Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counterpropagation Neural Network

Marjana Novič* and Jure Zupan

National Institute of Chemistry, SLO-61115 Ljubljana, Hajdrihova 19, Slovenia

Received May 12, 1994[⊗]

Two different artificial neural networks (ANNs) for infrared spectra analysis are presented: the self-organizing Kohonen ANN for mapping of the infrared spectra into a 2-D plane and the counterpropagation ANN for determination of the structural features of organic compounds based on their infrared spectra. The preliminary learning in the Kohonen ANN with **all** spectra from the collection yields the information of possible grouping. The preliminary grouping has been used for the separation of spectra into the training and into the test set containing 755 and 2529 "spectrum-structure" pairs, respectively. The counterpropagation ANN trained on the "spectrum-structure" pairs from the training set has the ability to predict, with an average prediction ability of 0.77 and an average reliability of 0.82, structural fragments of an unknown compound from its infrared spectrum. Additionally, the counterpropagation ANN offers the possibility to simulate the infrared spectra from the structure representation.

INTRODUCTION

The difficulty of establishing reliable correlations between the chemical structure and the spectrum of the corresponding compound has not been overcome completely. Especially the interpretation of infrared spectra is not straightforward and requires a lot of expertise. Therefore, the scope of automatic interpretation of infrared spectra is open for further investigation that draws the attention of experts from different fields in chemistry and informatics. The recently introduced method, namely artificial neural networks (ANNs), has shown to be promising in this challenging spectroscopic problem.[1−8] In the review[9] on the application of ANNs in different fields of chemistry it can be seen that the spectra-structure correlation problems took about 25% of all chemical applications. The reviewed papers mainly apply the so called "error-back-propagation" learning method[10−12] in a multilayer ANN.

The error-back-propagation is a supervised learning method. This means that for learning a set of input-target pairs $\{X_s,T_s\}$ is required. In the case of spectra-structure problem the input $X_s = (x_{s1},x_{s2},...x_{sj}...x_{sm})$ is a spectrum of the sth compound represented by m intensity values. The corresponding target $T_s = (t_{s1},t_{s2},...t_{sj}...t_{sn})$ is a binary vector which indicates the presence ($t_{sj} = 1$) or absence ($t_{sj} = 0$) of the jth substructure in the sth structure. Thus, the learning of an ANN means to force the ANN to respond for each input spectrum X_s in the training set with the output Out_s identical to the target T_s. This is achieved by the iterative feeding of all training spectra X_s to the ANN and correcting the weights of the neurons according to the differences between the targets T_s and actual outputs Out_s. The error-back-propagation learning requires that, regardless of a particular structural environment, the presence of a certain substructure is **always** signaled by an exactly predefined output neuron. This requirement is specifically hard to implement in the domain of infrared spectroscopy. Due to a large variety of structural environments which can cause completely different infrared spectra (different inputs), the assignment of **only one** output neuron to each substructure feature seems to give too little flexibility to the predictive scheme. In other words, a system with the one-to-one correspondence between the fragments and output neurons seems to be too rigid to be able to adapt to the "infrared spectra-structure" correlation problem.

Therefore, the use of an unsupervised learning strategy[3,7−9] like the Kohonen ANN,[13−15] offers another perspective to the solution of this problem. The results of the Kohonen ANN can be used for the selection of the training-test sets of spectra which can further on be used in the supervised learning method, called counterpropagation ANN[16−18]. In the present paper we will first, briefly describe both: the Kohonen unsupervised learning and the counterpropagation supervised learning method, and, second, we will show that this methods can yield satisfactory spectra-structure correlation.

METHOD

Learning Strategies. Let us consider for a moment the ANN as a black-box with m inputs and n outputs only. After the training, such a box should be able to achieve one of the following tasks: to yield for any given signal X_s $(x_{s1},x_{s2}, ... x_{sm})$ the predefined target vector T_s $(t_{s1},t_{s2}, ..., t_{sn})$, to activate for any object X_s, belonging to the class q, a neuron within the qth segment in the output layer of neurons, or to achieve a 2-D distribution (a map) of input vectors although they are not split in classes.

The "supervised" learning requires for training a set of input−output pairs (X_s,T_s) with X_s $(x_{s1},x_{s2}, ... x_{sm})$ being the m-variable input vector (spectrum, multicomponent analysis, protein sequence, etc.) and T_s $(t_{s1},t_{s2}, ..., t_{sn})$ the n-response output or set of n targets associated with each X_s. During the supervised training, the output vector Out_s is calculated for each individual input X_s and is compared to the target T_s. After comparison, a corrective measure is taken according to the particular ANN strategy to change the weights in

* Corresponding author: Laboratory of Chemometrics, National Institute of Chemistry, Hajdrihova 19, SLO-61115 Ljubljana, Slovenia. Phone: 386-61-1760-200; Fax: 386-61-1259-244.

[⊗] Abstract published in *Advance ACS Abstracts,* April 15, 1995.

0095-2338/95/1635-0454$09.00/0

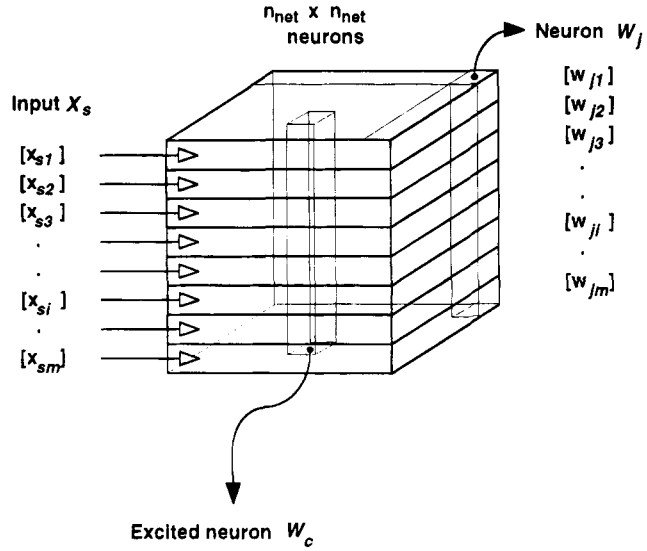

Figure 1. Kohonen ANN. The neurons are represented as columns and are arranged in a quadratic matrix ($n_{net} \times n_{net} = N$). The length of the columns (m) corresponds to the number of weights and consequently to the dimension of the input vector X_s (shown left). All the weights of the ith level receive the ith component of the input vector X_s simultaneously. The excited neuron W_c is chosen on the basis of the "best match" criterion, eq 1.

such a way that the corrections will assure better agreement between the Out_s and T_s.

The input of **all** p objects X_s to the network and the corrections of the weights p-times is considered to be **one** cycle (epoch) of the learning procedure. The learning of all objects is repeated until either the agreements between all T_s and the produced outputs Out_s are acceptable or until the number of iterations is exceeded.

On the other hand, the unsupervised learning requires only knowing the input vector X_s, while the associated category or class q to which the input vector X_s belongs is required merely for checking the results and **not** for the training. In the unsupervised learning, the category is implied by **the position** of the vector X_s in the measurement space of variables. The training of the Kohonen ANN is mapping the groups of objects X_s from the m-variable measurement space to the regions that are eventually formed within the 2-D plane (or 1-D array) of neurons.

Kohonen ANN. The Kohonen ANN resembles the biological NN probably most closely of all ANN architectures and learning schemes. As a rule, the Kohonen type of a ANN is based on a single layer of neurons arranged in a 2-D plane having a well defined topology (Figure 1).

A defined topology means that each neuron has a defined number of neurons in its first-, second-, third-, etc., order neighborhood. Because of the ease of programming, in most applications the quadratic neighborhood is regarded as having eight and not four nearest neighbors. The topology, i.e., the same number of neighbors at all points of the Kohonen ANN, can be considerably improved if the "toroid" boundary conditions are fulfilled. However, the toroid conditions in the correction algorithm decreases the available mapping-area in the same dimensional network. The maximal **topological** separation of two neurons is $n_{net}/2$ neurons in the $N = n_{net} \times n_{net}$ dimensional toroid ANN. In the case where a large area for mapping is needed a two times larger maximal possible topological distance between two neurons can be a predominant factor against the use of a toroid condition. Larger maximal topological distance between neurons offers better possibility for separation of clusters in the ANNs of the same size. However, it has to be pointed out that the maximal **topological distance** does not necessarily mean the largest **difference** between the neurons.

All neurons obtain the same multidimensional input. The most characteristic feature of the Kohonen ANN that actually makes it very similar to the biological NN is its implementation of a local feedback only. This means that the **output** of each neuron does not affect (*via* corrections) **all** neurons in the layer but only a **small** number of the entire ensemble: the ones that are **topological** close to it. Such local feedback of corrections causes that topological close neurons start acting similar if similar input signals are presented to the network.

The learning procedure in the Kohonen ANN is usually referred to as a *competitive learning*. It means that after an input is presented to the network only **one** neuron from the entire population of neurons in the layer is selected. Actually, the neurons are competing among themselves as to which one will be stimulated. The selection of the winning neuron c (c for *central*) from among N neurons ($N = n_{net} \times n_{net}$) is made on the comparison between all weight vectors W_j ($w_{j1}, w_{j2}, \ldots w_{jm}$) and the input signal X_s ($x_{s1}, x_{s2}, \ldots x_{sm}$)

$$\text{out}_c \leftarrow \min \{\sum_{i=1}^{m} (x_{si} - w_{ji})^2\} \quad j = 1,2,\ldots N \qquad (1)$$

After the neuron c which best satisfies the selection criterion (1) of the best match is found, its weights, w_{ci}, are corrected in such a way that the response the next time will be better or closer to the desired one than before.

