13 Quantitative Structure-Activity Relationships

learning objectives:

- the basis for Quantitative Structure-Activity Relationships (QSAR)
- identification of factors controlling anticarcinogenic activity of carboquinones related to the identities of substituents on the basic skeleton
- use of a neural net to associate biological activity with a given profile (set of values for the controlling factors)
- comparison of neural network approach with classical statistical approach
- representation of structures of different size with the same number of descriptors
- selection of appropriate descriptors for representing the input objects
- using a combination of unsupervised and supervised neural networks to study a data set
- application of a genetic algorithm for the reduction in the number of variables

13.1 The Problem

"QSAR" stands for Quantitative Structure-Activity Relationships, that is, quantitative relationships between a chemical structure and its physical, chemical, or biological activity. The search for such relationships is one of the most important applications of modeling techniques.

Correlating the chemical structures of drugs with their pharmacological activities is of particular interest. Because of the

large development costs of new drugs, a reliable quantitative prediction of activity **before** the compound is made is of great interest to synthesis laboratories.

Because neural networks can be developed into complex models they have gained large prominence in QSAR research. Particularly in pharmaceutical research and development many investigations on the relationships between structure and biological activity have been made. We will here investigate three different data sets, in the first and the third one, all compounds have the same skeleton and only the substituents on this skeleton have been varied. Such data sets are typical for many investigations that deal with the optimization of a pharmaceutical lead structure. The second data set encompasses a variety of structures having different skeletons. Such data sets are met in the search for a lead structure.

As a typical example, we will first discuss the work of Aoyama, Suzuki and Ichikawa (References 13-1 and 13-2). They chose a dataset that had previously been studied by Yoshimoto and coworkers (Reference 13-3) using traditional modeling techniques, (such as multilinear regression analysis (MLRA)), in order to compare those results with the results of a neural network. Aoyama and coauthors did their best to keep all variables, along with the selection of the training and test sets, as similar to the classical studies as possible.

The second data set comprises structures having different skeletons and different numbers of atoms in the molecules. Here, one has to face the problem of representing such a data set by a uniform set of descriptors, having the same number of variables for molecules of different size.

Neural networks, as any learning method, be it statistical or pattern recognition methods or neural networks, need the objects of a study to be represented by the same number of descriptors (variables).

The third data set is a collection of 55 biologically active flavonoid compounds, inhibitors of the enzyme tyrosine kinase. They all have the same skeleton on which different substituents are attached at different positions. Due to the initial 180-dimensional spectrum-like representation (for details on this representation see Chapter 21, Section 21.4) of their 3D structures the task to be solved in the example is the reduction of the 180 variable set into a smaller and more easily manageable set of variables which still contains the most relevant information about the biological activity of the flavonoids in question.

13.2 Dataset I

The dataset in this study involves modifications of the basic carboquinone skeleton shown in Figure 13-1. Many carboquinones exhibit varying degrees of *anticarcinogenic* activity.

This quantitative structure-activity relationship study was designed to predict the minimum dose of a drug required to produce a 40 percent extension of the lives of the test animals, BDF_1 mice that had been inoculated with lymphoid leukemia L-1210 cells.

This *minimum effective dose* depends on the concentration, C, of the substance necessary to give the desired effect, and is given as $\log(1/C)$. The more effective the drug is, the smaller will be the concentration necessary. (Since the required concentrations of different drugs vary over several orders of magnitude, it is more convenient to use the logarithm, $\log(1/C)$, as a measure of the effective dose.)

As expected, the anticarcinogenic activity depends on the identities of the substituents R^1 and R^2. In the standard multilinear analyses, this substituent is described by physicochemical variables that describe the combined influence of the substituents R^1 and R^2:

- the molar refractivity $\mathcal{MR}_{1,2}$
- the substituents' contribution to the hydrophobicity $\pi_{1,2}$
- the sum of the substituent constants for the field effect $\mathcal{F}$
- the sum of the substituent constants for the resonance effect $\mathcal{R}$

along with two local variables, describing only the influence of one substituent R^1:

- the molar refractivity $\mathcal{MR}_1$
- the contribution to the hydrophobicity π_1.

The assignment of substituents as R^1 and R^2 is based on their molar refractivities: $\mathcal{MR}_1 \leq \mathcal{MR}_2$.

The study used eleven different substituents R^1 (consisting primarily of short alkyl groups like methyl, ethyl, and propyl) and about 30 different substituents R^2 (mostly substituents having longer chains and bearing additional functionalities like $-CH_2CH_2OCH_3$ and $-CH(OCH_3)CH_2OCONH_2$). Two of the compounds are shown in Figure 13-2.

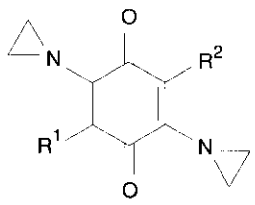
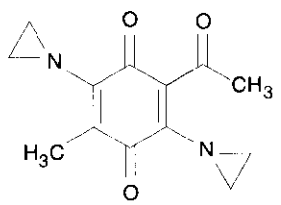
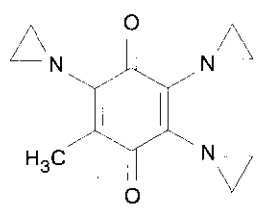


Figure 13-1: The characteristic structure of carboquinones, a class of compounds with anticarcinogenic activity. The substituents on the skeleton can be quite varied (see Figure 13-2).



a) **1**



b) **38**

Figure 13-2: The most (a) and the least (b) anticarcinogenic compound among the carboquinones.

Altogether, 35 different carboquinones were selected; for all of them, six variables, $\mathcal{MR}_{1,2}$, $\pi_{1,2}$, $\mathcal{MR}_1$, π_1, $\mathcal{F}$, and $\mathcal{R}$, were either measured ($\mathcal{MR}$) or taken from the literature (π, $\mathcal{F}$, $\mathcal{R}$), and the minimum effective dose was measured and given as log(1/*C*).

13.3 Architecture and Learning Procedure

Our inputs consist of variables ($\mathcal{MR}_1$, $\mathcal{MR}_{1,2}$, π_1, $\pi_{1,2}$, $\mathcal{F}$, $\mathcal{R}$) describing the structure, and our target data are values of log(1/*C*); thus, a supervised learning method should be used. In this example, we will try to find a model that can predict the minimum effective dose log(1/*C*) for each set of the six input variables, $\mathcal{MR}_{1,2}$, $\pi_{1,2}$, $\mathcal{MR}_1$, π_1, $\mathcal{F}$, and $\mathcal{R}$ for any carboquinone derivative. Hence, our network requires six input units and one output neuron.

As in most applications, one hidden layer turns out to be sufficient; after some trial and error, twelve neurons were placed into the hidden layer (Figure 13-3).

