Analytica Chimica Acta 348 (1997) 409-418

General Type of a Uniform and Reversible Representation of Chemical Structures[1]

Jure Zupan*, Marjana Novic

National Institute of Chemistry,Ljubljana, Hajdrihova 19, SLO-1115 Ljubljana, Slovenia,

Received: 1 June 1996, received in the revised form 4. December 1996; accepted 31. December 1996

Abstract

In any type of modelling (be classical or by artificial neural networks) involving chemical structures and their corresponding properties, the first problem encountered is the representation of chemical structures. A good structure representation should have different code for each 3-D structure (uniqueness), it should have the same number of variables for all structures, it should be reversible, and should be translation and rotational invariant. In the present contribution we are discussing a new method for representing chemical structures which, at least in principle and within limitations bound to the precision and resolution of the projection, fulfils all mentioned requirements with the exception (in some cases) of the rotational invariance. The discussed representation is based on the projections of atoms on the sphere with an arbitrary radius. The new structure representation of a molecule with N atoms is defined as n-dimensional vector $\boldsymbol{S} = (s_1,s_2,..s_i,...s_n)$ with each component defined as a cumulative intensity s_i, at a given point i on the circle with and arbitrary radius. The cumulative intensity s_i (the i-th point on the circle at angle φ_i.) is a sum of N contributions $I(i,\rho_j,\varphi_j)$ from each atom j in the molecule.

$$s_i = \sum_{j=1}^{N} I(i,\rho_j,\varphi_j) = \sum_{j=1}^{N} \frac{\rho_j}{(\varphi_i - \varphi_j)^2 + \sigma_j^2} \qquad \text{with} \qquad i=1...n$$

The intensity function $I(i,\rho_j,\varphi_j)$ can be any bell shaped function. In our case we have chosen the Lorentzian shape with maximum at the angle φ_j, maximal intensity proportional to the ρ_j, and having the width, σ_j, related to the type of the atom. The new proposed "spectrum-like" representation is additive with the respect to the constituent atoms of a given structure and can be easily decoded. Because the representation is additive it allows to *subtract* the a part of the "spectrum-like" representation which belongs to the structurally identical skeletons of all molecules in the study.

Keywords: Chemical code; Structure representation; Projection; Uniformity; Reversibility, Kohonen neural network

* Corresponding author: Tel:+386-61-176-02-79; fax: +386-61-125-92-44

[1] This paper is dedicated to Prof. Jean Thomas Clerc, our friend, a great chemist, and a believer that good chemical coding has yet to come.

1. Introduction

During the history of chemical structure coding many different ways and different strategies how to distinguish different compounds, solutions, and materials in a unique way can be encountered. Leaving chemical names which have being used through the entire human history aside, the beginning of the computer era has initiated many different structure representations for efficient computer handling: from different types of fragment codes (Wiswesser-Line-Notation or WLN (1)), atom connectivity tables and topology indices (2) to sophisticated coding enabling 3-D similarity searches (3). Unfortunately, the WLN and the connection table representations are not uniform, i.e. the representation cannot be expressed as a vector in the same *n*-dimensional space (fixed n for all structures in the study), while the topology indices, gnomonic projections and Kohonen maps are not reversible.

The most interesting, if not the most informative, aspect of structure coding history is that most of the attempts to find a new structure code, were driven by a specific property elucidation goal in mind. With other words, the representations are mainly not proposed or invented just for its own sake, but with an aim to explain a property, or a set of properties, for which it is believed that is related to the complete or parts of the structures. For example, the choices of fragments in a fragment code are always chosen in such a way that could be easily recognised by an appropriate analytical or spectroscopic method. Or, the hydrogen atoms in the connection tables are usually omitted because it is believed that in most applications they are not needed or can be easily added. In different spectroscopies, different structural features are important and thus encoded differently compared to the irrelevant structural parts. To stress this point it is worthwhile to cite the paper by Clerc and co-authors from more then 20 years ago(6) when they argued both witty and strongly that before any structure related property is modelled or correlated a good structure representation should be found.

A good structure representation for modelling (be classical *via* analytical functions or by means of artificial neural networks) should fulfil the following demands:

1. each compound should have one and only one code; with different codes for different structure (uniqueness),
2. each compound should be represented by the same number and type of variables, i.e. the code should be operational in the same problem space (uniform),
3. the structure should be possible to retrieve back from the code (reversibility), and
4. for rotated and/or translated structures the code should remain the unchanged (translation and rotational invariance).