At the beginning of learning, not only the excited neuron c but also its neighbors up to the pth neighborhood, $c - p$, $c - p + 1, \ldots c - 1, c, 1, \ldots c + p$, are stimulated. However, to which p and to which extent the neurons are stimulated, depends on the parameters of learning strategy (eq 2). During the learning the range p to which the neurons are still stimulated decreases. The correction of the weights Δw_{ji} of the neurons at the topological distance d around the selected neuron is calculated as

$$\Delta w_{ji} = [(a_{max} - a_{min})(p/n_{net}) + a_{min}][1 - d/(p+1)] \times (x_{si} - w_{ji}^{old}) \quad d = 0,1,2,\ldots p \qquad (2)$$

The entire expression in front of the term $(x_{si} - w_{ji}^{old})$ in eq 2 describes how the correction of the weights w_{ji} decreases with increasing learning time and increasing topological distance d between the jth and the central neuron c. The maximum distance p to which the correction (eq 2) is applied is shrinking, $p = (i_{tot} - i_{it})n_{net}/(i_{tot} - 1)$, during the learning procedure. At the beginning of learning $(i_{it} = 1)p$ covers the entire network ($p = n_{net}$), while at the end of the learning iteration steps $(i_{it} = i_{tot})p$ is limited only to the central neuron $c(p = 0)$. No matter, whether the difference $x_{si} - w_{ji}^{old}$ in eq 2 is positive or negative, i.e., whether x_{si} is greater or smaller than the weight w_{ji}^{old}, the w_{ji}^{new} will be closer to x_{si} than it was the w_{ji}^{old}. This means that if a variable x_{si} accessing the weight w_{ji} of the jth neuron has produced a positive difference, the weight w_{ji} is increased and *vice versa*; if the produced difference is negative, the weight is diminished (see eq 2).

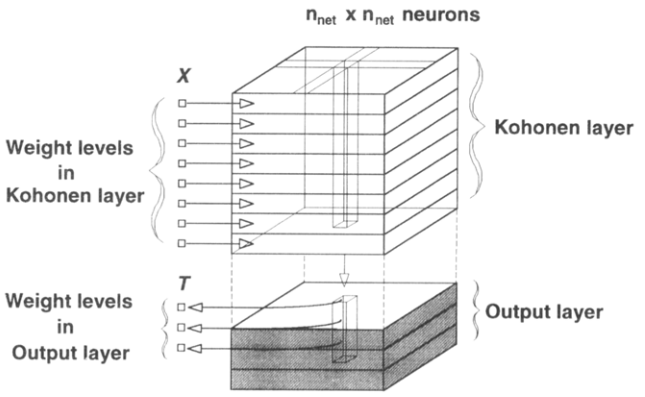

Figure 2. The counterpropagation ANN **after** the training. The excited neuron in the Kohonen layer defines the position of the neuron in the output layer, which carries the wanted "answer". During the training the target values T are input into the output layer—for the training the lower three arrows would be reversed.

The unsupervised training is usually carried out for a predefined number of training or iteration cycles, i_{tot}, although other stop criteria can be used like a minimal threshold for the total difference between all the input vectors and the corresponding winning neurons, or similar.

After i_{tot} number of training cycles is run through the ANN, the **complete** set of the training vectors X_s is run through the ANN again for checking the positions of neurons excited by the input vectors. If the neurons excited by the input vectors carrying certain information, for example, the presence of a feature we are seeking to classify, form topological well separated group(s), it can be concluded that learning has been successful and the procedure for prediction of the classification can be repeated with the test set.

Counterpropagation network. The counterpropagation ANN is a two layer ANN (Figure 2) consisting of a **Kohonen** and an **output** layer. The Kohonen layer in the counterpropagation ANN acts exactly in the same way as if it would act in the previously described Kohonen learning. The difference is the presence of the output layer which offers the possibility of training the network in the **supervised** manner. The output layer has similar layout of neurons as the Kohonen layer, but has a different function in the learning process. It serves for the storage of the answers (the contents of the look-up table) and for the dissemination of the response (target) values to the topological close neighbors.

The output layer has the same number and the same topological arrangement of neurons as the Kohonen layer has. Each neuron in the output layer has as many weights as there are responses (targets) associated with each input vector. The essential difference in the learning process between the Kohonen and the output layer is that the output layer **has no influence on the selection of the winning neuron.** The coordinates (the neuron's position) of the neuron in the output layer that has to be corrected are exactly the coordinates of the winning neuron in the Kohonen layer. At each learning step the neurons in the output layer are waiting for the winner to be selected in the Kohonen layer and only after this selection is made the correction of weights u_{ji} in the output layer (to distinguish them from the weights w_{ji} in the Kohonen layer) is performed in the following way

$$\Delta u_{ji} = [(a_{max} - a_{min})(p/n_{net}) + a_{min}][1 - d/(p+1)] \times (t_{si} - u_{ji}^{old}) \quad d = 0,1,2,...p \quad (3)$$

The parameters in eq 3 are same as in eq 2, the only difference being the substitution of the input vector component x_{si} with a component t_{si} of the target vector T_s. Step by step the weights w_{ji} in the neighborhood of the winning neuron and its counterparts u_{ji} in the output layer are corrected in such a way that they are becoming more and more similar to the components of the input object X_s and to the components of its accompanying target vector T_s, respectively.

It has to be emphasized again that the **position** of the winning neuron, and consequently the **position** where the corresponding target T_s $(t_{s1},t_{s2}, ..., t_{sn})$ is stored, is not influenced by the target itself. From this point of view, we could say that the counterpropagation learning strategy is a combination of supervised and unsupervised methods.

DATA SET

Representation of Spectra. The spectra used for the experiment were selected from the collection of infrared spectra of 3284 compounds.[8] The original spectra have 650 intensity points in the region from 4000 to 200 cm^{-1}. The spectral regions from 4000 to 2000 cm^{-1} and from 2000 to 200 cm^{-1} are covered by 200 (resolution of 10 cm^{-1}) and 450 intensities (resolution of 4 cm^{-1}), respectively. For our purpose 512 intensities in the range from 3500 to 550 cm^{-1} with the corresponding resolution was selected. On the sets of 512 intensity points the fast Hadamard transformation was applied.[19] Number 512 is the closest power of 2 number ($512 = 2^9$) required by any fast transformation. After the transformation on 512 intensities, the resulting sets of 512 Hadamard coefficients were truncated to 128 coefficients with the rest of the sets filled with zeros. All Hadamard coefficients were divided by the first one, and then the inverse Hadamard transformation was applied to get the 512 intensity points spectra back. Due to the 1:4 reduction of coefficients, the resulting 512 point spectra consist of groups of four equal intensities. Only one intensity point from each group is retained, hence, a 1:4 reduction of the original spectrum representation is achieved. The reduction reflects mainly in a lower resolution (high frequency terms).[19] The spectra reduced as described above were used throughout this work.

To illustrate the loss of information, the 512 intensity points spectrum is shown on Figure 3 in comparison with eight spectra obtained by the inverse transformation from different truncated sets of Hadamard coefficients. The increasing reduction of the coefficients causes an increasing loss of information. According to our preliminary studies,[19] the four-times reduced set is an acceptable trade-off between the loss of information and efficiency of handling large quantities of data. The speed and efficiency is specially required for the training of ANNs where thousands of iteration cycles are needed to achieve satisfactory results.

Representation of Structures. In order to have a simple and representative description of chemical compounds, for each structure in the study a 34 component vector determining the presence or absence of 34 structural fragments was determined. The 34 structural fragments were defined on the basis of their infrared absorption frequencies. We have based our selection of 34 representative structural fragments on the set of functional groups used previously.[1,2] Represented as the connection tables all 3284 compounds in the

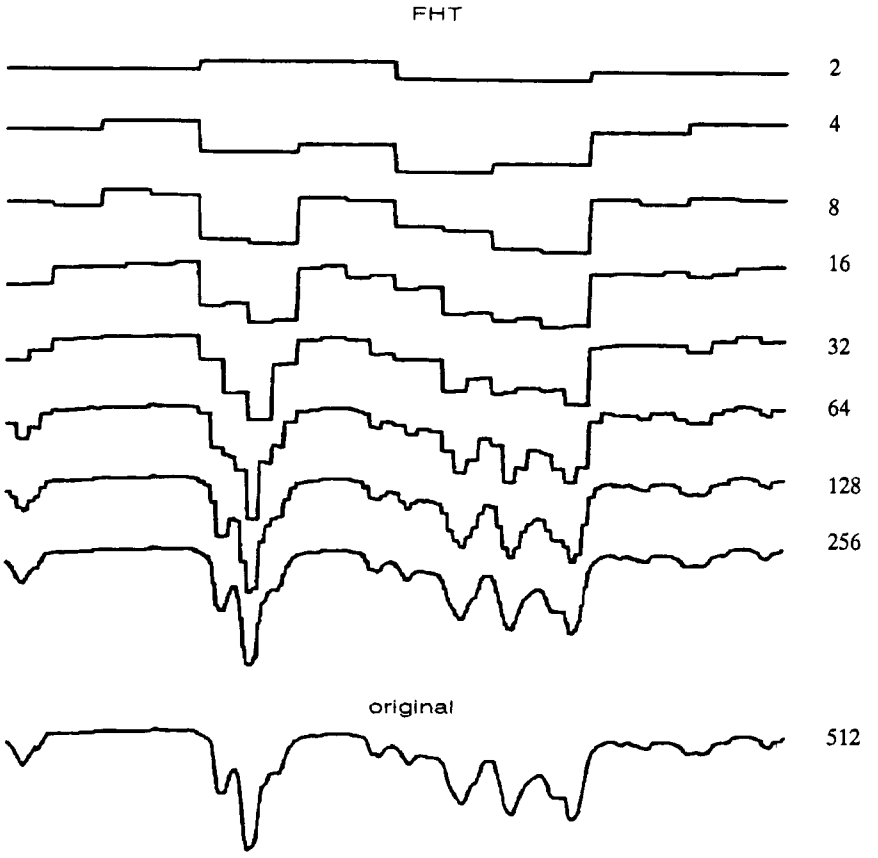

Figure 3. Original infrared spectrum of a compound containing an ether group, represented by 512 intensity values (bottom). After the fast Hadamard transformation (FHT) the coefficients are truncated, transformed back to 512 intensity values, and plotted. Eight back transformations using different truncations are shown above the original spectrum.

