21 Representation of Chemical Structures

learning objectives:

- coding of structure information of various degrees of sophistication
- consideration of physicochemical properties of atoms
- requirements for a good structure representation
- transformation of structure information into fixed-length representations
- molecular transformation of 3D structure information
- encoding of molecular surface properties
- structure encoding methods applicable to large data sets

21.1 The Problem

It has already been mentioned several times in this book that the most important key to success in the application of neural networks lies in the choice of the proper representation of information. In chemical applications, we have often to deal with relationships between chemical structure and physical, chemical, or biological properties. Therefore, the representation of chemical structures is of paramount importance.

The problem of – and solutions to – structure representation has been mentioned in various sections of Chapters 11, 13, 14, and 17–20. This chapter serves to provide an overview and to order the various types of structure representation into a coherent framework.

Chemists have developed a variety of methods for representing and communicating structure information. The most widely used,

international language is the structural formula; it is still the method of choice when representing chemical reactions. For a more in-depth analysis, three-dimensional molecular models are built, either by mechanical molecular model kits, or, increasingly by computer modeling. A variety of representations is available, from framework, through ball and stick, to space-filling models. An even more refined analysis of molecules, particularly when studying biological activity, has to consider molecular surfaces, surface properties, and molecular potentials and fields.

All these various representations of chemical structures have to be translated into a form amenable to computer manipulation. A further requirement set by the use of learning methods is that molecules have to be represented by the same number of descriptors, irrespective of their size, of the number of atoms in a molecule. Only then can data sets of different molecules automatically be processed by statistical or pattern recognition methods or by neural networks.

In the following, we will present various techniques for encoding these different forms that the chemists use for structure representation, from the constitution of a molecule, through 3D structures to molecular surfaces. Several of these structure representation have already been introduced in previous chapters. Nevertheless, we mention them here again to collect them in a concise overview. These different encoding methods have been developed for the different requirements made by the intended applications. Furthermore, what kind of coding method will be chosen, will also strongly be dictated by the size of the data sets that have to be studied. Data sets of hundreds of thousands or millions of structures have to rely on rather rapid encoding procedures in order to be handled in a reasonable amount of time.

21.2 Coding the Constitution

The structural formula (Figure 21-1) can be considered as a mathematical graph; graph theory has therefore played a major role in the computer handling of structure information. However, the representation of a molecule as a graph, as a list of atoms and bonds does not fulfill the requirement for a fixed number of descriptors, irrespective of the size of a molecule. In many applications, molecules are represented by lists of fragments or substructures, in the form of bit strings; the presence or absence of a certain functional group is

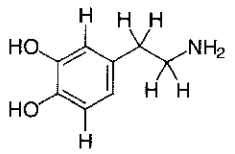

Figure 21-1: The constitution of a molecule.

indicated by a 1 or 0. Thus, there is a clear correspondence between the position of a bit and the substructure present or absent. Such representations of a structural formula are often called 2D descriptors. However, they do not carry any direct 2D information; they are only a reflection of the constitution of a molecule, and therefore should be called topological descriptors, at most.

With a predefined number of substructures one does, indeed, arrive at a structure representation with a fixed length. However, the choice of substructures to be considered will always be arbitrary because, in principle, the number of substructures in organic compounds is unlimited.

In order to be able to consider large sets of substructures an extension of this structure coding by fragments has been introduced, the so-called fingerprints. The occurrence of a large set of fragments is hash-coded into a bit-string representation in order to arrive at a more concise representation. Nevertheless, such a fingerprint representation may consist of a string of 500 – 1,000 bits. In contrast to the former representation no inference on the kind of substructures present can any more be made. Because of the hash-code algorithm no direct relationships between a bit position and a substructure exists any more.

We have sought for methods that allow one to encode various physicochemical properties of the atoms in a molecule, such as partial charges, polarizability, etc. Our approach rests on autocorrelation functions as outlined in Section 13.6 and Chapter 20 and given by Equation (21.1):

$$A(d) = \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} \delta_{ij}\, p(i)\, p(j) \qquad (21.1)$$

With a topological autocorrelation vector only the constitution of a molecule is considered.

A value for the autocorrelation function A, at a certain topological distance (number of bonds), d, is calculated by summation over all products of a certain property, p, of atoms i and j having the required distance, d.

A range of properties such as partial atomic charges, measures of the inductive effect, resonance, or polarizability effect were calculated by rapid empirical methods contained in the program package PETRA (Parameter Estimation for the Treatment of Reactivity Applications).