It is not difficult to fulfil each of the above four conditions separately, however, to fulfil all of them using the same representation is a very difficult task which has not yet being solved satisfactorily.

For modelling the quantitative structure-activity relationship (QSAR) it has been agreed upon that one of the major role has to be assigned to steric molecular (steri-mol) parameters (5), i.e. to the shape of the molecule. This means that position of atoms in 3-dimensional space are mandatory. Of course, besides the steric effects of the complete structure the electronic and hydrophobic effects have to be taken into the account for better modelling of biological activities (6).

However, it has been pointed out (5) that, before any general rule or prediction about the influence of the shape of molecules on the biological activities can be made, the separation of the steric and electronic effects should be made. This means that as a first step in the search for a good structure representation a stand alone representation of a 3-dimensional chemical structure should be sought. Chau and Dean (7), for example, have introduced for analysing a property on a molecular surface a so called gnomonic projection (projection from a central point) of a property to a surface of a sphere. Recently, Gasteiger and co-workers (8) have used the Kohonen type artificial neural networks (ANNs) to produce a fixed size 2-dimensional map for any 3-dimensional shape of the molecule. Although all this techniques have some advantages, a real common problem to all of them is the reversibility, i.e., they do not allow to retrieve the initial structure back from the representation.

In the present contribution we are discussing a new method for representing chemical structures which, at least in principle and within limitations bound to the precision and resolution of the projection, fulfil all of the above four requirements.

2. Structure representation by projection

By now from all structure coding systems, the ones that fulfil the second condition (uniform representation) are gnomic projection (7), Gasteiger's approach using ANNs (8) and several coding based on different topological indices. On the other hand the fragment coding systems (like WLN chemical notations (1), connection tables, etc. from which the structure can be retrieved are unfortunately not uniform. It is clear that the representation of the structure having N atoms described by N $[x_j,y_j,z_j]$ triplets does not fulfil the condition of uniformity either. Therefore, **a transformation** of the molecular representation defined by three co-ordinates for each of the N atoms in the molecule (i.e., $3N$ dimensional representation) into a unique n-dimensional vector representation which should be reversible is sought (9).

The new representation (coding of a 3-dimensional chemical structure) is based on the projection of constituent atoms onto an imaginary spherical surface large enough for a molecule under consideration to be accommoda-ted within it. The projection of molecule's atoms is not made onto the entire sphere, but only onto two perpendicular equatorial trajectories on the sphere. For the representation of a 3-dimensional molecule composed of N atoms, first, in all N triplets $[x_j,y_j,z_j]$ the z_j co-ordinates are set to zero $[x_j,y_j,0]$ and the projection of a (x,y)-"planar molecule" is made on the circle defined by the cross-section of the sphere and the (x,y) plane. Next, all y_i and finally all x_i co-ordinates are set to zero yielding $[x_j,0,z_j]$ and $[0,y_j,z_j]$ "planar molecules" and consequently the projections of atoms onto circles lying in the (x,z) and (y,z) planes are made.

The technique for obtaining all three projections is essentially the same. For a (x,y)-planar case each molecule containing N atoms is described by a set of N pairs $(x_2,y_1, x_2,y_2,...x_j,y_j,..,x_N,y_N)$.

In order to eliminate the difference due to the translation of molecules, the molecular co-ordinates are first transformed so that the origin of the co-ordinate system is set into the molecular centre of mass. The molecular centre of mass with the co-ordinates $T_c=(1/N)(\Sigma x_i,\Sigma y_i,\Sigma z_i)$ is also the centre of the circles $T_c=(0,0,0)$ on which border the atoms of the particular molecule will be projected. The radius R of the circle is arbitrary, the only condition set is that it must be larger or equal to the distance between the centre of mass and the most distant atom, r_{max}

The new structure representation is defined as n-dimensional vector $\boldsymbol{S} = (s_1,s_2,..s_i,...s_n)$. In fact, because three projections are required, the final representation has $3n$ variables! Each component s_i of the n-dimensional vector $\boldsymbol{S}$ is defined as a cumulative intensity, s_i, at a given point i (or better over the finite interval i) on the circle with arbitrary radius R. The i-th point (interval) is placed on the circle at angle φ_i (Figure 1). The cumulative intensity s_i is a sum of N contributions $I(i,\rho_j,\varphi_j)$ from each atom j in the molecule.

$$s_i = \sum_{j=1}^{N} I(i,\rho_j,\varphi_j,\sigma_j)$$
$$= \sum_{j=1}^{N} \frac{\rho_j}{(\varphi_i - \varphi_j)^2 + \sigma_j^2} \quad \text{for } i=1...n \qquad /1/$$

The intensity function $I(i,\rho_j,\varphi_j,\sigma_j)$ can be any bell shaped function. In our case we have chosen the Lorentzian shape (Figure 1) for describing the "atom projection" to the circle.