The (6 x 12 x 1) neural network, with one hidden and one output layer, was trained with 35 carboquinones by the back-propagation algorithm; afterwards the log(1/*C*) output values were compared with those obtained by multilinear regression analysis on the same set of 35 compounds.

The anticarcinogenic activity of 17 of the carboquinones is predicted with higher accuracy than in the multilinear regression analysis study, for six compounds the results are of about equal quality, and for 12 structures they are worse. Overall, the results of the neural network (NN) are significantly (but not dramatically) better than those obtained by multilinear regression analysis (MLRA).

Table 13-1 will give you an impression of the data and results for six of these structures.

Apparently, the problem under investigation is adequately handled by a linear model, but the neural network does lead to slight improvements. **Nonlinear** QSAR problems will show much larger improvements when modeled by neural networks.

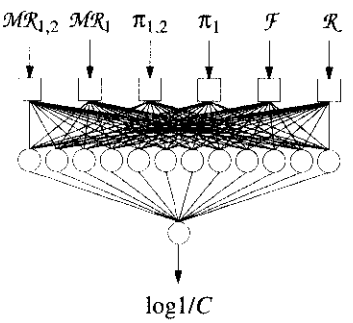


Figure 13-3: Network architecture for studying the anticarcinogenic activity of carboquinones.

13.4 Prospects of the Method

In investigations of the biological activity of a series of compounds, several different biological activities are often monitored.

	substitutents		variables						log(1/*C*)		
no.	R^1	R^2	$\mathcal{MR}_{1,2}$	$\pi_{1,2}$	$\mathcal{MR}_1$	π_2	$\mathcal{F}$	$\mathcal{R}$	exp.	MLRA	NN
1	CH_3	$COCH_3$	1.69	−0.05	−0.55	0.57	0.28	0.07	3.94	4.12	4.39
3	CH_3	$(CH_2)_3C_6H_5$	4.50	3.66	3.16	0.57	−0.08	−0.26	3.93	4.23	4.18
6	CH_3	$CH_2C_6H_5$	3.57	2.51	2.01	0.57	−0.12	−0.14	4.74	4.77	4.67
10	C_2H_5	C_2H_5	2.06	2.00	1.00	1.03	−0.08	−0.26	4.94	5.01	4.99
32	C_2H_5	$(CH_2)_2CONH_2$	3.09	0.95	−0.05	1.03	−0.08	−0.26	5.98	5.55	5.59
38	CH_3	$N(CH_2)_2$	2.13	0.68	0.57	0.18	0.06	−1.05	6.54	6.30	6.31

Table 13-1: Input and output variables of six compounds used in the QSAR study of anticarcinogenicity of carboquinones by Aoyama and coworkers.

For example, when investigating the anesthetic activity of a compound, one will also monitor its toxicity; or, in investigations of the carcinogenicity of a compound, two different types of carcinogenicity tests might be performed.

The two – or more – different biological activities quite often depend on the same types of structural variables, e. g., the value of the coefficient, log*P*, indicating the distribution of the compound between aqueous and lipid phases. In such cases, as we have mentioned in previous examples (see, for example Section 9.3) standard techniques develop **separate** modeling equations for the **two** biological activities, expressed as logarithms of the inverses of some threshold concentrations C_1 and C_2:

$$\log\frac{1}{C_1} = c_{01} + \ldots + c_{i1}\log P + \ldots \qquad (13.1)$$

$$\log\frac{1}{C_2} = c_{02} + \ldots + c_{i2}\log P + \ldots \qquad (13.2)$$

With neural networks, however, it becomes feasible to model **both** biological activities **simultaneously** in one network. Then, one output neuron will be used to output the first activity (expressed as $\log(1/C_1)$), whereas a second output neuron is used to indicate the second activity ($\log(1/C_2)$).

Figure 13-4 shows the part of a two-layer neural network that expresses the influence of log*P* on two biological activities. Thus, while MLRA requires two coefficients to express the influence of log*P* on the two activities, a neural network with one hidden layer containing, say, three neurons provides **nine** weights for expressing

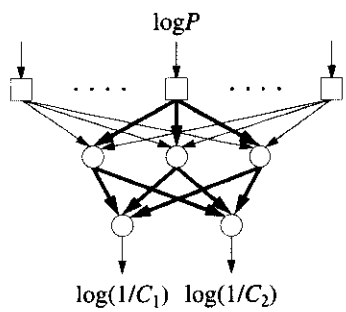


Figure 13-4: Propagation of the influence of one input variable (log*P*) on two different biological activities by a two-layer neural network with three hidden neurons.

these influences. With n neurons in the hidden layer, $3n$ weights are available to convey the influence of one input variable onto two activities. This indicates quite clearly the higher flexibility of a neural network compared to a statistical analysis.

13.5 Dataset II

The second data set comprises 31 steroids having different binding affinity to the *corticosteroid binding globulin* (CBG) receptor. Table 13-2 gives the full list of compounds with their binding affinity data and a classification into high, intermediate, or low affinity.

	CBG			CBG	
compd	affinity (pK)	activity class[a]	compd	affinity (pK)	activity class[a]
1	–6.279	2	17	–5.225	3
2	–5.000	3	18	-5.000	3
3	–5.000	3	19	-7.380	1
4	–5.763	3	20	-7.740	1
5	–5.613	3	21	-6.724	2
6	–7.881	1	22	-7.512	1
7	–7.881	1	23	-7.553	1
8	–6.892	2	24	-6.779	2
9	–5.000	3	25	-7.200	1
10	–7.653	1	26	-6.144	2
11	–7.881	1	27	-6.247	2
12	–5.919	2	28	-7.120	2
13	–5.000	3	29	-6.817	2
14	–5.000	3	30	-7.688	1
15	–5.000	3	31	-5.797	2
16	–5.225	3			

Table 13-2: *Corticosteroid binding globulin* (CBG) affinity data.
[a] 1, high; 2, intermediate; 3, low; this classification was obtained by dividing the data set into three classes of comparable size.

Figure 13-5 shows one structure each with high, intermediate, or low binding affinity to the CBG receptor, respectively. The full data set, with the structures encoded as connection tables, is contained on the web site for this book

(*http://www2.ccc.uni-erlangen.de/ANN-book/*).

See the Appendix for further information.

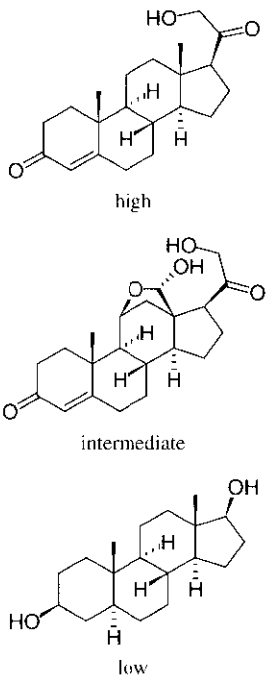


Figure 13-5: One steroid each with high, intermediate, or low binding affinity to the *corticosteroid binding globulin* (CBG) receptor.

