INTERNALLY DEFINED DISTANCES: NOVEL DESCRIPTORS FOR 3D-QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

Christian Th. Klein
Institut für Pharmazeutische Chemie
UZA II, Althanstrasse 14
1090 Vienna, Austria

With grid based methods a new generation of molecular descriptors was introduced in 3D quantitative structure-activity relationships (3D-QSAR), with the possibility to create virtual receptor sites (VRSs), thus giving insight into the interaction of active molecules with their biological targets, even if the spatial structure of the targets is unknown. The common philosophy of these methods is to place the superimposed molecules into a grid, and to calculate energy contributions – mostly steric and electrostatic – at each grid point. The contributions can be interaction energies between a probe, placed at the grid point, and the considered molecule(s), as in CoMFA (Comparative Molecular Field Analysis) or GRID. SOMFA (Self Organizing Molecular Field Analysis) uses a binary description of steric contributions. The electrostatic contribution is the molecular electrostatic potential (MEP) at a grid point, rather then a interaction energy with a probe. Moreover, the concept of mean centered activity is introduced, a sort of variable weighting. In COMPASS the interaction points are placed at 2 Å from the common molecular surface of the molecules.
A major problem of these methods is the high sensitivity to the molecular alignment of the compounds of interest. Methods which use the concept of similarity are less sensitive to superposition. CoMSiA (Comparative Molecular Similarity Analysis) uses Gaussian functions to evaluate the similarity at a grid point to a given probe, the resulting similarity fields being smoother, i.e. less affected by superposition, than the interaction fields. Similarity matrix methods also give good results: n x n similarity indices between all pairs of the n compounds considered are evaluated. To overcome the problem of molecular superposition CoMMA (Comparative Molecular Moment Analysis) calculates descriptors based on 3D structures without reference to a common orientation frame. Descriptors are the moments of inertia (shape), magnitude of dipole and principal quadrupole moment (electrostatics), and additional parameters which relate shape and charges.

In the following we present a conceptually simple but efficient method to generate 3D molecular descriptors. Alignment must not be performed explicitly, since an internal coordinate system is defined relative to molecular features, e.g. relative to positions of atoms, relative to centers of mass of certain substructures, or to the principal axes of inertia. Considering for example three points, **a**, **b**, **c**, defined by atoms of a molecule, an internal coordinate system (**u**, **v**, **w**) with the origin **a** can be defined using **a**, **b** and **c** in the following way:

$$\mathbf{u} = \mathbf{b} - \mathbf{a}, \quad \mathbf{w} = \mathbf{u} \times \mathbf{h}, \quad \mathbf{v} = \mathbf{u} \times \mathbf{w} \quad \text{with} \quad \mathbf{h} = \mathbf{c} - \mathbf{a} \tag{1}$$

From the origin of this system distances to the surface of the molecules at defined spherical coordinate angles, θ and φ, respectively are calculated: If $\mathbf{t} = \mathbf{t}(\theta, \varphi)$ is the unit vector in spherical coordinates defined in the new coordinate system, the distance to the van der Waals surface of the molecule is given by:

$$r_{\theta\varphi} = \mathbf{t}'\mathbf{k}_i + \sqrt{\mathbf{t}'\mathbf{k}_i - \mathbf{k}_i'\mathbf{k}_i + R_i} \qquad (2)$$

where $\mathbf{k}_i$ are the centers of the atoms and R_i are the van der Waals radii. " ′ " denotes the respective transposed vectors.
The points on the surface are associated with the corresponding electrostatic potential, $V_{\theta\varphi}$.
The $\{r_{\theta\varphi}, V_{\theta\varphi}\}$ matrix ($\theta \in [0, 180]$, $\varphi \in [0, 360)$) generated by varying θ and φ by fixed increments is translationally and rotationally invariant, since the (**u**,**v**,**w**) coordinate system is internally defined.
The descriptor matrix is correlated with biological activities by PLS (partial least squares). The approach – which will be called IDA (internal distances analysis) – turns out to be highly predictive. The descriptors are easy to interpret and the obtained models can be visualized as in the case of grid-based approaches.

The Table below shows results obtained with the benchmark steroid set, first analyzed by Cramer et al. The steroids were obtained from the Gasteiger group homepage. In this set bugs from previous versions are fixed. The structures were generated by the CORINA software and are consequently completely unaligned. To obtain charges for the electrostatic potential calculation, AM1 minimization was performed. Atoms 7, 8 and 9 (in the IUPAC numeration) were used for the definition of the internal coordinate system.
The benchmark set consists of 31 steroids with measured affinities to the corticosteroid binding globulin (CBG). As usual, the first 21 compounds were used as training set. For the remaining steroids the standard deviation of errors of prediction (SDEP) was calculated, which is a measure of the external predictivity of the obtained models.
The IDA approach, here illustrated with 472 and 1738 descriptors, respectively, is compared to other highly predictive 3D-QSAR methods. The IDA predictions are better than those performed with the other methods. In particular steroid 31 – fluorinated at the 9-α-position – which appears as a strong outlier with the other methods, is predicted more accurate with IDA.

	CoMFA	**Similarity matrix analysis**	**COMPASS**	**MS-WHIM**	**SOMFA**	**IDA (472)**	**IDA (1738)**
SDEP	0.716	0.640	0.705	0.662	0.584	0.391	0.487