Each atom j is thus represented ("projected") by one Lorentzian curve maximum of which is located at angle φ_j. The intensity at the maximum point ($\varphi_i=\varphi_j$) is proportional to the radii-vector ρ_j.

Parameter σ_j describes the width of the curve associated with each atom. Therefore, it can bear the information on the nature of each atom. It can be related to the atom type, to its Van der Vaals radius, to its charge, or any other property considered to be relevant to the problem.

Due to the fact that parameter σ_j can be specified for each atom the new proposed coding scheme is flexible enough to be adaptable to various types of problems for which "problem-specific" structure code is preferred. As pointed out by Professor Clerc

(10) many times, there are indeed, much more problems that require “problem-specific” structure notation as there are problems which do not have this requirement.

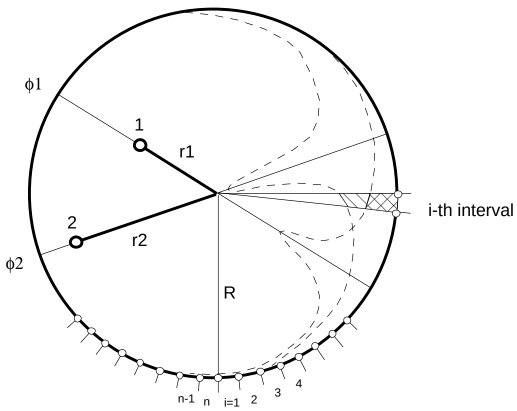

Figure 1. Contribution of atoms No. 1 and No. 2 (at the positions (r_1,φ_1) and (r_2,φ_2) to the intensity s_i at the interval (position) *i* on the circle with radius *R*, shown as shaded areas of the corresponding Lorentzian bell-shape functions. Circular trajectory with radius *R* to which the projections are made is divided into *n* intervals (360 in our example).

If only the space geometry and the shape of molecule should be described, the parameter s_j (eq. /1/) can be equal to one for all atoms. In the present discussion all parameters σ_j are taken equal to one.

There is of course a question how rotational invariance of the new code can be achieved. This aspect of the code will be addressed later on. If the modelling for which the code is set up concerns a class of compounds having identical skeletons or at least identical small substructure a group of atoms, the easiest way to achieve independence from the rotation is to bring all molecules under study into such position that the common substructure will be at the same (or almost the same) position in all compounds.

The dimension *n* of the new representation is determined by the number of equidistant points on the circles to which the atoms are projected. The dimension *n* should actually be independent from the number of atoms *N* in the molecule. Unfortunately, *N* and *n* are not entirely independent: the dimension *n* is related to the resolution of the representation and consequently to the quality of the inverse process of recovering the structure. The larger *n*, the larger number of atoms *N* in each molecule which can distinguished by the code.

In general *n* should be adapted to the number of atoms *N* in the *largest* molecule of the study. After the size n (the dimension or the number of variables) is chosen the new representation should be able to map each molecule, regardless of the number of its constituent atoms, into the same *n*-dimensional space. In many studies where only the orientation in space or direction of the substituents with the respect to the skeleton or to each other is sought the dimension of n can be as low as 36 or even 18.

3. Coding of structure

The starting point for coding are *{x,y,z}* co-ordinates of all atoms and consequently their projections into the *(x,y) (x,z), and (y,z)* planes. where:

$$\rho_i^2 = x_i^2 + y_i^2 \text{ and } cos^2\varphi_j(x,y) = x_i^2 / (x_i^2+y_i^2) \qquad /2/$$

or equivalent for the *(x,z)* and *(y,z)* projections:

$$\rho_i^2 = x_i^2 + z_i^2 \text{ and } cos^2\varphi_j(x,z) = x_i^2 / (x_i^2+z_i^2) \qquad /2a/$$

$$\rho_i^2 = y_i^2 + z_i^2 \text{ and } cos^2\varphi_j(y,z) = y_i^2 / (y_i^2+z_i^2) \qquad /2b/$$

