*Analusis*, **1998**, *26*, 333-341
© EDP Sciences, Wiley-VCH

# Comparison of chemical databases: Analysis of molecular diversity with Self Organising Maps (SOM)

P. Bernard[1], A. Golbraikh[1], D. Kireev[1], J.R. Chrétien[1] and N. Rozhkova[2]

[1]*Laboratory of Chemometrics, University of Orléans, BP. 6759, 45067 Orléans Cedex 2, France*
[2]*Plant Protection Chemical Research Institute, Ugrezhskaya 31, 109088 Moscow, Russia*

**Abstract.** Self Organising Map (SOM), also known as Kohonen Neural Network, is tested as a non supervised procedure for comparing molecular databases. Each chemical compound being represented by a point in the hyperspace of the molecular descriptors, SOMs was used to reflect the multidimensional hyperspace onto a two dimensional (2D) map while preserving the order of distances between the points, but in a non linear way. The aim of this work was to apply SOM to the study of the overlapping of two databases in order to obtain information about the extent of their differences in regard to their molecular diversity. Firstly, the ability of SOM to discriminate between two virtual databases was investigated. The positions of these two virtual databases were made to vary from non-overlapping to overlapping ones. In any considered cases, all the individuals of these two databases are processed simultaneously to give one SOM. From this map it is possible to analyse and understand the structure of the original data. Secondly two chemical databases are compared. The first chemical database deals with the commercially available organophosphorous pesticides (OPC), the second one deals with more than two thousand OPC tested as potent pesticides. Given the biological data known for each compound, the second database was shown to bring an interesting supplement to the structural information nested in the first database taken as a reference. Furthermore, the results obtained indicate that SOM can be used for the search of new leads among available databases and the exploration of new structural domains for a given biological activity.

**Key words.** Kohonen neural network – self organizing map – classification – chemical databases – virtual screening – pesticide – organophosphorous compounds.

## Introduction

Rapid developments in combinatorial chemistry, agrochemistry and medicinal chemistry have introduced new methods and significant improvements simultaneously in three complementary directions: (i) lab automation for High Throughput Screening (HTS) [1], (ii) analytical chemistry for microdetection and micropurification of the tested compounds or samples [2], and (iii) exploitation of large sets of structural databases based on computer aided strategies for finding new leads [3,4].

At present, most pharmaceutical or chemical companies possess their own chemical molecular databases which include hundreds of thousands of compounds. As it was already described [5], for pharmaceutical companies these databases are a real source of information which is used for systematic search of new leads in Drug Design strategies. So, these companies are constantly interested in acquiring new databases. Nevertheless, taking into account the cost of such databases and of experiment design based on HTS, decisions have to be justified in term of molecular diversity. Evidently, one solution could be the screening of all the compounds of all databases, but due to the time and the cost of such a procedure, this solution has to be rejected. An alternative solution is the study of molecular diversity by means of clustering and projection techniques to get a rational experimental design in HTS.

Self-Organizing Map (SOM) [6] also referred to as Kohonen Neural Network suggests new solutions to the problem of managing multidimensional data. SOM combines the advantages of both clustering and projection methods. Indeed, SOM divides a set of compounds into a number of clusters and visualises these clusters on a 2D map. This map allows the user to analyse the cluster distribution in the multidimensional descriptor hyperspace. SOM preserves the topology of the original hyperspace, i.e. the points located near each other in the original space remain neighbours on SOM.

SOM was already applied to such chemical problems as protein [7] or DNA [8] secondary structure mapping, molecular surface unfolding [9,10] or the modelling of IR and NMR spectra of organic compounds [11,12]. The first applications of SOM to manage large databases appeared recently. Thus, SOM was used as a tool to establish structure-activity relationships in a database containing more than 10 000 compounds extracted from the National Cancer Institute database [13]. In the recent work of Bauknecht et al. [14] SOM was used to discriminate dopamine from benzodiazepine agonists out of a training set of 172 compounds by using autocorrelation descriptors. Recently, we showed the capacity of SOM to classify biologically active compounds [15].

Its dual nature, clustering and projecting, provides SOM with interesting properties, most of which remain to discover and will require intensive investigation. Measuring overlap of two databases on their corresponding SOM allows to evaluate the structural complementarity between them. In this paper, first, virtual examples will be considered. They consist of two sets of points distributed within equal cubes, the distance between the cubes was regularly reduced and the cubes made to overlap at different degrees. In the second example, a small pesticide database corresponding to commercial organophosphorous compounds (OPC) was compared with a larger data base STRAC [16] including more

* Correspondence and reprints.

Received August 25, 1998; revised September 7, 1998; accepted September 15, 1998.

Article available at http://analusis.edpsciences.org or http://dx.doi.org/10.1051/analusis:1998182

than two thousands OPC. Comparison was performed in terms of chemical molecular diversity and of a more detailed data examination of the compounds according to their location in the cells of the corresponding SOM.

## Method

In order to explore the molecular diversity for comparing databases, the attention has to be focussed on three points: (i) the method of treatment of the information, (ii) the characteristics of the databases and (iii) the choice of the descriptors. In two previous papers [15,17] we showed, the advantages of SOM for the exploitation of molecular diversity in large structural databases in comparison with other techniques of classification and clustering. So, SOM was chosen as a strategic tool for analysing molecular diversity relative to different databases.

### *Self-organising maps*

SOM is a non-linear mapping technique, intended to give a representation of a given set of points in an *n*D multidimensional hyperspace by a set of points in a 2D map which directly corresponds to the human brain capacity to analyse a set of points. The representative set on a 2D map preserves the structure of the original set of points [18].

Let $M$ be the dimensionality of the original hyperspace, i.e. in our case $M$ is equal to the number of molecular descriptors. The SOM neural network consists of the input layer (which, in turn, consists of $M$ nodes which accept the input data, i.e. molecular descriptors) and the 2D Kohonen layer which are connected to each other, so that the data from each input layer node are connected to all the Kohonen layer, nodes. Each Kohonen layer node is characterized by $M$ adjustable weights corresponding to its connections to the input layer. Each point of the original data set is projected onto a cell of the map, the weights of which fit with its original co-ordinates, in the best way in terms of Euclidean distances. The training procedure consists in adjusting the weights, so that neighbouring Kohonen layer nodes would be characterized by close weights. Hence, the chemical compounds which are represented by close points in the multidimensional descriptor space are represented by close cells or even one cell on the Kohonen map. The SOM implementation, used in this study [19], includes some modifications of the original Kohonen algorithm. For example, a “conscience mechanism” recently suggested by DeSieno [20] to address the local minima problem was used. Another useful option applied was interpolation which increases the resolution of the 2D Kohonen map. Interpolation consists in calculating the co-ordinates of the representative point by taking into account the relative distances of the three closest points in the original space. An explicit technical description of the algorithm used in this study may be found in reference [19].

