PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher’s version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/29276

Please be advised that this information was generated on 2019-03-03 and may be subject to change.
Test Series Selection from Nonlinear Neural Mapping

Daniel Domine*, James Devillers
CTIS, 21 rue de la Bannière, 69003 Lyon, France

Dietrich Wienke and Lutgarde Buydens
Catholic University of Nijmegen, Laboratory for Analytical Chemistry, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

Abstract

A new nonlinear neural mapping (N2M) technique based on the combined use of Kohonen self-organizing map (KSOM), minimum spanning tree (MST), and nonlinear mapping (NLM) is introduced for optimal test series selection. With the N2M method, KSOM results are enhanced by the visualization of the actual distances between the loaded neurons from MST and NLM. N2M provides an easily interpretable and comprehensible graphical display which guides the selection of representative test series especially when the number of individuals is high. In addition, structure-activity relationships can be derived. The approach is open since any information useful for data interpretation can be plotted by means of graphical tools.

Key words: nonlinear neural mapping, Kohonen self-organizing map, minimum spanning tree, nonlinear mapping, selection of test series, structure-activity relationships.

1 Introduction

The research for new biologically active chemicals in medicinal chemistry requires inspection of several thousands of candidates for which only limited information is available. Since it is practically impossible to test all the possible candidates, a decision must be taken as where to start from. This implies the adoption of strategies for a rational selection of the most relevant compounds for preliminary biological testing and for the derivation of (quantitative) structure-activity relationships ((Q)SARs). This kind of strategies is known as optimal test series design. A lot of work has been directed towards this aim and numerous methods have been proposed [1-3]. Significant advances dealt with the use of linear multivariate methods [4] and linear graphical display methods [5-10] allowing selection of test series by simple visual inspection of 2-D maps summarizing the information content of a matrix of physicochemical properties. Even if these approaches have been successfully used, it was pointed out that from a practical point of view, none of them are completely satisfactory and there is a need for new tools providing complementary information or overcoming methodological problems [2, 3]. In this context, we recently proposed the use of an original graphical approach based on the nonlinear mapping (NLM) method [2, 3, 11]. We showed that it was possible to obtain easily interpretable nonlinear maps of aromatic and aliphatic substituent constants for the selection of test series and the derivation of SAR. However, it is obvious that when the number of individuals becomes too high, the interpretation of the maps is more difficult. In addition, the selection of optimal test series from nonlinear maps could still be improved if the substituents could be grouped into clusters containing substituents with similar properties. Thus, one could simply select substituent in the clusters in the different regions of a 2-D map. Under these conditions, we propose the use of a newly derived mapping technique called nonlinear neural mapping (N2M) [12] for solving the above problems. This new graphical approach combines the advantages of the Kohonen self-organizing map (KSOM), of the minimum spanning tree (MST), and of the nonlinear mapping (NLM) method.

2 Nonlinear Neural Mapping (N2M)

The N2M algorithm (Figure 1) combines three pattern recognition techniques. First, the Kohonen map forms an initial step for data reduction of p multivariate individuals defined in an n-dimensional space into v clusters (loaded neurons). However, information about the correct distances between the neurons disappears during the projection onto the 1, 2, or 3-D array of nodes. One possibility to overcome this distance problem is to add information that allows the simultaneous visualization of the distances between the loaded neurons in the map. The technique recently chosen by Wienke and coworkers [13-16] is a calculation of an MST [17, 18] between the loaded neurons of a trained KSOM in order to visualize the shortest distances between them. These two steps constituted the basis of the 3MAP algorithm designed by Wienke and coworkers [13-16]. As there remains information not represented about the correct distances between all the loaded neurons, in the third step of an N2M analysis, an NLM of the loaded neurons is performed in order to visualize all the distances separating them [12].