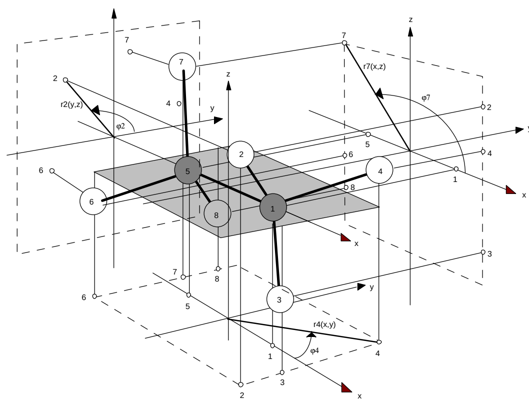

Figure 2 Structure of the ethane and parts of the *(x,y)*, *(x,z)*, and *(y,z)* projections are shown. For the sake of clarity, in each projection only one atom No. 4 (ρ_4,φ_4), No.7 (ρ_7,φ_7), and No. 2 (ρ_2,φ_2) is shown explicitly with radius vector and corresponding angle, respectively

As an example how the suggested representation $\mathbf{S} = (s_1,s_2,..s_i,...s_N)$ can be obtained from the structure's *{x,y,z}* co-ordinates, the entire procedure for calculating the variables s_i will be explained on a simple molecule of ethane (Figure 2). The *{x,y,z}* co-ordinates of all eight atoms are given in Table I. For each atom the position at which the Lorentzian bell-shaped curve has its maximum and the maximal intensity are given for all three projections.

At the beginning of the coding procedure, the molecule is put with its centre of gravity into the centre of the co-ordinate system. Then in all n points (intervals) on the circle the cumulative contributions, s_i, $i=1...n$, are calculated using equation /1/ for all eight atoms.

Two resulting "spectrum-like" representations: one projected onto circle in (x,y) and one onto the circle in (x,z) plane are shown in Figures 3a and 3b.

As many Lorentzian curves as there are atoms in the molecule have to be added together to obtain the "spectrum-like" representation.

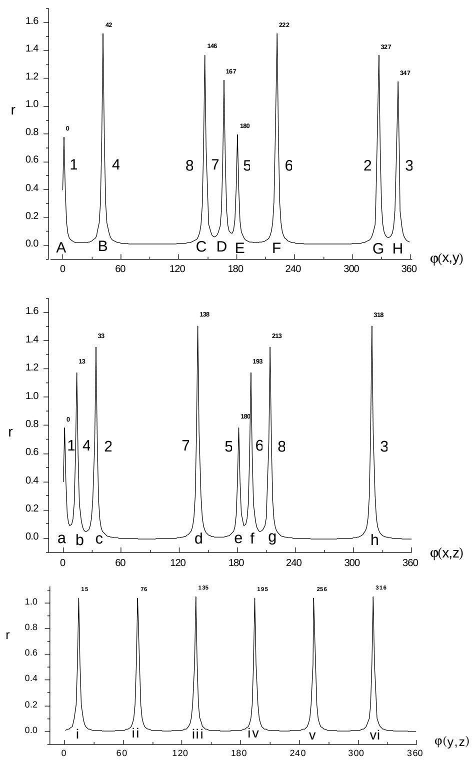

Figure 3 "Spectrum-like" structure representations of ethane projected in *(x,y)*, *(x,z)*, and *(y,z)* planes. The assignments of the peaks to 8 ethane atoms is shown with figures standing next to the peaks. The intensities and positions are listed in Table I.

The new "spectrum-like" representation is easy to obtain. By projecting any number of atoms onto a circular line of length 2π, the uniformity of the code is achieved. The actual resolution used for a specific application depends on the needs and computational capability of the user. The smaller σ/n ratio the larger the resolution of the representation. The more atoms has the largest molecule in the study, the finer should be the resolution of the "spectrum-like" representation (division of the 2π circle into n intervals).

The translation invariance of the code is assured by setting the co-ordinate origin of the sphere to which the projections are made into the centre of all $\{x,y,z\}$ co-ordinates of the molecule. In many applications a certain atom (for example, atom common to all structures) is set into a centre of co-ordinate origin.

If any atom is in the central position it has no peak in the "spectrum-like" representation (see (y,z) projection in the following example - Fig. 3c).