Various applications have shown the merit of such a representation. Of particular importance is the possibility of choosing such physicochemical properties of the atoms that are deemed responsible for the effect under investigation. The example given in Section 20.3 – 20.4 is a case in point. The representation of structures by topological autocorrelation of a variety of electronic properties of the atoms in a molecule allows one to distinguish structures having different biological activity, to limit the search space in lead compound search, and to compare different libraries of compounds.

21.3 Coding the 3D Structure I

The study of the relationships between biological activity and the 3D structure (Figure 21-2) of a molecule on a broad scale has been made possible by the advent of universal and efficient automatic 3D generators. Programs are available such as CORINA which can convert large databases such as the Beilstein file with about 7 million structures. With a 3D structure accessible for practically any organic molecule, the problem is then, how to encode the 3D structure under the restriction of having to come up with a fixed number of variables, independent of the number of atoms in a molecule. Clearly, again, autocorrelation of atomic properties as given by Equation (21.1), now inserting genuine spatial distances can be used.

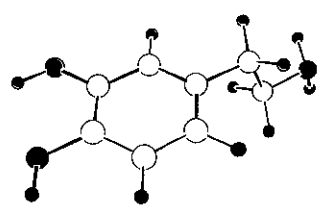

Figure 21-2: The 3D structure of the molecule shown in Figure 21-1.

In fact, useful applications of such a structure representation were made. We were, however, also seeking for a structure representation that offers the possibility of regaining the 3D structure from the molecular code. Building on equations used for obtaining the 3D structure of a molecule from electron diffraction experiments the encoding procedure embodied in Equation (21.2) was developed (cf. Section 18.7).

$$\boldsymbol{I}(s) = \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} a_i a_j \frac{\sin(sr_{ij})}{sr_{ij}} \qquad (21.2)$$

In this equation, $\boldsymbol{I}(s)$ is the intensity of the scattered electron beam at observation angles s, a_i and a_j are atomic properties such as atomic number, or partial charges, and r_{ij} is the distance between the atoms i and j; n is the number of atoms in the molecule.

In electron diffraction, the intensity is measured and the 3D structure as given by all distances r_{ij} is derived from the intensities on

the basis of Equation (21.2). In our approach, we have turned the equation around, inputting the 3D structure of a molecule in the form of the distances r_{ij} and calculating $\boldsymbol{I}(s)$. Furthermore, these values of $\boldsymbol{I}(s)$ are calculated only at discrete, equidistant values of s, providing a fixed, predefined number of values of $\boldsymbol{I}(s)$ which are then used as an encoding of the 3D structure of a molecule. This molecular representation was called 3D-MoRSE Code (3D Molecule Representation of Structures based on Electron diffraction).

This 3D-MoRSE code was mainly used for the simulation of infrared spectra. However, it has also be demonstrated that this code shows great promise for correlating structure with biological activity. Dopamine D1 agonists could be separated from dopamine D2 agonists on the basis of the 3D-MoRSE code by a Kohonen network.

In more recent work, a structure code quite similar to the 3D-MoRSE code was developed because it could be more easily transformed back into a 3D structure. This novel encoding scheme is based on atom radial distribution functions and is therefore called RDF code. Equation 21.3 gives the basis of this code:

$$\boldsymbol{G}(r) = \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} a_i a_j e^{-b(r-r_{ij})^2} \qquad (21.3)$$

As in Equation 21.2, a_i and a_j are arbitrary atomic properties, n is the number of atoms in a molecule, and r_{ij} is the distance between the atoms i and j. b is a factor, the so-called temperature factor, that determines the accuracy of position of atoms. r is the running variable and is made discrete, to arrive at a fixed length representation of a molecule by $\boldsymbol{G}(r)$.

Use of the RDF code allows the simulation of infrared spectra of similar quality than when using the 3D-MoRSE code. On top of this, it has become possible to transform this structure representation back into 3D space and, thus, develop a method to derive the 3D structure from an infrared spectrum.

21.4 Coding the 3D Structure II

To repeat once more, a good structure representation should satisfy four conditions that will make it universal:

- uniqueness – one and only one code for each compound and different codes for different structures,

- uniformness – each compound should be represented by the same number and type of variables,
- reversibility – the structure should be retrievable back from the code, and
- translational and rotational invariance – for rotated and/or translated structures the code should remain unchanged.

In the present Section we are discussing a method for representing chemical structures which is uniform, unique and reversible, but lacks the translational and rotational invariance. Although in general, the lack of the origin invariance may be regarded as a drawback it will be shown that exactly this property, namely dependence of the representation on the choice of coordinate system, might in special cases offer some advantages.