This data set was chosen because it had been selected for the introduction of the widely used CoMFA method and has also been studied by a variety of other methods. Although all compounds of this data set are steroids they, nevertheless, comprise different skeletons having A- and B-rings with or without double bonds or, in some cases, aromatic A-rings. Furthermore, the substituents at various positions differ quite extensively, and the number of atoms in the set of compounds also varies.

13.6 Structure Representation by Autocorrelation of the Molecular Electrostatic Potential

Neural networks, as many learning methods, need the objects of study to be represented by the same number of input descriptors. Thus, with a data set as just described, one is faced with the task of transforming the structure into a preset number of descriptors. In this book, we will present a variety of methods for such a mathematical transformation of chemical structure information into a set of descriptors (cf. Chapter 21). The choice on the structure encoding scheme should somehow take into consideration the factors that are thought to be involved in the property investigated.

The electrostatic potential on the surface of a molecule (cf. Figures 19-2, 19-13, and 19-14) is one of the most important factors for binding a ligand to its receptor. The question is then, how can this property be encoded into a fixed number of descriptors? This task was achieved by autocorrelation, as indicated in Equation (13.3)

$$A(d) = \frac{1}{m}\sum_{i,j} p(i)\ p(j) \qquad \text{with } d_l < d_{ij} < d_u \qquad (13.3)$$

First, the molecular electrostatic potential (MEP) is calculated for a set of points evenly distributed over the van der Waals surface of the molecule with a selected density. Then, the products of this property, p, (the MEP), at points i and j is calculated whereby the distance d_{ij} between these two points must be between a lower, d_l, and upper, d_u, bound (say, between 3Å and 4Å). All these products are collected into a single value of $A(d)$; in this case, $A(3)$. This value is normalized by the total number, m, of distances in this interval (Figure 13-6). With a series of distance intervals with different upper and lower bounds, a

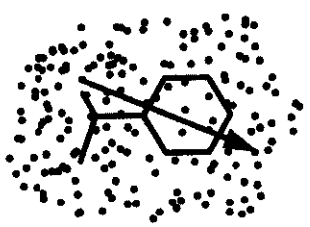


Figure 13-6: Autocorrelation of a property on a molecular surface (see Equation 13.3).

vector of autocorrelation coefficients is obtained. In our case, we have collected the products between 1Å and 2Å, all the way to 12Å and 13Å, thus providing an autocorrelation vector of length 12. The values of $A(d)$ are displayed at the center of each interval. Figure 13-7 illustrates this process of calculating an autocorrelation vector of the MEP using the steroid *corticosterone* as an example. In effect, the molecular electrostatic potential at the van der Waals surface was encoded into an autocorrelation vector of 12 values for each of the 31 steroids. Clearly, this encoding scheme is a drastic reduction of the information on the MEP. However, the goal, to encode a molecule into a preset number of descriptors (here 12), irrespective of the size of the molecule, was achieved.

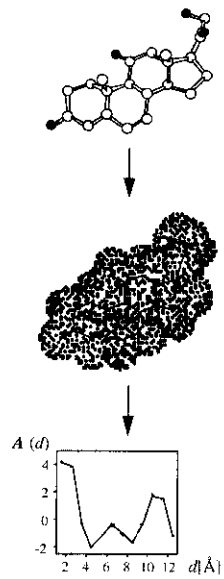


Figure 13-7: Calculation of the autocorrelation vector of the molecular electrostatic potential for *corticosterone*.

13.7 Verification of Structure Representation by Unsupervised Learning

The final goal of this study was to find a quantitative relationship between a structure encoding and the binding affinities to the CBG receptor by a back-propagation neural network. The back-propagation algorithm is such a powerful modeling technique that it will establish apparent relationships, albeit of low predictive power, even between input and output data that have only a small correlation. We have, therefore, found it highly recommendable to first establish whether the variables describing the objects are, in fact, significant for the property under investigation. In our case, the question is: Is there a relationship between the encoding of the MEP of a steroid by a 12-dimensional autocorrelation vector and the activity of binding to the CBG receptor?

This question can be answered by an unsupervised learning technique such as the one contained in Kohonen neural network learning. The 12-dimensional descriptor space was projected into a toroidal plane using a Kohonen network in order to visualize the distribution of the objects in the high-dimensional descriptor space. The projection into a Kohonen map was performed by training a network that consisted of 15 x 15 neurons, with each neuron having 12 weights corresponding to the 12-dimensional autocorrelation vector describing the MEP of a steroid. After projection of the data set of 31 steroids into this two-dimensional Kohonen map, the projection was visualized by marking those neurons having obtained a steroid with

high, intermediate, or low binding affinity with a filled square, an asterisk, or with a cross, respectively. (Recall, that this activity level was not used in the training of the Kohonen network).

The Kohonen network used had the topology of a torus, i.e., neurons at the left and the right side of the network, and neurons at the upper and the lower part of the networks are directly connected (cf. Figures 6-6 and 6-7). Therefore, four identical copies of the resulting Kohonen map were arranged like tiles (cf. Figure 6-14) in Figure 13-8 in order to better present the clusters formed by the steroids. The compounds of high, intermediate, and low activity form three clearly perceivable clusters in the Kohonen map as indicated in Figure 13-8. Only one compound, a steroid of intermediate activity, is not grouped together with compounds of the same activity class, but is surrounded by highly active compounds, instead.

The ability of the Kohonen network to here distinguish between compounds belonging to different activity classes shows, that the autocorrelation vector fulfills one of the prerequisites for a successful quantitative analysis: compounds that are similar to each other in the descriptor space exhibit similar biological activity. The visualization proved that the compounds group together in the descriptor space corresponding to their biological activity. Therefore, we were encouraged to quantitatively model the binding affinity with a feed-forward neural network trained by back-propagation as the next step.

We have also investigated the 12-dimensional descriptor space by another unsupervised learning method, a principal component analysis (PCA). Figure 13-9 shows the clustering of the steroids in a plot of the first against the second principal component. The compounds are by far not as well separated as in the Kohonen map of Figure 13-8. The principal component analysis performs a rotation of the coordinate axes of a high-dimensional space, trying to put as much variance as possible into the first few component. In our case, we are apparently left with more than two components and, thus, a plot of only two components cannot quite separate the compounds into their activity classes. The learning in a Kohonen network, on the other hand, knows from the very beginning that it has to end up with two dimensions and therefore places as much information as possible into these two dimensions.

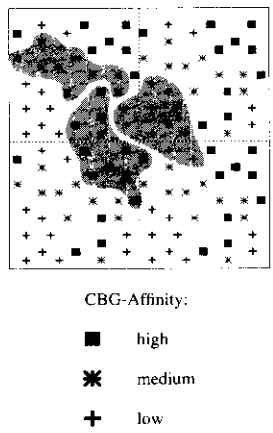


Figure 13-8: Arrangement of four identical Kohonen maps obtained from the 12-dimensional MEP autocorrelation space showing the separation of steroids of high (squares), intermediate (asterisks), and low activity (crosses).