data set were scanned to mark each compound if it has one (or more) of the 34 fragments. In order to distinguish between the ether- and ester group as well as between the amine- and amide group, the general substructure search program[20] was customized for this purpose. The resulting 3284 structure representations were thus 34-dimensional binary vectors. The ones and zeros correspond to the presence or to the absence of a particular fragment, respectively. The selected fragments, their labels, and the total number of compounds having these fragments in their structure are shown in Table 1.

MAPPING OF INFRARED SPECTRA USING KOHONEN ANN

Architecture of Kohonen ANN. The architecture of Kohonen ANN is determined by the number of neurons $N = n_{net} \times n_{net}$, by the number of weights in each neuron m, and by the topology of the neighborhood, i.e., by the layout of neighbors of each neuron. The number of neurons, arranged in a quadratic matrix $n_{net} \times n_{net}$, defines the **size** of the plane where the spectra are mapped.

As the method[21,22] for selecting the training set we have used the Kohonen self-organized map of 3284 spectra. Our intention was to select approximately about 20−25% (between 650 and 820) of all spectra for the training set and to retain the rest for testing. Based on the previous evidence,[18] the basic hypothesis for the choice of the Kohonen ANN size is that **after** the training at least 20% of neurons should **not** be excited, thus providing the necessary "free" space for accommodating the different unknowns and outliers. For a training size of about 800 spectra and with taking into account about 5−10% unavoidable hits into the same neuron, an ANN of the size of about 900 neurons (i.e., 30 × 30 layout) seems to be a good choice. The deciding measure for the final selection of 755 training vectors was an even distribution of **all** input vectors on the Kohonen map. Hence, we took one spectrum from each excited neuron. Regarding the distribution of spectra in the entire measurement space for the choice of the objects for training it is better to consider the **variety** of the compounds' structures whose spectra are in the data collection than the total **number** of spectra. The detailed description of how the selection of the training set was made is explained later on.

The number of weights in the neurons is equal to the dimension of the input vector. In our case the input vectors are sets of 128 intensity values obtained as described in paragraph "Representation of Spectra". Before the training has started, all weights in the ANN were initialized by randomization in the interval [0, 1]; the parameters used in the Kohonen ANN learning are given in Table 2 (column 2).

The number of weights in each neuron defines the number of the *levels* (maps) in the Kohonen network. Additionally to 128 weight maps, the zeroth or the *top-level map* was constructed in order to see the distribution of labels (map of labels). The map of labels is generated at the end of the training. For every input spectrum at the positions of the most excited neuron in the 30 × 30 neuron matrix the corresponding box in the 30 × 30 label map receives the structure label of the input spectrum (see Table 1 for the description of labels). Obviously, with one label the compounds containing more than one of the selected

Table 1. Description of the Data Set

fragment		compds in the data set	one character label
1	OH	630	H
2	alcohol	462	M
3	prim. alcohol	288	P
4	sec. alcohol	171	S
5	tert. alcohol	30	T
6	1,2 glycol	29	G
7	phenol	68	F
8	aryl-CH_2OH	31	R
9	NH	412	N
10	prim. amine	250	1
11	sec. amine	146	2
12	CN	842	0
13	tert. amine	170	3
14	C=O	1006	C
15	COOH	87	A
16	COO−	415	W
17	ester	409	E
18	aldehyde	120	Y
19	ketone	284	K
20	amide	65	D
21	benzene	1055	J
22	naphthalene	28	4
23	furan	25	5
24	thiophene	29	6
25	pyridine	81	7
26	NO_2	83	Z
27	aryl-NO_2	60	V
28	CO−	1456	8
29	ether	471	O
30	C−X	841	X
31	C−F	218	U
32	C−Cl	438	L
33	C−Br	214	B
34	C−I	42	I

Table 2. Parameters Used in the Study for the ANNs' Architecture and Learning Procedures

item	Kohonen NN	counterpropagation NN
no. of layers	1	2
$n_{net} \times n_{net}$		
first layer	30 × 30	30 × 30
second layer		30 × 30
Dimension of neurons		
m − first layer	128	128
n − second layer		34
Weights' initialization	random (0.0−1.0)	random (0.0−1.0)
no. of input objects	3284	755
layout of the top-map	30 × 30	none
labels in the top-map	34	
form of the correction function (eqs 2 and 3)	triangular	triangular
no. of iterations (i_{tot})	65 680	75 500
epochs in training	20	100
p_{max}	30 neighbors	30 neighbors
a_{max}	0.5	0.5
a_{min}	0.01	0.01
toroid condition	no	no

structural fragments cannot be properly assigned. In such cases, simply the label of the first fragment found was given. Therefore, the labels have to be regarded only as a partial information. Nevertheless, the complete information about the presence of multiple fragments is available for each structure.

Due to the fact that we had about four times as many spectra as there are neurons in the network, one neuron can be excited by different spectra. Additionally, the spectra which excite the same neuron do not necessarily carry the same label—therefore, all 3284 labels should be displayed in the 30 × 30 map. The graphic display of the top map with 3284 labels is quite incomprehensible and difficult to read, and, therefore, special software was designed for inspection of the top-map. In order to give the feeling of how such a 900 fields map could look like, in Table 3 the frequency of excitations per neuron is presented. For example, the last five items in Table 3 show that six neurons in the 30 × 30 network were excited by 18 and more different spectra.

Table 3. Frequency Distribution of Multiple Hits N_{hits} of Total 3284 Objects on Each of 900 Neurons in the 30 × 30 Map

N_{hits}	n_{neu}	N_{obj}	N_{hits}	n_{neu}	N_{obj}
0	145	0[a]	12	20	240
1	184	184	13	2	26
2	112	224	14	7	98
3	113	339	15	3	45
4	68	272	16	4	64
5	76	380	17	5	85
6	37	222	18	1	18
7	43	301	19	2	38
8	30	240	21	1	21
9	20	180	22	1	22
10	13	130	23	1	23
11	12	132			
				900	3284

[a] 145 neurons out of 900 were not excited by any object (spectrum).

To improve the flexibility of information extraction, the formation of the top-map was designed in such a way that frequency distribution of hits for each fragment (not label!) on the entire 30 × 30 map can be displayed separately. In this way it was possible to obtain 34 "partial" top-maps, i.e., one map for each structural fragment. In Figure 4 "partial" top-map of all compounds having an ester group is shown in two different ways.

We shall now inspect the clustering of spectra of compounds containing one small structural fragments in common but having different structural features attached to this fragment. Let us take the −OH fragment as an example. Obviously, the −OH fragment is in alcohols, acids, and phenols; each of which has a different neighborhood of the −OH fragment: we shall call all of these compounds the −OH set. Further on, the alcohols are divided into primary, secondary, and tertiary alcohols, and within the subset of primary alcohols the aromatic primary alcohols form another subset. The above hierarchical scheme of subsets having different neighborhoods of the −OH fragment can be written as follows.