2.1 Kohonen Self-Organizing Map (KSOM)

The Kohonen model of self-organization [19-21] is based on the idea that the brain tends to compress and organize sensory data, spontaneously [22]. A Kohonen network consists of two layers of neurons (i.e., input and output) (Figure 2). The input layer contains n neurons corresponding to the n variables describing the individuals. The output layer can be a 1, 2, or 3-dimensional geometrical arrangement of u neurons. In this study, a 2-dimensional array of neurons presenting a square shape was used. The u neurons of the input layer are all connected to each of the u neurons of the output layer so that these u neurons can be considered as u n-dimensional
weight vectors. The Kohonen algorithm consists in a repeated comparison of the $p$ $n$-dimensional individuals with the $u$ weight vectors using any distance metric such as the Euclidean distance. In the classical algorithm, the weight vectors $W_j$ of the neurons $j$ on the Kohonen map are initialized with small random numbers. In the present study, we used a slightly different method called convex combination [14, 15] in which all weights are initialized with a constant value. Then, during each comparison, a winner ($W_c$) among the $u$ weight vectors can be found. It corresponds to the weight vector having the highest degree of similarity (i.e., smallest distance) with the $i$-th considered individual $X_i$. Then, the $n$ weights connected to the winning neuron $W_c$ are updated in order to become closer to the considered individual using the Kohonen learning rule:

$$W_{c}^{new} = W_{c}^{old} + \eta (X_i - W_{c}^{old})$$

where $W_{c}^{new}$ and $W_{c}^{old}$ represent the weight vectors of the winning neuron after and before the update, $X_i$ is the $n$-dimensional vector of sample $i$, and $\eta$ is the learning rate with $0 < \eta < 1$. Together with the winning neuron, $W_c$, the nodes that are topologically close in the neuron array up to a certain distance (i.e., within a limited topological neighborhood of radius $R$) activate each other in order to learn from the same input sample. One sequential comparison of all the $p$ samples with all $u$ neural weight vectors and their modification is called one epoch. From one epoch to the next, the neighborhood $R$ and the learning rate $\eta$ are slightly decreased. They both tend to zero as the number of epochs increases. After repeating the training process over a large number of epochs, a self-organizing behavior of the $p$ samples in the low $m$-dimensional neural array can be observed. They form a visual topological structure. The final result of this training process is an aggregation of subsets of input vectors dedicated to a number $v$ of weight vectors in the neural array with $v \leq u$. Wienke and coworkers [13–16] named these $v$ neurons "loaded neurons". Additional information on KSOM can be found in numerous papers and books [19–34].

### 2.2 Minimum Spanning Tree (MST)

The theory related to the construction of minimum spanning trees has been described by different authors [18]. Briefly, it can be summarized as follows. A set of $p$ objects to be classified the ones in relation to the others, may be considered as a set of points in a space. The representation is classical if the objects are described by a series of $n$ parameters. In that case, one has $p$ points in the IR$^n$ space. More generally, if one has only the values of an index of dissimilarity, which does not necessarily satisfy the properties of a distance, one may represent the objects as points of a plane, for example. Each couple of objects is then connected by a continuous line to which the value of the dissimilarity index is linked. Thus, the set of objects and the index values are represented by a complete valued graph. However, if the number of objects is too important, this type of representation becomes inextricable.

One may then try to extract from this graph a partial graph having the same nodes but fewer edges. Among all the partial graphs that can be constructed, those that have a tree structure are particularly interesting, since they may be represented on a plane. Thus, a tree is connected graph (i.e., a line may connect any vertex couple). Moreover, it has no cycle (i.e., one may not have in that tree an edge coming from and converging on the same point without taking twice the same edge). One may consequently define, in the same way, a tree with $p$ vertices as being a connected graph without a cycle and with $p - 1$ edges. The length of a tree is the summation of the length of its edges. An MST is constructed when the length is minimal. Many different algorithms have been proposed for the construction of trees of a minimal length [18]. Readers are advised to consult the article of Devillers and Doré [18] for more information on the theoretical foundations of MST as well as examples of applications. For the N2M technique, we applied the Prim's algorithm [17]. In N2M, this algorithm finds the shortest possible connections between all $v$ loaded neurons with the constraint that no closed connections are allowed.

### 2.3 Nonlinear Mapping (NLM)

NLM [35, 36] allows to represent a set of points defined in an $n$-dimensional space by a human perceivable configuration of the data.
in a lower m-dimensional space \((m = \text{two or three})\) which we will call either display space or nonlinear map. NLM tries to preserve distances between points in the display space as similar as possible to the actual distances in the original space. The procedure for performing this transformation can be summarized as follows.

(i) Interpoint distances in the original space are computed. The Euclidean distance is the most widely used.

(ii) An initial configuration of the points in the display space is chosen. Most often, the coordinates of points in the display space are set in a random manner.

(iii) A mapping error \((E)\) is calculated from distances in the two spaces.

(iv) The coordinates of the points in the display space are iteratively modified by means of a nonlinear procedure so as to minimize the mapping error.

