20 Libraries of Chemical Compounds

learning objectives:

- separation of structures according to their biological activity
- autocorrelation as a mathematical transformation producing a fixed number of descriptors
- structure coding by autocorrelation considering the constitution of a molecule, different atomic properties, or molecular surface properties
- perception of similarity between chemical structures
- analysis of chemical structure space
- definition of similarity / dissimilarity of chemical compounds
- analysis of diversity and similarity of large chemical libraries
- search for new lead structures

20.1 The Problems

The development of a new drug or agrochemical presently requires, on average, the synthesis of approximately 40,000 new compounds, and this number is still rising. This underscores that the search for a new drug is quite often like the search for the needle in the haystack. On the other hand, biological test systems have become available that allow the testing of many compounds in a short time. The capacity for high-throughput screening (HTS) puts a lot of pressure on rapidly synthesizing many new compounds. These requirements have recently been met by the development of parallel synthesis and combinatorial chemistry that provide large collections, so-called libraries, of chemical compounds.

When synthesizing, or testing, a library of compounds one wants to be sure that the new library is different - dissimilar - to the one previously investigated. Furthermore, in the search for a new lead structure, first, libraries should be synthesized that span the chemical space as broad as possible - are highly diverse - in order to ensure that the kind of compounds that show the desired biological activity are contained in the libraries. In later stages of the search, focussed libraries that center on that part of the chemical space that contains those structures having the desired biological activity should be investigated. Thus, questions of similarity and diversity of chemical structures and libraries become important. Figures 20-1 and 20-2 illustrate the concept of similarity and diversity of chemical libraries.

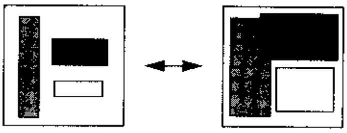

Figure 20-1: Visualization of the concept of diversity: The data sets on the left-hand side are not diverse enough to fill the entire chemical space.

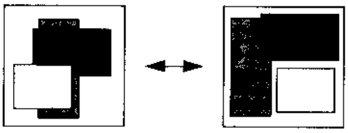

Figure 20-2: Visualization of the concept of similarity: The data sets on the left-hand side are too similar to fill the entire chemical space.

To answer those question, an appropriate structure coding has to be chosen, a structure coding that is somehow related to the biological activity under investigation. We will first discuss several methods for representing chemical structures and then investigate the capability of one such structure coding method to differentiate between compounds of different biological activity. Furthermore, we will show that this structure representation focusses the structures with the desired biological activity into a restricted part of the chemical space. We will then address questions of similarity and diversity of large chemical libraries as met in combinatorial chemistry.

20.2 Structure Coding

The question of similarity can only be defined for a specific purpose, in our case, biological activity: structures are considered similar if they carry similar biological activity. Thus, the structure representation has to consider those properties of a chemical structure that are deemed to be responsible for the biological activity under investigation. Furthermore, the structure coding scheme must produce the same number of descriptors, irrespective of the size of a molecule, the number of atoms in a molecule. For, any automatic learning method such as a neural network has to have a fixed number of input units and, therefore, requires the objects under investigation to be represented by a predetermined, constant number of variables.

Thus, the chemical structure has somehow to be transformed to produce a fixed number of descriptors. In this chapter we will present one such mathematical transformation, autocorrelation, and show how it can consider structure information of various degrees of sophistication, either the constitution only, or molecular surfaces. Furthermore, we will show how various physicochemical properties of atoms, or molecular surfaces can be introduced into the structure coding method. Structure coding by autocorrelation has already been used in Chapter 13. A more extensive discussion of various methods for structure representation is contained in Chapter 21.

The idea of using autocorrelation for the transformation of the constitution of a molecule into a fixed length representation was introduced by Moreau and Broto. The property, p, of an atom, i, is correlated with the same property on atom j and these products are summed over all atom pairs having a certain number of intervening bonds, a certain topological distance, d. An example for the definition of a topological distance is given in Figure 20-3. This gives one element of a topological autocorrelation function $A(d)$:

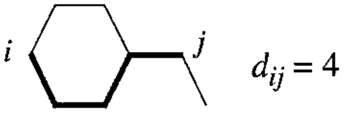

Figure 20-3: Definition of topological distance, d_{ij}, as the number of bonds between two atoms i and j.

$$A(d) = \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} \delta_{ij}\, p(i)\, p(j) \qquad (20.1)$$

with $\delta_{ij} = 1$ if $d_{ij} = d$, otherwise $\delta_{ij} = 0$

The following properties were calculated by previously published empirical methods contained in the program package PETRA for all atoms of a molecule: sigma charge, q_{σ}, total charge, q_{tot}, sigma-electronegativity, χ_{σ}, pi-electronegativity, χ_{π}, lone pair-electronegativity, χ_{LP}, and atom polarizability, α.