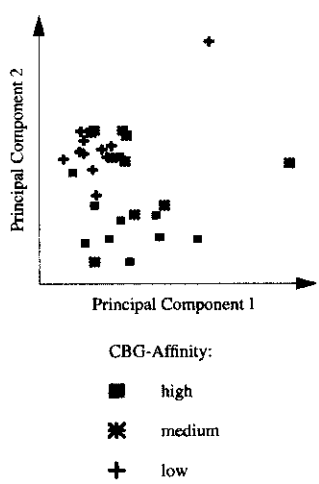


Figure 13-9: Plot of the first two components of a principal component analysis of the 12-dimensional descriptor space.

13.8 Modeling of Biological Activity by Supervised Learning

Projection of the 12-dimensional descriptor space by a Kohonen network had indicated that the encoding by autocorrelation of the molecular electrostatic potential has a relationship to the binding affinity to the CBG receptor. We, therefore, generated a quantitative model of CBG activity by a feed-forward neural network trained by back-propagation. The architecture of the network used was as follows: twelve input units corresponding to the twelve autocorrelation coefficients, two hidden neurons, and an output neuron (Figure 13-10).

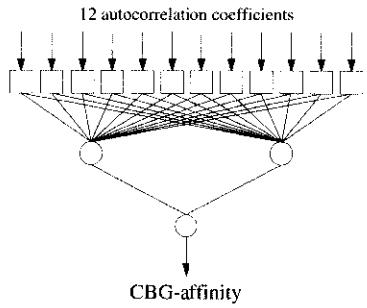


Figure 13-10: Architecture of the feed-forward network used for modeling CBG binding affinity from the 12 autocorrelation coefficients.

In order to estimate the predictive power of the approach, cross-validation following the leave-one-out scheme was performed. In 31 independent experiments, the network was trained with the data of 30 steroids. After training, the network was used to predict the activity of the 31st compound. This procedure was repeated 31 times, until the biological activity of each compound had been predicted by a neural network that had not included this compound in the training set. Figure 13-11 shows the results in the form of a plot of the predicted affinity values against the experimental ones. Although the predicted values show the correct trend, the quality of the predictions is not quite satisfactory, having a cross-validated correlation coefficient r^2 of 0.63. Especially one outlier (marked by a circle) can be identified – the very same outlier already identified in the Kohonen map of Figure 13-8. This compound is the only one in the entire data set bearing a fluorine substituent and, thus, apparently outside the structure space covered by the other compounds. After omitting this compound from the data set and repeating the cross-validation, a much better predictive power is obtained, with a cross-validated r^2 of 0.84. It is interesting to note that the Comparative Molecular Field Analysis (CoMFA) method that was introduced with this data set achieved only a cross-validated r^2 of 0.66, on a subset of 21 steroids.

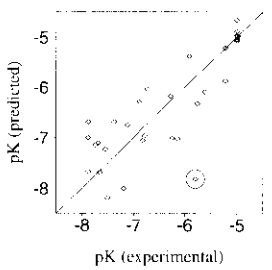


Figure 13-11: Crossvalidation of CBG activity of 31 steroids modeled by back-propagation.

13.9 Data Set III

The third set of data consists of 55 *flavonoid* derivatives which are low molecular weight substances found in most parts of all plants. Due to their broad variety of biological activities they are often called

"*bioflavonoids*". In this particular case, we shall explore flavonoid substances which are inhibitors of the *protein tyrosine kinase* (PTK) and are, therefore, important factors in cellular signal transduction. The skeleton of the flavonoids is given in Figure 13-13. The biological activity, *a*, in this study is the logarithm of the inverse experimental biological activity IC50, i.e., the molar concentration necessary for 50% of maximal inhibition of PTK in comparison with the experiment without the flavonoid. Table 13-3 gives the full list of 55 compounds and their corresponding activities. The compounds are described with their substituents and the positions at which they are bonded to the flavonoid skeleton. The data are taken from three papers by Cushman et al (References to Chapter 13).

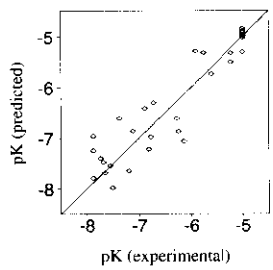


Figure 13-12: Crossvalidation of CBG activity of 30 steroids modeled by back-propagation.

13.10 Structure Representation by *Spectrum–Like* Uniform Representation

It has been explained many times that modeling either with artificial neural networks or any other method requires uniform representation of input data. Because different molecules are assembled from different number of atoms a direct 3D description by coordinate triplets $(x,y,z)_j$ of each of the constituent atom j does not fulfill the requirement of uniformness. One of the possible alternatives which is used in QSAR modeling quite often is a description of a molecule by several topological and electronic descriptors. In the present case a different approach featuring the so called *spectrum-like* structure representation will be used. In this Section only a brief explanation of the coding principles are given while in Section 21.4 a more detailed description how to calculate such a representation for any molecule is outlined.

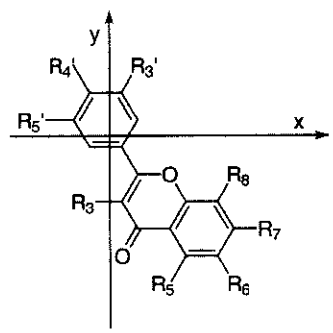


Figure 13-13: The flavonoid skeleton.

The main idea of the *spectrum-like* representation is to mimic a "light source" placed somewhere close to the molecule which casts "shadows" of atoms onto the surface of an imaginary sphere drawn around the light source (Figure 13-14). The positions and intensities of the shadows of the atoms on the surface of the sphere depend on the relative positions of the atoms and the light source. The complete shadow of all atoms on an arbitrary equator of the imaginary sphere resembles a "spectrum" (Figure 13-15), hence, the name of the representation.

