The challenge of molecular structure representation for property prediction

Johann Gasteiger

Résumé **Un défi de l'informatique chimique : la représentation des structures moléculaires pour la prédiction des propriétés**

Les méthodes de recherche des bases de données et des similarités structurales entre molécules sont insuffisantes pour modéliser les caractéristiques de composés chimiques comme l'activité biologique. De plus, elle ne prend pas en compte la nature 3D des molécules. ***La représentation 3D des molécules*** est réalisée à partir de générateurs automatiques, qui intègrent des règles et principes d'élaboration des composés organiques. On peut ainsi décrire des propriétés physico-chimiques de la molécule à partir de sa représentation 3D. Une nouvelle étape est franchie avec la détermination des surfaces de potentiel par la méthode d'autocorrélation du ***potentiel électrostatique moléculaire*** (MEP), qui permet de distinguer les sites récepteurs d'interaction entre molécules biologiques. L'élaboration d'un ***code de chiralité*** permet de prévoir l'énantiomère le plus favorable dans une réaction catalytique énantiosélective. Enfin, la flexibilité moléculaire, liée à la recherche des conformations bioactives, étape ultime dans la connaissance des molécules biologiques, a fait l'objet de nombreuses recherches, mais reste encore un sujet d'étude très ouvert.

Mots-clés **Empreinte digitale, sous-structure, topologie, structure 3D, fonction de distribution radiale, potentiel électrostatique moléculaire, code de chiralité, flexibilité moléculaire.**

Abstract Many methods used for searching in stucture databases and for defining structural similarities between molecules are insufficient for modeling properties of chemical compounds such as their biologic activity. Furthermore this does not take into account the three dimensional nature of molecules. ***A 3D representation of molecules*** can be achieved *via* automatic generators, which incorporate data and rules on the construction principles of organic compounds. One can thus describe molecular physico-chemical properties, starting with its 3D representation. A new stage is reached with the representation of properties on the ***surfaces of molecules*** through the autocorrelation of ***molecular electrostatic potential*** (MEP), which has been used to distinguish different biological receptor binding sites. The elaboration of a ***chirality code***, can now be used to predict the most favorable enantiomer in an enantioselective catalytic reaction. Finally, molecular flexibility, linked to the search for bioactive (biologically active) conformations, which is the ultimate phase in the knowledge of biological molecules, has given rise to various scientific developments, but remains a subject calling for further research.

Keywords **Fingerprints, substructure, topology, 3D structure, radial distribution function (RDF), molecular electrostatic potential (MEP), chirality code, molecular flexibility.**

A key challenge in drug design and indeed in other fields of chemistry is the understanding and modelling of the relationships between the structure of a molecule and its physical, chemical or biological properties. Many of these properties cannot be directly derived by purely theoretical calculations from the molecular structure. An indirect approach is necessary to find a relationship between the two: the structure of a molecule first has to be represented by structure descriptors, which are in turn used to model the property of interest. This second step is achieved by the application of inductive learning methods such as statistical or pattern recognition methods or artificial neural networks to establish a relationship between the structure descriptors and the biological activity.

In this article, we will concentrate on the representation of molecules. Starting in the 60s, Jacques-Émile Dubois pioneered and revolutionized the field by developing the DARC system, based on a representation of molecules into concentric layers around a focus corresponding to the principal function [1-2]. At times, when other groups were using line notations and fragment codes, Prof. Dubois emphasized that in structure representation all atoms and bonds of a molecule have to be explicitly considered. From the very beginning of my work more than 30 years ago, I have followed his visionary concept. Over time, a variety of methods has been developed to derive structure descriptors for a molecule [3-5].

Fingerprints of molecular structures

One approach has been to search molecules for the presence or absence of certain predefined functional groups and other substructures and compress this information into a bit string of given length. Such a representation is called a fingerprint of a molecular structure [6-7]. Initially, fingerprints were developed to enable a quick scan of structure databases to determine the presence or absence of certain chemical

structures. The same purpose is achieved by hashcodes [8], which are constructed very much like fingerprints.