In addition to these six electronic variables, the identity function, i.e., each atom being represented by the number 1, was used in Equation (20.1) to only account for the connectivity of the atoms.

The autocorrelation of these variables was calculated for seven topological distances (number of intervening bonds) from two to eight. The basic assumption thus was that the interaction of atoms beyond eight bonds can be neglected. Thus, the descriptor for representing molecular structures is given by Equation (20.2)

$$A(p_k,n) = \sum_{i,j \in M(n)} p_k(i)\, p_k(j) \qquad (20.2)$$

$$\text{with } M(n) = \{(i,j) \mid \#bond(i,j) = n\}$$

$p_k(i)$ is the k-th property on atom i, and #bonds (i,j) is the minimum number of bonds between atoms i and j.

With seven variables and seven distances an autocorrelation vector of dimension 49 was obtained for each molecule, irrespective of its size or number of atoms.

Ligands and proteins interact through molecular surfaces and therefore, clearly, representations of molecular surfaces have to be sought in the endeavor to understand biological activity. Again, we are under the restriction of having to represent molecular surfaces of different size, and, again, autocorrelation was employed to achieve this goal.

First, a set of randomly distributed points on the molecular surface has to be generated. Then, all distances between the surface points are calculated and sorted into preset intervals according to Equation (20.3).

$$A(d) = \frac{1}{m}\sum_{i,j} p(i)\, p(j) \qquad (20.3)$$

$$\text{with } d_l < d_{ij} < d_u$$

where $p(i)$ and $p(j)$ are property values at points i and j, respectively, d_{ij} is the distance between the points i and j, and m is the total number of distances in the interval $[d_l,d_u[$ represented by d. For a series of distance intervals with different lower and upper bounds, d_l and d_u, a vector of autocorrelation coefficients is obtained. It is a condensed representation of the distribution of the property on the molecular surface. This coding was also used in the example contained in Sections 13.6 – 13.8.

20.3 Separation of Benzodiazepine and Dopamine Agonists

In order to investigate the potential of topological autocorrelation functions for the distinction of biological activity, a data set of 112 *dopamine agonists* (DPA) and 60 *benzodiazepine agonists* (BDA) was studied. A Kohonen network of size 10 x 7 was used to project these 172 compounds from the 49-dimensional space spanned by these autocorrelation vectors into two dimensions. The results are shown in Figure 20-4.

It can be seen that the two types of compounds, DPA and BDA, are nearly completely separated in the Kohonen map, underscoring the potential of this molecular representation to model biological activity.

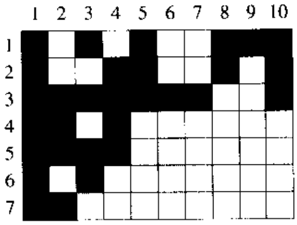

Figure 20-4: Kohonen map of size 10 x 7 neurons obtained for the data set of 112 *dopamine* (black) and 60 *benzodiazepine agonists* (light gray). The separation of the two types of biologically active molecules is nearly complete.

20.4 Finding Active Compounds in a Large Set of Inactive Compounds

To put this potential for comparing data sets of compounds and clustering of compounds with a desired biological activity to a more stringent test, this data set of 112 DPA and 60 BDA compounds was mixed with the entire catalog of a chemical supplier consisting of 8,323 commercially available compounds comprising a wide range of structures from alkanes to triphenylmethane dyestuffs.

The map of Figure 20-5 shows that both sets of compounds, DPA and BDA, occupy only limited areas in the overall map. Furthermore, the areas of DPA and BDA are quite well separated from each other, only one neuron with BDA, intrudes into the domain of DPA and only two neurons have conflicts, obtaining both DPA and BDA. Clearly, the areas of neurons with DPA and BDA are larger than one probably has hoped them to be. For, with the results obtained here the search for new active compounds or new lead structures in a data set of compounds of unknown activity will have to scan a fairly large area and, correspondingly, quite a few compounds. However, compared to the overall size of the network, the areas where DPA and BDA are to be found are distinctly smaller and quite concentrated. Closer analysis of the mapping shows interesting insights that are further discussed in the original publication.

