Managing molecular diversity

Juan J. Perez

Received 24th August 2004
First published as an Advance Article on the web 10th January 2005
DOI: 10.1039/b209064n

The present work provides an overview of the different methods used in molecular diversity analysis. Issues like identifying voids in proprietary databases, reducing the number of redundancies present in databases, or designing focused libraries by grouping compounds similar to a template with the aim to fine tune its properties, are potent diversity analysis tools that may be used to optimize molecules based on their properties and specifically, to speed up the process of lead discovery and optimization. The present work describes first methods that are used to describe molecular systems. This is followed by a section devoted to describe different measures of similarity between molecules, to finish with a description of different methods used to select subsets molecules according to the constraints imposed. The final section deals with the validation of these methods, based on different studies available in the literature.

Introduction

Diversity is a concept used in different contexts to measure the extent of objects with differential features in a set. Alternatively, diversity can be qualitatively used to assess the odds of finding new members with differential features in a set. The concept has traditionally been used in biology to indicate the number of different species living in an ecosystem. Indeed, biodiversity has been a topic of very wide concern for many years. Within this context, it can be considered that Noah used diversity criteria to fulfil his commitment to gather different animals to travel with him in the Ark, by selecting the least number possible of animals and at the same time, covering the maximum number of different species.

In chemistry, molecular properties as well as biological activities of compounds can be related to their molecular structure. Thus, molecular diversity methods focus on the evaluation of the extent of molecules with differential structural features in a library or database implicitly assuming, according to the structure–activity paradigm, that this process also distinguish the diversity of the library in terms of the properties of the species as well as on their biological activities. Accordingly, molecular diversity can be used in qualitative terms to differentiate among a bunch of databases of chemical compounds, which is the one that exhibits the larger variety of compounds. Moreover, diversity of a database could also be used in quantitative terms, provided that we can compare with a reference database, containing molecules with all possible features molecules could exhibit. If we had such a reference library, diversity analysis performed on a database would provide the characteristics of new compounds needed to be added to a database in order to increase its diversity. Construction of a universal library is of paramount importance in many areas. Specifically, in the field of bioactive compounds, screening a universal library may represent an enormous benefit in the drug discovery process, since such a library would have the capability of identifying a new hit, any time it is screened against a new therapeutic target. Unfortunately, there is no simple way to construct such a reference database exhibiting the highest diversity possible, also known as universal library, although approximations can be pursued. Intuitively, it can be thought that a universal library could be constructed by adding new compounds to a database until the addition of a new compound does not increase its diversity (*i.e.* saturation is reached). However, this procedure is foreseen unpractical, since there are more potentially useful molecules than there are atoms in the universe.[1] Accordingly, in order not to miss any useful compound with special properties, methods are required in order to carry out diversity analyses and to design tools to eliminate redundant molecules from them, as well as to provide guidance to the process of adding new compounds to enrich the set.

Practical solutions to diversity analysis come from the concept of chemical space. In geometrical terms, molecules can be represented by points in a space whose coordinates depend on the values of selected descriptors or features, and where the diversity of the set can be assessed by the way points are

Juan Jesus Perez

Juan Jesus Perez. Born in Madrid, 24-12-1955. Full Professor Physical Chemistry, Dept. of Chemical Engineering, Barcelona School of Engineering. Technical University of Catalonia, MSci, Physical Chemistry, University of Barcelona, 1977. PhD, University of Barcelona, 1982. Postdoctoral training: UMIST, UK, 1983; Max-Planck Institut fur Physik, Munich, 1983. Visiting scientist, IBM Labs, Kingston NY, 1987–88. Senior researcher, Molecular Research Institute, Palo Alto CA, 1991–93. Awarded in 1999 with the Salvat Foundation award for Scientific Achievement distributed in this space. Considering this geometrical picture of a database, a universal library can be viewed as a set of points evenly distributed, with no voids in it. Diversity analysis consists of a set of tools that permit to define chemical spaces (coordinate axes) and to measure the distance between points, providing a photograph of the database, useful for eliminating redundancies and selecting new compounds to increase the diversity of the set. Fig. 1 shows pictorially how a database is represented in a chemical space through the definition of the coordinate axes (features or descriptors). It should be pointed out though, that chemical spaces are defined once a procedure to describe molecules has been selected. Thus, upon the selection of descriptors used for this purpose being either structural or property-related, the chemical space can be different and diverse results from the diversity analysis can be obtained for the same database. Moreover, not only the definition of the space, but the use of different metrics to measure the distance between objects in a specific space can be used. Accordingly, the saturation requirements of the database can differ upon the way the space is defined and the metrics used. It should be stressed however, that due to the fact that the chemical spaces are geometrical models of databases, a saturated one represents only an approximation of a universal database within the chemical space selected.

In medicinal chemistry, diversity has shown to be a valuable concept that can be used to alleviate the process of designing new drugs.[2,3] The process of bringing a new drug onto the market is long and costly, taking an average time of 12 years with an associated cost of 650 M€.[4] The design of a new drug begins at the step of finding compounds that bind onto a specific biological target (process known as lead discovery), followed by the process of lead optimisation, where a selected compound is modified with the aim of improving its pharmacological profile, as well as to remove any possible toxicological effects. Afterwards, compounds are subject to clinical trials, where their safety and efficacy are thoroughly investigated. Diversity assessment methods can be used to characterise the limitations and capabilities of databases of compounds used for screening. Traditionally, the databases most widely used for screening against new targets have been proprietary databases, that consist of lists of compounds that every pharmaceutical corporation has accumulated during years, these are collections of products available from stock or that can be easily synthesised. However, more recently, combinatorial chemistry methods[5,6] represent a new source to design chemical libraries. In this procedure the same synthetic scheme is used to attach different moieties to a scaffold, enabling a parallel synthesis of the different compounds to be performed by taking advantage of the different reagents available. Following this methodology, if a scaffold has different substituent attachment points, the number of compounds that can be synthesised following this strategy grows rapidly. However, actual synthesis of all possible compounds requires an enormous effort that may not be necessary. In order to cut down the number of compounds selected for synthesis, procedures designed to assess the diversity of a database, provide a guide to select the least number of compounds necessary to cover the diversity of the whole library, eliminating *redundant* molecules. Two different types of combinatorial libraries can be designed of complementary usefulness at different stages of the drug discovery process. On the one hand, diverse libraries, aimed at identifying new leads, designed to contain the least number of compounds covering as many as possible different profiles. In this case, diversity methods are used to select a few compounds that represent the whole library, keeping the chances of finding new leads with a lower economical investment. On the other hand, focused libraries are useful in the process of lead