To the contrary of the translation invariance, the rotational invariance of the code is harder to ensure. For a set planar molecules, (in such cases only one projection S is sufficient), the rotational invariance can be achieved by calculating the power spectrum of the Fourier transformation (7), $FT(\boldsymbol{S})$ or even much simpler autocorrelation transformation $A(\boldsymbol{S}) = (a_0, a_1, \ldots a_{n-1})$ (11):

$$a_j = \sum_{j=1}^{n} s_i\, s_{mod((j+i),n)} \qquad for\ j=0,1,...n\text{-}1 \qquad /3/$$

However, for real 3-dimensional molecules the 3-d Fourier transformation together with corresponding 3-d power spectrum of all "spectrum like" representations, $S(xy)$, $S(x,z)$, and $S(y,z)$ should be calculated. At the moment we are exploring this possibility, but we do not have an explicit answer yet, because the solution is too computer-space demanding for coding large data sets.

It should be noted again that by applying the Fourier transform power spectrum or the autocorrelation operation (eq. /3/) to the new representation S the reversibility of the code is lost in the since of direct recalculation (inverse decoding). Nevertheless, using genetic algorithms or some other optimisation method, the structure can be retrieved from the power spectrum by iterations.

The iterations should start with an initial structure guess accompanied by the corresponding power spectrum and than changes of the structure guided by diminishing distance between the final and calculated power spectrum. The most important aspect of a uniform code is that it enables its use in any kind of direct or inverse modelling of structure-property relationships.

peaks two groups of x co-ordinate is obtained: 1.12 and -1.13Å For each of the two values three different *y* (1.01, -0.74, -0.26) and three different *z* co-ordinates (0.26, 0.74, -1.01) are

Table 1. The co-ordinates of eight ethane atoms are in columns 2-4. In columns 5 and 6 are ρ(x,y) and φ_i(x,z). Radii ρ(x,y) and angles are calculated according to equations: $\rho(x,y)=(x^2+y^2)^{1/2}$ and $\cos(\varphi_i(x,y))=x/\rho(x,y)$, respectively, and taking into account that $\varphi_i(x,y)=360 - \varphi_i(x,y)$, if y co-ordinate is negative. For calculations of φ_i(x,z) the y is replaced by z.

No.	Atom	x [Å]	y [Å]	z [Å]	ρ(x,y)	φ_i(x,y)	ρ(x,z)	φ_i(x,z)	ρ(y,z)	φ_i(y,z)
1	C	0.77	0.00	0.00	0.77	0.	0.77	0.	0.00	
2	H	1.13	-0.74	0.74	1.35	327.	1.35	33.	1.05	135.
3	H	1.13	-0.27	-1.00	1.16	347.	1.51	318.	1.05	256.
4	H	1.13	1.00	0.27	1.51	42.	1.16	13.	1.05	15.
5	C	-0.77	0.00	0.00	0.77	180.	0.77	180.	0.00	-
6	H	-1.13	-1.00	-0.27	1.51	222.	1.16	193.	1.05	195.
7	H	-1.13	0.27	1.00	1.16	167.	1.51	138.	1.05	76.
8	H	-1.13	0.74	-0.74	1.35	146.	1.35	213.	1.05	316.

4. Decoding the "spectrum-like" representation

The reversibility of the "spectrum like" representation is not evident because neither matching of peaks in the spectra, i.e. identifying triplets of peaks from all three "spectra" belonging to the same atom, nor the assignation of peaks are straightforward. Due to the fact that at the locations of peak positions ($\varphi_i=\varphi_j$, eq. /1/) the *x* and *y* co-ordinates can be calculated from the peak intensities $I(\varphi_i=\varphi_j)$ exactly using:

$$x = I(\varphi_i=\varphi_j)\cos(\varphi_i) \text{ and } y = I(\varphi_i=\varphi_j)\cos(\varphi_i) \quad /4/$$

The assignments of the z co-ordinates calculated from equations /2a/ and /2b/ can be easily made from the previously fixed *x,y* pairs. This procedure is shown in Tables 2 and 3. The values of *y* and *z* co-ordinates can be calculated from the peak intensities. In the presented example we have intentionally chosen the orientation of a molecule for which, due to its symmetric position, several possibilities for assignments of *x,y* pairs to the *z* co-ordinate arises.