With the advent of combinatorial chemistry and the ensuing need to represent large sets of compounds, fingerprints became a way of representing chemical structures to model their biological activity. They are, however, inadequate in their representation of molecular details.

Fragment codes and substructures

Whereas fingerprints compress the presence or absence of certain substructures or fragments into a concise representation that no longer allows the identification of individual fragments, fragment codes explicitly retain the information about the presence or absence of a certain substructure. A predetermined set of fragments is used and each fragment corresponds to a certain position of a bit string having a length equal to the number of fragments in the predefined set. Fragment codes have been used to model a variety of properties, like predicting biological activities or simulating infrared spectra.

Fragment codes are also often used to define the similarity of structures by calculating the Tanimoto index I, from the number n_A of substructures present in structure A, but not contained in B, the number n_B of substructures present in structure B but absent in A, and the number n_C of substructures in common between structure A and structure B as expressed:

$$I = \frac{n_C}{n_A + n_B + n_C} \qquad (1)$$

On the other hand, fragment codes and lists of fragments only report the presence or absence of certain fragments and provide no information on the arrangement of these fragments in a given molecule. The two structures shown in *figure 1a*, for instance, have a high similarity based on the Tanimoto index, although, clearly, for many types of problems, such as questions on how to synthesize them, the two molecules are topologically quite different.

Another major issue is that while fragment codes are powerful in telling us whether a certain substructure is present in a molecule, they remain mute as to the distance between two substructures (*figure 1b*), a vital piece of information when modelling the biological activity of a compound, since the distance between two atoms will be a critical factor when it comes to determining whether a ligand can bind to one or two sites of the receptor.

Topological distances and atomic properties

Let us therefore, in our structure representation, consider the distance between the atoms of a molecule. In the simplest case, we can use the topological distance, which corresponds to the number of bonds between two atoms. One must consider not only the distances between two atoms, but also the identity of these atoms, in particular their physicochemical properties, such as partial charges or hydrogen bonding potentials.

One approach to simultaneously considering atomic properties and distances between atoms is topological autocorrelation as expressed in eq. 2 [9-10]:

$$A_{top}(d) = \sum_{j=i+1}^{N} \sum_{i=1}^{N-1} a_i a_j \cdot \delta(d - d_{ij}) \qquad (2)$$

In this equation, a_i and a_j are properties of atoms i and j, respectively, and d_{ij} is the topological distance between atoms i and j. δ is the delta-function with a value of 1 when the running variable, the distance d, is equal to the distance d_{ij} between the two atoms, and a value of zero when this is not the case. The summation is made over all combinations of atoms i and j.

As atomic properties, a_i, any property of an atom, such as atomic number or its mass, can be used. To represent the electronic properties of atoms, we have however developed methods to compute such important physicochemical effects as partial charges [11-12], inductive [13], resonance [12] or polarizability effects [14]. These methods are based on simple and rapid algorithms that allow the processing of large sets of molecules comprising hundreds of thousands or even millions of structures.

The benefits of topological autocorrelation of electronic properties of atoms have been shown in studies distinguishing molecules with different biological activities [10], in order to find new lead structures, for lead hopping, and for comparing libraries of compounds. This kind of structure representation by topological autocorrelation is able, as for instance in *figure 2a*, to perceive the similarity of two structures, both being dopamine agonists in this case [10]. For all its success, topological autocorrelation still only considers the constitution of a molecule, its set of atoms and how they are bonded (*figure 2b*).