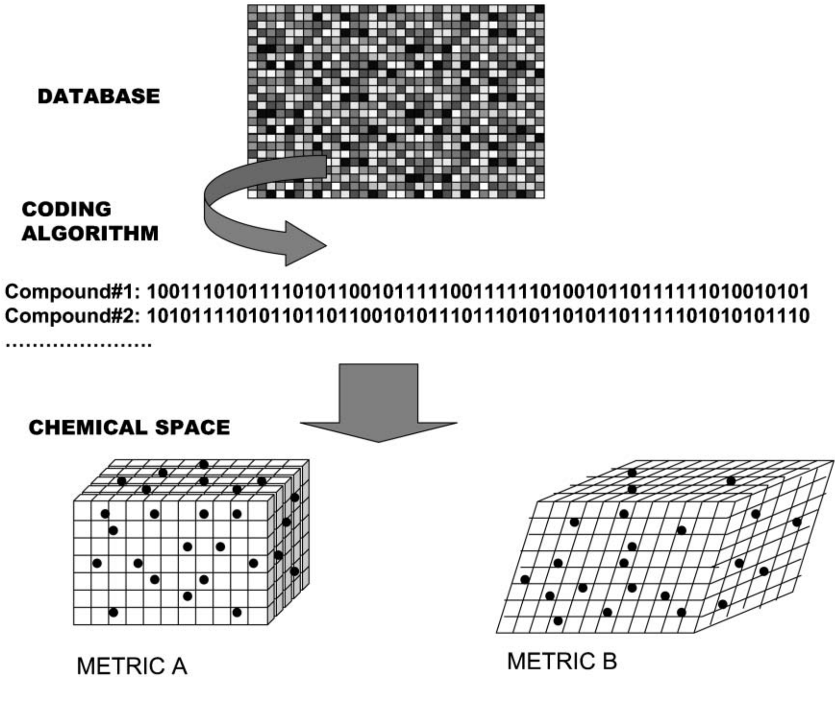
DATABASE
CODING
ALGORITHM
Compound#1: 10011101011110101100101111100111111010010110111111010010101
Compound#2: 1010111101011011011001010111011101011010110111110101010101110
....
CHEMICAL SPACE
METRIC A
METRIC B

Fig. 1 Process of transforming a database into a chemical space.

optimisation. In this case diversity is used to select molecules from the library that group together and consequently are candidates that bind to the same target.

Measuring the diversity of a set of compounds involves computation of the (dis)similarity between all the pairs of molecules in the set. Measures of similarity can be carried out either by comparing the properties of the molecules or by comparing their structures. The average value of the dissimilarity of the members of a set can be used as a measure of the diversity of a set:

$$div(\mathrm{A}) = \frac{\sum_{\mathrm{i}=1}^{N} \sum_{\mathrm{j}=1,\mathrm{i}\neq \mathrm{j}}^{N} diss(\mathrm{i,j})}{N(N-1)}$$

where *div*(A) is a measure of the diversity of a set A, and *diss*(i,j) is a measure of the dissimilarity between objects i and j. Dissimilarity is a complementary measure of similarity, so that if a measure of similarity between two objects is defined between 0 and 1, its dissimilarity is simply defined as (1 − similarity). According to this definition, the closer *div*(A) is to 1, the more diverse is the set investigated.

In geometrical terms the process assessment of the diversity of a set consists in understanding how far points are from each other, according to a similarity criterion or metric. Moreover, measuring the distance between molecules in a chemical space permits to group them according to neighborhood criteria and cconsequently according to their expected properties. Depending on the analysis required, after grouping the molecules the result can be used to select only a representative molecule of each group, to generate the smallest most diverse set within the limitations of the diversity of the initial set, or simply selecting all the members of the same subgroup if we are seeking molecules with common properties.

There is though an additional issue associated with any measure of diversity. This concerns the different possible ways to describe a molecular system. This largely influences the metrics used to measure similarity between molecules and also, the nature of the chemical space in which molecules are represented.[7] A molecule can be unequivocally represented by its electron density or by the coordinates and atomic numbers of the constituent atoms. However, molecules can also be represented by the properties they exhibit. Accordingly, instead of describing molecules by their essence, they are described by listing their own attributes. Molecular properties used to describe molecules and further, to discriminate them in a set, are called descriptors. The simplest descriptors of a molecule are its bulk properties. These, so-called one-dimensional descriptors, may include either structural or physicochemical properties of the molecules like: the octanol/water partition coefficient (logP(o/w)), molecular weight, molecular refractivity, dipole moment, polarisability, *etc*… Rationale for using this type of descriptors comes from the experience acquired in the last fifty years on QSAR studies.[8] A more sophisticated way to describe molecules can be done using structural fragments or topological indices as descriptors. These are called two-dimensional descriptors since they can be deduced from the chemical formula of a molecule. These descriptors are molecular fragments that can be either predetermined or generated from the analysis of the molecular structure that is being inspected. There are also three-dimensional descriptors providing molecular information in the context of the 3D distribution of chemical groups, described either using molecular field analysis or pharmacophoric features. Comprehensive discussion about molecular descriptors used in diversity analysis has been discussed elsewhere.[9,10] A set of molecular descriptors defines the chemical space, whose dimension corresponds to the number of descriptors included. For feasibility reasons of handling large sets of compounds, it is desirable to consider the minimum number of descriptors, selecting only the most significant in order to have the analysis procedure as simple a possible. Principal Component Analysis (PCA) is usually the technique used to reduce the number of variables, by selecting the minimum set necessary to provide an adequate description of the molecules.