From Table 2 it is evident that only two atoms are, i.e., two peaks (A,E and a,e) from the first two spectra, can be assigned. From other six possible. Altogether this would amount to 27 possibilities. Due to the fact that only the combinations containing all six values are possible only six possibilities for assigning three peaks remain: Bb,Gc,Hh; Bb,Gh,Hc; Bc,Gb,Hh; Bc,Gh,Hb; Bh,Gb,Hc: and Bh,Gc, Hb. This can be easily resolved by considering the third spectrum. The calculated values from the *S(y,z)* are given in Table 3. The signs of *z* co-ordinate in the 6-th column are resolved by the size of angle φ.

The *{x,y,z}* triplets shown in Table 3 are in fair agreement with the original co-ordinates of six hydrogen atoms shown in Table 1. Small discrepancies (about ±0.01 Å) originate from the fact that the peak positions are taken as integers on the 0-360° scale.

5. Applications

For simple tasks or applications, such as structure retrieval or identification of structures, the uniformity of the structure code is not required. These tasks can be well executed by chemical names, connection tables, WLNs or similar non-uniform coding systems. On the other side, the applications of a uniform 3-dimensional structure code are many.

The most important aspect of a uniform code is that it enables its use in any kind of direct or inverse modelling of structure-property relationships.

$Y(y_1,..y_n)$) can only be achieved from a space having well defined metrics, i.e., the same number of defined variables (axes), *m*, for each object *X*:

Table 2 Decoding procedure when peak intensities and peak positions are known.Peaks A-H and a-h correspond to the S(x,y) and S(x,z) ´spectrum-like´reprtesentation shown in Fig. 2. The alternative possibility of signs at ys and zs is resolved by the size of angle φ: it is smaller than 180^0 tahn the sign is positive, otherwise negative.

Peak	Int Intensity[Å]	φ [deg]	cos φ	x Int (cosφ)	y $\pm(Int^2-x^2)^{1/2}$	z $\pm(Int^2-x^2)^{1/2}$
A	0.77	0	1.000	0.77	0.00	
B	1.51	42	0.743	1.12	1.01	
C	1.35	146	-0.829	-1.12	0.75	
D	1.16	167	-0.974	-1.13	0.26	
E	0.77	180	-1.000	-0.77	0.00	
F	1.51	222	-0.743	-1.12	-1.01	
G	1.35	327	0.839	1.13	-0.74	
H	1.16	347	0.974	1.13	-0.26	
a	0.77	0	1.000	0.77		0.00
b	1.16	13	0.974	1.13		0.26
c	1.35	33	0.839	1.13		0.74
d	1.51	138	-0.743	-1.12		1.01
e	0.77	180	-1.000	-0.77		0.00
f	1.16	193	-0.974	-1.13		-0.26
g	1.35	213	-0.839	-1.13		-0.74
h	1.51	318	0.743	1.12		-1.01

There is an immense variety of the problems of this kind: from various spec-trastructure elucidation, spectra simulation, to quantitative structure activity relation-ship (QSAR) problems, to mention only the few most important ones.

$$Y(y_1,..y_n) = \mathbf{A}[X(x_1,x_2,...x_m)] \qquad /4/$$

so that the distances or similarities between various objects X^s in this space can be calculated. The symbol A labels any system (set of equations, architecture of neurons) containing parameters, a_{ji}, called coefficients, weights, pointers, or similar, that is able to perform the required mapping from the *m*-dimensional space of structure representations X^s to a *n*-dimensional space of properties Y^s.

Table 3. Peak intensities and positions in the S(y,z) spectrum. The values of x co-ordinates are average values taken from Table 2.

Peak	Int Intensity [Å]	φ [deg]	cos φ	x [from Table II]	y Int (cosφ)	z $\pm(Int^2-y^2)^{1/2}$
i	1.05	15	0.966	1.13	1.01	0.28
ii	1.05	76	0.242	1.13	0.25	1.01
iii	1.05	135	-0.707	1.13	-0.74	0.73
iv	1.05	195	-0.966	-1.13	-1.01	-0.28
v	1.05	256	-0.242	-1.13	-0.25	-1.01
vi	1.05	316	0.719	-1.13	0.75	-0.72

For most of such problems the uniformity of the structure code is mandatory. The reason for this is that models (achieved either by the analytical functions or by artificial neural networks) require uniform code of objects. In formal words: mapping of chemical structures into the space of sought property (or properties