	R3	R5	R6	R7	R8	R3'	R4'	R5'	activity *a*	class
1	OH	OH		OH		OH	OH		4.88	6
2	OH			OH		OH	OH		4.86	6
3		OH		OH			OH		4.83	6
4		OH					OH		4.80	6
5			OH			OH			4.80	6
6		OH		OH					4.71	6
7		OH		OH		OH	OH		4.46	5
8				OH		OH			4.41	5
9			OH			OCH_3	OCH_3	OCH_3	4.22	5
10	OH	OH		OH		OCH_3	OH	OCH_3	4.16	4
11	OH	OH		OH		OH		OH	4.00	4
12			OH				OH		3.93	4
13				OH	OH	OCH_3	OH	OCH_3	3.92	4
14			OH				OR		3.92	4
15			OH			OCH_3	OH	OCH_3	3.89	4
16				OH			OH		3.78	3
17				OH	OH	OH			3.75	3
18	OH	OH		OH					3.53	3
19		OH		OCH_3			OH		3.55	3
20		OH				OH			3.50	3
21				OH	OH				3.50	3
22				OH					3.47	3
23			OH			OCH_3	OR	OCH_3	3.43	3
24				OH	OH	OCH_3	OCH_3	OCH_3	3.40	3
25				OH			OR		3.01	2
26				OH		OCH_3	OH	OCH_3	2.90	1
27				OH		OCH_3	OR	OCH_3	2.82	1
28			OH				NH_2		5.92	9
29		OH		OH			NH_2		5.13	7
30						OCH_3	OH	OCH_3	4.57	5
31				OH			NH_2		3.86	4
32							NH_2		3.68	3
33	COOMe						OH		3.36	2
34							OH		3.30	2
35	COOMe						NH_2		3.09	2
36	COOH			OCH_3			OH		2.99	1
37	COOH						OH		2.80	1

Table 13-3: Flavonoid compounds with activities *a*. The substituents are marked according to the assignment of atoms in Figure 13-13. The activity values are divided into nine classes. The substituent OR stands for $OSi(Me)_2C(Me)_3$.

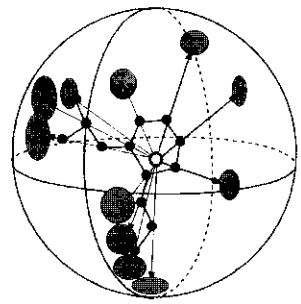


Figure 13-14: Shadows of atoms on a spherical surface with an arbitrary radius.

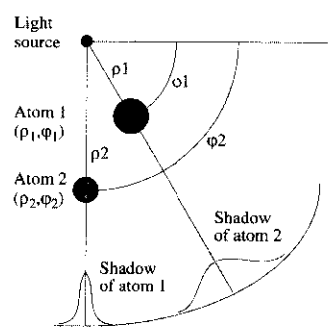


Figure 13-15: "Spectrum" of two atoms' shadows on the equator.

	R3	R5	R6	R7	R8	R3'	R4'	R5'	activity *a*	class
38		NH_2	OH	NH_2			NH_2		4.74	6
39		NH_2	OH	NH_2		NH_2			4.34	5
40			OCH_3		NH_2	NH_2			4.25	5
41			NH_2				NH_2		3.99	4
42			NH_2		NH_2		NH_2		3.97	4
43			OH		NH_2		NH_2		3.93	4
44					NH_2		NH_2		3.91	4
45			NH_2	OH			NH_2		3.85	4
46			NH_2			NH_2			3.70	3
47		OH	NH_2				NH_2		3.65	3
48		OH			NH_2		NH_2		3.49	3
49				OH	NH_2		NH_2		3.48	3
50			OCH_3		NH_2		NH_2		3.42	3
51			NH_2	OH		NH_2			3.30	2
52			NH_2	OH	NH_2		NH_2		3.12	2
53				OH					2.81	1
54		OCH_3			NH_2		NH_2		2.79	1
55				OH	NO_2		NO_2		2.73	1

Table 13-3: (continued) Flavonoid compounds with activities *a*. The substituents are marked according to the assignment of atoms in Figure 13-13. The activity values are divided into nine classes. The substituent OR stands for $OSi(Me)_2C(Me)_3$.

In order to compare *spectrum-like* representations of several compounds their structures must be aligned in the same manner. For example, to compare the flavonoids all their skeletons must be oriented in the same direction and superimposed onto each other. In other words the coordinate origin of the "light source" must be placed at exactly the same relative position of the internal coordinate system of the skeletons. In the present example the coordinate origin (or the "light source") for all representations is placed at the point (–1.7Å, 3.9Å, –0.6Å relative to atom no. 2 in the benzopyran ring system) within the benzene ring as shown in Figure 13-13.

The intensity s_{ij} of the shadow of an atom j on the equator at point i is described by the Lorentzian function (Equation 13.4). The Lorentzian function is chosen because of its simplicity. It could be any other appropriate function. In order to acquire information of the entire 3D structure, the "shadows" of the atoms are projected onto three mutually perpendicular circles. Figure 13-15 shows how for

each atom j its shadow's intensity s_{ij} depends on the angle φ_i on one circle:

$$s_{ij} = \frac{\rho_j}{(\varphi_j - \varphi_i)^2 + \sigma_j^2} \tag{13.4}$$

The position of the atom j is determined by the polar coordinates (ρ_j,φ_j) in the internal coordinate system of the "light source". The peak width parameter σ_j of the Lorentzian function (Equation 13.4) is used to describe any individual property of the atom j (atomic or ionic radius, atomic number, ionization energy, electron affinity, charge, etc.). Throughout this example for all atoms in all molecules the parameters σ_j are set to $\sigma_j = 1 + \textit{charge on atom } j$. If the charge on atom j is negative σ_j is less than *1*, otherwise it is larger than *1*.

If more atoms j are taken into account the intensities s_{ij} at all positions φ_i are clearly additive. The cumulative formula for any variable s_i of the additive *spectrum-like* representation of the whole structure consisting of n atoms can be written as:

$$s_i = \sum_{j=1}^{n} \frac{\rho_j}{(\varphi_j - \varphi_i)^2 + \sigma_j^2} \tag{13.5}$$

with φ_i running from φ_1 to φ_{360}

The number of variables s_i in each representation depends on the number of angles φ_i which divides the equator around the molecule – the finer the division, the more precise the description. If the resolution of 1° radial degree is chosen for the projection on each equator, one spectrum has 360 intensities. Hence, a complete *spectrum-like* representation of each molecule (projections of its structure into three perpendicular equators) has 1080 intensity values.

As can be easily understood, such representations are too large for most applications. In the present example the interval of division on each circle is 6°. Therefore, the spectrum on one circle is composed of 60 intensities, what means that the entire *spectrum-like* representation of any flavonoid has 180 intensities. Although much smaller, even this number is too high for handling 55 compounds and must therefore be reduced.

13.11 Selection of the Most Important Variables Using a Genetic Algorithm

A genetic algorithm is one of the most effective optimization methods for problems involving large number of variables. Its idea is to mimic the optimization by natural selection of living organisms in real life. The three main factors governing the natural selection are:

- **survival of the fittest**,
- changing the individual genetic material by **cross-over** of chromosomes, and
- changing of individual genetic material by **mutation** of genes.

All three mentioned factors are implemented in the computer simulated optimization called genetic algorithm, or GA for short. Before explaining these three factors in detail a few more parallelisms between living organisms and the objects to be optimized must be explained.

It is assumed that properties of each living being in nature are determined by genes "stored" in chromosomes. The presence or absence of genes that might be beneficial to or dangerous to the survival of an individual influences the chance on whether the subject will live long and have many offsprings or whether it will die having few or no offsprings at all.

