Intelligent Pharma

Your Computational Partner In Drug Discovery

Flexible QSAR: functional machine learning in computational chemistry

Ignasi Belda, PhD
CEO

24th September 2010
ECML PKDD 2010
Barcelona

Intelligent Pharma

Intelligent Pharma is a biotechnology company dedicated to developing, commercializing and using new computational technologies for drug discovery carrying out fully integrated research projects.

Business lines

Drug Discovery

We develop technologies based on computational chemistry to give the best in-silico services in drug discovery .

ICT for Life Science

We offer advanced software solutions for companies and institutions in the life science field such as pharmaceutical, biotechnology, nutritional or cosmetic companies.

Intelligent Pharma

Vision: to be the European leaders in computational chemistry in the next three years.

Mission: to be your partner in drug discovery, especially during the early stages of your projects by helping you in being more efficient and competitive, reducing time-to-market, risks and costs of your research.

SERVICES

- Finding non-structural mimetic compounds:
  - To identify new active compounds.
  - To identify more synthesizable or scalable compounds.
  - To identify "back-ups" in early drug discovery stages.
- Drug Reprofiling.
- Finding mechanism of action.
- To extend patent protection.
- Determination of inhibitors.
- To optimize the activity finding an analog.
- ADME/Tox prediction.
- Computational chemistry projects …

Research and Innovation

PREMIO EMPRENDEDOR XXI IV EDICIÓN

"We perform research in computational chemistry and computer science to supply the best *cutting edge* technologies and innovations to our customers"

- Artificial intelligence procedures
- Virtual screening technologies
- QSAR methodologies
- Distributed computing technologies

Drug Design & Drug Discovery

Challenging engineering problem where:

- Few thinks are known about the system and its rules.
- Constrains are usual, diffuse and of different nature.
- Noise is always present as well as subjectivity.
- Several challenging tasks must be performed, including multidimensional optimization and prediction.

- The full process costs around a $Billion (with **B** of Barcelona).

Hit-to-Lead Stage

A Hit is a molecule that works (<$2M).

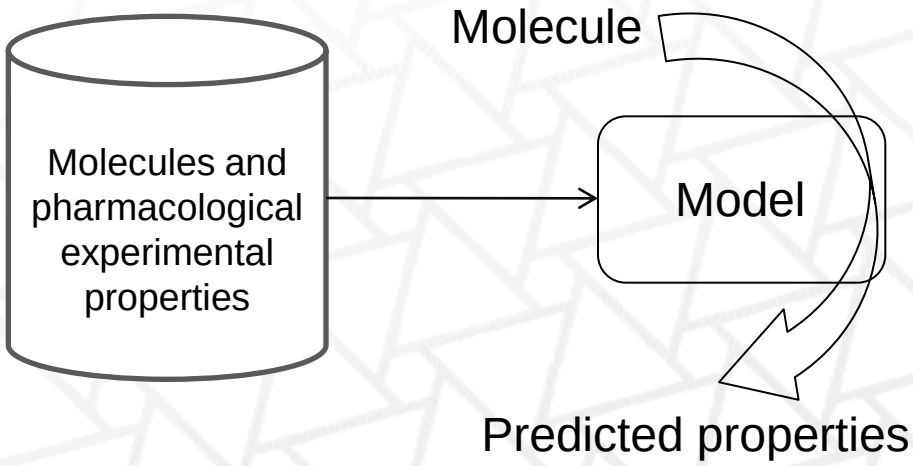
Now you need a molecule that can be absorbed, presents low toxicity, and can be easily eliminated from the body (~$5M-10M).

Before optimizing the hit to get a lead, you need to find the **best hit** (~$2M-5M).

Hit optimization

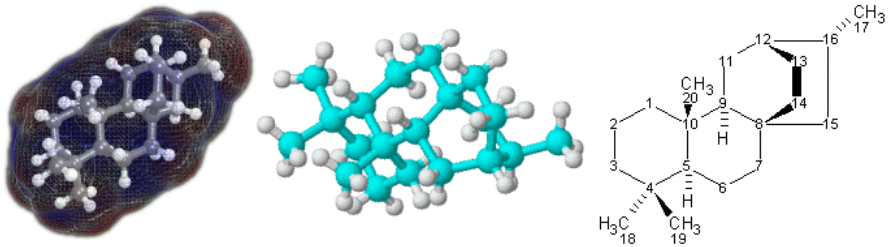
One of the key points in drug discovery…

…where, usually, **Machine Learning** is used.



QSAR

Quantitative Structure-Activity Relationship



Properties:

- Physico-chemical properties, ej. molecular weight, total charge, etc.
- pseudo-3D structural properties, ej. Number of rings, presence of certain chemical groups, etc.

QSAR prediction rates

All typical statistical and machine learning tools are normally used at QSAR. However …

… the accuracy in test sets is seldom higher than 70%.

And this is the main reason why *drug designers* do not rely too much in QSAR.

QSAR and the accuracy

Proper and accurate QSAR tools could save millions to the pharmaceutical industry:

- Not only in the speed-up of the hit-to-lead optimization stage,
- but also in the prediction of future issues.

For such a reason, at Intelligent Pharma , we carry out intensive research in this field.

Our working hypothesis

The attributes (physico-chemical properties) used to describe the instances (molecules) in QSAR do not reflect the dynamic nature of a chemical entity.