−OH
alcohol
prim. alcohol
$ArCH_2OH$
sec. alcohol
tert. alcohol
acid
phenol

By inspecting the partial top-map for the compounds of the −OH set (Figure 5a) four distinctive groups of neurons excited by these compounds can be seen. If the partial top-map of the alcohol subset is displayed (Figure 5b) the three groups of the former map remain virtually nonaltered, while one distinctive group in the upper left corner and a rather nondistinctive group (low frequency group) in the foreground

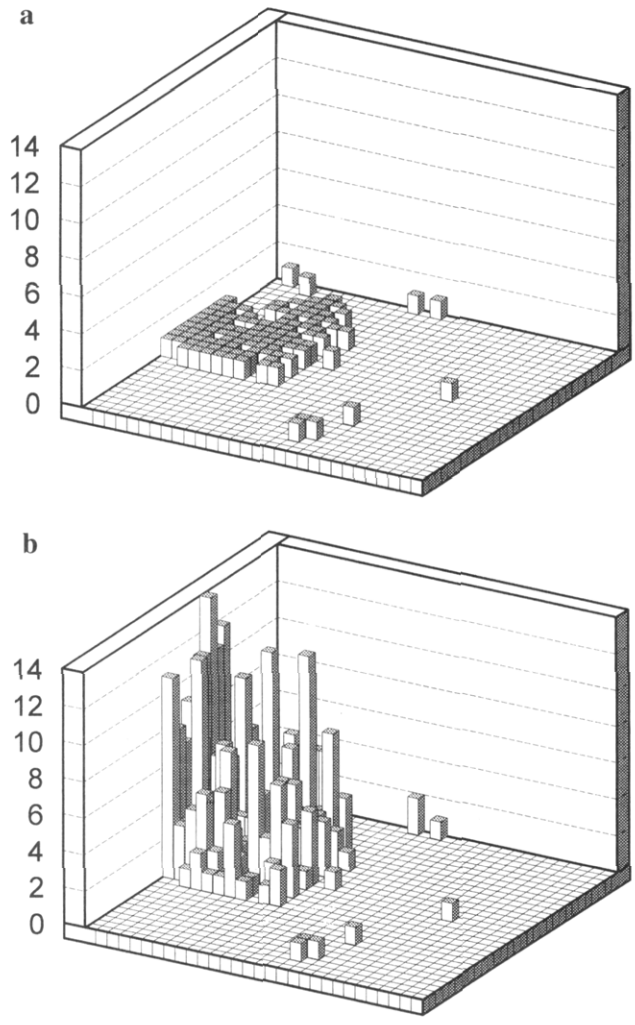

Figure 4. Partial top-map for the ester structural fragment distribution in the Kohonen ANN. The bars of length one (a) are marking all the neurons excited by at least one spectrum of a compound having ester fragment(s). The actual heights of the bars (b) show how many such spectra are exciting each specific neuron.

associated with the acids and with the phenols, respectively, disappeared almost completely. The next partial top-map (Figure 5c) for the primary alcohols subset shows only quantitative difference with the partial top-map of all alcohols (Figure 5b) indicating that distinguishing between the primary and other alcohols on the basis of IR spectra is rather difficult. As shown in Figure 5d none of the aromatic primary alcohols has excited the neurons in the left two distinctive groups of Figure 5 (parts b and c).

This indicates that the predictions of structural features can be linked to the spectral features that have produced the partial top-maps. During the learning in the Kohonen ANN the spectral features directly influence (change) the weight levels of the **input** neurons and thus decide which neurons will be excited. The fact that a good overlap between contours in the partial top-map for fragment forming clusters in the top-map and the contours in the maps of weights at certain levels of the Kohonen layer was found confirms that Kohonen and counterpropagation ANN offers the a tool for extracting the correlations between structural features and IR spectral absorption regions.

The sections across the neurons at different levels described as maps in the Kohonen ANN offer 30 × 30 contour plots of infrared absorptions. Because input spectra are reduced to 128 intensities, there are 128 maps in the Kohonen ANN. Each map (the cross-section at a given weight level) corresponds to one frequency region in the infrared spectrum.

Any cluster of **common** structural fragment(s) found in the partial top-map can be projected to **all** weight maps in the Kohonen ANN. (Remember: the weight maps are actually spectral intensity maps of selected spectral regions). It may happen that in a certain weight map the contour line of the high intensity region, coincides notably with the boundary of the top-map cluster. Two examples of such intensity maps and the corresponding overlapping top-map clusters are shown on Figure 6.

Figure 6 presents the two maps or intensity distributions of all spectra in the frequency region 2620−2580 cm^{-1} (Figure 6a) and 1036−1020 cm^{-1} (Figure 6b), respectively. The cluster of acids in the top-map of the Kohonen ANN (Figure 6c) overlaps with the maximal intensity region in the upper left corner of the contour plot (Figure 6a). However, at slightly lower intensity regions in the right part of the contour plot (Figure 6a) no acids were found which means that in the map-area the spectra of compounds with **different** structural fragments that **also** absorb in the frequency region 2620−2580 cm^{-1} must be found. Indeed, after checking all 34 partial top-maps it can be seen that alcohols (Figure 5b—lower-right part) and amines are clustered in this regions. Similarly, the two clusters of ethers (Figure 6d) overlap with the high intensity regions in the lower left and middle right area of the contour plot (Figure 6b). The third high intensity region in the middle left part is associated with the esters (compare Figure 6b with Figure 4b).

With this method, the fragments that **do not have** characteristic frequencies in infrared, but only influence the characteristic frequencies of the groups to which they are attached, **do not** produce well defined clusters and consequently no good correlations between their partial top-map and any of the intensity contour maps can be found.

PREDICTION OF STRUCTURAL FRAGMENTS WITH COUNTERPROPAGATION ANN

Selection of Spectra for the Training Set. As shown in Table 3 (first line), 145 neurons out of all 900 neurons from the 30 × 30 network were **not excited** by any of all 3284 objects in the final test. The remaining 755 neurons were excited by at least one object (spectrum). In order to assure a balanced presence of as many **different fragments** as possible and not to perpetuate the heavily biased **frequency distribution of the fragments** in the data base itself, only one spectrum from the list of spectra exciting each neuron was randomly selected to form the training set. Therefore, by picking up one spectrum from each excited neuron 755 spectra (close enough to the desired one fourth of the collection) distributed over the entire 30 × 30 map was obtained. Comparison of the ratio between the frequencies of the dominant structural feature (fragment no. 28 (C−O single bond) and the least frequent one (the furan fragment, no. 23) in the selected training set with the entire collection shows a decrease from 58 to 37 times. The distribution for the medium frequent fragments in the training set is balanced much better than it is in the entire collection. For example: the acid/furan ratio has been improved from 3.5 in the entire collection to 0.9 in the training set.

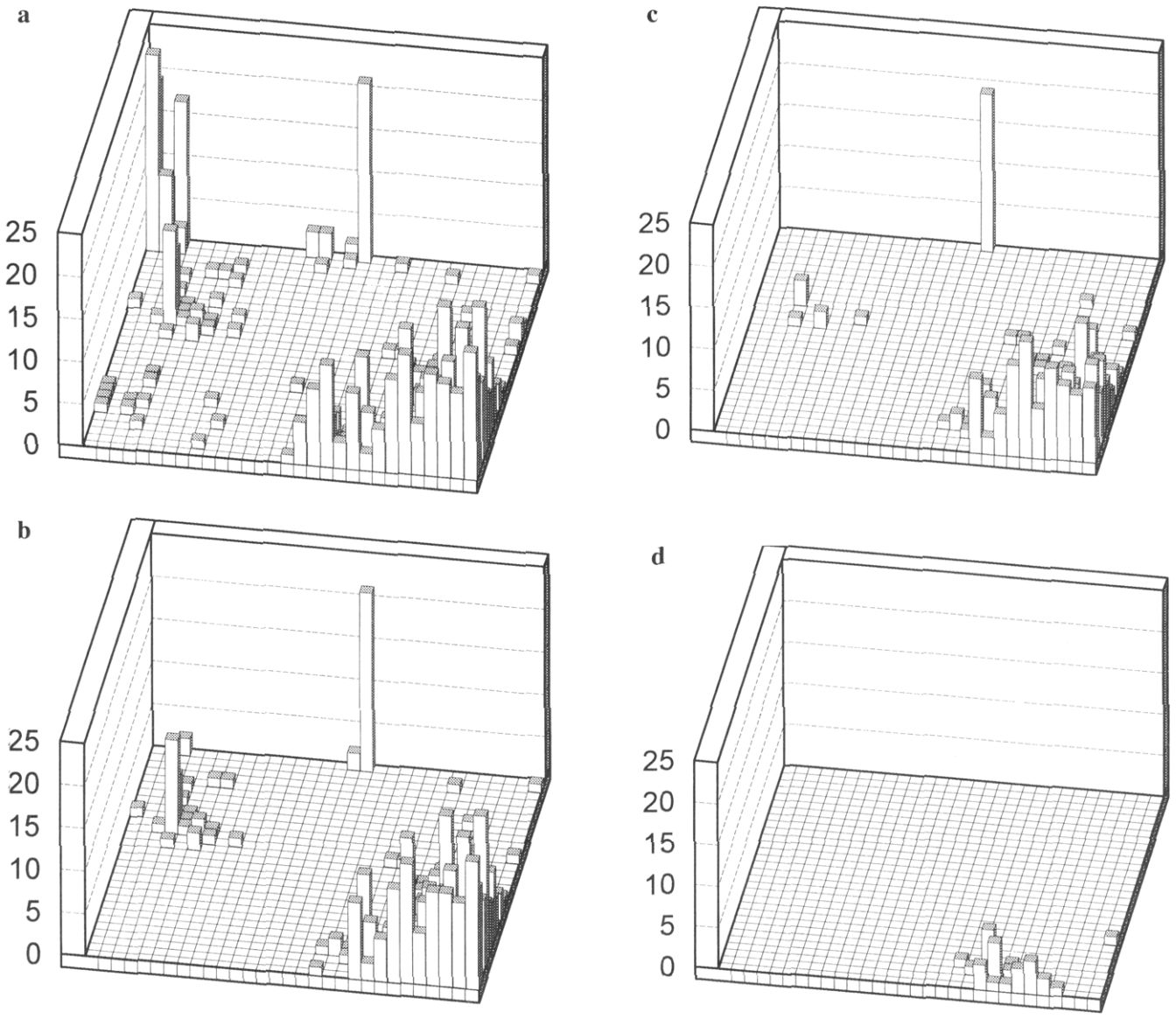

Figure 5. Partial top-maps for spectra of −OH set (a), of the alcohols (b), of the primary alcohols (c), and of the aromatic primary alcohols (d), i.e., fragment nos. 1, 2, 3, and 8, respectively. In each partial top-map of the Kohonen ANN (a−d) the neurons excited by the objects of the four mentioned types are marked by a bar with a height proportional to the number of objects exciting them.