The reliability of projecting multidimensional data onto a 2D map was evaluated using the Root Mean Square (RMS) distance. This approach to the SOM validation introduced in our previous work [15,17] is based on a reverse operation consisting in projecting the data from the 2D map onto the original multidimensional data space. This RMS value accounts for the distances between the back-projected and original data points in the original data space and is expressed by the following formula:

$$\mathrm{RMS} = \frac{\sqrt{\sum_N \left(\bar{\vec{R}}_i - \vec{R}_i\right)^2}}{\sqrt{N \cdot M}} \tag{1}$$

where $\bar{\vec{R}}_i$ is the co-ordinate vector obtained by means of the reverse projection for the *i*-th data point, $\vec{R}_i$ is the vector of the initial co-ordinates for the *i*-th data point, $N$ is the number of data points, $M$ is the number of descriptors used to generate the hyperspace.

## Data

In this study, two types of data were used: virtual databases (*VD*) and real databases (*RD*) based on chemical structures.

### *Virtual data*

For demonstrating the potentialities of SOM, two sets of 200 points in a 3D space, simulating a multidimensional hyperspace, were randomly distributed with a uniform distribution function within two equal cubes with edges equal to 2. They were used to simulate a virtual database $VD_n$ constituted by the union of two virtual databases $VD_{n,a}$ and $VD_{n,b}$:

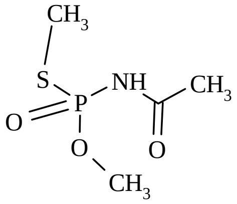
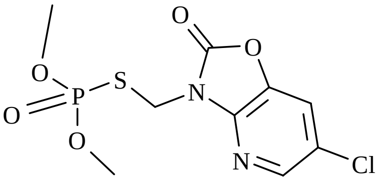
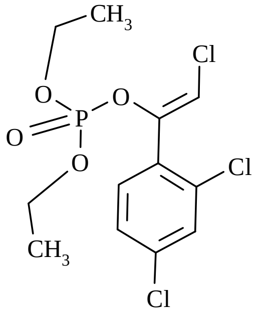
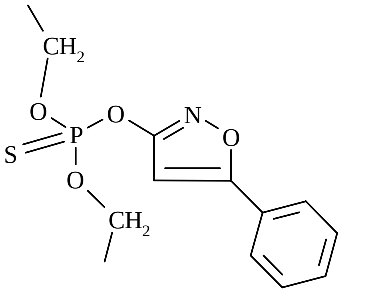
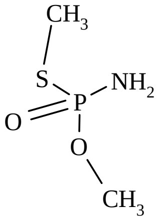
$$VD_n = VD_{n,a} \cup VD_{n,b} \; . \tag{2}$$

The SOM procedure was applied to the $VD_n$ database. $VD_{n,a}$ and $VD_{n,b}$ were positioned according to n different relative positions to explore the capability of SOM to analyse the structure of the data in the original hyperspace, or more simply, in this tutorial example, to explore the ability of SOM to separate potent clusters and/or to know to what degree they are overlapping.

### *Real data*

Two real databases were taken in order to simulate the real case which consists of the evaluation of the necessity to buy a new database for a company. The real databases used in this study were as follows:

- *Database 1*, $RD_1$, was a set of 98 commercial pesticides taken from the pesticide manual [21] molecular diversity of these compounds is illustrated by schema 1. This database included only organophosphorous compounds. Some common structures are given table I. This database formed by commercial pesticides was considered as a reference in order to evaluate the molecular diversity and then the degree of originality of another real database $RD_2$.
- *Database 2*, $RD_2$, is a part of STRAC [16] a large database of about 50 000 compounds tested as potent pesticides, with the same biological test. In order to facilitate the comparison between $RD_1$ and STRAC, two types of compounds were extracted from STRAC. Thus, database 2 consisted of two subsets of STRAC. The first subset, $RD_{2,a}$, a set of 94 compounds, represented the active OPC which, when applied in a well defined dose, caused more

| *Name* | *Molecular structure* |
|---|---|
| **Acephate** |  |
| **Azemethiphos** |  |
| **Chlorfenvinphos** |  |
| **Isoxathion** |  |
| **methamidophos** |  |

**Schema 1.** Example of molecular diversity of commercial pesticides.

than 85% of mortality for at least two of the six insect species tested. The second one $RD_{2,b}$ was exclusively formed by organophosphorous compounds inactive on five of the six insect species. This series contained 1059 compounds. Database 2 contained about 8% of active compounds, which can be considered as a correct ratio of active/inactive compounds in a database.

Kohonen Neural Network presents an apparent drawback, resulting from the randomly initialised 2D maps, which prevents a direct comparison of two SOMs even if issued from the same database represented in the same hyperspace. To avoid this drawback the comparison of two databases is done with help of a unique and complete calculation of global database RD in which the two previous ones are merged so that:

$$RD = RD_1 \cup (RD_{2,a} \cup RD_{2,b}) \ . \qquad (3)$$

All the 1215 compounds from *RD* were projected and visualised on the same SOM map. Then it was possible to compare the molecular diversity of commercial pesticide database $RD_1$ with that of active and inactive subsets $RD_{2,a}$ and $RD_{2,b}$ from STRAC.

### Molecular descriptors

While virtual databases were expressed in terms of points in a three dimensional space, the molecular structures forming the real databases were represented by points in a 45 multidimensional hyperspace derived from the 45 different molecular descriptors selected. This set of 45 descriptors includes topological, physico-chemical and electronic descriptors. Molar refraction, molar volume and molecular weight were used as size descriptors. The shape features of the molecules were characterised by these topological indices which account for the ramification degree, the oblong character, etc. In this study, 20 connectivity indices (${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^3\chi_C$, ${}^3\chi_P$, ${}^4\chi_P$, ${}^4\chi_{PC}$, ${}^5\chi_P$, ${}^5\chi_C$, ${}^6\chi_P$, ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi^v{}_C$, ${}^3\chi^v{}_P$, ${}^4\chi^v{}_P$, ${}^4\chi^v{}_{PC}$, ${}^5\chi^v{}_P$, ${}^5\chi^v{}_C$, ${}^6\chi^v{}_P$), 7 information content descriptors (**$I_V^D$**, $IC_N$, $SIC_N$, $CIC_N$, $N = 0, 1$), the Wiener index (*W*), the centric index of Balaban, the numbers of paths of lengths 1 to 4 and the numbers of nodes with 1 to 4 nearest neighbours were used. These descriptors were reviewed in detail by Basak [22] and Rouvray [23]. A lipophilicity descriptor represented by the octanol/water partition coefficient (log $P_{\text{oct/water}}$) was calculated using the Hansch and Leo method [24]. One more group of descriptors was derived from atomic electronegativity values (*nv*) according to Sanderson [25]. These descriptors were min(*nv*), max(*nv*) and mean(*nv*) in the molecule, and the variation of *nv* over the molecule.