Molecular representation

Actual information regarding values of the different descriptors of the molecule, can be conveniently stored in a bit string, usually called *fingerprint* since it represents a code that identifies the molecule. A bit string stores information of 1D, 2D, 3D descriptors or a combination of the different categories. The process of encoding the information into a bit string can be done by binning property value ranges of the different descriptors selected in intervals. In this way a bit is associated to a small range of a property and the information encoded simply indicates the presence or absence of a specific feature or value range, that can be either structural or a bulk property. For example, suppose that we are encoding the hydrophobicity index logP. One possibility is to bin the property in different intervals: [0,1), [1,2), [2,3), [3,4), [4,5), 5 or more. To each of these intervals a different bit of the string is assigned, in such a way that if the molecule exhibits a logP of 4.5, the first, second, third, fourth and sixth bits will be '0', whereas bit number 5 will be '1'.

A completely different scheme encodes actual molecular features into bit strings using an algorithm to compact the information. This encoding procedure, called hashing, is designed in such a way that similar structures exhibit similar hashed fingerprints, and consequently, a large enough number of bits needs to be considered in order to avoid dissimilar structures exhibiting the same fingerprint. This way of processing information has its origins in the representation of chemical formulas either using systematic nomenclature like the IUPAC or using line notation, like the popular SMILES,[11] where molecules are described in a line including all the constitutive atoms and connectivities.

A third approach to encode a molecular structure into a bit string consists in enumerating the actual molecular features using an expert system, where encoding is done without hashing. This procedure uses larger amounts of computer resources, and it is advised for medium size data sets.[12]

There are several fingerprints proposed in the literature to describe molecular systems, most of them make use of 2D descriptors. These classes of descriptors are defined exclusively on the information available in a chemical formula, and consequently describe features that an expert will deduce from its inspection. In a rapid look at a chemical formula an expert retrieves information about the chemical groups present in the molecule. Accordingly, key descriptions used in the literature to describe a molecular structure consist of analysing the presence or not of different chemical groups from a dictionary. This procedure can be easily translated to a bit string, and examples of its implementation are the MACCS keys or in the BCI fingerprints. More sophisticated is the information stored in the MDL keys. In this case, stored information concerns atoms and their environment in the form of topological features, like the distance in number of bonds between two different groups.

There are more sophisticated ways to describe a molecule using 2D descriptors. Indeed, from the mathematical point of view, a chemical formula can be viewed as a graph with nodes (atoms) and edges (bonds). This information can be stored in the form of a connection table that lists the characteristics of the different atoms as well as their connections. Accordingly, there have been proposed procedures to describe a molecule by enumerating all bond paths through it.[13] Starting with paths of zero length (atoms), paths of length one (one atom and a bond), up to paths of length seven. This information can also be stored in a bit string providing a fingerprint of the molecule, however, in this case, information is encoded through a hashing procedure and consequently the same bit of two different molecules can account for distinct structural features. Examples of these type of fingerprints are the Daylight or the UNITY fingerprints.

3D descriptors involve storing properties associated with the 3D atomic distribution of the molecule. One of these approaches involves encoding the information obtained by molecular field analysis. CoMFA (comparative molecular field analysis) is an appealing way to compare molecular fields.[14] In this procedure steric, electrostatic and hydrophobic fields are compared using partial least square analysis to identify the characteristics and distribution of the portions of the molecular fields that contribute to the activity of the compounds in regard to a target. One of the most serious problems of this procedure concerns the alignment of molecules. In the case where there is a common core in the series, then it can be used to superimpose the different structures. However, if this is not possible, the use of the molecular moments of inertia or quadrupole moments are alternative procedures. Another problem concerns molecular flexibility. There are several procedures proposed in the literature to deal with this problem. One of these procedures consists of averaging the molecular field of a set of conformations before comparing the molecular fields of two molecules. Another procedure consists of identifying the conformations of both molecules that produce a molecular field most similar to that of the other molecule to which the comparison is to be done. The most successful procedure is to project the four-dimensional conformational space into a three-dimensions by using a rule-based algorithm to generate characteristic conformations of the molecule. In the case of assessing the diversity of a set of compounds, it is interesting to compare common substructures. This is called topomeric CoMFA analysis.[15] This type of comparison provides the possibility of comparing a wide range of structurally diverse compounds that locally exhibit similar molecular fields, and are then supposed to exhibit similar activities. Specifically, this is a rational way to recognize the role played by bioisosteres.[16]

3D molecular information can also be described in the form of pharmacophoric features.[17] A pharmacophore is a schematic representation of steric and chemical features of a molecule that may be relevant for its recognition by a receptor. Pharmacophoric features include hydrogen accepting centres, hydrogen donor centres, basic and acid centres, aromatic centroids, lipophilic regions. All possible two, three or four pharmacophores can be coded into a bit string, and each molecule is stored with a '1' in the corresponding position if the pharmacophore is fulfilled and with a '0' if it is not. In this procedure, pharmacophores are coded according to the distance between pharmacophoric features using a binning procedure of discretisation. For example, in the ChemDiverse/DiR method distance between two pharmacophoric features is binned as follows: assign the first bit if distance is shorter than 1.7 Å; between 1.7 and 3.0 Å assign a bit to each distance from 1.7 in increments of 0.1 Å; between 3.0 Å and 7.0 Å assign a bit to each distance from 3.0 in increments of 0.5 Å; between 7.0 Å and 15 Å assign a bit to each distance from 1.0 Å in increments of 1.0 Å; finally assign the next bit if the distance is larger than 15 Å. Combining all possible distances with the possible pharmacophoric features defines all the possible pharmacophores. In this procedure conformational analysis is required, to search whether a molecule fulfils different pharmacophores from the different conformations attainable. Fig. 2 shows pictorially the different ways used to describe a molecular system.