Let us consider the case that in a world of fixed and limited resources there is a pool of living organisms that have only one chromosome with exactly 180 genes. This means that there are only 180 properties which can be important for their lives. For increasing the chance of survival some properties are good, some bad, and some irrelevant. Hence, the best suited individual for the given world would be the one whose chromosome would have only the genes assuring the good properties and none of those causing the bad ones. The crucial question in the optimization is how to find out which genes are responsible for the good and which for the bad properties. Unfortunately, we do not know this; all we know is only the individual's behavior in the defined world – its overall performance. If the individual has many good genes it will survive and have a lot of offsprings. The number of offsprings the individual "produces" is therefore influenced by the selection criterion which ranks each individual. For objects to be optimized this criterion is called a **fitness**

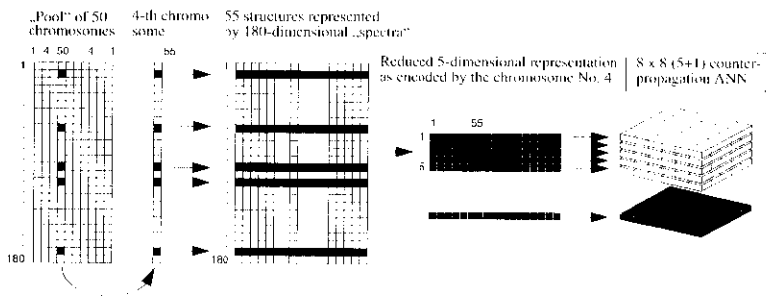


Figure 13-16: Pool of 50 chromosomes (left part). The 4-th chromosome is taken out and 55 flavonoid structures are represented by 5 intensity representation. With the set of 55 five-dimensional spectra an (8x8)-counter-propagation network model is generated and the RMS on the output layer evaluated (right part).

function. Definition of a good fitness function is always the crucial task in any optimization process.

One more parallelism between the natural selection of living organisms and the problem of selection of the most relevant variables in terms of a GA is that reduced *spectrum-like* representations can be described by one chromosome – like the above mentioned living organism. Each reduced representation features different set of intensities (genes) picked out from the 180 possible ones. One reduced representation might have only three intensities, another one fifty, still another one fourteen, and so on. How to decide which is the best?

The best, the fittest, or the optimized, is the reduced *spectrum-like* representation that would yield the best model. So, all we have to do is to represent all 55 objects with one of the possible reduced representations, make a model, test it; then represent the 55 objects again with another reduced representation, make a new model, test it and compare both results. By consecutive testing of various reduced representation with the generated models one can gradually find better and better representations.

The unpleasant part of such testing, however, is that a chromosome having 180 genes (bits) offers 2^{180} ($\approx 10^{54}$) possible reduced structure representations! Because there is no real chance to check even a minute part of this huge number of possibilities, we will try to use a genetic algorithm which promises to find, if not the very best, at least a very good one.

First, we set up a pool of 50 chromosomes each with 180 bits randomly turned to either ones or zeros. These strings of bits are "genetic codes" for 50 different reduced representations (Figure 13-16, left part). By considering the first "generation" of 50 reduced representations (out of the 10^{54} possible ones) we shall explain how

the testing is carried out. Once more, all 55 *flavonoids* are encoded 50 times, each time with a different reduced representation as suggested by the corresponding 180-bit-string – if the bit is turned to one, the intensity is taken into the account and otherwise not.

With each reduced representation a small 8 x 8 x (5 + 1) counter-propagation network (Figure 13-16, right part) is trained. The number of weights in the Kohonen layer in each of the 50 networks depends on the number of bits turned to one in the chromosome suggesting the tested reduced representation. In a small network of only 64 neurons on the average about 20-30% of neurons is expected to be excited by two or more different objects. The fitness function, *ff*, and, thus, a measure of quality of the tested representation is the RMS value (cf. Equation 7.6) or the square root of the sum of squared differences between the experimental activities of compounds exciting the same neuron and the response given by this neuron:

$$ff_k = \sqrt{\frac{\sum_{i=1}^{n_e}\sum_{j=1}^{n_i}\left(a_i - w_j^{out}\right)^2}{n_e \cdot n_i}} \qquad (13.6)$$

for $k = 1...50$ reduced representations

The first summation runs over all n_e neurons that are excited at least twice, while the second summation runs over all n_i experimental biological activities a_i of compounds that have excited the j-th neuron having in the output layer the weight (output activity) w_j^{out}. The smaller the fitness function (Equation 13.6), the better the reduced representation. Once all 50 different counter-propagation models are made, their rank list can easily be made by sorting the corresponding outcomes of the fitness functions $\{ ff_k \}$.

The second step of a GA is natural selection. This step involves the selection of the best chromosomes for mating, allowing them to have offsprings, and omitting ("killing") the others with low values of fitness function. From the ranked list of fitness functions $\{ ff_k \}$ sixteen best ones, approximately one third of all, i.e., 50/3 ≈ 16, are selected and the rest is ignored. By allowing to mate each of these sixteen chromosomes to three randomly selected partners 48 chromosomes of the new generation are produced. Additionally, the very best chromosome mates once more (four times altogether) and its clone (the identical chromosome) is added to the new generation. In this way 50 new chromosomes (reduced representations) are obtained.

In the GA, the offsprings of the mating process are generated by the **cross-over** procedure. For each mating pair, a random gene position is determined and from that position the cross-over (twisting) procedure is applied. The resulting two new chromosomes are obtained by exchange of the twisted parts (Figure 13-17). Each of the two offsprings has one part (lower or upper one) of the gene sequence from one parent and the other part from the second parent.

In a relatively small population of only 50 chromosomes it may well happen that all genes at a given position have the same value ("0" or "1") in all of them. Such gene position can never be changed by the cross-over procedure only. Therefore, a process called **mutation** is applied. Mutation means random switching of a small percentage of bits to its opposite value. If chosen for a mutation, the gene with value "0" turns to "1" and *vice versa*. In order to not disturb the improvements due to the cross-over breeding process too much, the mutation procedure ought to be applied with care. The probability at which a gene is subject to mutation should be low (usually below 1%).

Once all three steps of a GA – natural selection, cross-over and mutation – have been applied, new testing or fitness function evaluation of all 50 chromosomes of the new generation can start again. Due to the fact that the clone of the best chromosome from the previous generation is always present in the next one, constantly increasing values of the fitness function are assured. The number of generations required to obtain an optimum varies from case to case. In the present example 500 generations each consisting of 50 chromosomes have been tested (Figure 13-18). In the entire GA process, 25,000 times all 55 structures were encoded into different reduced representation and 25,000 counter-propagation networks were built and tested. The improvement of the RMS value as the fitness function in the GA process is shown in Figure 13-18.

After 500 generations the best fitness function as defined by Equation (13.15) on the 8 x 8 x (5 + 1) counter-propagation network was RMS = 0.167. This value was obtained using a reduced representation of 18 variables only.