The advantage of these descriptors is their ability to take into account not only the main structural features of each molecule, but also their global behaviour.

## Results and discussion

### Virtual data

For demonstrating the clustering and visual representation capabilities of SOMs eight virtual databases were created in a 3D space. The eight virtual databases correspond to: $VD_X = (VD_{X1} \cup VD_{X2})$ where *X* **equals** *a* to *h* (Fig. 1). $VD_{X1}$ and $VD_{X2}$ are two subsets. Each $VD_X$ database consisted of 400 points distributed randomly and uniformly in two subsets of 200 points within two equal cubes. In all examples, the edges of the cubes are equal to 2. The centre of one of the cubes corresponding to $VD_{X1}$ was taken as the reference point of the cartesian co-ordinates. All the faces of the two cubes were parallel to the corresponding co-ordinates. The centre of the second cube corresponding to $VD_{X2}$ was located at different points of the *Oz* axis, so that the distance between the centres of the two cubes decreased accordingly, as seen in the following examples (Fig. 1). In the first example of figure 1a the distance, *d*, between the closest parallel planes of the cubes equals 1, thus making up 50% of the cube size. In the second and third examples, given in figures 1b and 1c, this distance, *d*, equals 0.2 and 0.02, respectively making up 10% and 1% of the cube size, whereas in

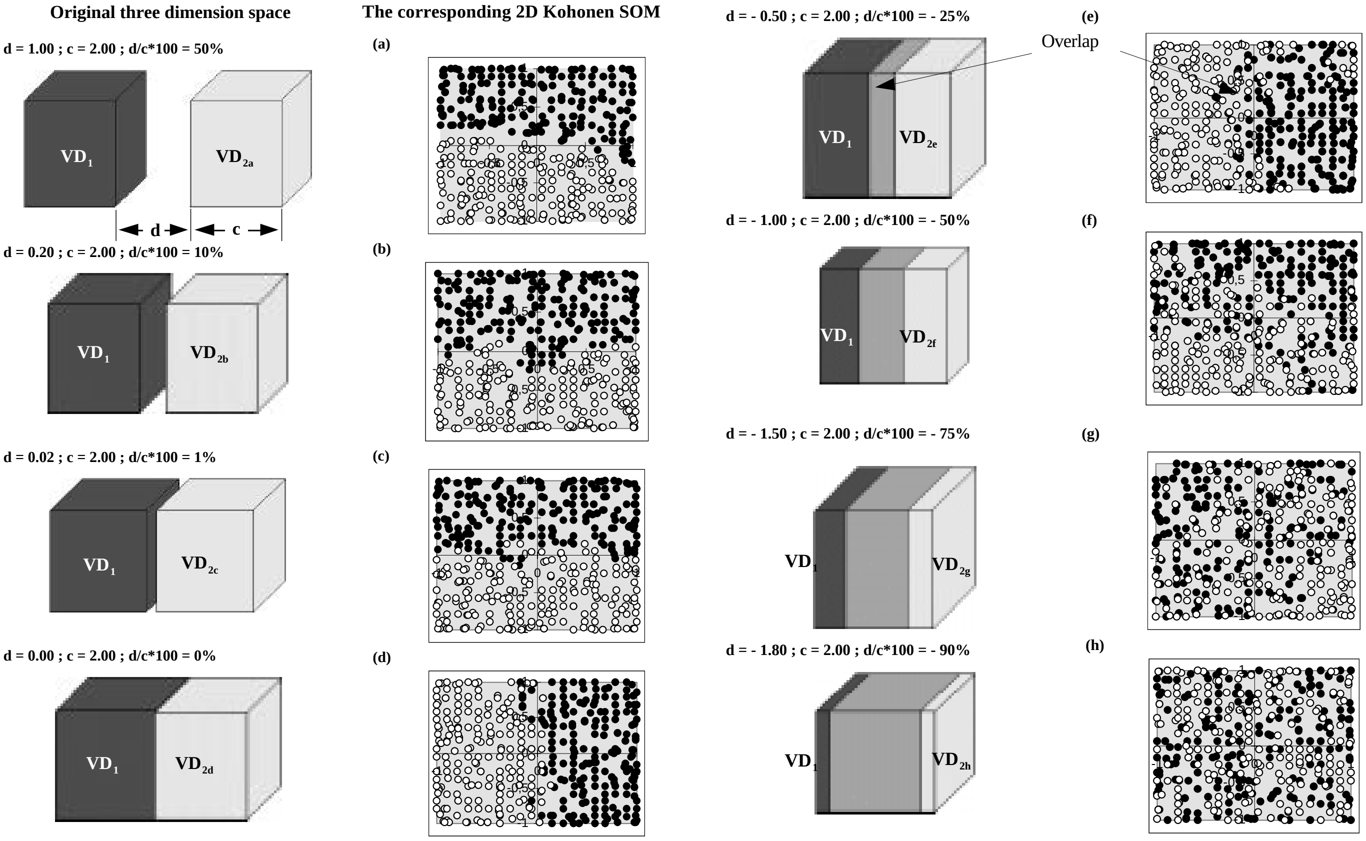

Original three dimension space
The corresponding 2D Kohonen SOM
d = 1.00 ; c = 2.00 ; d/c*100 = 50%
(a)
VD1
VD2a
d
c
d = 0.20 ; c = 2.00 ; d/c*100 = 10%
(b)
VD1
VD2b
d = 0.02 ; c = 2.00 ; d/c*100 = 1%
(c)
VD1
VD2c
d = 0.00 ; c = 2.00 ; d/c*100 = 0%
(d)
VD1
VD2d
d = - 0.50 ; c = 2.00 ; d/c*100 = - 25%
(e)
Overlap
VD1
VD2e
d = - 1.00 ; c = 2.00 ; d/c*100 = - 50%
(f)
VD1
VD2f
d = - 1.50 ; c = 2.00 ; d/c*100 = - 75%
(g)
VD1
VD2g
d = - 1.80 ; c = 2.00 ; d/c*100 = - 90%
(h)
VD1
VD2h

**Fig. 1.** Kohonen SOMs for two clusters of randomly distributed points within cube sets ranging from non-overlapping ones to overlapping ones. Each cube contains 200 points randomly distributed.

the fourth example (Fig. 1d) the distance, *d*, equals 0, thus the two cubes were adjoining each other. In the second series of examples given in figures 1e to 1h, the cubes were partially made to overlap. The overlap volume equals 25%, 50%, 75% and 90% of one of the cubes, in the examples selected. The corresponding SOMs are shown on the right side of figure 1. Filled circles on the maps correspond to the points distributed within the black cube, open circles correspond to the second cube. The overlapping areas of the cubes are shown in grey. It is clearly seen that two clusters on the SOM are separated by a gap, if there is a sufficiently large distance between the clusters of the original set (Figs. 1a and 1b). If there is only a small (about 1% of their size) gap between the original clusters, or if the two original clusters adjoin each other, there is a small overlap rather than a gap between their representative clusters on the map (Figs. 1c and 1d). When original clusters overlap, there is a corresponding overlap on the SOM. The larger the overlap is in the original hyperspace, the greater it is in the projection on the 2D map obtained with the SOM procedure (Figs. 1e-h). Other more complex virtual examples not presented here have been constructed in spaces with higher dimensionalities and have given similar results.