In an attempt to reduce the dimensionality of the chemical space, there have also been described in the literature fingerprints combining 1D, 2D or 3D descriptors. Recently, the BCUT descriptors (Burden–CAS–University of Texas, after the origin of their definition), have been developed in the context of receptor-relevant subspace concept, and with the constraint of generating descriptors that work in a low dimensional chemical space.[18] Combining the three classes of descriptors, each BCUT condenses a large amount of molecular structure and property information into a single number. The properties integrated are relevant to receptor affinity including, atomic polarisability, atomic charge, and atomic hydrogen-bond donor and receptor ability. BCUTs are the highest and the lowest eigenvalues of square matrices, including property information in the diagonal elements and distance-related information in the off-diagonal elements. Various scaling factors are also incorporated for both the diagonal and off-diagonal components. Generally, many BCUT descriptors are calculated for a set of compounds, and the subset of BCUTs that provides the best separation between the compounds is selected using a chi-squared algorithm. This subset, usually 4–6 BCUT descriptors, defines a low dimension space.

Measures of similarity

Different procedures have been described in the literature to measure molecular similarity, choosing one or the other depends largely on the way molecules are described.[19] There are measures known as *distance measures*, where similarity is

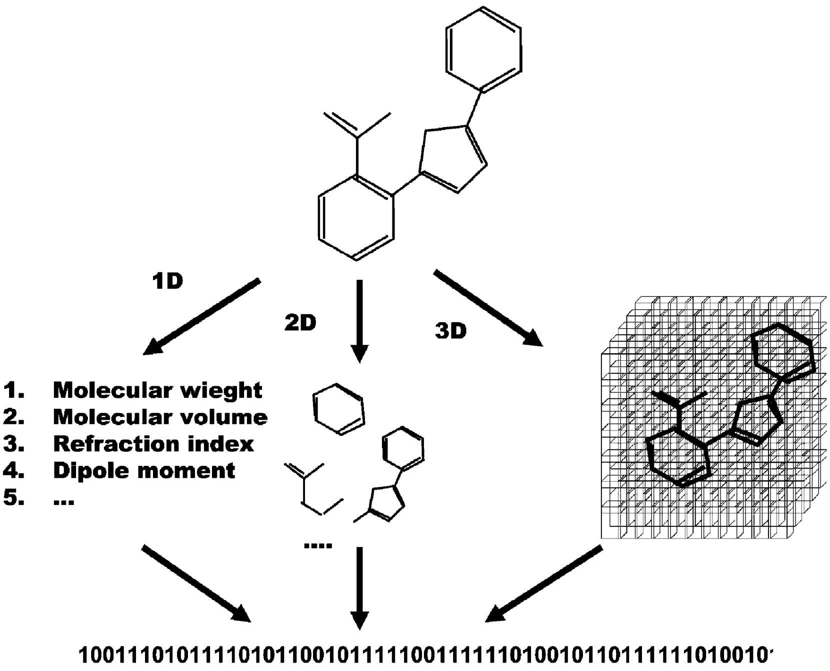

Fig. 2 Diverse procedures to encoding molecules.

expressed as the Euclidean distance between a pair of objects in a chemical space. Other similarity measures include similarity indices, based on a comparative analysis of the presence or absence of certain features between the two molecules. These are known as *association measures*. Other measures are based on the computation of the statistical significance of correlations between two sets of variables. These are known as *correlation measures*. Finally, measures based on the occurrence of observed features in datasets are known as *probabilistic measures*.

3D molecular similarity

A molecule can be described by the coordinates and atomic numbers of all constituent atoms. This information can be used to compute an approximate wavefunction or electron density and from it, all the properties of a molecule. Accordingly, the different features of the molecules are embedded in its electron density and consequently, similarity between molecules can be measured from its direct comparison. Similarity can be measured by computing the overlap integral between the electron densities (ρ) of the two molecular systems and in general, between the values of any property P measured on both molecules: PA and PB. This is in fact a measure of dissimilarity, since the more similar two molecules are, the larger is the overlap integral. Furthermore, since the overlap integral has the properties of an internal product between vectors, a similarity index can be defined. The most popular of the indices described in the literature is the Carbo similarity index (CAB),[20] computed by dividing the overlap matrix by root square of the density matrices of the molecules compared:

$$C_{\mathrm{AB}} = \frac{\int \rho_{\mathrm{A}}\rho_{\mathrm{B}}\mathrm{d}\tau}{\left[\int \rho_{\mathrm{A}}^{2}\mathrm{d}\tau \int \rho_{\mathrm{B}}^{2}\mathrm{d}\tau\right]^{1/2}}$$

The Carbo similarity index is, as a matter of fact, the cosine of the angle between the electron densities of the two molecules compared. Computing the similarity between two molecules has two difficulties associated, the first concerns the approximate nature of the electron density used to perform such calculations, and second, the dependence of the overlap matrix on the way the two molecules are superimposed.

Use of the overlap integral offers the advantage of using continuous measures, however the measure can be done through a discretisation procedure of the space. This can be useful when properties need to be calculated using a grid of points, like molecular fields. A correlation procedure to measure the similarity between molecules includes the Spearman rank correlation coefficient, being mostly applied for studies comparing the molecular electrostatic potential or the accessible surface:

$$S_{\mathrm{AB}} = 1 - \frac{6\sum_{\mathrm{i}=1}^{n} d_{\mathrm{i}}^{2}}{n^{3}-n}$$

where d_{i} is the difference in the property ranking at point i of the two structures, and n is the total number of points over which the property is measured.

Another useful way to compare molecules is through the measurement of the Euclidean distance, that is the root-mean square deviation of a property P, that is compared between molecules A and B, summed to all points of the grid:

$$rmsd = \sqrt{\frac{\sum\left(P_{\mathrm{Ai}}^{2} - P_{\mathrm{Bi}}^{2}\right)}{N}}$$

Specifically, comparisons regarding the similarity of molecules can be carried out using the atomic coordinates only. Obviously, this can only be applied when all the atoms are present in the two molecules, and consequently this method is particularly suitable to compare different conformations of a molecule. However, this method can be very useful if we want to compare the similarity between two objects in regard to their fulfilling the requirements of a specific template, for example a pharmacophore, or to compare how similar are two molecules in regard to a substructure. Pharmacophores provide a reasonable framework to measure the similarity between two molecules in regard to their performance at being recognized by a target receptor. The basic idea is to assume that two molecules that exhibit a similar binding behavior towards a target receptor fulfil a common pharmacophore that defines its characteristic features. However, some caution must be taken into account, since in some cases different compounds eliciting the same action on a receptor bind differently, and consequently they do not necessarily fulfil the same pharmacophoric requirements. Calculation of the rmsd can also be extended to a ppropherty *P* that is compared between molecules A and B at different points of the 3D space.