The six electronic factors and the connectivity of a molecule, their encoding into topological autocorrelation vectors and their projection into a two-dimensional map by the self-organizing capability of a Kohonen network provide a powerful means for the detection of similarity in the structure of organic molecules. *Dopamine agonists* can be separated from *benzodiazepine receptor agonists* and this

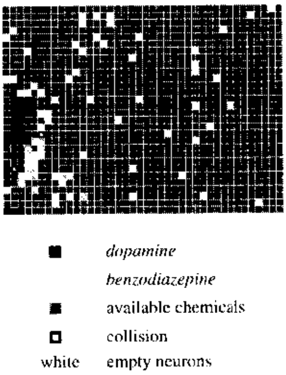

Figure 20-5: Kohonen map of 40 x 30 neurons obtained by training with 112 *dopamine* (DPA), 60 *benzodiazepine agonists* (BDA), and 8,323 commercially available compounds. Only the type of compounds mapped into the individual neurons is indicated. Black identifies DPA, light gray BDA, and dark gray the compounds of unknown activity. Empty neurons are shown in white; the two neurons marked by a black frame indicate conflicts where both DPA and BDA are mapped into the same neuron.

separation is maintained when these two types of compounds are embedded in a larger set of structures.

This opens the way for searching for compounds with a desired biological activity and for discovering new lead structures in large databases of compounds.

Furthermore, this approach can be used for the comparison of libraries of compounds in order to decide whether a commercially offered compound library is distinctly different from the inhouse compound collection.

20.5 Diversity and Similarity of Combinatorial Libraries

The merit of the autocorrelation of molecular surface properties such as the molecular electrostatic potential for the classification and the modeling of biological activity has already been shown in Section 13.6. Here, we will show how this structure representation can be used for the analysis of large combinatorial libraries.

The methods introduced in the previous sections have the advantage that they allow for a rapid visualization of high-dimensional descriptor spaces. The importance of this feature has increased with the advent of the large compound collections that can be generated by combinatorial chemistry and related techniques: small data sets comprising tens or hundreds of compounds can be analyzed using almost any method without reaching the limits of currently available computer hardware. Special techniques, however, are needed for the handling of data sets of hundreds of thousands of compounds. To demonstrate the merits of Kohonen networks and spatial autocorrelation descriptors in handling large data sets, we analyzed three combinatorial libraries that together comprise more than 87,000 compounds.

Rebek et al. published the synthesis of two combinatorial libraries of semi-rigid compounds that were prepared by condensing a rigid central molecule functionalized by four *acid chloride* groups with a set of 19 different *L-amino acids*. This process is summarized in Figure 20-6. In addition to the two published libraries we included a third, hypothetical library with *adamantane* as central molecule into our study.

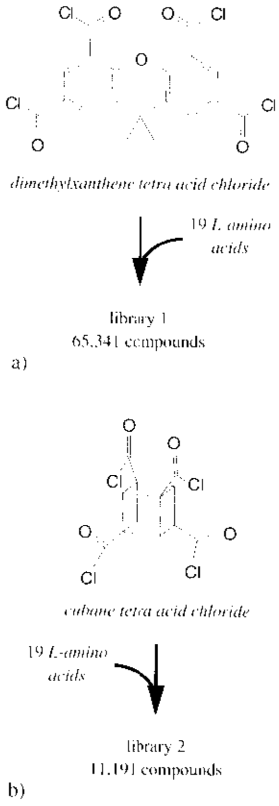

Figure 20-6: Preparation of the *xanthene* and the *cubane* libraries by reaction of four *acid chloride* substituents on these skeletons with four (identical or different) *amino acids* from a set of 19 amino acids. The more symmetric *cubane* skeleton gives less compounds.

A Kohonen network with 50 x 50 neurons was trained with the combined descriptors of the *xanthene* and the *cubane* libraries, each molecule represented by 12 autocorrelation values calculated from the electrostatic potential on the molecular surface by equation 20.3. The resulting map is shown in Figure 20-7. The neurons are colored according to the most frequent central molecule that is mapped into them. All 2,500 neurons of the map are occupied. The compounds of the cubane library form a cluster in the center of the map that is separated from the compounds of the *xanthene* library. The neural network can clearly separate the two libraries quite well - they both cover different parts of chemical space - only 3 per cent of the neurons obtain both *xanthene* and *cubane* derivatives. Consequently, they are remarkably different and, thus, both worthwhile to be considered in a screening program.