The representation of a structure (having n atoms) described by n $[x_j,y_j,z_j]$ triplets does not fulfill the condition of uniformity. The most important feature of the described transformation from a 3N-dimensional representation into a unique m-dimensional *spectrum-like* representation is its reversibility. The new representation is based on the projection of constituent atoms onto an imaginary spherical surface large enough for a molecule under consideration to be accommodated within it. The projection of atoms is not made onto the entire sphere, but onto three perpendicular equatorial trajectories of this sphere. For the representation of n triplets $[x_j,y_j,z_j]$ the coordinates z_j, y_i and x_i, of all atoms are set to zero in sequence, hence, defining three "planar molecules" described by the $3n$ two-plets $[x_j,y_j]$, $[x_j,z_j]$, and $[y_j,z_j]$. The obtained three planar molecules are projected onto circles in the (x,y), (x,z), and (y,z) planes, respectively, forming three equivalent sets of the new representation $\boldsymbol{S}$.

Each component of the representation $\boldsymbol{S}$ in one of the planes is defined as a cumulative intensity, s_i, at a given point i on the circle at angle φ_j (Figure 21-3) as a sum of n contributions from each atom j in the molecule.

$$s_i = \sum_{j=1}^{n} \frac{\rho_j}{(\varphi_j - \varphi_i)^2 + \sigma_j^2} \qquad (21.4)$$

for $i = 1...m$

The contribution under the sum associated with each atom j can be any bell-shaped function. In the case of a *spectrum-like*

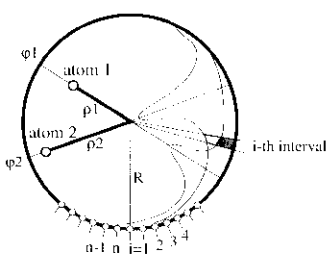

Figure 21-3: Contribution of atoms no. 1 and no. 2 (at positions (ρ_1,φ_1) and (ρ_2,φ_2) to the intensity s_i at the interval (position) i on the circle with radius R, shown as shaded areas of the corresponding Lorentzian bell-shape curves. The circle is divided into m intervals.

representation, the Lorentzian shape was chosen for this purpose. For the explanation of the evaluation of the Lorentzian shapes of the projection of atoms onto the circle, the polar coordinates are more plausible than Cartesian ones. In the actual calculations when the atoms are described by two-plets of Cartesian coordinates, Equation (21.4) is used in the rewritten form with Cartesian coordinates (cf. Equations (21.5) and (21.6)).

Each atom j is represented by one Lorentzian curve with the maximum located at angle φ_j. The intensity at the maximum point ($\varphi_i = \varphi_j$) is proportional to the radii-vector ρ_j. The parameter σ_j describes the width of the bell-shaped curve and can bear any information about the nature of the atom (atomic number, van der Waals radius, charge, etc.). If the geometry and the shape of a molecule is to be described only, the parameter σ_j should be set to one for all atoms.

The dimensionality of the new representation is determined by the number of equidistant points on the circles to which the atoms are projected. In principle, the division of the circle does not depend on the number of atoms n in the molecule. However, because the division of the circle determines the resolution, i.e., the quality of the representation and consequently the quality of the inverse process of recovering the structure, this is not entirely true. The larger the division, the more precise is the description of the atoms in the molecule. In general, the division of the circles should be adapted to the number of atoms n in the *largest* molecule of the study. After the division m is chosen, the new representation should be able to map each molecule, regardless of the number of its constituent atoms, into the same $3m$-dimensional space. In many studies where only approximate positions of the substituents with respect to the skeleton are sought m can be as low as 36 or even m=18. On the other hand, in cases when small differences in space positions of atoms are important, or for precise recoveries of structures back from the *spectrum-like* representations m can be as large as 720 (division by 0.5°), thus making the full *spectrum-like* representation 2160 intensities (variables) long.