3D structure representation

Molecules are, however, three-dimensional objects and any in-depth representation of a molecule should take into account its 3D structure and metric. The first step is to gather 3D information on the molecular structure. To date, the 3D structure of about 250 000 organic and organometallic molecules has been determined by X-ray diffraction or NMR

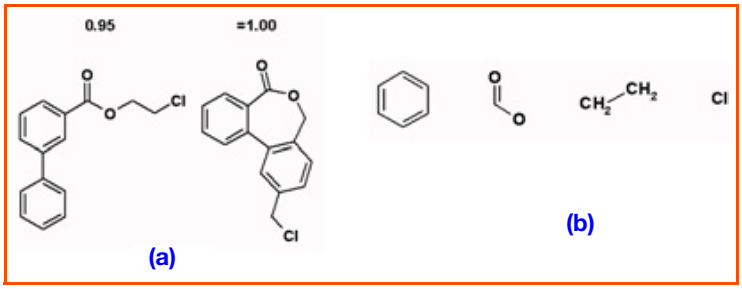

Figure 1 – a) Two structures with a high Tanimoto similarity index; the one on the left is biphenyl-4-carboxylic-acid-2-chloro-ethylester. b) The fragmentation of chemical structures.

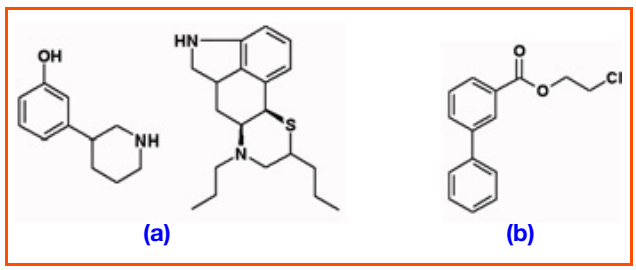

Figure 2 – a) Two dopamine agonists having different number of atoms: 28 atoms for structure on the left and 50 atoms for structure on the right. b) The topology of molecules as expressed by the relative arrangement of the atoms in a molecule; the compound is biphenyl-4-carboxylic-acid-2-chloro-ethylester.

studies and has been stored in the Cambridge Structure Database (CSD). Although this number may seem large, it is in fact almost negligible when one considers that the number of known compounds exceeds 30 million. The question becomes: can we draw enough rules from the known 3D structure of organic compounds to enable us to predict the 3D structure of the remaining 99% of organic compounds? The answer is yes. Several automatic 3D structure generators capable of generating a 3D molecular model from information on the constitution of a single molecule have been developed [15]. Our own group has developed the CORINA 3D structure generator [16]. CORINA incorporates data and rules on the construction principles of organic compounds, which in turn allow the generation of a 3D model for basically any organic molecule [15, 17]. Thanks to this, the publicly available database of the National Cancer Institute, containing 250 251 structures, could automatically be converted into 3D molecular models in a single run requiring 1.1h on a PC (1.6 GHz, Linux) and providing 3D models for 99.4% (248 795) of its structures. CORINA produces a single low energy conformation of a molecule. Comparison with experimental 3D structures from X-ray structure determination has shown the high quality of the 3D structures [18].

With automatic 3D structure generators able to produce 3D molecular models for basically any organic molecule the question becomes: how can the 3D structures be represented for data analysis methods requiring the same number of descriptors, irrespective of the size and number of atoms in a molecule? Clearly, the Cartesian coordinates cannot be used, as the number of descriptors would be directly related to the number N of atoms in a molecule requiring 3N coordinates. A fixed-length representation of the 3D structure can again be obtained by autocorrelation in an analogous manner to that shown by eq. 2 with the distance d_{ij} being binned into ranges.

As an alternative, radial distribution functions (RDF) originating in powder X-ray diffraction or electron diffraction studies for the representation of the 3D structure of molecules can be used as shown in eq. 3 [19]:

$$g(r) = \sum_{j=i+1}^{N} \sum_{i=1}^{N-1} a_i a_j \cdot e^{-b(r-r_{ij})^2} \qquad (3)$$

In eq. 3, the radial distribution function *g(r)* is obtained from the product of the properties a_i and a_j of atoms *i* and *j* and considering the distances r_{ij} between those two atoms. The parameter *b* is the so-called temperature factor, fuzzifying the distances. The value of *r* is a distance and is the running variable of the function.