2D molecular similarity

Measures of similarity can be also defined regarding the topology of the molecules. For example, a probabilistic similarity measure can be defined by considering the number of times that a certain bond type (connectivity) appears in the two molecules. If N_i, N_j are the number of times that a bond type appears in molecules i and j, and N_ac is the number of times it appears in the maximal totally connected subgraph identified by comparing structures i and j, a distance (S_ij) can be define as:[21]

$$S_\mathrm{ij} = \frac{1}{2}\left(\frac{N_\mathrm{i}}{N_\mathrm{ac}} + \frac{N_\mathrm{j}}{N_\mathrm{ac}}\right)$$

For example, the distance between benzene and naphthalene, considering the aromatic bond is: S_ij = 0.5 (6/11 + 11/11) = 0.77.

In the case of fingerprints, distance measures can be done using a distance similarity or association measures. Distance measures are more suitable for physical property data. When the comparison is to be made between bit strings whose bits are assigned to a specific feature being observed or not , the presence or absent of that bit can give us an idea of the similarity between the two molecules when this is summed up to all the bits of the fingerprints. One of the most widely used measure is the Tanimoto index, defined as follows:

$$d_\mathrm{AB} = \frac{C}{A+B-C}$$

where *A* is the number of '1' in the bit string representing molecule A, *B* is the number of '1' in the bit string representing molecule B, and *C* is the number of bits that are filled simultaneously for molecules A and B. Other similar measures include the Hamming distance, defined as the number of bits which are different between the two bit sets. For binary keys the Euclidean distance is the square root of the Hamming distance. Fig. 3 shows pictorially the procedure for measuring the distance between molecules A and B using different metrics.

Classification methods

Selecting the subset of molecules of size *n* from a database of size *N* requires the evaluation of the combinations of *n* elements chosen from a larger database of *N* elements:

$$\frac{N!}{n!(N-n)!}$$

a number big enough to design methods to select the compounds without evaluating all the possible subsets (to select a subset of 10 compounds from a database of 100 compounds there are about 20 billion compounds). To classify molecules in different classes according their similarity, three different strategies have been

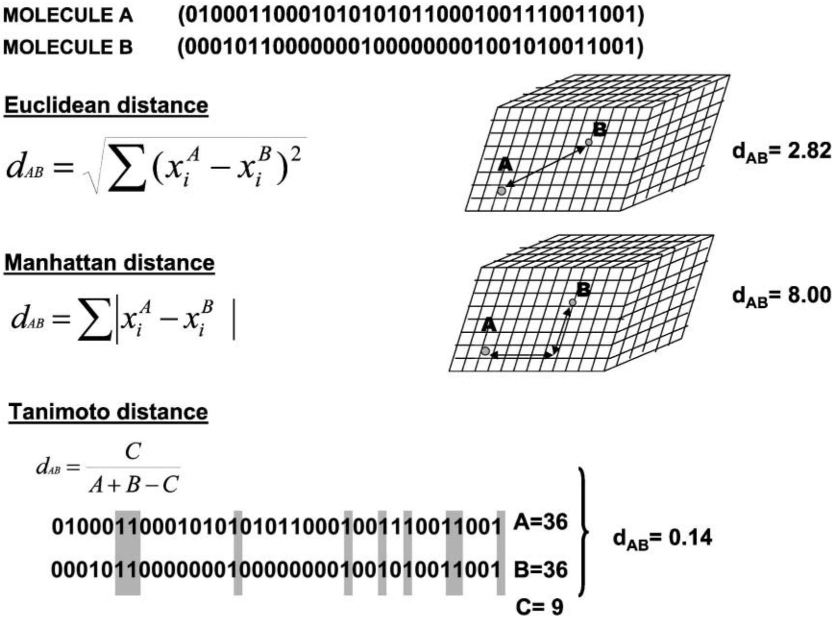
MOLECULE A (010001100010101010110001001110011001)
MOLECULE B (000101100000000100000000100101001 1001)
Euclidean distance
$d_{AB} = \sqrt{\sum (x_i^A - x_i^B)^2}$
A
B
d_{AB}= 2.82
Manhattan distance
$d_{AB} = \sum \left| x_i^A - x_i^B \right|$
A
B
d_{AB}= 8.00
Tanimoto distance
$d_{AB} = \frac{C}{A+B-C}$
010001100010101010110001001110011001 A=36
000101100000000100000000100101001 1001 B=36
C= 9
d_{AB}= 0.14

Fig. 3 Different procedures for measuring the distance between two molecules.

proposed: cluster-based selection, partition-based selection, dissimilarity-based selection methods and optimisation techniques.[22]

Cluster-based methods

Cluster analysis is a process for dividing a set into subsets or clusters, where objects share a certain degree of similarity and at the same time elements in different clusters are dissimilar. There are several procedures to carry out the clustering process. Basically, the procedures can be classified into hierarchical methods and non-hierarchical methods. The use of different methods depends basically on the size of the database, as will discussed later.

Hierarchical methods are bottom-up procedures, and are iterative procedures where the two nearest clusters are joined at each step to form a single, larger cluster. Initially, each of the m molecules in the set is treated as an individual cluster, and after n iterations, the number of clusters created is $m - n$. The procedure can be followed up to fit all the molecules in one cluster. In the clustering process a dendogram can be constructed to display the structure the of the data set. These kind of methods are called hierarchical because they classify the molecules in a bottom-up process, in such a way that dendograms can be cut in a prefixed threshold value to set the number of clusters of the set, requiring a termination criterion to stop after an appropriate number of clusters has been created. There are statistical tests that measure the probability for the existence of any particular number of clusters, but frequently no clear-cut optimum can be determined. A more efficient procedure consists of adding a training set of compounds to the initial set, to act as tracers of the clustering process. This can be as simple as adding a few compounds, some of them active and some inactive, for a target. The number of clusters can be determined at the time these two sets are discriminated.

