The ACCAMBA Project: what is the proper way to describe molecules in order to explain and predict their bioactivity using machine learning methods?

Samia Aci[1], Samuel Wieczorek[1], Caroline Barette[2], Mirta B.Gordon[3], Dragos Horvath[4], Eric Maréchal[2], Gilles Bisson[3], Laurence Lafanechère[2], Sylvaine Roy[1]

[1]*Laboratoire de Biologie, Informatique, Mathématiques*
Département Réponse et Dynamique Cellulaire
CEA Grenoble 17, rue des Martyrs 38054 Grenoble
sylvaine.roy@cea.fr
[2]*Centre de Criblage pour des Molécules Bio-Actives – Grenoble*
[3]*Equipe Apprentissage – Laboratoire Leibniz-IMAG – Grenoble*
[4] *Structures et Fonctions des Biomolécules, UMR8525, Institut de Biologie de Lille*.

Automatic screening is being used in the pharmaceutical world for about 20 years to discover drug candidates from chemical libraries. In an academic context, this approach can be used to select new compounds, bearing a biological activity on original cell phenotypes. The characterized compounds are expected to be innovative research tools for biology and/or drug candidates. The benefits derived from using intact, living cells for compound screening (called *phenotypical screening*) include:

- the isolation of functionally selective molecules, i.e. compounds which are active within the context of the subcellular "working environment",
- the isolation of compounds active on a cell function without any prior knowledge on the involved molecular and/or regulatory processes.

In such a context, the real target can remain unknown, so *docking methods* are not relevant to study structural activity relationship.

The ACCAMBA project (http://accamba.imag.fr/) is a collaborative multidisciplinary project supported by the French ministry of research, involving biologists, chemists and computer scientists. Its main objective is to develop tools to analyse chemical libraries and to model screening results using Machine Learning approaches. The accuracy of these models depends on the relevance of the language used to describe the studied objects, molecules in our case.

In this poster, we present a general overview of the ACCAMBA approach and we focus on the *representation problems* involved by the description of the molecules. First, the libraries of compounds that we use are of various academic and commercial origins; it is therefore necessary to harmonize the chemical associated data, prior to any *in silico* study. Second, it is crucial to define the best way to represent molecules since a single molecule representation is not always sufficient to capture all its potential chemical functionalities. For example, a chemical compound can exist in several tautomer forms and we need to represent all of them in order to discover which one is active on a given biological target.

We are currently studying these problems in the frame of the ACCAMBA project. Our goal is to propose a reliable and efficient representation of the molecules for the Machine learning approach, allowing to build pertinent bioactivity models.