy mixing, *Molecules* **1996**, *1*, 57 – 67.
[112] D. K. Agrafiotis, On the Use of Information Theory for Assessing Molecular Diversity, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 576 – 580.
[113] R. D. Cramer, D. E. Patterson, R. D. Clark, F. Soltanashahi, M. Lawless, Virtual Compound Libraries: A New Approach

to Decision Making in Molecular Discovery Research, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1010 – 1023.
[114] Y. C. Martin, R. D. Brown, M. G. Bures, Quantifying diversity, in: E. M. Gordon, J. F. Kerwin Jr. (Eds.), *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*, Wiley, **1998,** pp. 369 – 385.
[115] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, Weinberger Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors, *J. Med. Chem.* **1996**, *39*, 3049 – 3059.
[116] R. D. Clark, R. D. Cramer, Taming the combinatorial centipede, *CHEMTECH* **1997,** *27*, 24 – 30.
[117] W. M. Meylan, P. H. Howard, R. S. Boethling, D. Aronson, H. Printup, S. Gouchi, Improved Method for Estimating Bioconcentration / Bioaccumulation Factor from Octanol/ Water Partition Coefficient, *Environ. Toxicol. Chem.* **1996**, *18*, 664 – 672.
[118] M. Skvortsova, I. Baskin, Molecular Similarity. 1. Analytical Description of the Set of Graph Similarity Measures, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 785 – 790.

Received on April 3, 2003; Accepted on July 31, 2003