Hierarchical clustering methods differ mainly in the criterion used to create the clusters. In the single linkage method, the similarity criterion selects the shorter distance between objects. Alternatively, if the longest distance between two objects is the criterion used, the method is called the complete linkage method. Finally, if the average similarity criterion is used, the method is called the group average linkage method. Fig. 4 shows schematically the flux diagram of this procedure.

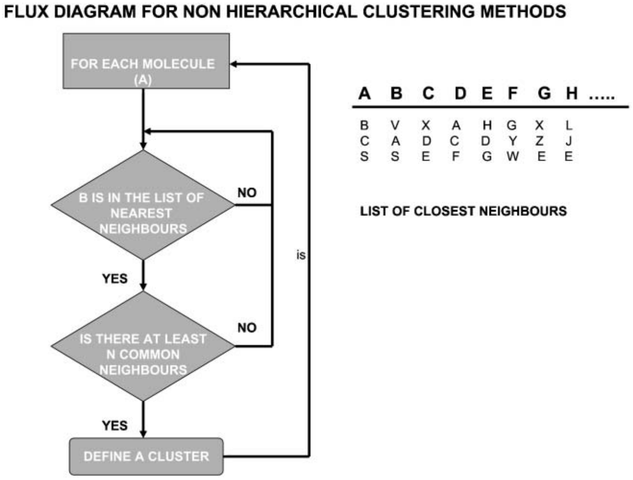
FLUX DIAGRAM FOR NON HIERARCHICAL CLUSTERING METHODS
FOR EACH MOLECULE (A)
B IS IN THE LIST OF NEAREST NEIGHBOURS
NO
YES
IS THERE AT LEAST N COMMON NEIGHBOURS
NO
YES
DEFINE A CLUSTER
is
A B C D E F G H .....
B V X A H G X L
C A D C D Y Z J
S S E F G W E E
LIST OF CLOSEST NEIGHBOURS

Fig. 5 Flux diagram of a non-hierarchical clustering method.

Non-hierarchical methods do not generate a tree structure. Generally, automatic determination of the cluster boundaries is a major advantage of these methods, compared to the hierarchical methods. Nevertheless a non-trivial parameter setting is usually required that reflects some prior knowledge of the space. Nearest neighbours methods are commonly used under this title, the Jarvis–Patrick clustering method being a typical representative. The method consists in grouping all the nearest elements of the set, and clustering proceeds in such a way that only mutual neighbours are grouped together to form a cluster. Accordingly, for every element in the group one has to list all the elements that are at a distance less than a certain threshold. Two elements belong to the same cluster if they are in the neighbour list of each other, and if comparison of the lists of nearest neighbours of both elements permits the identification of a few common elements. One of the main difficulties associated with this method is the impossibility of specifying the number of clusters required. Fig. 5 shows schematically the flux diagram of these kind of methods.

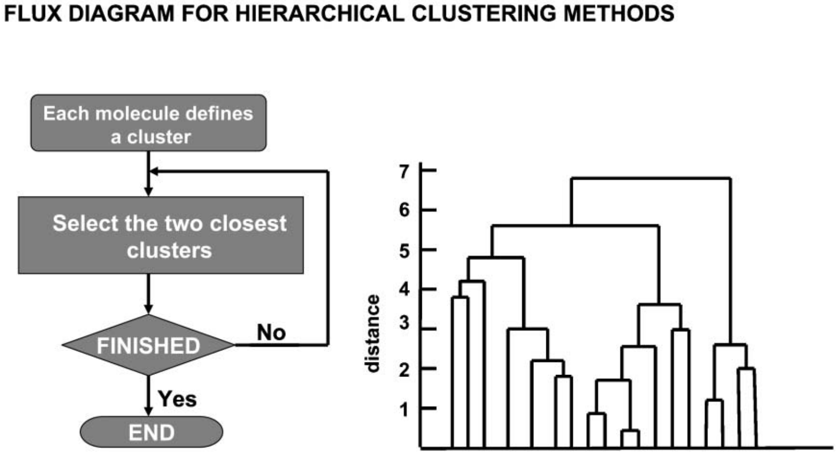
FLUX DIAGRAM FOR HIERARCHICAL CLUSTERING METHODS
Each molecule defines a cluster
Select the two closest clusters
FINISHED
No
Yes
END
distance
7
6
5
4
3
2
1

Fig. 4 Flux diagram of a hierarchical clustering method.

Clustering methods reveal a natural partitioning of the dataset. They are appropriate for high dimensionality data, but limited to the treatment of small databases. Adding new compounds to the subset requires initialising the procedure from scratch.

Partitioning methods

These methods represent a natural procedure of partitioning the chemical space. As mentioned before, each dimension of the chemical space is a descriptor, and the individual compounds are points in this space whose coordinates are the values of each of the different descriptors. Each dimension is divided into bins according to the range of the property. This binning further defines a grid of bins or cells within the chemistry space. Several cell-based diversity measures have been proposed in the literature, ranging from simple occupancy counts to entropy measures, χ^2 values, and other metrics. The smallest most diverse subset is then selected by choosing one molecule of each cell. Partition-based methods are especially useful for comparing different compound populations and useful for identifying diversity voids, *i.e.* cells not occupied.

Partitioning methods have the advantage to identify voids in a database. New compounds can be easily added in the analysis of the set. However, a strong limitation regards the arbitrariness of cell boundaries. It is a fast method, but restricted to low elements of a low dimensional space.