Counterpropagation Learning. Since the training set was selected with the 30 × 30 Kohonen ANN we have deliberately chosen the counterpropagation ANN to have its Kohonen layer architecture identical to the previously used stand-alone Kohonen ANN. The other parameters used the counterpropagation learning are given in the third column of Table 2, shown above.

After the training with 755 objects (spectra and corresponding 34-structure descriptors) has been completed, the 900 neurons in the Kohonen layer have been adapted to all 755 spectra with an overall discrepancy of 5% per weight. The discrepancy of 5% is a total root mean square (RMS)[9,17] error calculated for almost 100 000 weights (755 * 128 = 96 640). At the same time the 30 600 (34 * 900) weights in the output layer have been changing iteratively using eq 3 with the 34-dimensional target vector. The resulting 34-dimensional output neurons are therefore the best fit to the 755 **binary** target vectors distributed over the network.

In the present study we have interpreted the output layer of 30 600 weights in terms of 34 maps. Each map of output weights u_{ji} contains 30 × 30 "probabilities" for a particular structure to be mapped on a given location. Each weight u_{ji} of the 34-dimensional jth output neuron carries a probability value between $0.0 \leq u_{ji} \leq 1.0$. Two output maps are given in Figure 7.

The actual span of u_{ji} varies due to the different absolute number of fragments **and** due to the number of clusters the spectra are forming on the map. The u_{ji} value at each location in the map is influenced by the u_{ji} values of neighboring locations; therefore, a single u_{ji} gives the ratio of hits of objects with vs objects without the ith fragment only in the actual neighborhood and not in the entire ANN. This is especially true when a set of objects having the fragment i is split into two or more groups. For alcohols (fragment No. 2) the range of u_{ji} is between 0.0 and 1.0, while for glycols (fragment no. 6—a subgroup of alcohols) the probability values u_{ji} are between 0.0 and 0.5. For these two fragments the results of predictions based on the probability values of u_{ji} taken from the corresponding two output maps (output map no. 2 and output map no. 6, respectively) are given in Table 4 (Figure 8).

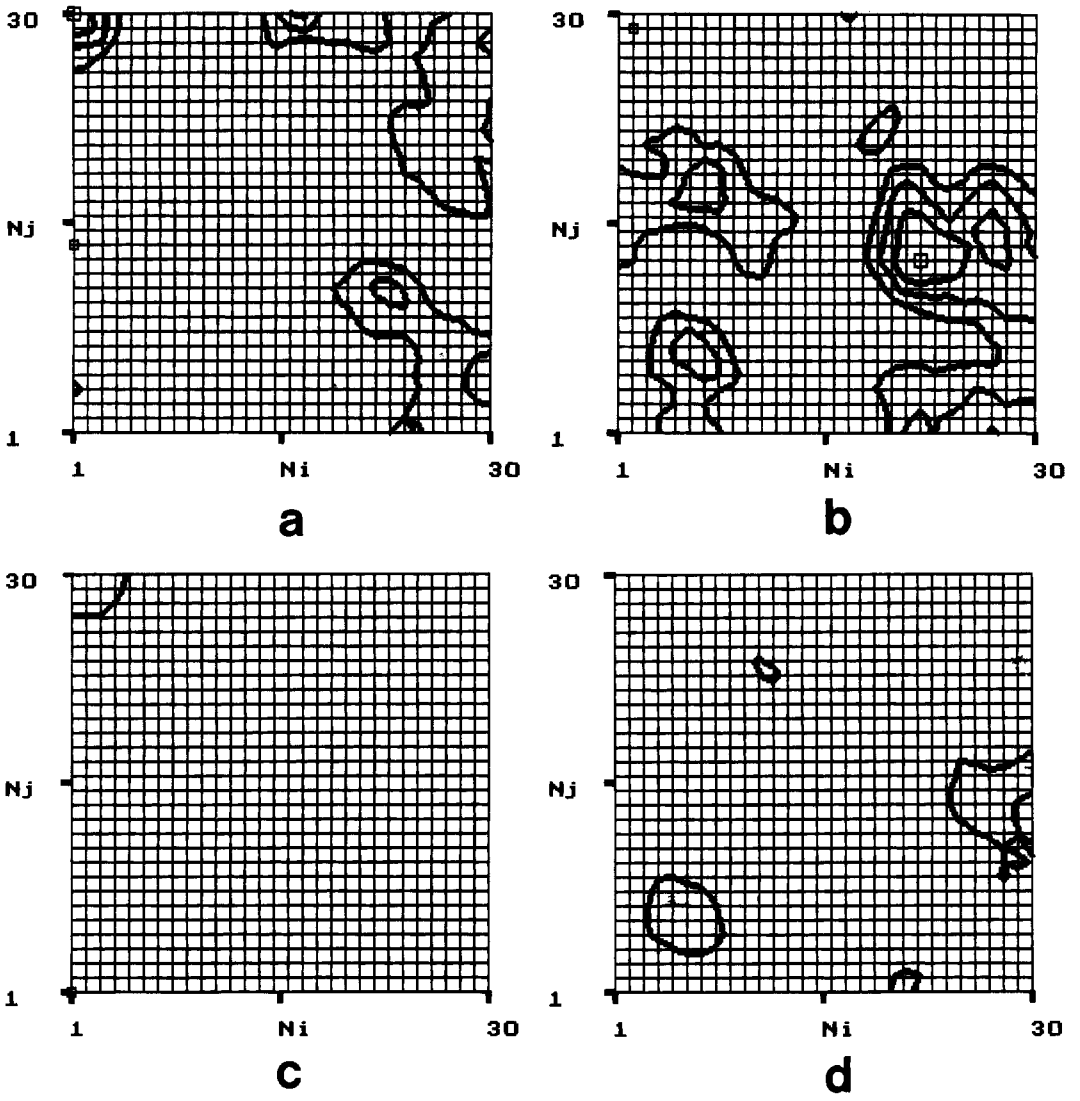

Figure 6. The contour plots of the weight maps, where the neurons in the Kohonen ANN are sliced through at the levels 23 (a) and 98 (b). The corresponding frequency regions of the maps are 2620−2580 cm^{-1} (a) and 1036−1020 cm^{-1} (b). The clusters of neurons excited by over 90% of acid compounds (cluster in upper left corner of the map (c)) and by about 80% of ether compounds from the data set (two clusters in lower left and middle right side of the map (d)), best overlap with the maxima of contour plots (a) and (b), respectively.

Table 4 shows that not a single spectrum (object) of the compound having the fragment no. 2 (alcohol) has excited neurons on the position where probability values in the output map no. 2 was lower than 0.2. On the other hand, not a single spectrum (object) of the compound **not being** an alcohol has excited a neuron in the regions where the probability values in the map no. 2 were higher than 0.8 (see two last boxes in Figure 7). For each fragment such limits can be found in the corresponding map.

Within certain reliability limits the resulting counterpropagation ANN is able to predict the probability presence and/or absence of any of the 34 structural fragments. For any test object the answer is found in the output layer exactly below the *c*th Kohonen (excited) neuron. Because the probability values for each fragment are spread around the entire output map, for each fragment the threshold value above which a positive (affirmative) prediction about the presence of the fragment should be made. The threshold u^{thr}_i is determined as the probability value at which the number of hits produced by the objects **having** the fragment *i* exceeds the number of hits of objects **not having** it (Figure 8). In the same way as shown in Figure 8, all 34 threshold values u^{thr}_i were determined (Table 5, column 3). The threshold values were determined on the basis of ratios taken from 34 tables (see Table 4 for two of them). It is worthwhile to emphasize that the threshold value u^{thr}_i for each fragment *i* is obtained on the basis of the 755 objects in the training set. These u^{thr}_i are further used to predict whether an object from the test set (2529 objects) has the *i*th fragment or not. For an affirmative or true-positive decision about the presence of the *i*th structural fragment the u_{ci} value (the *i*th weight on the output neuron exactly below the excited neuron *c* in the input layer) must be larger or equal to the u^{thr}_i determined for the output map *i*.

The **prediction ability** of the constructed counterpropagation ANN was obtain by testing 2529 spectra-structure pairs. Not a single pair of the 2529 ones was used in the training phase. In Table 5 the results of the predictions of the presence for each of the 34 structural fragments and the **reliability** of each prediction, if affirmative, are given.