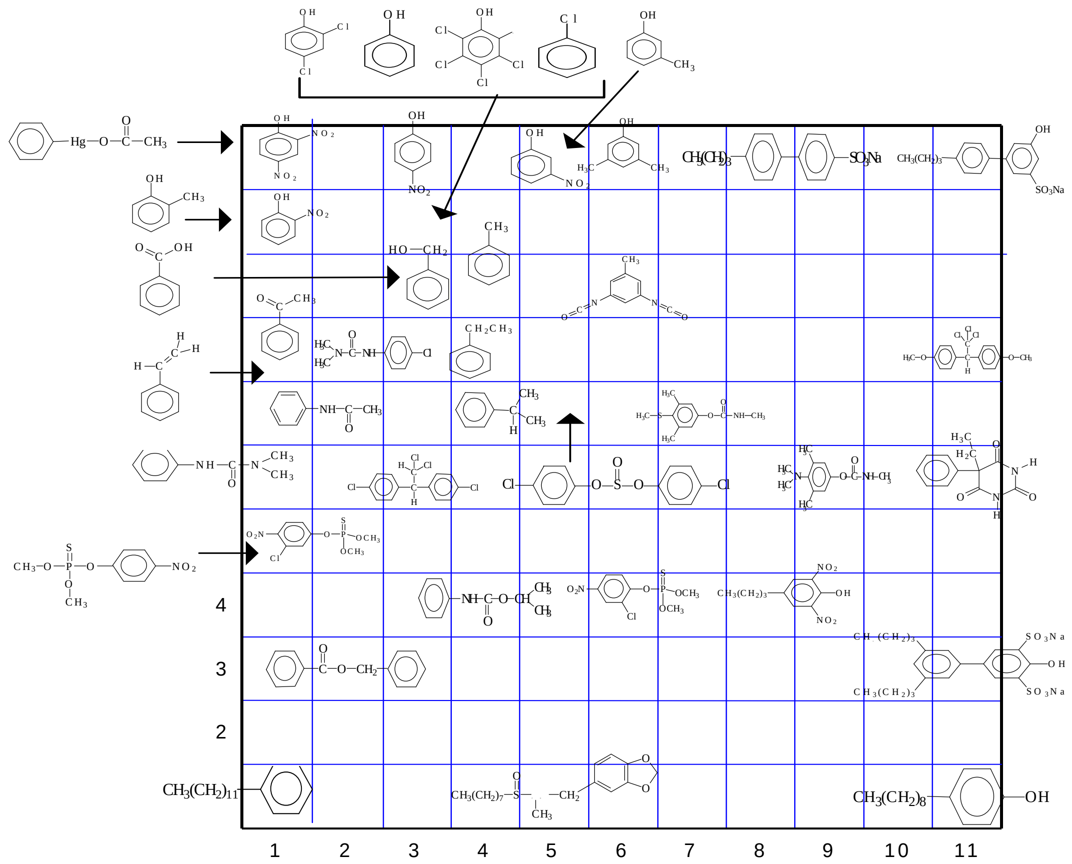

Figure 4. Distribution of 41 compounds on the 11x11 Kohonen neural network. Each compound is shown on the position of the neuron which was excited by its 54-dimensional “spectrum-like” representation.

One purpose of modelling (predicting) properties of chemical compounds based on their 3-dimensional structure descriptions is to obtain knowledge how different structural parts (substituents, radicals, fragments, sub-structures, skeletons, etc.) influence the properties of interest. However, this is not the only goal of the structure-property studies. In many cases most information is gained not from the observation of one molecule (i.e. from the type and positions of substituents with the respect its skeleton), but from the comparison of the relative positions of substituents and their spatial distribution with the respect to their own **and** to other molecules’ skeletons in the investigated set of compounds.

In the type of problems where the property of interest is influenced by the shape and spatial position of the molecule restricted by the outside factors the rotational invariance of the code is undesirable. Rotational invariance means that for each molecule an unique internal co-ordinate system is established according which the code is calculated. Such internal co-ordinate system is invariant to any outside influence and therefore not compa-rable to internal co-ordinate systems of other mole-cules. If from the problem’s point of view not all directions are equally preferable, the best way is to orient (rotate) all molecules in the study according to this direction and then perform the coding. In this way non of the 3-dimensional features (positions, distan-ces, shapes, angles, etc.) within the individual molecules nor the relative spatial features to the outside direction and consequently to other molecules are lost.