Maximum dissimilarity-based selection

The maximum dissimilarity-based selection procedure is based on the identification of a subset of compounds comprising the *n* most dissimilar molecules in a database containing *N* molecules (where, typically, $n \ll N$).[23] The procedure in a first step initialises the subset by transferring to it a compound from the database. In a second step the procedure computes the dissimilarity between each remaining compound in the database and the compounds in the subset. In a third step, the compound from the database that is most dissimilar to the subset is selected and accepted. To finish, the algorithm returns to step 2 if there are less than *n* compounds in the subset. There are different versions of the procedure depending on how steps 1 or 3 are implemented. Thus, the first compound can be selected at random, by choosing the most dissimilar in the database or choosing the molecule that is in the centre of the database. On the other hand, for choosing the rest of the molecules different algorithms have been suggested in the literature. Thus, MaxMin chooses compounds with maximum distance to its closest neighbour in the subset, whereas MaxSum chooses compounds according to maximum sum of distances to all the compounds in the subset. Other dissimilarity-based procedures include the sphere exclusion algorithm.[24] In this procedure, starting from a seed molecule together with a predetermined threshold, generates a sphere around the compound. The following molecules are incorporated to the sphere if the similarity index is lower than the radius of the sphere and then is taken; if not the molecule nucleates a new sphere. Finally, there are other algorithms like D-optimal design, taken from the methods used in experiment design. These methods are based on maximising the determinant of the covariant design matrix that implies the minimisation of the prediction error of a possible regression model.[25]

These methods are suited for high dimensionality spaces. These are fast procedures, although tend to select outliers, *i.e.* compounds with extreme values of some property.

Stochastic methods

These procedures attempt to select the most dissimilar subset of *n* molecules, that are optimally diverse and representative in the descriptor space.[26,27] This is carried out through the minimisation/maximisation of the diversity of the possible subsets of dimension *n* from the database of dimension *N*, using a diversity index. Procedures used for diversity optimisation are typically, genetic algorithms and simulated annealing.

The function to be optimised may include large dimensionality spaces. However, to use these procedures the diversity index should be easy to calculate.

Other separation methods are based on information theory. In this procedure, compounds are selected based on the assumption that diversity design attempts to maximize the information content of the resulting subset.[28,29]

Validation of the different approaches

Assessment of the performance of different methods available in the literature for diversity measurements requires the establishment of specific criteria for comparison purposes. As mentioned before, diversity assessment can be carried out to recognise possible voids in a chemical database, or in other words, to understand whether a certain family of compounds can increase the diversity of a database in order to be closer to saturation. On the other hand, general diverse libraries are designed to exhibit the minimum number of compounds that represent the whole chemical space of the original set of compounds. A third possibility includes the design of focused libraries aimed at optimising lead compounds, that requires the selection of a subset of active compounds from a larger database. Accordingly, the different methods published in the literature for diversity assessment and library design are applied with different goals. Methods are also applied depending on the size of the database analysed. Accordingly, it should be borne in mind that the selection of the various methodologies used depends specifically on the circumstances of the library analysed. Selection of a classification procedure is much influenced by the choice of molecular descriptors and, consequently, they have to be considered together to determine their overall performance.[30]

In the case of designing focused libraries, there are several studies concerning the process of clustering active molecules from those that are not. Thus, in a classical study[31] the authors found that 2D sub-structural keys performed better than 2D fingerprints and 3D structural descriptors. More specifically, the authors concluded that MACCS keys, together with a hierarchical clustering algorithm performed the best for discriminating between active and non-active compounds. However, using 2D fingerprints as molecular descriptors together with the use of cluster analysis requires the choice of a distance measure. In a recent study,[12] the performance of the Tanimoto and Euclidean distances to discriminate between active and non-active compounds from a database was compared, concluding that the former performs better than the latter. Other studies suggest that statistical methods, including linear discriminant analysis and recursive partitioning, perform better that hierarchical clustering methods.[32] Whereas different studies performed in the past suggest that 2D descriptors are enough for diversity assessment, analysis of the performance of different descriptors for its capacity to identify isosteres, demonstrated that a combination of 2D and 3D molecular field descriptions performs better than a set of 2D descriptors.[16] Similarly, the use of BCUT descriptors, intended to handle small dimension chemical spaces, together with cell-based partitioning methods has demonstrated a high capability to cluster together active molecules involving different scaffolds.[33]

Libraries designed following strategies to select the least number of molecules representing the diversity of a database, avoiding redundancies in the chemical database, are used in lead discovery. Cell-based methods are specially suited for this purpose, since selection of a representative from each box provides a subset containing the whole diversity of the original set. Comparison of different binning schemes by their ability to provide an even distribution, suggests that non-linear binning performs better than linear binning.[27] Diversity-based compound selection methods appear to be superior for this task.[34] Comparison of different versions of these methods, suggest that these methods increase their efficiency in regard to random selection of compounds with size of the database.[23] Furthermore, the work also concludes that they appear to be the most effective in selecting compounds associated with a range of activities. Moreover, in a very interesting study, Potter *et al.*[35] stressed that the use of the maximum-dissimilarity algorithm, together with 2D fingerprints and the Tanimoto index, provide a better performance than a random compound selection even for a small database of 1,300 compounds. Cluster methods can also be used for this purpose. Using an agglomerative cluster centre method, Potter *et al.* demonstrated the superiority of the procedure for selecting a subset of representative active compounds in regard to a random selection. Using a small database of about 300 compounds, the authors found that 40 compounds (12% of database size) were enough to cover all the biological target classes of the database by at least one hit per target. Comparison with a random selection procedure showed that even selecting 80 compounds randomly covered only 65% of the biological targets of the database.

In the case of the assessment of the diversity of a database, only a few studies have been published in the past. Thus, Voigt *et al.*[36] compared the diversity of different commercial and public databases, including the National Cancer Institute (NCI), the Available chemical directory (ACD), Chem ACX, the Maybridge Catalog, the Asinex database, the Sigma-Aldrich catalog, the World Drug Index (WDI) and the organic part of the Cambridge Structural Database (CSD). The authors used two stochastic selection procedures. On the one hand, the optimal dissimilarity selection method[25] and on the other hand, the stochastic clustering algorithm. The authors selected the smaller subset of compounds representing the diversity of the different databases, finding similar results with the two methods. The CSD appears as the most diverse with a 39% (depending on the procedure used to assess it), and the least diverse appears the Asinex database with a 10% diversity. The CSD also appears as the most diverse chemical database in different studies.[37]

Conclusions

The concept of diversity of a database of compounds can be used in the selection of a subset of compounds representing all the differential molecular features of the original set. This is a useful concept since it is not necessary to synthesize and test all the molecules of the set, but only those that represent the whole set. This can be done in practice by grouping the molecules of the set by their similarity and selecting only the representative of each group. This requires defining a chemical space, where molecules are represented by points, followed by a similarity measure that permits knowing the nearest neighbours of each of them. Neighbourhood measures permit classification of molecules into subsets. On the other hand, diversity can be used to find that subgroup of molecules that exhibits similar properties. In this case, it is necessary to group the molecules of the database, and then select those molecules of the group expected to exhibit similar physicochemical properties. Diversity can also be used to compare two databases and to assess the degree of saturation of a database This procedure could also be used to speculate about the size of a database in order to be universal.