### Real data

An interpolated and non-interpolated 10 × 10 SOM for the *RD* database including 1251 congeneric organophosphorous compounds were obtained after about 40 000 iterations. The global map, with its non linear procedures, demonstrates the property of SOM to maximise the areas occupied by the representative points, thus increasing the resolution power in comparison with other projection techniques. The resulting interpolated and non-interpolated maps were divided into three maps corresponding to the three data sets, namely commercial pesticides database $RD_1$, the active compounds from STRAC $RD_{2a}$ and the inactive compounds from STRAC $RD_{2b}$. The three maps obtained using the interpolate option are shown in figure 2. In addition, figure 3 also presents the more usual non-interpolated maps and the corresponding number of compounds included in each Kohonen cell, also called node. The validation of these Kohonen maps were determined with the root mean square value (RMS*)*. All the maps possessed a RMS value not exceeding 0.06 which is very satisfactory relatively to similar cases studied elsewhere.

The interpolated map allows to visualise the distribution of different sets of compounds on the Kohonen map. Commercial compounds occupy only one part of the map, the active compounds from STRAC have a larger distribution and the inactive ones are distributed over the whole map. Nevertheless, in order to quantify the overlap and the distribution of compounds for analysing and comparing the three data sets, it is easier to work with the non-interpolated map.

### Comparison of commercial databases

#### SOM analysis

The commercial pesticides of $DB_1$ occupy 33% of the nodes among which, 14 nodes are occupied only by one

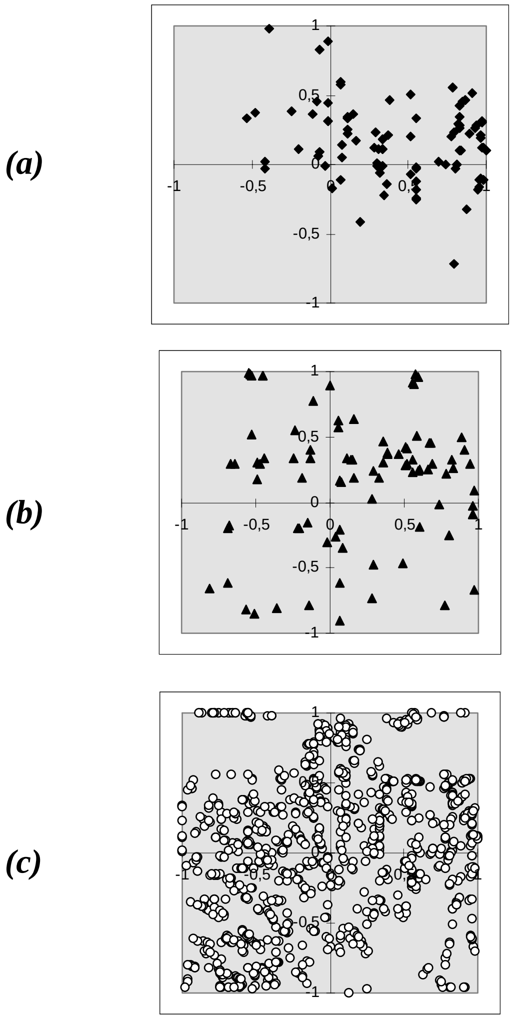


**Fig. 2.** Organophosphorous compounds database projected on the Kohonen map with the interpolate option: (a) Commercial pesticide compounds $RD_1$, (b) Active compounds from STRAC $RD_2$ and (c) Inactive compounds from STRAC $RD_3$.

compound. The STRAC active compounds in $DB_2$ occupy 46% of the nodes with 23 nodes occupied by only one compound. Thus, the active compounds from STRAC cover a larger area than the commercial pesticides. Nevertheless, among the active compounds, 63 out of 94 were found to fall into the same nodes as the commercial compounds. Like the commercial compounds, the STRAC active compounds are essentially concentrated around the upper right quarter of the map. Outside this region, some active compounds form several other independent clusters in other areas of the map. These compounds are probably active; but due to their high toxicity they are bad candidates for commercial compounds.

### Structural analysis

Two categories of areas were found on the map i.e. areas containing only STRAC active compounds and areas containing active compounds from STRAC and commercial pesticides.

