descriptors used in QSAR programs such as CODESSA.[94] These include those that can be observed experimentally, such as dipole moments, and those that cannot, such as partial atomic charges. Clark and coworkers have recently used AM1-based descriptors to distinguish between drugs and nondrugs and to understand the relationship between descriptors and their physical properties.[95]

Most descriptors are calculated at the semiempirical level of theory using programs such as AMPAC or MOPAC. However, with computer speed increasing steadily the use of ab initio and DFT methods are becoming increasingly common. These methods allow the descriptors to be calculated from first principals. Yang and coworkers examined various DFT-based descriptors to generate models for a series of protoporphyrinogen oxidase inhibitors. It was shown that the DFT-based model outperformed the PM3-based model.[96]

FIELD-BASED METHODS: CoMFA

Comparative molecular field analysis (CoMFA)[97] and CoMSIA (comparative molecular similarity indices analysis)[98] are field-based or grid-based methods where all the compounds in the data set are aligned on top of one another and steric and electrostatic descriptors are calculated at each grid point using a probe atom. As a result there are many more descriptors than molecules; therefore, a PLS data analysis is used to generate linear equations. A study by Weaver and coworkers compares different field-based methods for QSAR, including CoMFA and CoMSIA, finding that field-based methods provide a robust tool to aid medicinal chemists.[99] Absent from the traditional MFA approaches are quantum mechanically derived descriptors of electronic structure. QMQSAR is a relatively new technique where semiempirical QM methods are used to develop quantum molecular field-based QSAR models.[100] Placing the aligned training set ligands into a finely spaced grid produces quantum molecular fields, where each ligand is characterized by a set of probe interaction energy (PIE) values. A PIE is defined as the "electrostatic potential energy obtained by placing a positively charged carbon 2s electron at a given grid point and summing the attractive and repulsive potentials experienced by that electron as it interacts with the field of the ligand L":

$$\mathrm{PIE} = -\langle s_i s_i \,|\, V(L) \rangle = \int_{r_1} \chi^*_{s_i}(r_1)\chi_{s_i}(r_2) \times \left\{ \sum_{\alpha=1}^{N_{\mathrm{atoms}}} \left[\frac{z_\alpha}{|r_1 - r_\alpha|} - \sum_{\mu\in\alpha}\sum_{\mu'\in\alpha} P_{\mu\mu'} \int_{r_2} \frac{\chi^*_\mu(r_2)\chi_{\mu'}(r_1)}{|r_1 - r_2|} dr_2 \right] \right\} dr_1.$$

The nuclear charge z_α is simply the number of valence electrons on atom α and the notation $\mu \in \alpha$ indicates the set of valence atomic orbitals centered on atom α. Density matrix elements $P_{\mu\mu'}$ are given by the following sum over the occupied MOs:

$$P_{\mu\mu'} = 2\sum_{k=1}^{N_{\mathrm{occ}}} c_{\mu k} c_{\mu' k}.$$

When applied to data sets containing corticosteroids, endothelin antagonists, and serotonin antagonists, linear regression models were produced with similar predictability compared to various CoMFA models.

SPECTROSCOPIC 3-D QSAR

The spectroscopic QSAR methods include EVA (vibrational frequencies),[101] EEVA (MO energies),[102] and CoSA (NMR chemical shifts).[103] It is a requirement of 3D QSAR that all compounds that are being studied contain the same number of descriptors. However, none of the above techniques provides this necessarily. The number of vibrational frequencies is dependent on the number of atoms, N, in a molecule (3N-6 or 3N-5 if linear). The number of NMR chemical shifts depends on N while the number MOs also depends on basis set size. A solution of this problem is to force the information onto a bound scale using a Gaussian smoothing technique, where the upper and lower limits of this scale are consistent for all compounds in the data set. A Gaussian kernel with a standard deviation of σ is placed over each calculated point, EVA, EEVA, or NMR chemical shift. Summing the amplitudes of the overlaid Gaussian functions at intervals x along the defined range results in the descriptors for each molecule, $f(x)$:

$$f(x) = \sum_{i=1}^{3N-6} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-f_i)^2/2\sigma^2}.$$

These descriptors contain a wealth of structural information when we consider the physical basis of the methods. Infrared spectroscopy provides information concerning the arrangement of molecular functional groups and NMR chemical shifts are highly dependent on substituents effects in a congeneric series of compounds. However, MO energies give the electronic structure of the molecule such as the HOMO/LUMO energies that play an important role in the binding process.

The choice of theory used to calculate these descriptors depends on the number of compounds in the data set and the accuracy that is required; all can be calculated using semiempirical or ab initio methods. The QSAR results also depend on the choice σ and x in the above equation.

These methods have provided predictive models for a number of data sets and have an advantage over the field-based methods because they are "alignment-free"; in other words there is no need to superimpose the structures in the data set. Asikainen and coworkers provided a comparison of these methods in a recent article where they studied estrogenic activity of a series of compounds.[104]