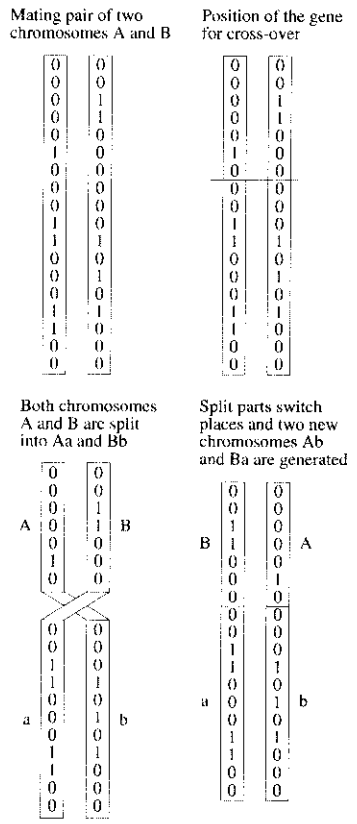


Figure 13-17: Cross-over procedure for making two new offsprings Ab and Ba from two mating chromosomes Aa and Bb.

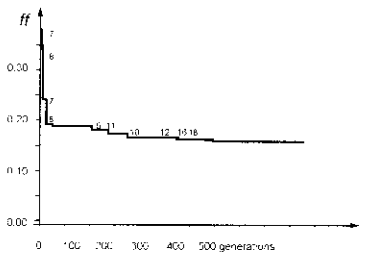


Figure 13-18: GA optimization evolution through 500 generations. The lengths of the reduced representations has increased from the starting 7 to the final 18 intensities.

13.12 Cross-validation of the Counter-Propagation Model Obtained by the Optimal Reduced Representation

The first task in the modeling process was to obtain the short uniform representation consisting of 18 variables for all 55 *flavonoids*. Out of the 180 intensities of the complete *spectrum-like* representation, the following 18 intensities were selected: 30th (180°), 32nd (192°), 33rd (198°), 44th (264°), and 46th (276°) from the first part (x,y-projection), 83rd (138°), 93rd (198°), 98th (228°), 108th (288°), 111th (306°), and 117th (342°) from the second part (x,z-projection, intensities from 61 to 120) and finally, 127th (42°), 129th (54°), 133rd (78°), 143rd (138°), 145th, (150°), 146th (156°), and 166th (276°) of the y,z-projection intensities from 121 to 180 (Figure 13-19). In parentheses, the corresponding angle in radial degrees is given. Because these selected variables represent defined space windows of 6° radial degrees it is easy to conclude that the presence or absence of substituent atoms in these eighteen directions comprises the most influential factor in the biological activity of flavonoids.

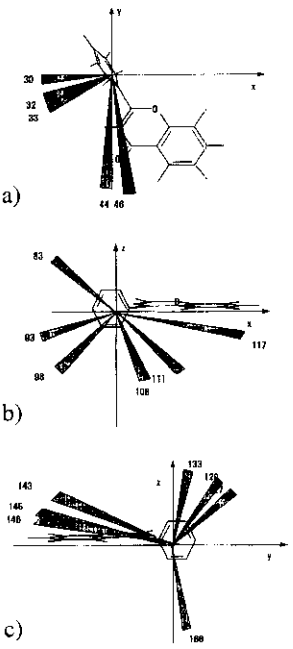


Figure 13-19: Each of the selected 18 variables define a 6° degree wide window in the space where the most relevant substituents are lying. There are 5 directions in the (x-y) plane a), 6 in the (x-z) plane b), and 7 directions in the (y-z) plane c).

It was mentioned in Section 13.11 that the RMS value of the recognition results on the 8 x 8 counter-propagation neural network model is applied as fitness function. At this place two more questions have to be elaborated in more detail. The first one is why the counter-propagation network was chosen for modeling and not the error back-propagation, and, second, why the choice was made on a 8 x 8 network and not something else, let us say a 7 x 7 or 10 x 10 network?

Both answers are relatively simple. Because in the Kohonen layer of the counter-propagation network the formation of clusters on the basis of input representations is achieved, one can evaluate the formation of the optimized reduced representation that discriminates between the compounds in question better than with the error back-propagation which delivers only the model and no internal information about the representations. The other argument for the preference of counter-propagation over the error back-propagation model is the number of training epochs necessary for the networks to converge. The convergence rate of the former is two orders of magnitude better than that of the latter one.

The choice of the 8 x 8 layout of the counter-propagation network (Figure 13-20) is based on the following reasoning. In a 49 neurons

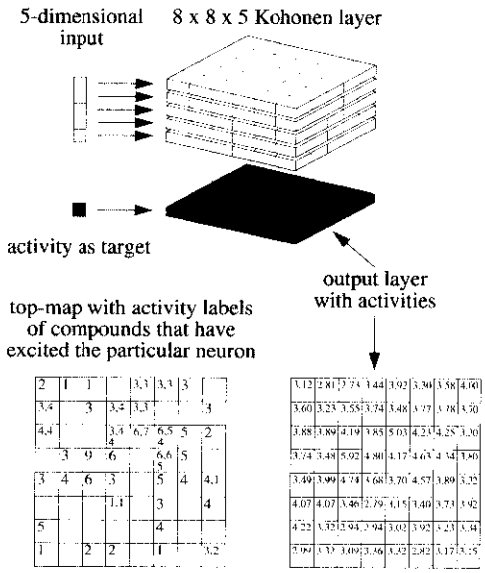


Figure 13-20: Counter-propagation network used in the GA procedure and for the final model (above). The distribution of 55 hits on the output map (lower part left). Double and triple hits are marked by the numbers 2 or 3, respectively. The final output layer with resulting activities (lower part right).

network (i.e., 7 x 7 layout) there is not even a theoretical chance that each of the 55 objects would excite its own neuron. On the other hand in a large network of, let us say, 400 = 20 x 20 neurons, each object would almost inevitably excite its own neuron, hence no good quantitative measure of how structurally similar *vs.* dissimilar and at the same time biologically more-active *vs.* less-active compounds would be clustered. Therefore, the first larger network of 64 neurons (8x8 layout) that theoretically allows exciting of 55 different neurons is a reasonable choice because it enables the theoretical possibility of the formation of a network which would have an RMS value equal to zero (each object would excite its own neuron). However, in the optimization process, while testing different representations, the RMS values will undoubtedly vary quite significantly, thus allowing to make quantitative comparisons of the representations. Even more, the final output map of the counter-propagation network will show the distribution of all objects depending on the optimized representation.

Figure 13-20 (lower part) shows the final output map of the 8 x 8 counter-propagation network of 55 flavonoids. Nine neurons were excited twice and three of them three times. This counter-propagation model generated with all 55 flavonoids represented by 18 variables in 900 epochs has a correlation factor of estimates a_i^{esi} *vs.* experimental biological activities a_i of 0.95.