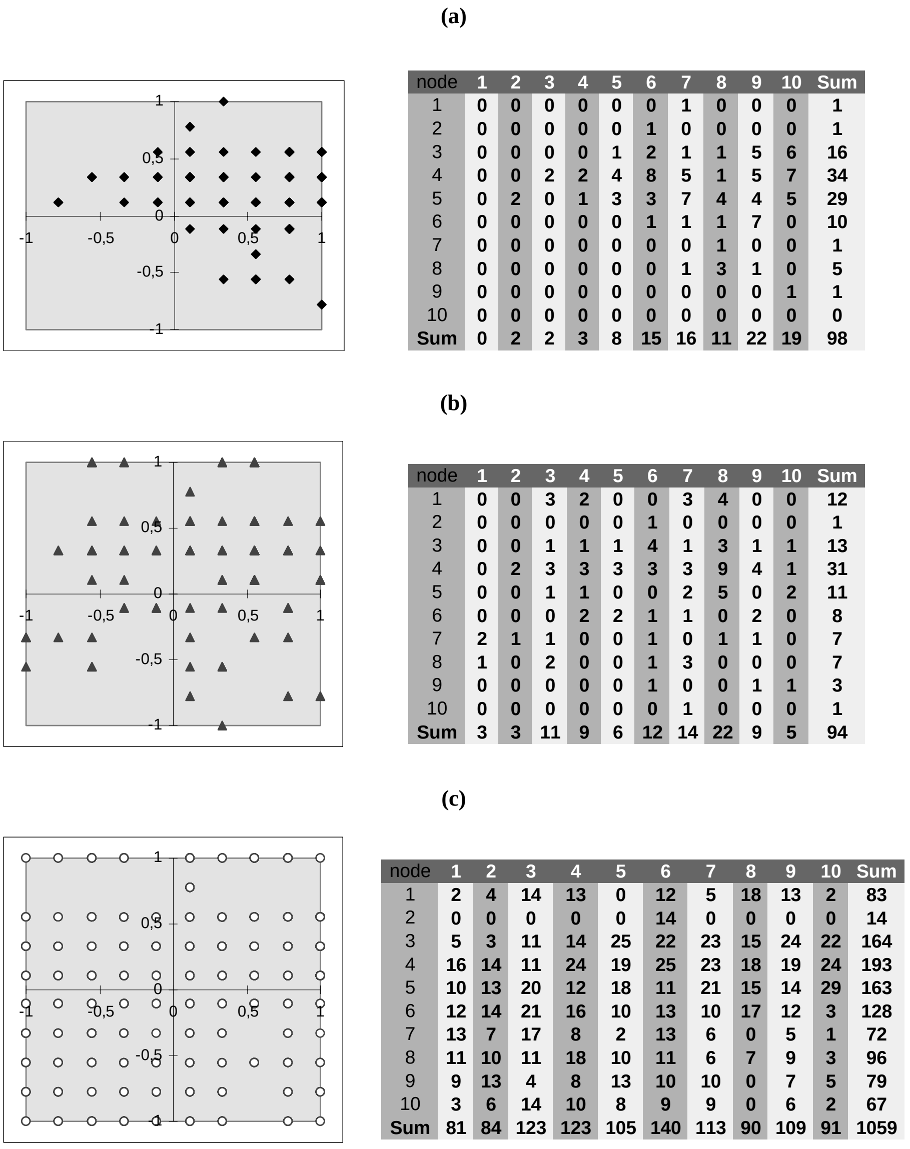
The first category consists of three areas including only the active compounds from STRAC. These areas and some included molecular structures are shown in figure 4. The structural analysis of the compounds from these areas was performed. It was shown that only organophosphorous compounds with two or three phosphorus are located in these areas, whereas among all commercial pesticides only one compound is of this type. This first result suggests a larger molecular diversity for the active compounds from STRAC than for the active commercial compounds. Analysing this map, the active compounds found in cluster 3 which present a methyl non-leaving group plus a P-S bond which exhibits a good leaving group (Fig. 4). This leaving group is known to play a major role in the biological activity organophosphorous compounds (OPC). Several authors, by means of different methods such as mutagenesis [26], Nuclear Magnetic Resonance [27], Molecular Modelling [28] studies, showed that OPC interact with acetylcholinesterase in mammals, insects, and other species, by irreversibly blocking the catalytic site of the enzyme. Moreover, it was shown that highly toxic organophosphorous compounds possess a small non-leaving group, a methyl group for example, and a good leaving group at the P-S, P-F bonds, etc. Such compounds were used as warfare agents like sarin, soman, VX [29,30], etc. These remarks suggest that some active insecticide compounds from STRAC, like the compounds included in cluster 3, were excluded from commercial clusters because they are too toxic for other organisms in the environment. So, such compounds are bad candidates for new commercial insecticide compounds.

The second category of areas contains both active compounds from STRAC and commercial ones. Some molecular structures of commercial and STRAC active compounds are shown in figure 5. Cluster 3 includes compounds with halogen atoms. Clusters 1 and 2 include only aromatic compounds. The three active compounds from STRAC found in the nodes together with commercial ones have either (i) a good leaving group with a bad non leaving group, e.g. the active compounds from cluster 3, or (ii) a bad leaving group but with a good spatial arrangement of the substituents around the phosphorus atom, as shown in the STRAC active compounds from clusters 1 and 2. However, it must be underlined that in these clusters 1 and 2 no compounds will exhibit both of the above mentioned favourable features which generally lower the toxicity for the environment species, but not for pests. The variations in the homology between the acetylcholinesterases of different species and the phenomenon of transport into the organisms are responsible for the behaviour of these compounds. In addition, cluster 4 contains only the one active compound from STRAC that is also present in the commercial database. It is coherent with the robustness of the classification method, a compound belonging simultaneously to the two data bases can be found twice and in the same node. Nevertheless, no other STRAC active compound was found in this node which contains five commercial compounds, which underlines the differences in redundancies between the two databases.

## Comparison of the commercial pesticide database with the inactive compounds from STRAC

### SOM analysis

As far as commercial pesticides are concerned, 82 out of 98 of them are distributed within the domain occupying 25% of



(a)

| node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 0 | 1 | 2 | 1 | 1 | 5 | 6 | 16 |
| 4 | 0 | 0 | 2 | 2 | 4 | 8 | 5 | 1 | 5 | 7 | 34 |
| 5 | 0 | 2 | 0 | 1 | 3 | 3 | 7 | 4 | 4 | 5 | 29 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 7 | 0 | 10 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 1 | 0 | 5 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sum | 0 | 2 | 2 | 3 | 8 | 15 | 16 | 11 | 22 | 19 | 98 |

(b)

| node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 3 | 2 | 0 | 0 | 3 | 4 | 0 | 0 | 12 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 1 | 1 | 1 | 4 | 1 | 3 | 1 | 1 | 13 |
| 4 | 0 | 2 | 3 | 3 | 3 | 3 | 3 | 9 | 4 | 1 | 31 |
| 5 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 5 | 0 | 2 | 11 |
| 6 | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 2 | 0 | 8 |
| 7 | 2 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 7 |
| 8 | 1 | 0 | 2 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 7 |
| 9 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 3 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| Sum | 3 | 3 | 11 | 9 | 6 | 12 | 14 | 22 | 9 | 5 | 94 |

(c)

| node | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 14 | 13 | 0 | 12 | 5 | 18 | 13 | 2 | 83 |
| 2 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 14 |
| 3 | 5 | 3 | 11 | 14 | 25 | 22 | 23 | 15 | 24 | 22 | 164 |
| 4 | 16 | 14 | 11 | 24 | 19 | 25 | 23 | 18 | 19 | 24 | 193 |
| 5 | 10 | 13 | 20 | 12 | 18 | 11 | 21 | 15 | 14 | 29 | 163 |
| 6 | 12 | 14 | 21 | 16 | 10 | 13 | 10 | 17 | 12 | 3 | 128 |
| 7 | 13 | 7 | 17 | 8 | 2 | 13 | 6 | 0 | 5 | 1 | 72 |
| 8 | 11 | 10 | 11 | 18 | 10 | 11 | 6 | 7 | 9 | 3 | 96 |
| 9 | 9 | 13 | 4 | 8 | 13 | 10 | 10 | 0 | 7 | 5 | 79 |
| 10 | 3 | 6 | 14 | 10 | 8 | 9 | 9 | 0 | 6 | 2 | 67 |
| Sum | 81 | 84 | 123 | 123 | 105 | 140 | 113 | 90 | 109 | 91 | 1059 |