Prediction ability, $^{+}p_i$, for the fragment i as provided by the counterpropagation ANN is given in the sixth column of Table 5. The prediction ability $^{+}p_i$ is fraction of the true-

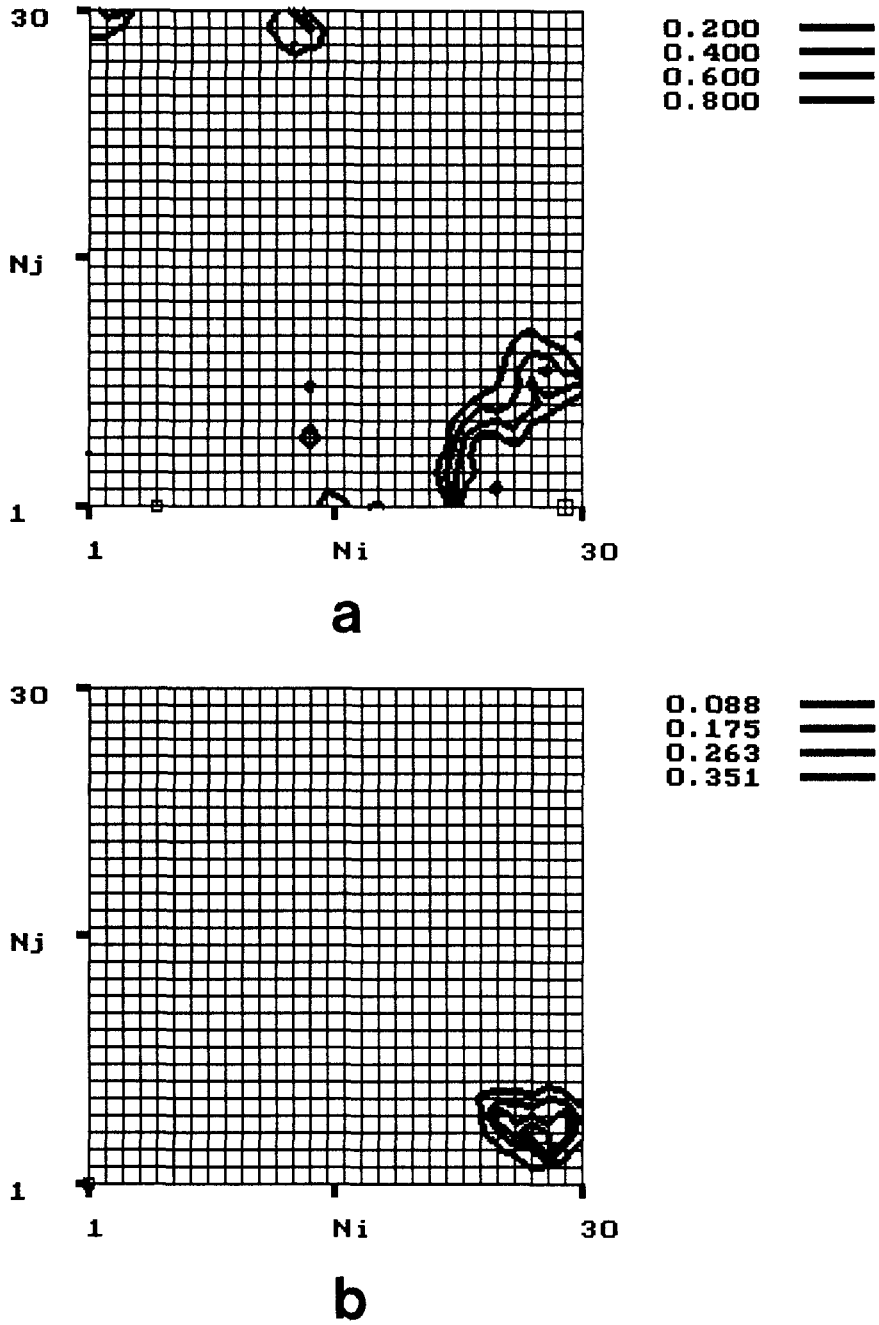

Figure 7. The map of the second (a) and the sixth (b) weight level of the output layer as obtained at the end of the training of the counterpropagation ANN with 755 objects. Note that the maximal contour level in the lower map (b) is 0.351 encompassing the highest probability value of 0.439 (cf. Table 4—the fifth interval).

positive decisions for the fragment i

$$^{+}p_i = {}^{+}n_i/N_{fri} \tag{4}$$

where $^{+}n_i$ is the number of correct predictions and N_{fri} is the number of compounds in the test set with fragment i. Evidently, an exactly equivalent equation holds for the false negative decisions $^{-}p_i$

However, the prediction ability alone does not give the information as to how reliable such predictions are. If, for example, the prediction ability $^{+}p_i = 1$, meaning that **all** compounds from the test set having the fragment i are predicted **correctly**, contains no information if it is obtained by the rule "any compound has fragment i". This rule would, of course, classify wrongly **all** objects **not having** the fragment i **as if they had** this fragment, which would make the 100% correct predictions upon those who actually have fragment i completely useless, i.e., not reliable.

Therefore, an estimation of the fraction of false-positive decisions ^{-}p in the test set is necessary. The false-positive decision is a prediction in which the system classifies an unknown compound as if it had the fragment i although it does not have it. For all fragments the fractions of false-positive decisions are given in the ninth column of Table 5.

From the two predictions, the true-positive ^{+}p and the false-positive ^{-}p one, the reliability R_i for predictions of the ith fragment can be calculated in the following way:

$$R_i = \frac{2(^{+}p_i - {}^{-}p_i)}{1 + |^{+}p_i - {}^{-}p_i|} \tag{5}$$

The reliability value R_i ranges from 1 to -1. In the best case when **all** compounds having fragment i are correctly predicted as such and **none** of the "not-haves" is predicted as to have the fragment i, the **reliability** of the ith prediction

Table 4. Weight Values u_{ci} of the Excited Neurons c at the Corresponding Levels i for Hits in the Final Counterpropagation ANN for Compounds With, n, and Without, ¯n, Fragment No. 2 (Alcohol) and Fragment No. 6 (Glycol)

fragment i $i = 2$ and 6	Number of hits	Weight at the ith level u_{ci} 0.0−0.1	0.1−0.2	0.2−0.3	0.3−0.4	0.4−0.5	0.5−0.6	0.6−0.7	0.7−0.8	0.8−0.9	0.9−1.0	N_{fri}
	$^{+}n_2$[a]	0	0	8	4	6	1	5	1	4	56	85
alcohol	cumulative[b] %	0	0	9.4	14.1	21.2	22.4	28.2	29.4	34.1	100	
(no. 2)	$^{-}n_2$	626	16	21	2	2	0	2	1	0	0	670
	cumulative %	100	6.6	4.2	1.0	0.8	0.5	0.5	0.2	0	0	
	$^{+}n_6$	0	0	0	4	2	0	0	0	0	0	6
glycol	cumulative %	0	0	0	66.7	100	100	100	100	100	100	
(no. 6)	$^{-}n_6$	735	10	1	3	0	0	0	0	0	0	749
	cumulative %	100	1.9	0.5	0.4	0	0	0	0	0	0	

[a] Number of spectra of compounds exciting the particular weight u_{ci} having or not having structural fragment i are indicated by $^{+}n_i$ or $^{-}n_i$, respectively. [b] The cumulative percentages of hits are cumulative ratios $(1/Nfr_i)\Sigma^{+}n_i$ for the hits on the weight values u_{ci} between zero and the higher limit from the left side of the u_{ji} range, or $(1/Nfr_i)\Sigma^{-}n_i$ for the hits between one and the lower limit from the right side of the u_{ji} range, respectively.

Table 5. Prediction Abilities and Reliability of Predictions for 34 Structural Fragments with the Counterpropagation NN as Obtained on 2529 Test Objects[a]

i	Fragment	u^{thr}_i	$+Nfri$	$+n_i$	$+p_i$	$-Nfr_i$	$-n_i$	$^{-}p_i$	R_i
1	OH	0.3	519	473	0.91	2010	71	0.04	0.93
2	alcohol	0.3	377	349	0.93	2152	59	0.03	0.95
3	prim. alcohol	0.5	239	144	0.60	2290	61	0.03	0.73
4	sec. alcohol	0.3	132	100	0.76	2397	118	0.05	0.83
5	tert. alcohol	0.2	24	15	0.63	2505	2	0.00	0.77
6	1,2 glycol	0.3	23	10	0.43	2506	27	0.01	0.59
7	phenol	0.3	53	41	0.77	2476	30	0.01	0.87
8	aryl-CH_{2-OH}	0.3	26	10	0.38	2503	27	0.01	0.54
9	NH	0.3	316	268	0.85	2213	80	0.04	0.90
10	prim. amine	0.3	202	168	0.83	2327	28	0.01	0.90
11	sec. amine	0.3	105	75	0.71	2424	86	0.04	0.80
12	CN	0.4	633	463	0.73	1896	184	0.10	0.77
13	tert. amine	0.3	123	74	0.60	2406	53	0.02	0.73
14	C=O	775	743	0.96	1754	40	0.02	0.97	
15	COOH	0.3	78	73	0.94	2451	10	0.00	0.97
16	COO−	0.3	319	284	0.89	2210	50	0.02	0.91
17	ester	0.3	316	281	0.89	2213	50	0.02	0.91
18	aldehyde	0.3	92	42	0.46	2437	39	0.02	0.59
19	ketone	0.4	211	159	0.75	2318	87	0.04	0.83
20	amide	0.3	49	25	0.51	2480	16	0.01	0.67
21	benzene	0.5	786	616	0.78	1743	153	0.09	0.82
22	naphthalene	0.2	18	10	0.56	2511	16	0.01	0.71
23	furan	0.3	16	5	0.31	2513	5	0.00	0.47
24	tiophene	0.2	21	4	0.19	2508	25	0.01	0.31
25	pyridine	0.3	55	27	0.49	2474	42	0.02	0.64
26	NO_2	0.3	68	54	0.79	2461	24	0.01	0.88
27	aryl-NO_2	0.4	50	33	0.66	2479	18	0.01	0.79
28	CO-	0.5	1117	979	0.88	1412	143	0.10	0.88
29	ether	0.4	340	210	0.62	2189	132	0.06	0.72
30	C−X	0.4	623	410	0.66	1906	245	0.13	0.69
31	C−F	0.3	161	97	0.60	2368	97	0.04	0.72
32	C−Cl	0.3	323	167	0.52	2206	244	0.11	0.58
33	C−Br	0.3	162	49	0.30	2367	131	0.06	0.39
34	C−I	0.4	35	0	0.00	2494	2	0.06	−0.13
			$\Sigma(Nfr_i{}^{+}p_i)/\Sigma Nfr_i$		0.77		$\Sigma(Nfr_iR_i)/\Sigma Nfr_i$		0.82

