[P28] Predicting likelihood of organic molecules by structure based Markov random field

Grzegorz Skoraczyński[1], Mateusz Kitlas[1], Błażej Miasojedow[1], Sara Szymkuć[2], Ewa Gajewska[2], Bartosz Grzybowski[2], Anna Gambin[1]

[1]*Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Banacha 2, 02-097 Warszawa, Poland*
[2]*Institute of Organic Chemistry of the Polish Academy of Sciences, Kasprzaka 44/52, 01-224 Warszawa, Poland*

Computer aided retrosynthesis is a forward-looking method of automating the process of chemistry development. [1] The underlying methodology benefits from modern statistical techniques yielding increasingly effective tools. The challenging open problem is to restrict a space of potential reactants to those which are feasible. There are many solutions proposed. One of possible approach is to develop a bunch of measures to efficiently compare computationally generated molecules with the existing ones.

We focus here on assessing the probability of a given organic molecule based on its structure. Our goal is to create an oracle tool which can answer whether a given molecule is possibly existing, e.g. is stable and there exist molecules with similar structure. To this aim we employed a set of over 200 motifs schemes, which were carefully and manually curated by experts. Pattern is usually a generic schema of functional group or layout of groups. Moreover, we created a set of descriptors characterizing a molecule by a specific motifs layout. It includes the existence of specific motifs, its numbers and pairwise distances.

To grasp a structure of interactions between descriptors (e.g. coexistence in one molecule and feasible minimal distance) we applied a Markov Random Field (MRF) model. [2] The model infers the probability distribution over graphs of interactions between descriptors. The structure of graph and the strength of interactions between vertices are learned by sparse Machine Learning techniques.

We estimated parameters of the model from over 6 million entries database. Thanks to this model we can easily and fast evaluate a score of molecule with given structure. This can help to restrict the size of space explored during computational retrosynthesis by early rejecting the improbable molecules candidates.

Bibliography:

[1] Szymkuć S., Gajewska E.P., Klucznik T., Molga K., Dittwald P., Startek M., Bajczyk M., Grzybowski B.A. Angew. Chem. Int. Ed. Engl. 55, 20 (2016) 5904-37.
[2] Kindermann R., Snell J.L., Contemp. Math. 1 (1980).