**Fig. 3.** Organophosphorous compounds projected on the Kohonen map and the distribution of compounds in each Kohonen node with the non-interpolate option: (a) Commercial pesticide compounds alone, (b) Active compounds from STRAC alone and (c) Inactive compounds from STRAC alone.

the whole map. This result confirms the relatively low molecular diversity among all the commercial compounds due to the great attention brought in selecting those with high activities on pests and low toxicities on the environment. The inactive compounds from STRAC occupy the whole map, even the nodes containing commercial compounds, 32 nodes contain both commercial and inactive compounds. These nodes contain 98 commercial compounds and 525 inactive compounds. In the extreme case, with a 10 × 10 map and with a uniform distribution of compounds, each node would contain 1 commercial compound and about 10 inactive compounds, thus making the average number of commercial compounds equal to 9% of all compounds (100 out of 1000 + 100 = 1100). If only the nodes containing commercial compounds were taken into account, their number in these nodes would, on average, rise to about 16% of all the compounds in these nodes (98 out of 525 + 98 = 623). Moreover, some nodes contain more than 20% of commercial compounds. All these ratios are represented in figure 6 where it is shown that the number of commercial compounds for some nodes arises to 37%. Thus, in these nodes, the probability to find a potential commercial compound is higher. The richness, *R*, in commercial compounds, of a non-empty node, can be defined as follows:

$$R^{ij} = N_X^{ij} \;/\; (N_X^{ij} + N_Y^{ij}) \times 100 \qquad (4)$$

where *i* and *j* correspond to the co-ordinates of the node on the Kohonen map (in this case *i*, *j* = 1, ..., 10); $N_X^{ij}$ is equal to the number of commercial compounds in a node with the co-ordinates *ij*; $N_Y^{ij}$ is equal to the number of STRAC inactive compounds in a node with the same co-ordinates *ij*.

The enrichment, *E***,** of a node by commercial compounds can be defined as follows:

$$E^{ij} = R^{ij} \;/\; R^0, \qquad (5)$$

where $R^0$ is the richness of each node on the map with uniformly distributed compounds (in our case $R^0 \approx 9\%$). *R* and *E* values can be useful for estimating molecular diversity when comparing large databases containing mainly active compounds (for a given map, *R* and *E* values are proportional to each other).

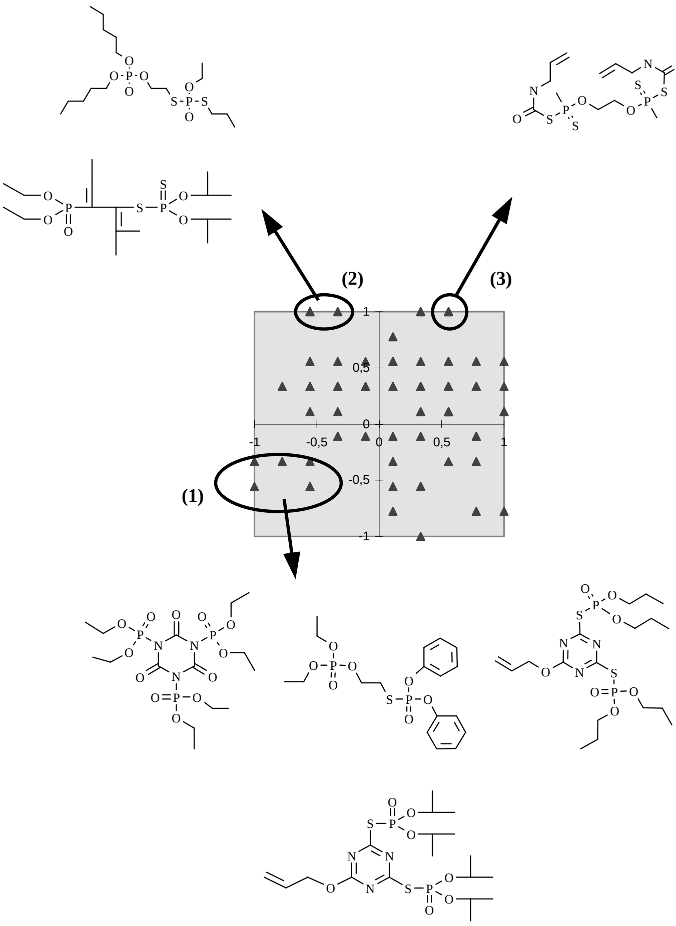


**Fig. 4.** Examples of molecular structures found in specific areas of active compounds belonging only to STRAC database.