In order to check the reliability of both the model and the selected reduced representation, a cross-validation test is performed. The cross-validation leave-one-out test is one of the possibilities how to simulate real-life conditions in shortage of a test set. The leave-one-out test (known also as jack-knifetest) requires to make 55 models by the same modeling procedure with the same representation, but each time with one object omitted from the modeling procedure. The test of the prediction of this model obtained on 54 objects only is executed by input of the left-out object into it as an "unknown". The correlation coefficient r between 55 predictions $a_i^{est\ by\ CV}$ in the cross-validation procedure and the actual experimental activities a_i gives a fair estimate how the actual model obtained by the same procedure on 55 objects will perform when encountered with really unknown objects of the same type. The cross-validation correlation factor r obtained in our study was 0.86. The table of actual biological activities *vs.* the predictions obtained by the cross-validation predictions is given in Table 13-4.

Sample No.	a_i	$a_i^{est\ CV}$	Sample No.	a_i	$a_i^{est\ CV}$	Sample No.	a_i	$a_i^{est\ CV}$
1	4.88	4.74	21	3.50	3.56	41	3.99	4.02
2	4.86	4.73	22	3.47	3.49	42	3.97	3.57
3	4.83	4.98	23	3.43	3.22	43	3.93	3.89
4	4.80	4.46	24	3.40	3.87	44	3.91	3.81
5	4.80	4.47	25	3.01	3.22	45	3.85	3.90
6	4.71	4.12	26	2.90	3.53	46	3.70	3.66
7	4.46	4.66	27	2.82	3.37	47	3.65	3.78
8	4.41	4.61	28	5.92	4.88	48	3.49	3.74
9	4.22	3.75	29	5.13	4.97	49	3.48	3.77
10	4.16	3.53	30	4.57	4.06	50	3.42	3.70
11	4.00	3.98	31	3.86	3.82	51	3.30	3.59
12	3.93	3.71	32	3.68	3.64	52	3.12	3.14
13	3.92	4.25	33	3.36	3.23	53	2.81	3.21
14	3.92	3.50	34	3.30	4.05	54	2.79	2.80
15	3.89	4.23	35	3.09	3.23	55	2.73	3.10
16	3.78	3.83	36	2.99	3.61			
17	3.75	3.63	37	2.80	3.32			
18	3.53	4.12	38	4.74	4.10			
19	3.55	3.64	39	4.34	4.02			
20	3.50	3.63	40	4.25	3.98			

Table 13-4: Comparison of the experimental biological activities a_i and activities obtained by the cross-validation process $a_i^{est\ CV}$. The correlation factor r between these two series is *0.86*. The estimated $a_i^{est\ model}$ as yielded by the complete model on 55 objects are even better giving r=*0.95*.

13.13 References and Suggested Readings

13-1. T. Aoyama, Y. Suzuki and H. Ichikawa, "Neural Networks Applied to Quantitative Structure-Activity Relationships", *J. Med. Chem.* **33** (1990) 905 – 908.

13-2. T. Aoyama, Y. Suzuki and H. Ichikawa, "Neural Networks Applied to Quantitative Structure-Activity Relationship (QSAR) Analysis", *J. Med. Chem.* **33** (1990) 2583 – 2590.

13-3. M. Yoshimoto, H. Miyazawa, H. Nakao, K. Shinkai and M. Arakawa, "Quantitative Structure-Activity Relationships in 2,5-Bis(1-aziridinyl)-p-benzoquinone Derivates against Leukemia L-1210", *J. Med. Chem.* **22** (1979) 491 – 496.

13-4. R. D. Cramer, III, D. E. Patterson, and J. D. Bunce, "Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins", *J. Am. Chem. Soc.* **110** (1988) 5959 – 5967.

13-5. G. Moreau, and P. Broto, "Autocorrelation of Molecular Structures: Application to SAR Studies", *Nouv. J. Chim.* **4** (1980) 757 – 764.

13-6. M. Wagener, J. Sadowski and J. Gasteiger, "Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks", *J. Am. Chem. Soc.* **117** (1995) 7769 – 7775.

13-7. E. A. Coats, "The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods" in *3D QSAR in Drug Design*, Vol. 3, Eds.: H. Kubinyi, G. Folkers and Y. C. Martin, Kluwer / ESCOM, Dordrecht, NL, 1998, pp. 199 – 214.

13-8. J. Zupan and M. Novic, "General Type of a Uniform and Reversible Representation of Chemical Structures", *Anal. Chim. Acta* **348** (1997) 409 – 418.

13-9. R. T. Burke, "Protein - Tyrosine Kinase Inhibitors", *Drugs of the Future* **17** (1992) 119 – 131.

13-10. M. Cushman, D. Nagarathnam and L. R. Geahlen, "Synthesis and Evaluation of Hydroxylated Flavones and Related Compounds as Potential Inhibitors of the Protein-Tyrosine Kinase p56", *J. Nat. Products* **54** (1991) 1345 – 1352.

13-11. M. Cushman, D. Nagarathnam, L. D. Burg and L. R. Geahlen, "Synthesis and Protein-Tyrosine Kinase Inhibitory Activities of Flavonoid Analogues", *J. Med. Chem.* **34** (1991) 798 – 806.

13-12. M. Cushman, H. Zhu, L. R. Geahlen and J. A. Kraker, "Synthesis and Biochemical Evaluation of a Series of Aminoflavones as Potential Inhibitors of Protein-Tyrosine Kinases p56, EGFr, p60", *J. Med. Chem.* **37** (1994) 3353 – 3362.

13-13. M. Novic, Z. Nikolovska-Coleska and T. Solmajer, "Quantitative Structure-Activity Relationship of Flavonoid p56lck Protein Tyrosine Kinase Inhibitors. A Neural Network Approach", *J. Chem. Inf. Comput. Sci.* **37** (1997) 990 – 998.

13-14. B. Hibbert, "Genetic Algorithm in Chemistry", *Chemom. Intell. Lab. Syst.* **19** (1993) 277 – 293.

13-15. D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, New York, USA, 1989.

13-16. G. Jones, "Genetic and Evolutionary Algorithms", in *Encyclopedia of Computational Chemistry*, Eds.: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III and P. R. Schreiner, Wiley, Chichester, UK, 1998, pp. 1127 – 1136.

13-17. V. Venkatasubramanian and A. Sundaram, "Genetic Algorithms: Introduction and Applications", in *Encyclopedia of Computational Chemistry*, Eds.: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III and P. Schreiner, Wiley, Chichester, UK, 1998, pp. 1115 – 1127.

13-18. J. Devillers (Ed.), *Neural Networks in QSAR and Drug Design*, Academic Press, London, UK, 1996.

13-19. J. Devillers (Ed.), *Genetic Algorithms in Molecular Modeling*, Academic Press, London, UK, 1996.