(v) The algorithm terminates when no significant decrease in the mapping error is obtained over the course of several iterations.

In the present study, an Euclidean distance was selected and the steepest descent procedure was used as minimization method [35, 36].

Additional information on the practical aspects of the NLM method, a review of its uses in QSAR studies and original examples of applications can be found in previous papers [2, 3, 11, 36-40].

### 3 Experimental Section

An N2M analysis was performed on a set of 103 aliphatic substituents (Table 1) [41] described by means of five substituent constants encoding their hydrophobic, steric, and electronic effects. These parameters were respectively, the hydrophobic constant for aliphatic substituents \(F_r\), the molar refractivity (MR), H-bonding acceptor (HBA) and donor (HBD) abilities, and the inductive parameter \(F\). For the N2M analysis, \(F_r\), MR, and \(F\) were centered (i.e., zero mean) and reduced (i.e., unit variance). HBA and HBD were not centered and the “1” were replaced by a value yielding a unit variance. This matrix was chosen for its relatively large number of individuals but also because comparison could be made with a recently published NLM analysis of the same data matrix [3].
Different N2M analyses were performed. Indeed, several sizes of networks were assayed and run for 10000 epochs. During learning the topological radius was regularly decreased during 8000 epochs and trained for further 2000 epochs. The learning rate was also decreased.

Calculations were performed by means of the 3MAP and STATQ-SAR packages [42].

4 Results and Discussion

4.1 KSOM Results

Figure 3 shows the 5 × 5 Kohonen map with a final radius of 0. It is worth noting that by adjusting the size of the network, it was possible to obtain more or less clusters (results not shown). In the present study, a 5 × 5 array of neurons provided 19 clusters (Figure 3). From Figure 3, it would be possible to select optimal test series by taking one or several substituents in each of the clusters formed. For comparison purposes, we have represented on the KSOM, an example of substituent selection performed in a previous study [3] (underlined bold numbers in Figure 3). It is noteworthy that the results obtained are in accordance with those previously obtained since there is only one (or two) substituents selected per neuron. There remain neurons with no selected substituents. These neurons can be used if a larger set of substituents is required. Figure 4 which shows the projection of the KSOM clusters on the nonlinear map recently derived [3] clearly indicates the complementarity of the two methods. Indeed, KSOM allows to define clusters which are coherent with the nonlinear map. This may help to divide the nonlinear map into a given number of zones from which selection of optimal test series can be made by picking up one or several substituents depending on the population of the clusters and the repartition of the substituents on the map.

Figure 3. KSOM of aliphatic substituents. Underlined bold numbers represent the substituents selected in a previous study [3].

Figure 4. Representation of the KSOM clusters on the nonlinear map [3]. See Figure 3 for correspondence between the letters and the clusters. This figure has to be compared with Figure 2, page 982 of the article of Domine et al. [3].

4.2 N2M Display

Even if a map with 132 substituents remains relatively easy to inspect and interpret, it is obvious that there are limitations in the number of points which can be displayed on a 2-D map. Therefore, the N2M display, which is a nonlinear mapping of the loaded

Figure 5. N2M display of aliphatic substituents. The values reported are the actual distances in the MST. See Figure 3 for correspondence between the letters and the clusters. See text for the significance of the arrows.
neurons on which an MST has also been projected, allows to reduce
the number of points to a few number of clusters (loaded neurons)
and take into account the distances separating the different clusters.
Inspection of the N2M display (Figure 5) indicates that the topology
of the Kohonen map is kept. Thus, for example, it can be noted that
the alignment of neurons A to E is preserved on Figure 5. One can
also note that neuron F is located between neurons C and K. The
same remark can be made for all other neurons. The only slight
divergence is observed for neuron I. Indeed, it can be noted that
I is no longer between neurons A and N on the N2M display. In
this case, the map which preserved the actual distances between
the neurons lost the correct topology. However, it must be noted
that this loss does not concern three adjacent neurons and that
the alignment of the three adjacent neurons I, J, and K for example
is kept. Projection of the MST allows to underline the shortest dis-
tances between the loaded neurons.