With their ability to encode the entire 3D structure of a molecule and thus model the vibrations of a whole molecule, both of individual bonds and of the entire skeleton [19-21], RDF codes have successfully been used for the simulation of infrared spectra. Their fairly clear physicochemical interpretation holds out a bright future for them in studies of the impact of the 3D structure on biological activity. An RDF code, for instance, has recently been used to analyze the NF-κB binding affinity of a series of sesquiterpene lactones [22]. Valuable as 3D structure codes are for the representation of molecules in modelling their biological activity, they still only enable us to represent the skeleton of molecules (*figure 3*).

Molecular surface properties

But molecules have both shapes and surfaces and interact with their environment through their surfaces and the

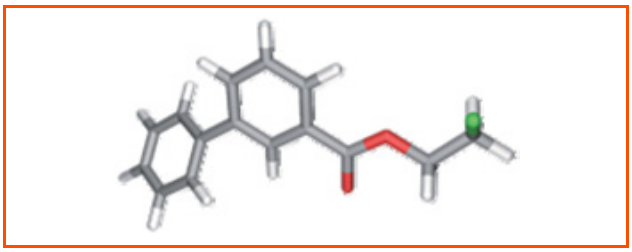

Figure 3 - A molecular 3D skeleton of the previous molecule.

properties on these surfaces. Here again, we are faced with the task of representing the properties on the surfaces of a series of molecules with a fixed-length vector, with the same number of descriptors, irrespective of the size of the molecule. Here again, autocorrelation can be used as expressed:

$$AC_{surf}(d) = \sum_{j=i+1}^{N} \sum_{i=1}^{N-1} p_i p_j \cdot \delta(d_{ij}, d_l, d_u)$$

$$\delta = \begin{cases} 1 \forall d_l \le d_{ij} < d_u \\ 0 \forall d_{ij} \le d_l \vee \ d_{ij} \ge d_u \end{cases} \qquad (4)$$

In this case, properties *p* of points *i* and *j* taken from the molecular surface with a certain sampling density will be used and the distance *d* will be binned between a lower d_l and an upper bound d_u, $d_l \le d \le d_u$.

It has been shown that, through autocorrelation of the molecular electrostatic potential, a representation is obtained that can model the binding affinity of a series of 31 steroids to the corticosteroid binding globuline receptor [23]. Autocorrelation of the molecular electrostatic potential has been used to define the similarity and diversity of combinatorial libraries consisting of amino acids attached to xanthene, cubane, and adamantane scaffolds [24].

In another attempt to represent molecular surface properties, two-dimensional maps of molecular surfaces have been produced by a non-linear mapping procedure utilizing a self-organizing neural network [25]. In this approach, the Cartesian coordinates of points sampled from a molecular surface are used to train a self-organizing (Kohonen) neural network. The mapping of the surface points into the neurons of the network can be visualized by any property these points had on the surface, e.g. the molecular electrostatic potential (MEP). *Figure 4a* shows the MEP on the surface of a molecule. As this is a linear projection, only part of the surface can be shown.

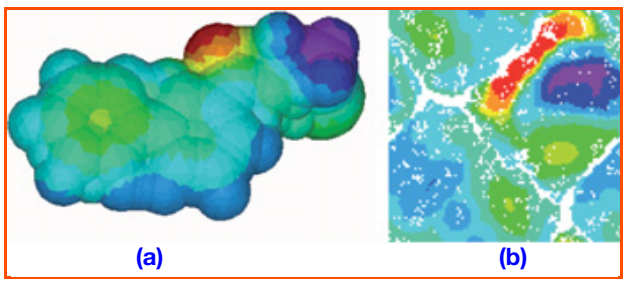

Figure 4 - Representation of the molecular electrostatic potential (MEP) of biphenyl-4-carboxylic-acid-2-chloro-ethylester (blue: positive potentials; red: negative potentials). (a) MEP on the surface of the molecule; (b) 2D self-organizing map of the entire molecular electrostatic potential into a single plane.

Figure 4b, on the other hand, shows the self-organizing map of the entire MEP, as this method is a non-linear projection method able to map the entire molecular surface into a single plane. It has been shown that such maps of the MEP can be used to distinguish compounds that bind to the muscarinic receptor from those that bind to the nicotinic receptor [26].