To show a mapping of molecules coded with the new “spectral like” representation into a 2-dimensional plane a simple example using 41 benzene and phenol compounds was made.

The molecules were first all oriented in such a way that all benzene rings were aligned in the same direction (all benzene atoms have been moved to approximately the same *(x,y,z)* co-ordinates) and then coded into three 18-dimensional "spectra". The complete 54 (3 x 18 = 54) "spectrum-like" representations were then mapped using Kohonen artificial neural network (ANN) (12) into the 2-dimensional map consisting of 121 neurons. The resulting 11 x 11 neuron map with corresponding structures that excited the particular neurons is shown in Figure 4.

Extremely low-dimensional representation, i.e. 18 intervals on the entire 360^o circle, allows to distinguish only the atoms separated in the space for more than 20^o. Nevertheless, as shown in Figure 4 the representation is able to cluster 41 compounds nicely into groups with short and long chains, groups with one and several chains, groups with one and with two rings etc.

6 Conclusion

The described "spectral-like" representation of the structures is important from two reasons. First, it offers a fixed-dimension representation in which a wide variety of different structures can be uniformly coded and second, it offers the possibility for decoding of the structure from the representation. Most of the contemporary fixed-dimension structure representation are based on a group of several topological, shape/form, electronic, hydrophobic and other single variable properties. From such representations decoding of the structures is practically impossible.

We are aware that with this representation still some problems remain, notably, the problem of rotational invariance for a general open system use. If the proposed representation is used on a set of structures that are previously oriented the "rotational invariant" code in a sense that the structures' orientations can be easily compared to each other is obtained. Even more, if the structures in question are aligned to the same external co-ordinate system the feed-back information from the model about the parts of the structures and their spatial distributions responsible for the modelled activity can be deduced.

Additionally, for coding the molecules all having the same skeleton or the same backbone structure (group of derivatives, analogues, etc.), all peaks in the new representation obtained for the atoms of the common substructure can be simply subtracted from the representation. The subtraction of common peaks makes the representation more sensitive to those parts of the structures that are actually relevant to the study. Thus, by orienting the common skeleton along the pre-defined co-ordinate system the code which is in a previous sense rotational invariant and at the same time reversible is obtained. Due to the fact that most of the QSAR and other structure-property relationships, notably spectra-structure relationships are done for the families of compounds, this representation can be an excellent tool for such purpose.

Acknowledgement

The financial support of the Copernicus CP94-1029 EST Project and the support of Ministry of Science and Technology of Slovenia within the Project J1-5014 is gratefully acknowledged.

References

1. W. J. Wiswesser, A Line-Formula Chemical Notation, T.Y. Crowell, New York, 1954; and E. G. Smith, The Wiswesser Line-Formula Chemical Notation, McGraw Hill, New York, 1968,
2. As review of different coding systems see for example: J. E. Ash, W. A. Warr, P. Willett, Chemical Structure Systems, Ellis Horwood, New York, 1991,
3. See for example: Concepts of Molecular Similarity, M. A. Johnson and G. M. Maggiora, Eds., Wiley Interscience, New York, 1990,
4. R. W. Taft, J. Am. Chem. Soc.,**74**, (1952), 3120
5. J.T. Clerc, P. Naegeli, J. Seibel, Artificial Intelligence, Chimia, **27**, (1973), 12,
6. C. Hansh, A. Leo, Exploring QSAR, ACS Professional Reference Book, ACS, Washington, D.C., 1995, Chapter 3.
7. P. L. Chau, P. M. Dean, J. Mol. Graphics, **5**, (1987), 97-100,
8. R. N. Bracewell, The Fourier Transformation and its Application, McGraw Hill, New York, 1986, p.381
9. M. Novi~, J. Zupan, Proceedings of the X-th CIC Symposium, (Computers in der Chemie), Hochfilzen, November 1995., Ed. J. Gasteiger, (in press)
10. See for example: C. Affolter, K. Baumann, J.T. Clerc, A New Strategy and Representation of Chemical Structure-Spectra Correlation, Plenary lecture, 9-th Symposium Spectroscopy in Theory and Practice" Bled, April 10-13 1995, Slovenia
11. J. Zupan, Algorithms for Chemists, J. Wiley & Sons, Chichester, 1989.
12. J. Zupan, J. Gasteiger, Neural Networks for Chemists: An Introduction, VCH, Weinheim, 1993.