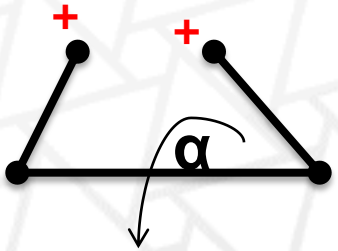
Molecules are flexible and dynamic objects and that provokes variations in the main physico-chemical properties.

Therefore, QSAR uses non-descriptive/non-accurate attributes to describe the molecules.

Functional data analysis

Our aim would be to use attributes that change with the same "dynamicity" that the molecule that they describe.

Molecules are flexible and dynamic BUT do not prefer all conformations in the same way.



$$\Pr ob(\alpha = 0^{o}) \approx 0$$

$$\Pr ob(\alpha = 180^{o}) > 0.9$$

First challenge

To sample several hundred thousands of conformations for each molecule, measure their **internal energies**, and their variable physico-chemical properties.

The measured properties (functional descriptors) are:

- Van der Waals Volume.
- DASA: Water accessible surface area of atoms charged positively.
- DCASA: Absolute difference in charge-weighted area.
- Module of the Dipolar Moment.
- GLOB: molecular globularity.

The constant descriptors

Classical QSAR molecular descriptors:

- Number of rotatable bonds.
- Number of rings.
- Number of hydrogen bond acceptors.
- Number of hydrgoen bond donors.
- Topological Surface Area.
- logP Coefficient= log(octanol(mol/vol)/water(mol/vol))).

Density estimation

The functional part is the density of functional descriptors.

Assumptions

- We have a molecule with *n* rotable angles α=(α1,..,αn). Every α is called a conformation, and each conformation defines a microstate.
- We have a sample of a functions that are the energy and the functional descriptors, that give a real value for each microstate. The range of values is different for each function.
- The Probability Density Function of Microstates is the Boltzmann Distribution:

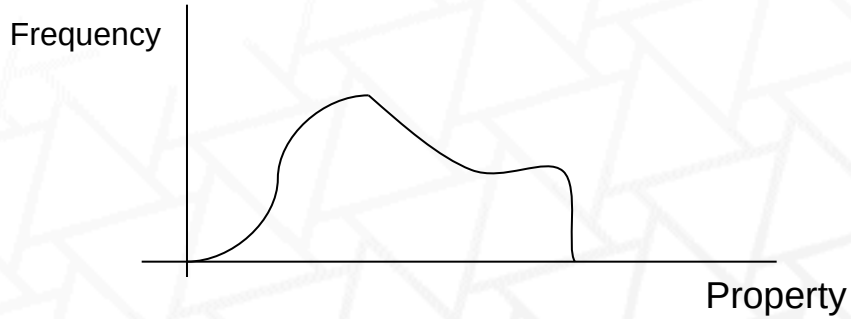
$$f(\alpha) \propto r^{-E(\alpha)/(RT)}$$

- We need Bayes rule to calculate Density of Descriptors:

$$f(y) = \int f(ycondx)\, f(x) dx$$

Functional data post-processing

After some mathematical manipulation we obtain:



- Gaussian smoothing
- Noise removal
- Range normalization

Second challenge

To project a finite number of samples represented through an infinite function space into a set of orthogonal basis functions.

- Principal Components:
  - Discrete data (conventional PCA)
  - Multivariant (b-splines basis function)
  - Minimization of likelihood with EM algorithm
- Chebyshev basis ($T_n(t) = \cos(n \cdot ar\cos t)$ $c_i = \langle f, b_i \rangle \langle b_i, b_i \rangle$)
- Wavelets:
  - Mexican wavelet
  - Haar wavelet
  - Derivative gaussian

Machine learning

10-fold cross validation was used in all experiments.

Support Vector Machines:

- Polynomial kernel
- Radial basis kernel
- Gaussian radial basis kernel
- ANOVA polynomial kernel

Experimentation and Results

Comparison among:

- Constant descriptors.
- Constant descriptors + Mean of Functional descriptors
- Constant descriptors + Mean Functional of descriptors + Functional descritptors

In two different scenarios.

Conclusions

1. The ANOVA polynomial kernel is the best performing kernel for the flexible QSAR case.

2. Experiments where functional information was taken into account were significatively better than experiments where this information was not used.

3. The smallest prediction error is 0,16 using all the functional information *vs* 0,21 using only the mean of the functional descriptors.

CONTACT

Parc Científic de Barcelona (España)
C/ Baldiri Reixac, 10
08028 Barcelona
T: +34 934 034 551
Fax: +34 934 034 551

Technologie Park Heidelberg (Alemania)
Im Neuenheimer Feld 582
69120 Heidelberg
Germany
T: +49 (0) 6221 5025716

Dr. Ignasi Belda – CEO — ibelda@intelligentpharma.com
Dr. Jascha Boblel – Product Manager — jblobel@intelligentpharma.com
Elena Gumà – Sales Manager — eguma@intelligentpharma.com

Comunicación y prensa — communication@intelligentpharma.com
RRHH — rrhh@intelligentpharma.com
Información general — info@intelligentpharma.com

www.intelligentpharma.com

Intelligent Pharma

Your Computational Partner In Drug Discovery

Thank you for your attention