Chirality codes

All proteins are chiral and therefore many receptors and enzymes respond differently to enantiomers [27]. Correspondingly, about 70% of all drugs are chiral. There is a strong tendency in the pharmaceutical industry to bring pure enantiomers to the market. Any more detailed modelling of the effects of structure on biological activity therefore has to represent chirality. In distance space, enantiomers cannot be distinguished. Thus, enantiomers will obtain the same 3D autocorrelation vectors or RDF codes. We have however developed both a conformation-dependent and a conformation-independent chirality code that is based on the 3D structure of a molecule and considers all the atoms of the ligands around a chiral center or chiral axis [28-29].

Such chirality codes have been shown to successfully predict the major enantiomer in an enantioselective reaction caused by a chiral catalyst [28]. Furthermore, chirality codes were used to predict the first eluted enantiomer in enantioselective chromatography [29]. The path is thus clear for using chirality codes in modelling the biological activity of different enantiomers.

Molecular flexibility and the generation of bio-active conformations

All structure representations mentioned so far have assumed rigid molecules, whereas most molecules are quite flexible, having single bonds that allow rotation yielding different torsional angles, and thus provide different conformations. The quest for the biologically active conformation therefore becomes key. Lack of knowledge about the biologically active conformation is also the reason why, in quite a few situations, topological or 2D descriptors outperform 3D descriptors in modelling biological activity. Clearly, molecules are three-dimensional and thus 3D descriptors should perform better than their 2D counterparts. However, as soon as the 3D structure of a molecule is considered, the problem of finding the right conformation becomes imminent.

The generation of conformations is fairly easy. Even the generation of low energy conformations is not that difficult. However, because of the large number of potential settings for torsional angles, one might soon find oneself with too many conformations to be handled. The challenge then becomes how to avoid generating too many conformations while still maintaining the biologically active conformation. Two approaches are conceivable: a constrained generation of conformations, in order to generate fewer conformations, or a direct search for the biologically active conformation. Attempts along both lines will be presented.

An analysis of the distribution of torsional angles around single bonds in X-ray structures showed clear preferences and provided a statistical distribution of the incidences of torsional angles [30]. Such distributions are taken by the program ROTATE to preferentially generate those conformations that have a high incidence in the Cambridge Crystallographic Structure Database (CSD). In addition, conformations with small deviations in torsional angles are collected into families and each family is represented by a single conformation. This allows the generation of a limited, but quite diverse set of conformations [31-32]. These sets of conformations also contain a conformation that is quite close to the receptor-bound, biologically active conformation.

The attempt to get direct access to the bioactive conformation rests on the idea that a set of ligands binding to the same receptor must have common spatial features. Thus, a search for the maximum common three-dimensional substructure (3D-MCSS) of a set of ligands is initiated by superimposing the 3D molecular models of these ligands to maximize the number of atoms of the different ligands that can be superimposed. In this process, rotations around single bonds of the ligands are allowed, thus introducing conformational flexibility. In order to manage this optimization problem, a genetic algorithm like a stochastic optimizer is used [33]. *Figure 5* shows the superimposition of three nicotinic allosterically potentiating ligands emphasizing their 3D structural similarity.

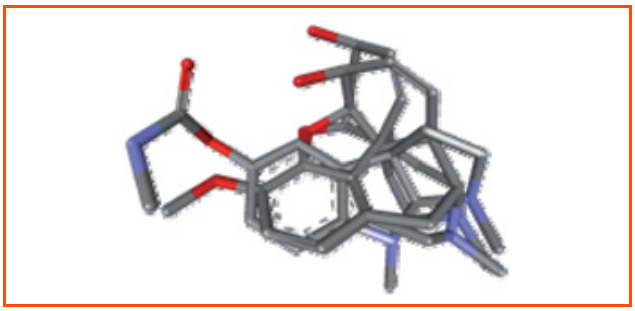

Figure 5 - Superimposition by GAMMA of the three nicotinic allosterically potentiating ligands galanthamine, codeine and physostigmine, emphasizing their 3D structural similarity.