For actual evaluations of spectral intensities from Cartesian coordinate two-plets, the following transformations are substituted into Equation (21.4)

$$\rho_j = \sqrt{x_j^2 + y_j^2} \quad \text{and} \quad \cos\varphi_j = \frac{x_j}{\sqrt{x_j^2 + y_j^2}} \tag{21.5}$$

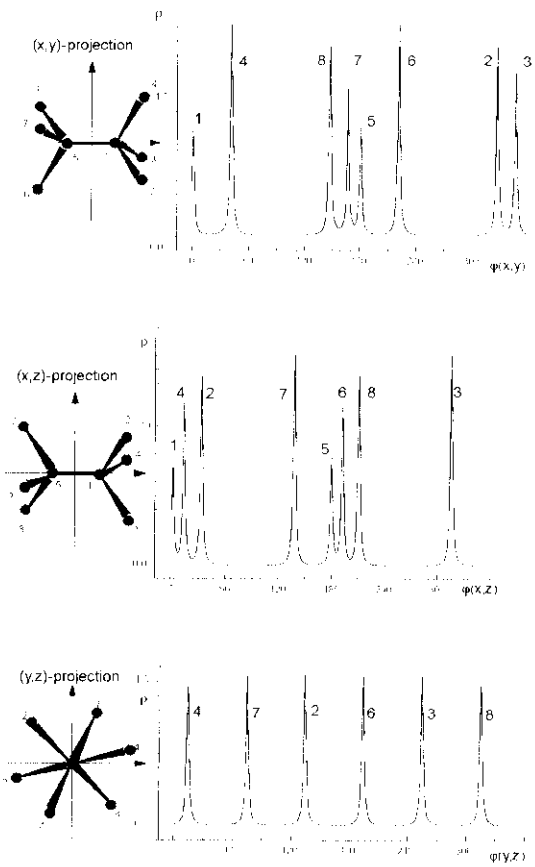

Figure 21-4: Structure of *ethane* in the three perpendicular projections together with the corresponding *spectrum-like* representations in the *(x,y)*, *(x,z)*, and *(y,z)*. The assignments of peaks to the eight *ethane* atoms is shown with figures standing next to the peaks. Atoms at the coordinate origin **do not** produce peaks.

hence:

$$s_i = \sum_{j=1}^{n} \frac{\sqrt{x_j^2 + y_j^2}}{\left[\operatorname{acos}\left(\frac{x_j}{\sqrt{x_j^2 + y_j^2}} \right) - \varphi_i \right]^2 + \sigma_j^2} \qquad (21.6)$$

and equivalent forms for the *(x,z)* and *(y,z)* projections. Figure 21-4 shows how a *spectrum-like* representation of all three projections of the *ethane* molecule is obtained.

The *spectrum-like* representations as described by Equation (21.6) depend on the choice of the coordinate origin. Therefore, such

description of molecular structures can be used only for comparative studies on **sets** of structures that can somehow be aligned to each other either by overlapping the skeletons or some other larger structural parts. The liberty to chose the coordinate origin from which the *spectrum-like* representations for an entire set of structures is calculated offers a possibility to search for a position of the origin from which the representation is most useful for a given task, i.e., the position from which the most relevant structural features could be characterized or "seen" best. In the example given in Section 13.10 the optimal coordinate origin for a set of flavonoids was determined (the point –1.7, 3.9, –0.5 in the benzene ring, relative to the atom 2 on the benzopyran ring system) in a preliminary screening by selecting the broadest variance distribution among all representations calculated from each of the tested origins.

21.5 Coding Molecular Surfaces

Molecules interact with each other at molecular surfaces (Figure 21-5). This is particularly true for the interaction of a ligand binding to its receptor. The investigation of molecular surfaces, the coding of surface properties, is therefore of primary importance.

Again, autocorrelation can be used to encode surface properties. Equation (21.1) is now modified such that the properties, p, are sampled on a molecular surface; for the distance parameter, d, all distances within a certain range, e.g., between 3 and 4 Å are collected in one autocorrelation value (cf. Equation 13.3).

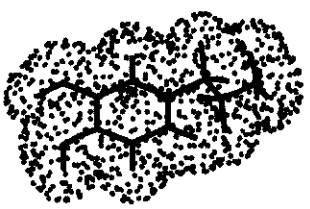

Figure 21-5: The van der Waals surface of the molecule shown in Figure 21-1.

We have shown in Section 13 that autocorrelation of the electrostatic potential on the van der Waals surfaces into 12 descriptors provides excellent descriptors for modeling the affinity of steroids for binding to the corticosteroid binding globulin (CBG) receptor. In a similar manner, autocorrelation of the hydrophobicity potential of a series of 78 polyhalogenated aromatic compounds can quantitatively model the binding to the cytosolic Ah receptor. The same encoding method, autocorrelation of the molecular electrostatic potential (MEP) into 12 descriptors, was used for the definition of diversity and similarity of combinatorial libraries (Section 20).