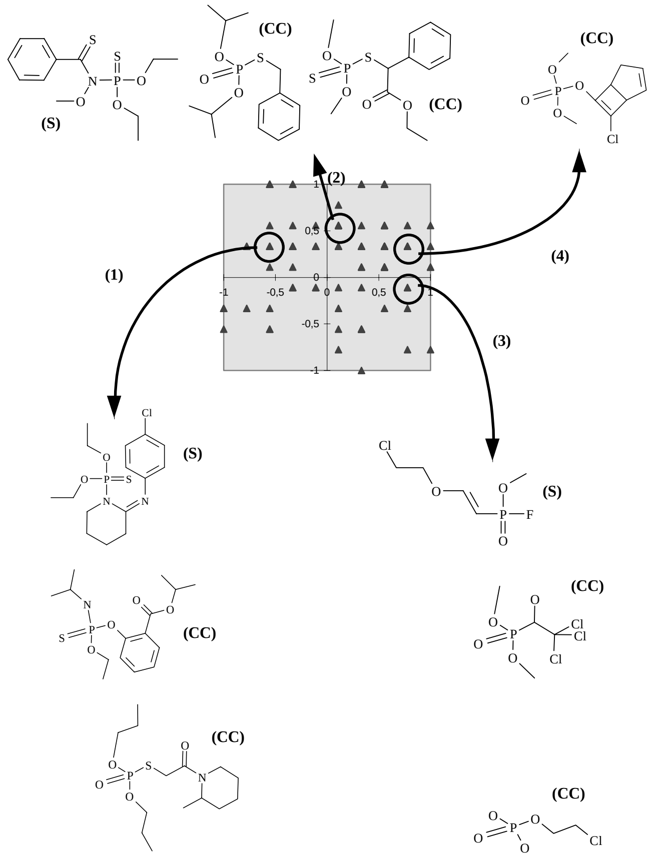


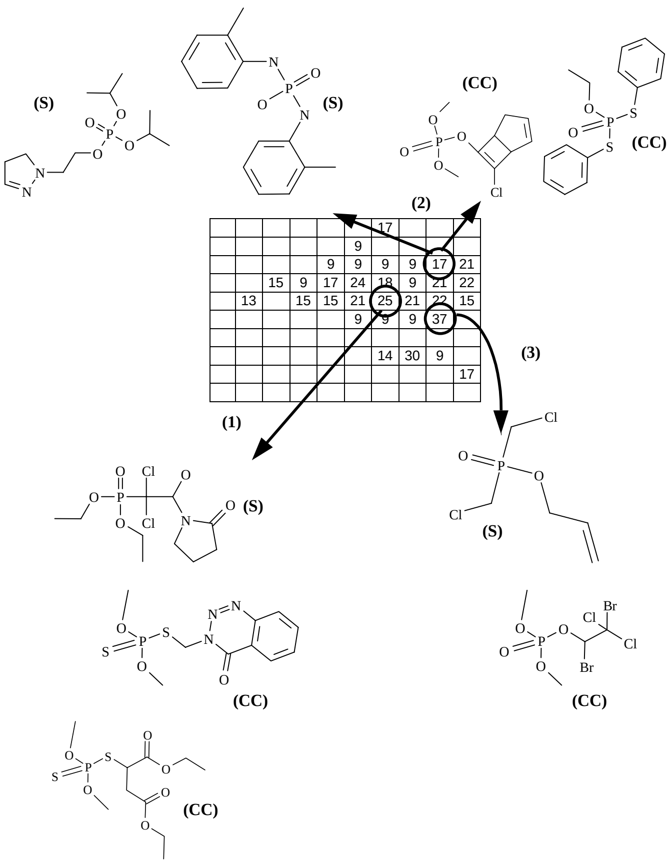
**Fig. 5.** Examples of molecular structures of active compounds from STRAC database (S) found into nodes occupied by commercial pesticides. The molecular structures of some commercial compounds are shown (CC).

### *Structural analysis*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 17 | | | |
| | | | | | 9 | | | | |
| | | | | 9 | 9 | 9 | 9 | 17 | 21 |
| | | 15 | 9 | 17 | 24 | 18 | 9 | 21 | 22 |
| | 13 | | 15 | 15 | 21 | 25 | 21 | 22 | 15 |
| | | | | | 9 | 9 | 9 | 37 | |
| | | | | | | | | | |
| | | | | | | 14 | 30 | 9 | |
| | | | | | | | | | 17 |
| | | | | | | | | | |

**Fig. 6.** Enrichment level of commercial compounds compared to inactive compounds. The values included in each Kohonen node express the ratio commercial compounds/inactive compounds from STRAC.

The structural analysis of clusters containing both STRAC inactive compounds and commercial compounds (Fig. 7) allowed to draw three conclusions: (i) Cluster 1 contains a majority of commercial compounds which, in addition to a good leaving group at the P-S bond, possess a carbonyl group in fourth position from the phosphorus atom. For this cluster the richness of the nodes in commercial compounds equals 25%. It was shown that such compounds are highly toxic on fish [28]. Thus, these commercial compounds are not well optimised. This means that these compounds could be considered as non commercial compounds. This result could explain also the presence in the same node of a STRAC inactive compound with a similar structure but without a good leaving group, as shown in figure 7. (ii) The STRAC inactive compounds from clusters 2 and 3 that are presented here, compared to commercial compounds, possess either bad leaving groups or bulky non leaving group thus inducing problems to enter the catalytic site of the enzyme. (iii) This last point concerned methodological aspects. The overlap between commercial compounds and STRAC inactive compounds could finally be explained by the limited discriminating power of the molecular descriptors chosen here. Other descriptors, more sensitive to the biological activity of acetylcholinesterase inhibitors, would have to be imagined by molecular modelling at the level of the catalytic site [28] and 3D QSAR studies of OPC and added to the list of the descriptors used or applied separately as a new set of descriptors. The other explanation could be the way these compounds were synthesised. In fact, synthesis of compounds is generally oriented, and this is the other reason why a great number of STRAC inactive compounds have molecular descriptors similar to those of some



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 17 | | | |
| | | | | | 9 | | | | |
| | | | | 9 | 9 | 9 | 9 | 17 | 21 |
| | | 15 | 9 | 17 | 24 | 18 | 9 | 21 | 22 |
| | 13 | | 15 | 15 | 21 | 25 | 21 | 22 | 15 |
| | | | | | 9 | 9 | 9 | 37 | |
| | | | | | | | | | |
| | | | | | | 14 | 30 | 9 | |
| | | | | | | | | | 17 |
| | | | | | | | | | |

**Fig. 7.** Examples of the molecular structures of inactive compounds from STRAC database (S) found in nodes occupied by commercial pesticides. A few molecular structures of some commercial compounds are shown (CC).

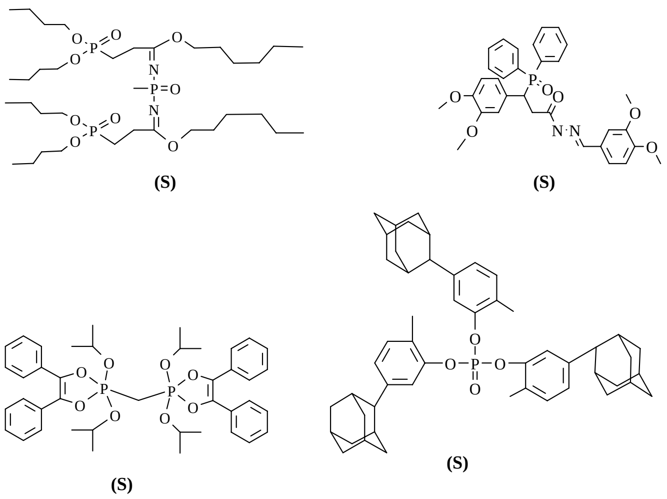


**Figure 8.** Molecular diversity of STRAC inactive compounds. These isolated molecular structures from STRAC must be compared to those of figure 7.

commercial ones. A last observation can be made as regards to the distribution of STRAC inactive compounds on the Kohonen map. The chemical separation of the data set containing only STRAC inactive compounds is evident. In fact, the compounds projected onto the upper part of the SOM map in figure 3c, and isolated from the others, have structures different from those of other compounds. Some of them were visualised in figure 8. Even if many inactive compounds fall into the same clusters as the commercial ones, there is, nevertheless, a wide molecular diversity among the inactive compounds of this data set of STRAC. In figure 8 compounds are shown whose high molecular weights are due to several phosphorus centres and bulky substituents. Since the catalytic site of acetylcholinesterases is known [28,32], it is possible to explain the cause of their inactivity. They are too big to be able to enter the active site and their activity is not optimised, e.g. compare the structures of STRAC active compounds from figures 4 and 5 and those of STRAC inactive compounds from figures 7 and 8.

## Conclusion

Kohonen Neural Networks, originally presented as Self Organising Maps (SOM) exhibit interesting projecting and visualising properties. From a methodological point of view, the most distinctive feature of SOM Neural Network is its ability to represent the original multidimensional set of points on a two dimensional map while preserving the order of distances between the points but in a non linear way.