4.3 Graphical Interpretation and Chemical Significance of the
N2M Display

To facilitate the interpretation of the N2M results and also the selec-
tion of optimal sets, graphical tools can be used. Thus, for ex-
ample, it is possible to represent the weight vectors of the neurons
by means of collections of graphs where the larger the squares the
higher the values of the weights for each variable and the larger the
circles the lower the values of the weights (Figure 6). In Figure 6,
for example, we can underline that the neurons located in the bot-
tom left-hand part of the map (i.e., neurons P, Q, and R) are asso-
ciated with large weight values for the HBD descriptor (Figure 6.3)
and low weight values for the Fr (Figure 6.1) and MR (Figure 6.4)
descriptors. From a general point of view, the collection of graphs
displayed in Figure 6 explains the repartition of the clusters on the
map. Indeed, as the weight vectors are generally representative of
the properties of the substituents included in each cluster (i.e.,
loaded neuron), Figure 6 gives a fair view of the repartition of
the clusters with regards to the physicochemical properties of
the substituents they contain. It is obvious that a better description
could be obtained by plotting the actual properties but the weight
vectors generally represent a good first approximate.

In order to give a full description of the N2M display in terms of
structural information and underline structure-property relation-
ships for the 103 aliphatic substituents under study, various struc-
tural features have been reported on the N2M display (Figure 7).
Thus, a series of graphs was drawn on which was represented
the presence of substituents containing given functional groups.
Figure 7 shows the relative position of the clusters as well as the
coherence of the map from a chemical point of view. Figure
7.1 shows that substituents containing primary or secondary amine
groups are all located in clusters P, Q, and R. Comparison with
Figure 6 reveals that these clusters are characterized by low weight
values for Fr (Figure 6.1) and high weight values for HBD ability
(Figure 6.3). Compared with the nonlinear map (results not shown),
it is however not possible to separate the substituents with general
formula NHR with R = H, alkyl, or phenyl (no. 9, 29, 46, 64, 80,
and 89) from the other substituents containing a carbonyl group (no.
20, 23, 38, 39, 54–56, 74, 75, and 96).

In Figure 7.2, we have represented in black the clusters containing
substituents having CN or NO₂ groups in their structure. All the
substituents containing these groups (no. 5, 13, 14, 32, 33, and
48) are located in the same region of the N2M display in clusters
A, B, and C. Comparison of Figure 7.2 with Figure 6.1 shows that
this location can be explained by low weights for Fr. In addition,
these substituents have the ability to accept but not to donate H-
bonds.

Figure 7.3 shows that the distribution of the clusters containing
substituents with an -S- group is less obvious and presents slight
differences with the distribution observed on the nonlinear map
[3]. Even if we still observe in separated clusters substituents
no. 8 and 90 (clusters N and H, respectively), the distribution of
the other substituents is less obvious. The cyclic substituents no.
66 and 67 (i.e., 2- and 3-Thienyl) can be found in a same cluster
(S) together with SC₃H₇ (no. 63). SCH₃ (no. 28) and SC₃H₅ (no. 44)
are in cluster I.

Figure 7.4 represents the presence of substituents containing C=O
or C=S groups in the clusters. These clusters can be separated into
two groups, i.e., clusters A, B, C, D, E, G, and H at the top of the
Figure 5. Representation of functional groups on the N2M display (Figure 5). Presence of substituents with a given structural feature is indicated in black or grey. See text for interpretation.

Figure 7. Representation of functional groups on the N2M display (Figure 5). Presence of substituents with a given structural feature is indicated in black or grey. See text for interpretation.

Figure 8. Inclusion of the 12 substituents in only four clusters (Figure 8) indicates that the selection made by the authors [43] was not represented in the starting group. It is noteworthy that the results obtained are in accordance with classical linear multivariate methods. Compared with classical NLM, the advantages of the clustering method and the display technique simplify the map and guide the selection.