[a] The first and the second column contain the sequential number and description of a fragment, the corresponding threshold value u^{thr}_i is in the third column, the fourth and the fifth columns show the number of objects in the test-set containing that fragment and the number of correct answers, the sixth column gives the fraction of true positive answers, the seventh and the eighth columns give the numbers of objects not having the particular fragment and fragments predicted wrongly as having it, while in the ninth column the fraction of false-positive prediction is given. The tenth column contains the reliability of each prediction according to eq 4.

is maximal (equal to 1). It is interesting to note that in the case when a false-positive prediction is zero ($^{-}p = 0$) the reliability R_i is always higher than the true-positive predictions. This means that if the system predicts the presence of a fragment such prediction is very reliable even if this happens seldom.

On the other extreme side, if all compounds having a fragment i are predicted as not having it ($^{+}p = 0$) and all "not-haves" are predicted as having it ($^{-}p = 1$), the reliability R_i is equal to −1 signaling that such decision is extremely reliable: because **all** of them are wrong one must only reverse the sign of the decision to obtain the correct one. For the remaining two extreme cases the system gives the **same answer for all** objects, i.e., either all objects have the fragment i ($^{+}p = 1$, $^{-}p = 1$) or none of the objects has it ($^{+}p = 0$, $^{-}p = 0$); the reliability is, of course, zero.

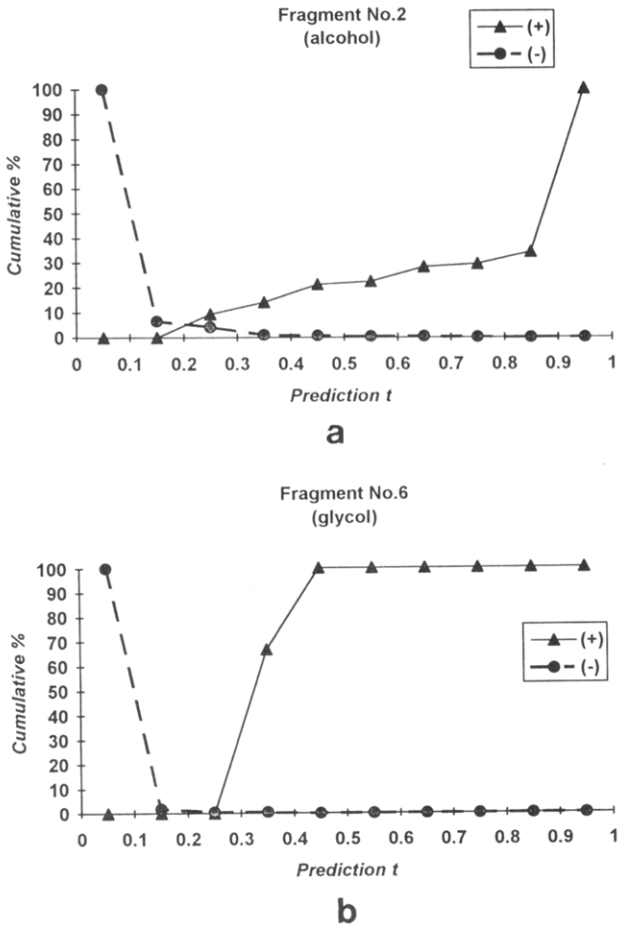

Figure 8. Cumulative percentages of hits for the fragment no. 2—alcohol (a) and the fragment no. 6—glycol (b), in the recall results of the counterpropagation ANN. The predictions' u_i region is divided into 10 intervals in the range from 0.0 to 1.0. The cumulative percentages of hits are calculated from the number of objects with and without the ith fragment (+ and −, respectively) having the predicted values u_i either between zero and the higher limit from the left side of the t-range (+), or between one and the lower limit from the right side of the t-range (−).

DISCUSSION

Several structural fragments were predicted correctly in more than 90% of cases; the −C=O group is predicted correctly even in 96% of cases. It should be emphasized that the reliability of these predictions are very high too. Good performance, i.e., the ability to predict the presence of a specific fragment in more than 80% cases, was found for the fragments nos. 1, 2, 9, 10, 15, 16, 17 and 28 (see Table 1 for the description of fragments). However, poor prediction results of less than 35% were obtained for fragment nos. 23, 24, 33, and 34 (furan, tiophene, C−Br, and C−I, respectively). The most characteristic from this group is fragment no. 33 (C−Br). In spite of the fact that there are 52 compounds with this fragment in the training set (compared to 18 compounds having the −COOH group yielding ${}^{+}p_i = 0.94$ and $R_i = 0.97$) the achieved results are quite poor (${}^{+}p_i = 0.30$ and $R_i = 0.39$). It can be concluded that the vibrations caused by this and similar fragments are not significant in infrared or their influences on the spectrum are lost among other, more dominant spectral features. For better predictions of such fragments, more investigation would be necessary in the direction of changing the spectral representation, subtracting the elementary characteristic spectral lines and emphasizing the variations of the characteristic spectral lines. Since the infrared spectrum reflects the whole chemical structure and the correlation of narrow spectral regions to only one structural fragment is not easy, prediction of structural features is a very complex problem and has not yet been solved satisfactorily.

At the end, we would like only to comment on the fact that the counterpropagation ANN has a potential for solving the reverse structure elucidation problem, i.e., the simulation of the infrared spectra from their structure. The detailed study of this problem is now under way and will be published in the forthcoming paper.

The simulation of infrared spectra is a hard problem, except for the simple molecules it was never brought up to the state of a practical use.[23] The simulation of infrared spectra by the proposed scheme is uncomplicated indeed. The roles of the input and the output data are simply reversed. Instead of entering the 128-dimensional spectral representation to the Kohonen layer the 34-dimensional binary (yes-no) structural description based on the presence or absence of considered structural fragments is **input** to the **output** layer of the existing (trained) counterpropagation ANN.[24] The central (excited) neuron c is obtained in the output, **not** in the Kohonen layer, by applying the criterion 1 on the structural representations in the output neurons (substitute w_{ji} with u_{ji} in eq 1). Once the central neuron c is determined, the simulated spectrum is found in the Kohonen layer, in the neuron exactly above the neuron c of the output layer.

However, we would like to point out that using large counterpropagation ANNs (of order 1000 × 1000 × [512 + 100]) able to map most of the available computer-readable 512-dimensional infrared spectra onto the map of 1 000 000 neurons could provide a reliable simulation of infrared spectra of complex compounds described by 100-dimensional structural representation.

Just to emphasize this assertion two simulation of infrared spectra not identical to any of the 3284 ones used in the study, are shown in Figure 9. With these two examples we do not want to give the implication that the existing 30 × 30 × [128 + 34] counterpropagation ANN can be used for actual simulations of infrared spectra. We would rather emphasize the fact that the simulations of infrared spectra can be actually achieved in this way. As said above, this work is in progress, and for the simulation of real world infrared spectra much larger counterpropagation ANN are needed.

CONCLUSION

The 128-dimensional spectral representation used in this study was a compromise between the desired spectral resolution and the acceptable dimension of the ANN, i.e., the time need to train the network. The choice of 34 structural fragments was again a compromise. This time the trade-off was between the selection already tested in the previous works[1,2] and the size of manageable ANN. The requirement to keep the representations (spectral and structural) as detailed as possible has made the ANN already rather large (30 × 30 × [128 + 34] = 145 800 weights) to work with in our computational resources. The training was performed on the μVAX. About 24 h were needed to train the network in 100 epochs, i.e., for the input of 755 objects 100 times.