In order to make use of the concept of diversity, different computational tools have been developed. First, procedures to define the chemical space according to the way molecules can be described, from a collection of 1D, 2D or 3D descriptors embedded in fingerprints to the use of the molecular coordinates and/or the electronic density of the molecular systems. Second, procedures to measure the (dis)similarity between the objects of the chemical space, from the use of an Euclidean distance to the use of more elaborated procedures according to the way molecules are described. Finally, a procedure to group the molecules according to their similarity. For this purpose different methods have been developed including cluster methods and binning methods. All these techniques require robust algorithms for a useful data handling and database mining.[38]

As a final remark, we consider that diversity analysis methods permit the extraction of relevant information about large databases of compounds, providing an analysis of the features both internal, about the coherence of the contents, as well as the features before specific external constraints. Due to the difficulties in creating an experimentally well balanced database and the need to speed up the process of finding new molecules with specific characteristics, diversity analysis methods will continue offering a tool to select compounds with specific features in the least time possible and saving costly experimental procedures. For this purpose there is a constant need to contrast the different tools and procedures available.

Juan J. Perez
Dept. d'Enginyeria Quimica, Technical University of Catalonia (UPC), ETSEIB, Av.Diagonal, 647, 08028 Barcelona, Spain.
E-mail: juan.jesus.perez@upc.es; Fax: 34 93 4016210; Tel: 34 93 401 6111

References

1 Y. C. Martin, *Perspect. Drug Discovery Des.*, 1997, **7**, 8, 159.
2 *Molecular diversity in drug design*, P. M. Dean and R.A. Lewis (Eds.), Kluwer, 1999.
3 J. Bajorath, *Nat. Rev. Drug Discovery*, 2002, **1**, 882.
4 J. A. Dimasi, R. W. Hanse and H. G. Grabowski, *J. Health Econ.*, 2003, **22**, 151.
5 R. E. Dolle, *J. Comb. Chem.*, 2002, **4**, 369.
6 M. D. Burke and S. L. Schreiber, *Angew. Chem. Int. Ed.*, 2004, **43**, 46.
7 P. Willett, *Similarity and clustering in chemical information systems*, Research studies Press, John Wiley and Sons, Inc, New York, 1987.
8 G. R. Famini and L. Y. Wilson, in *Reviews in Computational Chemistry*, Vol **18**., K. B. Lipkowitz and D. B. Boyd (Eds.), John Wiley and Sons, New York, 2002, 211.
9 R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Weinheim, Wiley-VCH, 2000.
10 R. D. Brown, *Drug Discovery Des.*, 1997, **7**, 8, 31.
11 D. Weineger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31.
12 X. Chen and C. H. Reynolds, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1407.
13 V. J. Gillet, P. Willet and J. Bradshaw, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 338.
14 R. D. Cramer III, S. A. DePriest, D. E. Patterson and P. Hecht, In *3D QSAR in drug design*, H. Kubinyi (Ed.), ESCOM, Leiden, 1993, 443.
15 R. D. Clark, A. M. Ferguson and R. D. Cramer, *Perspect. Drug Discov. Design.*, 1998, **9**, 10, 213.
16 A. Schuffenhauer, V. J. Gillet and P. Willett, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 295.
17 Y. C. Martin, M. G. Bures, E. A. Danaher, J. DeLazzer, I. Lico and P. A. Pavlik, *J. Comput.-Aided Mol. Design.*, 1993, **7**, 83.
18 R. S. Pearlman and K. M. Smith, *Perspect. Drug Discov. Design.*, 1998, **9**, 10, 339.
19 N. Nikolova and J. Jaworska, *QSAR Comb. Sci.*, 2003, **22**, 1006.
20 R. Carbo, L. Leyda and M. Arnau, *Int. J. Quantum Chem.*, 1980, **17**, 1185.
21 A. M. M. Jorgensen and J. T. Pedersen, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 338.
22 P. Willett, in *Molecular Diversity in drug design*, Ed. P. M. Dean and R. A. Lewis, Kluwer Academic Publishers, London, 1999, pp 115–140.
23 M. Snarey, N. K. Terrett, P. Willett and D. J. Wilton, *J. Mol. Graphics Modell.*, 1997, **15**, 372.
24 V. J. Gillet, P. Willett, in *Combinatorial library design and evaluation*, A. K. Ghose, and V. N. Viswanadhun, (Eds.), Marcel Dekker Inc., New York, 2001, 279.
25 V. V. Federov, *Theory of optimal experiments*, Academic Press, New York, 1972.
26 R. Clark, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1181.
27 D. K. Agrafiotis, *Mol. Diversity*, 2000, **5**, 209.
28 S.-K. Lin, *Molecules*, 1996, **1**, 57.
29 J. I. Miller, E. K. Bradley and S. L. Teig, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 47.
30 D. Gorse and R. Lahana, *Curr. Opin. Chem. Biol.*, 2000, **4**, 287.
31 R. D. Brown and Y. C. Martin, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1.
32 S. L. Dixon and H. O. Villar, *J. Comput.-Aided Mol. Des.*, 1999, **13**, 535.
33 J. S. Mason and M. A. Hermsmeier, *Curr. Opin. Chem. Biol.*, 1999, **3**, 342.
34 Y. Tominaga, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 867.
35 T. Potter and H. Matter, *J. Med. Chem.*, 1998, **41**, 478.
36 J. H. Voigt, B. Bienfait, S. Wang and M. C. Nicklaus, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 702.
37 M. P. Bradley, *Mol. Diversity*, 2000, **5**, 175.
38 N. Adams and U. S. Schubert, *J. Comb. Chem.*, 2004, **6**, 12.