Summary and conclusions

In this article, we mainly focused on the geometric aspects of structure representation. The proper consideration of physicochemical effects exerted by the atoms in a molecule is however of equal importance [4, 34]. The equations presented here allow their transparent incorporation into the various structure coding methods, from the constitution through the 3D structure to molecular surface properties (*figure 6*). These methods, combining molecular geometry of increasing resolutions with physicochemical properties, have been integrated into the package ADRIANA.Code (Automated Drug Research by Interactive Application of Non-linear Algorithm) [35].

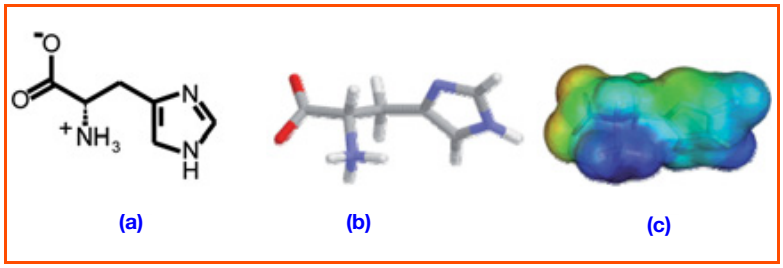

Figure 6 - A hierarchy of structure representation: a) 2D model; b) 3D model; c) molecular surface.

Although we have mostly focused on drug design issues, the methods for the representation of molecular structures can be used in all areas of chemistry. In fact, given the need to predict a wide range of physical, chemical or biological properties of compounds, the use of structure coding methods in many fields of chemistry can only increase.

Despite advances in the area of molecular structure representation, particularly in the 3D arena, since the pioneering work of Jacques-Émile Dubois in the sixties-eighties, there is still significant room for progress. The quest for the biologically active conformation remains a challenge calling for new ideas and approaches. It is our belief that the development of new structure representations should rest both on clearly defined levels of resolution of the geometry of molecules and on considerations of a variety of physicochemical effects.

Acknowledgements

Our work in this area has benefited from the input of a number of able and dedicated co-workers mentioned in the references. The German Ministry of Science and Technology (BMFT and BMBF), the German Research Council (DFG) and several pharmaceutical companies, Pfizer in particular, have actively supported our research. Last but not least, I want to acknowledge that Jacques-Émile Dubois was always a source of inspiration and I was lucky enough to enjoy his friendship. We have lost a great scientist, a wonderful person and friend.

