Overview

- GRIND - GRid-INDependent Descriptors
  - Idea
  - Construction
  - Results
- How good is QSAR really?

Problems in 3D QSAR

- Core problem in 3D QSAR studies is the correct alignment of the correct conformations
  - Which conformers should be aligned?
  - Same problems as for structure comparison
  - The larger the number of degrees of freedom in the molecule, the more difficult it gets to align the molecules properly

Bultnick et al., p. 589
Problems in 3D QSAR

- Core problem in 3D QSAR studies is the correct alignment of the correct conformations
  - Meaningful alignments can only be constructed if there is structural knowledge of the binding mode
  - It is often not clear, which property is most relevant for the superposition (hydrophobic interactions? electrostatics?)
  - Wrong alignments lead to wrong energies in the grids and thus to incorrect QSAR models

GRIND/ALMOND

- GRID-INDependent descriptors for 3D QSAR aim at alleviating the alignment problem

  - Idea
    - Compute virtual description of the receptor in the form of interaction points
    - Encode the geometry of the interaction points in a translation- and rotation-invariant representation
    - Use this representation as descriptors for PLS
  - GRIND is implemented in the (commercial) software package ALMOND

VRS - Virtual Receptor Sites

- GRIND uses the GRID potentials of six different probe groups to build a virtual representation of the receptor
- From this grid-based representation a sparse representation based on interaction points is computed
- Filter algorithms thin out this point set to about 100 points per probe group
MIFs

- Compute MIFs - Molecular Interaction Fields, i.e. GRID potentials for various probe groups around the ligand
- In CoMFA: grid entries are directly used as matrix entries -> alignment issues
- Instead: reduce grid values to interaction points (similar to what LUDI does, see lecture Drug Design 1)
- Filtering reduces number of interaction points further

VRS - Virtual Receptor Sites

- Areas in the MIFs with significant negative energies are interesting
- We search for a sparse representation of these areas
- Points in these regions can be considered places, where the receptor might have interaction sites
- They are thus called virtual receptor sites (VRSs)

In contrast to the position of a VRS, the distance between pairs of VRSs is rotation- and translation-invariant.

- We use a representation based on distances alone
- Compute distribution of distances between certain VRSs
- In analogy to a radial distribution function
- Disadvantages
  - Loss of Information
  - Not invertible
Recap: 3D Descriptors - RDF

- The distribution of all intramolecular distances is captured by the radial distribution function (RDF).
- Because the distances are typically discrete, they are turned into a density function:
  \[ g(r) = \sum_{i<j} e^{-B(r-r_{ij})^2} \]
  \( B \) is a constant describing the width of the Gaussian contribution to the density function.
- RDF is a one-dimensional description of the three-dimensional geometry of a molecule.

Recap: 3D Descriptors - RDF

- Isomers and even conformers (which, by definition, yield identical 2D descriptors) can be distinguished using the RDF.
- Disadvantage: RDF is a vector, not a single number.
- Similar to RigFit, physicochemical properties can also be integrated into RDFs:
  \[ g(r) = \sum_{i<j} p_i p_j e^{-B(r-r_{ij})^2} \]
  where \( p_i \) and \( p_j \) are the respective values for the property of atoms \( i \) and \( j \).
GRIND - Correlations

- GRIND applies the RDF idea to interaction points
- Instead of correlating atom coordinates, we use the interaction points
- Correlations can be computed just like RDFs
- Where RDFs multiply properties, GRIND multiplies the respective interaction energies

Auto- and Cross-Correlations

- ALMOND typically uses six different correlograms that are concatenated
- This results in a vector of scalar values used to encode the molecule
- Similar interaction options in similar distances will result in similar correlograms
Results
- Set of 10 superimposed correlograms (N1-N1) of glucose analogs
- The compounds are inhibitors of glycogen phosphorylase
- Points are color-coded according to biological activity (active = red, intermediate = white, inactive = green)
- Interaction energies correlate with activity!

Results
- Several cpds with serotoninergic (5-HT2A) activity
- GRIND model based on PLS
- Model:
  - 3 LV
  - $R^2 = 0.93$
  - $Q^2 = 0.81$
  - Average error: 0.35

Benchmarking Descriptors
- It is often very hard to decide, what type of descriptors should be chosen
- There are also very few studies systematically comparing different descriptor sets on large independent datasets
- One study was published in 2006 by Gedeck et al.
- Datasets
  - 944 in-house datasets (Novartis)
  - 570,000 data points in total
  - 143,000 different compounds
**Benchmarking Descriptors**

- **Descriptors**
  - GRIND
  - AlogP atom counts
  - Avalon fingerprints (hashed, graph-based fingerprints)
  - Pipeline Pilot fragment counts (hashed structure-based fingerprints)
  - Hologram QSAR (graph based fragment counts)
  - MDL public keys (binary structural fingerprint, 166 bit)
  - Similog descriptors (pharmacophoric triplet counts)

**Comparing Descriptors**

- **Full datasets**

**Results**

- **Graph**
Caveats

• So far no consistent single set of descriptors has been found
• Most of these studies yield contradictory results
• Descriptor choice also varies with the target/scaffold: descriptors good on one target might be poor on another
• Also, careful data preparation plays a key role:
  Gedeck et al. write: “Interestingly, the computationally most involved model based on the Almond descriptors, which uses the three-dimensional structure of the compound for the prediction, performed poorest. This might be due to the fact that the description of the three-dimensional structure of compounds is difficult, in general [...], and that the protocol to generate the structures in the present study was inappropriate, in particular.”

References

Books
• Bultnick et al. (Hrsg.), Computational Medicinal Chemistry for Drug Discovery, Marcel Dekker Inc., New York, 2004

Papers