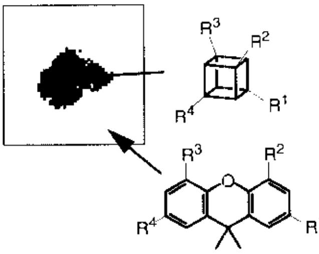

Figure 20-7: Kohonen map obtained from the two libraries of 65,341 *xanthene* derivatives and the 11,191 *cubane* compounds. The area into which the xanthene derivatives are mapped is colored grey whereas the area into which the *cubane* derivatives are mapped is colored black.

In a second experiment we trained the same network with the combined data set of the three libraries of *xanthene*, *cubane* and *adamantane* compounds. This resulted in the Kohonen map shown in Figure 20-8. Again, a distinct cluster that is clearly separated from the *xanthene* derivatives can be seen in the center of the map. The *cubane* and *adamantane* derivatives, on the other hand, cannot be distinguished by the neural network. They are tightly mixed in the central cluster, even more than can be inferred from Figure 20-8 as 88% of the cubane and adamantane compounds are mapped into common neurons.

The *cubane* and *adamantane* libraries, thus, cover the same part of the chemical space - they are so similar to each other that considering both of them in a screening program is a waste of resources and time. The *xanthene* library is evidently different from the other two libraries so that the *xanthene* and one of the *cubane* or *adamantane* libraries should be used for screening.

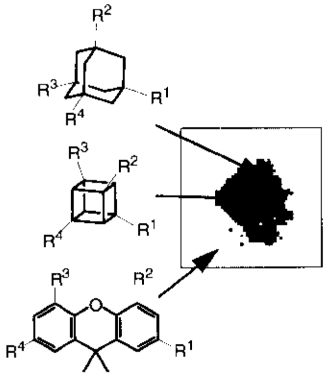

Figure 20-8: Kohonen map obtained from three libraries of 65,341 *xanthene*, 11,191 *cubane* and 11,191 *adamantane* derivatives. The center cluster shows the area into which the *cubane* (colored black) and *adamantane* (colored dark grey) compounds are mapped.

20.6 Deconvolution of Xanthene Sublibraries

Rebek et al. used their libraries to screen for novel *trypsin inhibitors*. Only the *xanthene* library showed significant *trypsin* inhibition, so that they concentrated further efforts on this library. In the next round of screening they divided the *xanthene* library into six sublibraries by using subsets of only 15 *amino acids* for the generation of the libraries. These subsets were generated by omitting three *amino acids* in turn from a set of 18 *amino acids* (Figure 20-9).

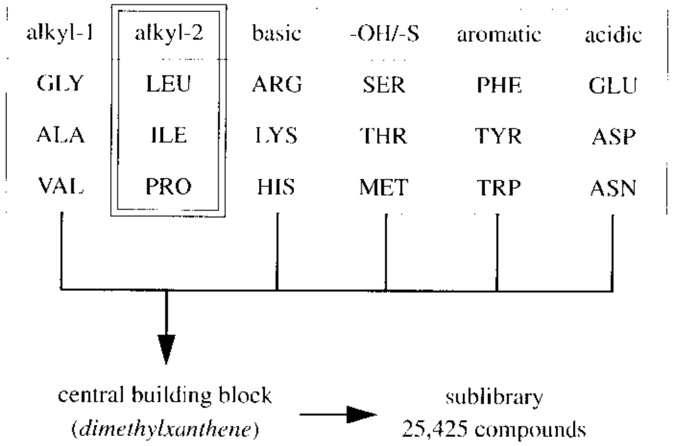

Figure 20-9: Building six sublibraries by reacting 15 *amino acids* with *dimethylxanthene* having four *acid chloride* group (see Figure 20-6). The three *amino acids* that are omitted in the present case are shown in a double frame. Each one of the six sublibraries contains 25,425 compounds.

This process resulted in six sublibraries each with 25,425 compounds, that were tested for their *trypsin* inhibition.

To study the diversity of the six sublibraries we first trained a network with the complete *xanthene* data set resulting in a map with all neurons occupied. Then, each one of the sublibraries was sent through this template network of the complete library, altogether giving six maps, one each for each sublibrary. In these maps, we then marked, for each sublibrary, only those neurons containing compounds of the respective sublibrary.