In the presented approach to the spectra-structures correlation problems, we have first used the Kohonen ANN to

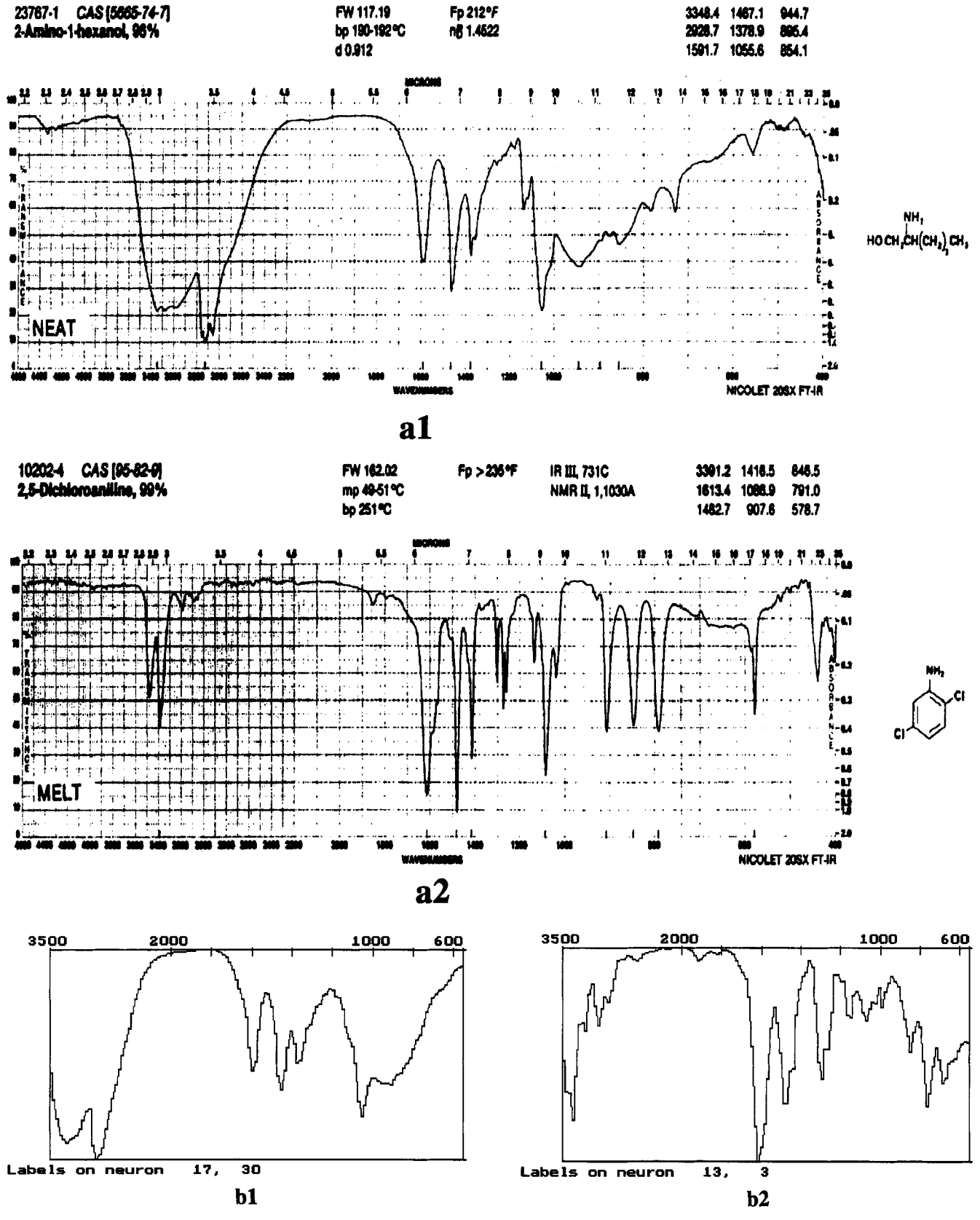

Figure 9. Comparison of two simulated spectra with their originals. The first IR spectrum is 2-amino-1-hexanole from literature (a1). The simulated spectrum from the counterpropagation NN (b1) is equal to the neuron at the position (17, 30) in the Kohonen layer. The second IR spectrum is 2,5-dichloroaniline (a2). The simulated spectrum from the counterpropagation NN (b2) is equal to the neuron at the position (13, 3) in the Kohonen layer. The wavelength scale changes at 2200 and 1000 cm^{-1} in spectra from literature (a1, a2), while in simulated spectra (b1, b2) the scale changes only at 2000 cm^{-1}. This causes some apparent dissimilarity of spectra in low wave number regions.

map the spectra into the plane for the **selection** of a balanced training set. The training set was used in the counterpropagation learning strategy to generate an ANN capable of predicting structural fragments from the unknown infrared spectrum. From the recall of 755 spectra-structure pairs in the training set, the threshold values u^{thr}_i were determined

for each of the 34 fragments. On the basis of the u^{thr}_i values and the predictions of structural features of the 2529 spectra-structure pairs in the test set the prediction abilities $^{+}p_i$ and $^{-}p_i$, together with the reliability R_i of each prediction i, were determined.

Most of the previous models, expert systems, procedures, or learning schemes are built so that the presence of a certain substructure regardless of its structural environment is always signaled by **one** output, on **one** position, or in **one** way. Due to a large amount of different structural environments which can cause completely different infrared spectra the assignment of only **one** output neuron to each **single** substructure feature[1,2] (as in the case of error back-propagation learning strategy) is simply not flexible enough. This requires a scheme where a **single** structural feature can be detected on **different** outputs or on different positions on the map. We believe that the generation of two-dimensional maps in the counterpropagation ANN seems to be a suitable tool to achieve this.

ACKNOWLEDGMENT

The authors acknowledge the financial support of the German Federal Ministry for Science and Technology (Bundesministerium für Forschung und Technologie—BMFT) and the Ministry for Science and Technology of Slovenia (Ministrstvo za znanost in tehnologijo—MZT) for partial financial support of this work. The authors also thank to Dr. Simona Bohanec for making several modifications of the substructure search routines needed for this work in the KIBK CNMR software package.

REFERENCES AND NOTES

(1) Robb, E. W.; Munk, M. E. A Neural Network Approach to Infrared Spectrum Interpretation. *Mikrochim. Acta [Wien]* **1990**, *I*, 131−155.
(2) Munk, M. E.; Madison, M. S.; Robb, E. W. Neural Network Models for Infrared Spectrum Interpretation. *Mikrochim. Acta [Wien]* **1991**, *II*, 505−514.
(3) Novič, M.; Zupan, J. 2-D Mapping of Infrared Spectra Using Kohonen Neural Network. *Vestn. Slov. Kem. Drus.* **1992**, *39*(2), 195−212.
(4) Weigel, U. M.; Herges, R. Automatic Interpretation of Infrared Spectra: Recognition of Aromatic Substitution Patterns Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 723−731.
(5) Meyer, M.; Meyer, K.; Hobert, H.; Neural Networks for Interpretation of Infrared Spectra Using Extremely Reduced Spectral Data. *Anal. Chim. Acta* **1993**, *282*, 407−415.
(6) Gasteiger, J.; Li, X.; Simon, V.; Novič, M.; Zupan, J. Neural Nets for Mass and Vibrational Spectra. *J. Mol. Struct.* **1993**, *292*, 141−160.
(7) Smits, J. R. M.; Schoenmakers, P.; Stehman, F.; Sijstermans, F.; Kateman, G. Interpretation of Infrared Spectra with Modular Neural-Network Systems. *Chemom. Intell. Lab. Sys.* **1993**, *18*(*1*), 27−40.
(8) Melssen, W. J.; Smits, J. R. M.; Rolf, G. H.; Kateman, G. Two-dimensional Mapping of IR Spectra Using a Parallel Implemented Self-Organising Feature Map. *Chemom. Intell. Lab. System* **1993**, *18*, 195−204.
(9) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or Just a Passing Phase? *Anal. Chim. Acta* **1991**, *248*, 1−30.
(10) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Propagation. In *Microstructures of Cognition*; Rumelhart, D. E., McClelland, J. L., Eds.; MIT Press: Cambridge, **1986**; Vol. 1, pp 318−362.
(11) Lippmann, R. P. An Introduction to Computing with Neural Nets. *IEEE ASSP Magazine*; 1987; Vol. 4.
(12) Kirchner, W. Neuronales Backpropagation-Netz zum Selbermachen, c't Heft, 1990; Vol. 11, p 248.
(13) Kohonen, T. Self-Organization and Associative Memory; Springer-Verlag: Berlin, 1988.
(14) Kohonen, T. The Self-organising Map. *Proc. IEEE* **1990**, *78*(9), 1464−1480.
(15) Ritter, H.; Martinez, T.; Schulten, K. Neuronale Netze, Eine Einführung in die Neuroinformatik selbstorganisierender Netzwerke; Addison-Wesley: Bonn, 1990.
(16) Hecht-Nielsen, R. Counter propagation Networks. *Appl. Optics* **1987**, *26*, 4979−4984.
(17) Dayhof, J. Neural Network Architectures, An Introduction; Van Nostrand Reinhold: New York, 1990.
(18) Zupan, J.; Gasteiger, J. Neural Networks for Chemists: An Introduction; Verlag Chemie: Weinheim, 1993.
(19) Razinger, M.; Novič, M. Reduction of the Information Space for Data Collections, PCs for Chemists; Zupan, J., Eds.; Elsevier: Amsterdam, 1990; pp 89−103.
(20) Bohanec, S.; Zupan, J. Structure Generation of Constitutional Isomers from Structural Fragments. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 531−540.
(21) Simon, V.; Gasteiger, J.; Zupan, J. A Combined Application of Two Different Neural Network Types for the Prediction of Chemical Reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148−9159.
(22) Zupan, J.; Novič, M.; Li, X.; Gasteiger, J. Classification of Multicomponent Analytical Data of Olive Oils Using Different Neural Networks. *Anal. Chim. Acta* **1994**, *292*, 219−234
(23) Affolter, Ch.; Clerc, J. T. Prediction of Infrared Spectra from Chemical Structures of Organic Compounds Using Neural Networks. *Lab. Inf. Manage.* **1993**, *21*, 151−157.
(24) Zupan, J.; Novič, M.; Gasteiger, J. Neural Networks with Counter-propagation Learning Strategy Used for Modelling. *Chemom. Intell. Lab. Syst.* Accepted for publication.

CI940054Q