The main features of SOM are presented here with simple virtual data bases which help us to understand how the distribution of the points in nodes of a 2D map can reflect the structure of the complex original hyperspace, via a non linear procedure. This SOM procedure is applied also to a real case: the comparison of two data bases. They deal respectively with the commercial organophosphorous pesticides and with a large number of organophosphorous compounds (OPC) tested as pesticides. This comparison is based on the analysis of molecular diversity by starting from the complex hyperspace established by a set of 45 molecular descriptors to arrive to an easily understandable 2D map. It must be underlined that to apply the SOM projection procedure the choice of an appropriate set of descriptors remains fundamental to obtain a proper hyperspace exhibiting a good discrimination, i.e. allowing a good appraisal of molecular diversity, among all the compounds of the considered chemical data bases.

The following salient features can be deduced from this study:

(1) *For comparing two databases on organophosphorous compounds as potent pesticides, it was shown that SOM Neural Network could be used as a powerful and reliable tool for exploring and comparing molecular databases of different origins*. The development of specific tools for the characterisation of these Kohonen 2D maps makes it possible to establish classical statistics in order to quantify the different aspects of the chemical diversity between and inside the compared data bases.

(2) *For the first time SOM was applied to the comparison of large series of congeneric compounds*. So despite the fact that the two data bases considered have only 1% of identical active compounds, it was possible to interpret the structural effects on the congeneric compounds, in an understandable way, on chemical and biological bases.

More generally, this study suggests new strategies and complementary tools in experimental drug design in connection with High Throughput Screening, for the search of new leads, and for the risk assessment of new OPC. Work is underway to develop both these points [33].

## References

1. Kansy, M.; Senner, F.; Gubernator, K. *J. Med. Chem.* **1998**, *41*, 1007-1010.
2. Zhao, Y. Z.; van Breemen, R. B.; Nikolic, D.; Huang, C. R.; Woodbury, C. P.; Schilling, A.; Venton, D. L. *J. Med. Chem.* **1997**, *40*, 4006-4012.
3. Pötter, T.; Matter, H. J. *J. Med. Chem.*, **1998**, *41*, 478-488.
4. Cummins, D. J.; Andrews Webster, C.; Bentley, J. A.; Cory, M. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 750-763.
5. Taylor, R. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59.
6. Kohonen, T. Self-Organization and Associative Memory, Springer-Verlag, Berlin, 1988.
7. Ferran, E. A.; Pflugfelder, B.; Ferrara, P. *Protein Sci.* **1994**, *3*, 507.
8. Arrigo, P.; Giuliano, F.; Milanesi, L. *Adv. Mol. Bioinf.* **1994**, 159.
9. Gasteiger, J.; Li, X.; Uschold, A. *J. Mol. Graph.* **1994**, *12*, 90.
10. Barlow, T. W. *J. Mol. Graph.* **1995**, *13*, 24.
11. Melssen, W. J.; Smit, J. R. M.; Rolf, G. H.; Kateman, G. *J. Chemo. Intell. Lab. Syst.* **1993**, *18*, 195.
12. Novic, M.; Zupan, J. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454.
13. van Osdol, W.W.; Myers, T. G.; Paull, K. D.; Kohn, K. W.; Weinstein, J. N. *J. Natl. Cancer Inst.* **1994**, *86*, 1853.
14. Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. *J. Chem. Inf. Comp. Sci.* **1996**, *36*, 1205-1213.
15. Kireev, D. B.; Ros, F.; Bernard, P.; Chretien, J. R.; Rozhkova, N. In Computer-Assisted Lead Finding and Optimization, van de Waterbeemd, H.; Testa, B.; Folkers, G. Eds., Verlag Helvetica Chimica Acta, Basel and Wiley-VCH, Weinheim, 1997; pp 255-264.
16. STRAC, Pesticide Structure-Activity Database, Moscow Plant Protection Institute, Moscow.
17. Kireev, D. B.; Bernard, P.; Chretien, J. R.; Ros, F. *SAR QSAR Environm. Res.* **1998**, *8*, 93-107.
18. Kohonen, T. *Procs. IEEE* **1990**, *78*, 1464.
19. Neural Computing: A Technology Handbook for NeuralWorks Professional II/Plus and NeuralWorks Explorer, NeuralWare, Inc, Technical publications Group, 202 Parc West Drive, Pittsburgh, PA 15275.
20. DeSieno, D. Adding a Conscience to Competitive Learning, in, Proc Int Conf on Neural Networks, v1, IEEE Press New York, 1988.
21. The Pesticide Manual, 7th ed, Worthing, C. R. Ed., The British Crop Protection Council, Lavenham, Suffolk, GB, 1983.

22 Basak, S. C. A nonempirical approach to predicting molecular properties using graph-theoretic invariants, in Practical Applications of Quantitative Structure-Activity Relationships QSAR in Environmental Chemistry and Toxicology, Karcher, W.; Devillers, J. Eds., Kluwer Academic Publishers, Dordrecht, 1990; p 83.

23. Rouvray, D. H. *Stud. Phys. Theor. Chem. Appl. Topol. Graph. Theory* **1983**, *28*, 159.
24. Hansch, C.; Leo, A. Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley Interscience Publication, 1979; p 13.
25. Sanderson, R. T. Chemical bonds and bond energy, Acad Press, N-Y, 1976.
26. Ordentlich, A.; Barak, D.; Kronman, C.; Ariel, N.; Segall, Y.; Velan, B.; Shafferman, A. *J. Biol. Chem.* **1996**, *271*, 11953-11962.
27. Segall, Y.; Waysbort, D.; Barak, D.; Ariel, N.. Doctor, B.P.; Grunwald, J.; Ashani, Y. *Biochemistry* **1993**, *32*, 13441-13450.
28. Bernard, P., Kireev, D. B., Chretien, J. R. *J. Mol. Model.* 1998 (under press).
29. Benschop, H. P.; Konings, C. A. G.; van Genderen J.; De Jong, L. P. A. *Toxicol. Appl. Pharmacol.* **1984**, *72*, 61.
30. Albaret, C.; Lacoutiere, S.; Ashman, W.P.; Froment, D.; Fortier, P.L. *Proteins: Struct. Funct. Genet.* **1997**, *28*, 543.
31. Handbook of Organophosphorus Chemistry, Engel, R. Ed., New York, 1992.
32. Sussman, J.L.; Harel, M.; Frolow, F.; Oefner, C.; Goldman, A.; Toker, L.; Silman, I. *Science* **1991**, *253*, 872-879.

33 Chrétien, J.R.; Kireev, D. B.; Bernard, P.; Fortier, P. L.; Coppet, L. 8th International Workshop on QSARs in the Environmental Sciences, Baltimore (USA), 16-20 May 1998.