The six maps are reproduced in Figure 20-10. They show remarkable differences: some of them are nearly completely filled, some of them exhibit large white areas representing neurons that no compound was mapped into. The larger these white areas are, the less the corresponding sublibrary covers the chemical space of the original *xanthene* library. For example, the omission of the basic or acidic *amino acids* has led to a decreased diversity as shown by the large number of empty neurons. On the other hand, the omission of the larger alkyl *amino acids* or the -OH and -S- substituted *amino acids* from the *xanthene* library does not lead to a remarkable decrease in diversity as there are only small white areas in the corresponding maps.

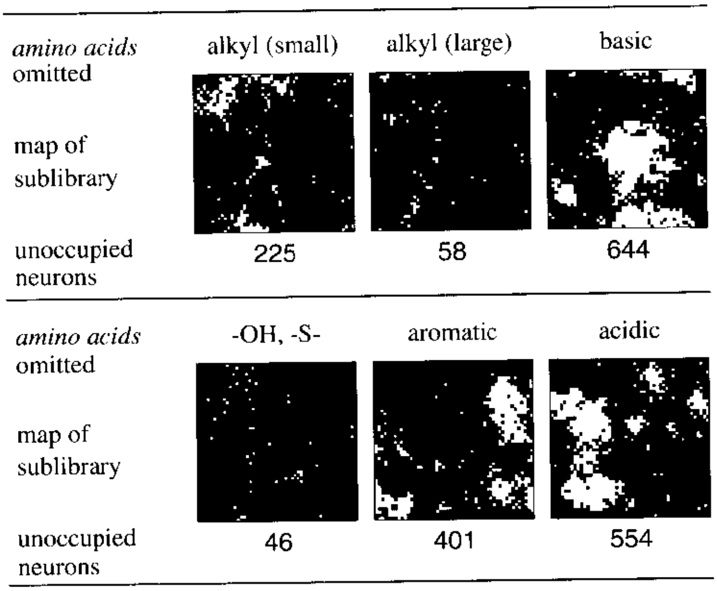

Figure 20-10: Kohonen maps of the six sublibraries built according to Figure 20-9 using the Kohonen network of the complete *xanthene* derivatives library as a template.

On the basis of these Kohonen maps of such sublibraries, strategies for the deconvolution of combinatorial libraries can be developed.

Clearly, these statements about diversity are valid only within the chosen structure representation, i.e., on the basis of the molecular electrostatic potential. However, for compounds containing aminoacids the molecular electrostatic potential is an important factor influencing biological activity.

20.7 References and Suggested Readings

20-1. G. Moreau and P. Broto, "Autocorrelation of Molecular Structures: Application to SAR Studies", *Nouv. J. Chim.* **4** (1980) 757 – 764.

20-2. For further information on PETRA see: *http://www2.ccc.uni-erlangen.de/software/petra/*

20-3. CORINA can be accessed on the internet: *http://www2.ccc.uni-erlangen.de/software/corina/*

20-4. H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski and J. Gasteiger, "Locating Biologically Active Compounds in Medium-Sized Heterogeneous Data sets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists", *J. Chem. Inf. Comput. Sci.* **36** (1996) 1205 – 1213.

20-5. T. Carell, E. A.Wintner, A. Bashir-Hashemi and J. Rebek, Jr, "A Novel Procedure for the Synthesis of Libraries Containing Small Organic Molecules", *Angew. Chem. Int. Ed. Engl.* **33** (1994) 2059-2061; *Angew. Chem.* **106** (1994) 2159-2162; T. Carell, E. A.Wintner, A. Bashir-Hashemi and J. Rebek, Jr, "A Solution-Phase Screening Procedure for the Isolation of Active Compounds from a Library of Molecules" *Angew. Chem. Int. Ed. Engl.* **33** (1994) 2061-2064; *Angew. Chem.* **106** (1994) 2162 – 2165.

20-6. J. Sadowski, M. Wagener and J. Gasteiger, "Assessing Similarity and Diversity of Combinatorial Libraries by Spatial Autocorrelation Functions and Neural Networks", *Angew. Chem. Int. Ed. Engl.* **34** (1995) 2674 – 2677; *Angew. Chem.* **107** (1995) 2892 – 2895.