4.5 Deriving Structure-Activity Relationships from the N2M Display

N2M is particularly suitable for underlining relationships between the structure of the molecules and their biological response(s). For illustrative purposes, an example dealing with the antifungal activity (i.e., \( P.I. = \log \left( \frac{1}{I_{50}} \right) \) where \( I_{50} \) values express the molar concentration of chemical for 50% inhibition of mycelial growth) of 1-(3,5-dichlorophenyl)-3-substituted-2,5-pyrolidinediones against Botrytis cinerea is presented (Figure 8) [43]. From a set of 22 \( p_{100} \) values [43], we kept the 12 values corresponding to substituents contained in our set. It has to be noted that the remaining ten substituents did not carry much more information and that the 12 selected substituents were representative of the set. The N2M clusters containing these 12 substituents are shown in Figure 8. Inclusion of the 12 substituents in only four clusters (Figure 8) indicates that the selection made by the authors [43] was not representative of the whole physicochemical space. Despite this, Figure 8 underlines relationships between the structure and the antifungal activity of the studied chemicals. Thus, the least active substituents both containing an S = O group are included in cluster A. Clusters 1 and 1 contain alkyl and thiocarbonyl substituents which are more active. The most active substituents being alky1s, phenyl, and thiocarbonyls with a size larger than the alkyls and thiocarbonyls found in clusters 1 and 1, are included in cluster S. Comparison of Figure 8 with Figure 6 shows, for example, that the antifungal activity increases with hydrophobicity and MR (Figures 6.1 and 6.4, respectively). Decreasing F values (Figure 6.5) also seem to increase the antifungal activity. These results confirm those published in the literature [43] which showed, from classical regression equations, a relationship between the antifungal activity of 1-(3,5-dichlorophenyl)-3-substituted-2,5-pyrolidinediones and log P. A new examination of these data revealed that both the size of the substituents and their dipole moment were also important [44].
5 Concluding Remarks

The purpose of our study is not to advance that N2M is the new panacea for solving the complex task of substituent selection. We only show that N2M is a powerful tool in the kit of chemometric methods providing useful information for selecting substituents. Indeed, in drug design, it is essential to select test series with high information content in order to reduce the costs in research by maximizing the information content obtained from each molecular probe in a set of congeners. In the same way, for the derivation of valuable QSAR from reduced sets of compounds, selection of sets presenting a maximum meaningful variety is necessary. Therefore, a lot of work have been directed towards this aim and many authors have proposed different methods. Among them, factorial design and D-optimal design have proved to be successful in many applications [e.g., 1, 45, 46]. However, practical problems arise when the dimensionality of data matrices becomes too high since visual inspection of the different factorial planes is a long and tedious task. In addition, when the number of individuals becomes high, the maps appear as galaxies of points hardly interpretable. N2M addresses these two problems since the use of a nonlinear technique allows one to obtain a single easily interpretable map and the use of a clustering technique prior the mapping allows one to obtain easily readable maps with fewer numbers of "points". In the present paper, to stress the heuristic potency of N2M, an example containing 103 substituents described by five physicochemical variables has been presented but the technique has also been successfully applied to a data matrix containing 604 chemicals described by nine variables [47]. It is also noteworthy that our approach is open and allows the graphical representation of any information susceptible to help in the interpretation of the results.

![Figure 8. Representation of the antifungal activity of twelve 1-(3,5-dichlorophenyl)-3-substituted-2,5-pyridinediones against Botrytis cinerea. The measured activity was the concentration inducing 50% inhibition of mycelial growth (10⁻⁵₀) [43]. The cluster in black contains two substituents having an SO group in their constitution (i.e., no. 26 (SOCH₃) and 62 (SOC₂H₅)) whose values of 10⁻⁵₀ (µmol l⁻¹) are superior to 5. The clusters with striped lines contain substituents no. 6 (H), 22 (CH₃), 28 (SCH₃), and 44 (SC₂H₅) whose values of 10⁻⁵₀ values (µmol l⁻¹) range from 2.09 to 3.24. The cluster in light grey contains larger substituents (i.e., no. 40 (CH₂CH₂), 57 (CH(CH₃)₂), 58 (C₂H₅), 63 (SC₂H₅), 76 (C₂H₅) and 86 (C₂H₅) whose values of 10⁻⁵₀ (µmol l⁻¹) range from 1 to 2.00.)

3MAP and STATQSAR are homemade software. Information on 3MAP can be obtained from Dr. D. Wienke at the University of Nijmegen. STATQSAR is commercially available upon request to CTIS, 21 rue de la Bannière, 69003 Lyon, France. However, note that N2M combines three algorithms which are widely described in the literature and directly available as source code on the Internet. Thus, for example, the best way for obtaining the KSOM is to use the following FTP connection: cochlea.hut.fi (130.233.168.48). An NLM program is also available at this address. In addition, the complete NLM algorithm with formulas of parameters specifically designed for QSAR studies is provided in Domine et al. (J. Chemometrics 7, 227–242 (1993)). Last, the MST algorithm can be obtained at the following address: http://biomserv.univ-lyon1.fr/ADE-4.html. In addition, numerous information on the different MST algorithms available in the literature can be found in Devillers and Doré (Ecotoxicol. Environ. Safety 17, 227–235 (1989)).


Received on April 7th 1996; accepted on June 18th 1996.