References

[1] Dubois J.-E., *The History and Heritage of Scientific, Technological Information Systems, Proceedings of the **2002** Conference*, W.B. Rayward and M.E. Bowden (eds), p.149. See www.chemheritage.org/pubs/asist2002/proceedings.html and www.chemheritage.org/pubs/asist2002/13-dubois.pdf
[2] a) Dubois J.-E., Hennequin F., Chastrette M., *Bull. Soc. Chim. France*, **1966**, *11*, p. 3568 ; b) Dubois J.-E., Hennequin F., *Bull. Soc. Chim. Fr.*, **1966**, *11*, p. 3572 ; c) Dubois J.-E., Schutz G., Normant J.-M., *Bull. Soc. Chim. France*, **1966**, *11*, p. 3578.
[3] Todeschini R., Consonni V., *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2001**.
[4] Gasteiger J., *Handbook of Chemoinformatics. From Data to Knowledge*, J. Gasteiger (ed), Wiley-VCH, Weinheim, Germany, **2003**, p. 1034.
[5] *Chemoinformatics – A Textbook*, J. Gasteiger , T. Engel (eds), Wiley-VCH, Weinheim, **2003**.
[6] Attias R., *J. Chem. Inform. and Comp. Sci.*, **1983**, *23*, p. 102.
[7] Willett P., Barnard J.M., Downs G.M., *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, p. 983.
[8] Ihlenfeldt W.-D., Gasteiger J., *J. Comput. Chem.*, **1994**, *15*, p. 793.
[9] Moreau G., Broto P., *Nouv. J. Chim.*, **1980**, *4*, p. 359.
[10] Bauknecht H., Zell A., Bayer H., Levi P., Wagener M., Sadowski J., Gasteiger J., *J. Chem Inf. Comput. Sci.*, **1996**, *36*, p. 1205.
[11] Gasteiger J., Marsili M., *Tetrahedron*, **1980**, *36*, p. 3219.
[12] Gasteiger J., Saller H., *Angew. Chem.*, **1985**, *97*, p. 699 and *Angew. Chem. Int. Ed. Engl.*, **1985**, *24*, p. 687.
[13] Hutchings M.G., Gasteiger J., *Tetrahedron Lett.*, **1983**, *24*, p. 2541.
[14] Gasteiger J., Hutchings M.G., *J. Chem. Soc. Perkin*, **1984**, *2*, p. 559.
[15] Sadowski J., Gasteiger J., *Chem. Reviews*, **1993**, *93*, p. 2567.
[16] Gasteiger J., Rudolph C., Sadowski J., *Tetrahedron Comput. Method*, **1992**, *3*, p. 537.
[17] CORINA can be used at www2.chemie.uni-erlangen.de/software/corina and is available from Molecular Networks GmbH, Erlangen, Germany (www.mol-net.com).
[18] Sadowski J., Gasteiger J., Klebe G., *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, p. 100.
[19] Hemmer M.C., Steinhauer V., Gasteiger J., *Vibrat. Spectroscopy*, **1999**, *19*, p. 151.
[20] Schuur J.H., Selzer P., Gasteiger J., *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, p. 334.
[21] Selzer P., Gasteiger J., Thomas H., Salzer R., *Chemistry - A European Journal*, **2000**, *6*, p. 920.
[22] Wagner S., Hofmann A., Siedle B., Terfloth L., Merfort I., Gasteiger J., *J. Med. Chem.*, **2006**, *49*, p. 2241.
[23] Wagener M., Sadowski J., Gasteiger J., *J. Am. Chem. Soc.*, **1995**, *117*, p. 7769.
[24] Sadowski J., Wagener M., Gasteiger J., *Angew. Chem.*, **1995**, *107*, p. 2892 and *Angew. Chem. Int. Ed. Engl.*, **1995**, *34*, p. 2674.
[25] Gasteiger J., Li X., Rudolph C., Sadowski J., Zupan J., *J. Am. Chem. Soc.*, **1994**, *116*, p. 4608.
[26] Gasteiger J., Li X., *Angew. Chem.*, **1994**, *106*, p. 671 and *Angew. Chem. Int. Ed. Eng.*, **1994**, *33*, p. 643.
[27] Dubois J.-E., *The Chemical Applications of Graph Theory*, A.T. Balaban (ed), Acad. Press., **1976**, p. 336.
[28] Aires de Sousa J., Gasteiger J., *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, p. 369.
[29] Aires de Sousa J., Gasteiger J., *J. Mol. Graphics Model.*, **2002**, *20*, p. 373.
[30] Klebe G., Mietzner T., *J. Comput.-Aided Mol. Des.*, **1994**, *8*, p. 583.
[31] Schwab C.H., *Handbook of Chemoinformatics*, J. Gasteiger (ed), Wiley-VCH, Weinheim, **2003**, p. 262.
[32] Renner S., Schwab C.H., Gasteiger J., Schneider G., *J. Chem. Inf. Model.*, **2006**, *46*, p. 2324.
[33] Handschuh S., Wagener M., Gasteiger J., *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, p. 200.
[34] Gasteiger J., *Mini Rev. Med. Chem.*, **2003**, *3*, p. 789.
[35] ADRIANA.Code is available from Molecular Networks GmbH, Erlangen, Germany at www.mol-net.com

Johann Gasteiger
is professor at the Institut für Organische Chemie, Erlangen University (Germany)*.

* Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, 91052 Erlangen (Germany).
E-mail: gasteiger@chemie.uni-erlangen.de