Furthermore, we have shown in Chapter 19 that a Kohonen network can directly be used to encode a molecular surface and produce maps of surface properties such as the molecular electrostatic potential.

It was shown that such maps of the electrostatic potential on a molecular surface can be used to distinguish between muscarinic and nicotinic agonists. The Kohonen network, in effect, stores the three-dimensional coordinates of points on the molecular surface. Such a network can, therefore, be used as a template for quantifying shape similarities in a series of compounds.

21.6 A Hierarchy of Representations

Approaches to the encoding of molecular structures have been developed that allow the investigation of data sets of diverse molecules by learning methods. These structure representations form a hierarchy of increasing sophistication. At the lowest level, only the constitution of a molecule is taken into account. Next, the 3D structure is considered. For more sophisticated applications, properties of molecular surfaces have to be encoded. The level used will largely be dictated by the size of the data set to be investigated. Representations of the constitution will be applied to data sets comprising millions of structures, whereas representations of molecular surface properties can still be chosen for data sets comprising 100,000 and more structures. Even with large data sets these methods are rapid enough to be performed on small workstations with computation times of a few hours.

21.7 References and Suggested Readings

21-1. G. Moreau and P. Broto, "Autocorrelation of molecular structures: Application to SAR studies", *Nouv. J. Chim.* **4** (1980) 757 – 564.

21-2. J. Gasteiger and M. Marsili, "Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges", *Tetrahedron* **36** (1980) 3219 – 3228.

21-3. J. Gasteiger and H. Saller, "Berechnung der Ladungsverteilung in konjugierten Systemen durch eine Quantifizierung des Mesomeriekonzeptes", *Angew. Chem.* **97** (1985) 699 – 701; J. Gasteiger and H. Saller, "Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept", *Angew. Chem. Int. Ed. Engl.* **24** (1985) 687 – 689.

21-4. M. G. Hutchings and J. Gasteiger, "Residual Electronegativity - An Empirical Quantification of Polar Influences and its Application to the Proton Affinity of Amines", *Tetrahedron Lett.* **24** (1983) 2541 – 2544.

21-5. J. Gasteiger and M. G. Hutchings, "Quantification of Effective Polarisability. Applications to Studies of X-Ray Photoelectron Spectroscopy and Alkylamine Protonation", *J. Chem. Soc. Perkin 2* (1984) 559 – 564.

21-6. For further information on PETRA see: *http://www2.ccc.uni-erlangen.de/software/petra/*

21-7. 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.

21-8. J. Gasteiger, X. Li, C. Rudolph, J. Sadowski and J. Zupan, "Representation of Molecular Electrostatic Potentials by Topological Feature Maps", *J. Am. Chem. Soc.* **116** (1994) 4608 – 4620.

21-9. S. Anzali, J. Gasteiger, U. Holzgrabe, J. Polanski, J. Sadowski, A. Teckentrup and M. Wagener, "The Use of Self-Organizing Neural Networks in Drug Design", in *3D QSAR in Drug Design,* Vol. 2, Eds.: H. Kubinyi, G. Folkers and Y. C. Martin, Kluwer/ ESCOM, Dordrecht, NL, 1998, pp. 273 – 299.

21-10. J. Sadowski and J. Gasteiger, "From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders", *Chem. Reviews* **93** (1993) 2567 – 2581.

21-11. J. Sadowski, J. Gasteiger and G. Klebe, "Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures", *J. Chem. Inf. Comput. Sci.* **34** (1994) 1000 – 1008.

21-12. CORINA can be accessed on the internet: *http://www2.ccc.uni-erlangen.de/software/corina/*

21-13. J. H. Schuur, P. Selzer and J. Gasteiger, "The Coding of the Three-dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure - Spectra Correlations and Studies of Biological Activity", *J. Chem. Inf. Comput. Sci.* **36** (1996) 334 – 344.

21-14. M. Hemmer, V. Steinhauer and J. Gasteiger, "Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra", *Vibrat. Spectroscopy*, **19** (1999) 151-164.

21-15. As review of different coding systems see for example: J. E. Ash, W. A. Warr and P. Willett, *Chemical Structure Systems*, Ellis Horwood, New York, USA, 1991.

21-16. M. A. Johnson and G. M. Maggiora (Eds.), *Concepts of Molecular Similarity*, Wiley Interscience, New York, USA, 1990.

21-17. J. Zupan and M. Novic, "General Type of a Uniform and Reversible Representation of Chemical Structures", *Anal. Chim. Acta* **348** (1997) 409 – 418.