CQB - Day 2013

July 4th

Book of Abstracts

Centro de Química e Bioquímica CQB Faculdade de Ciências da Universidade de Lisboa

FCT Fundação para a Ciência e a Tecnologia

CQB-Day 2013

July 4th

Faculdade de Ciências – Universidade de Lisboa

Centro de Química e Bioquímica CQB Faculdade de Ciências da Universidade de Lisboa

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Improving QSPR models for predicting standard enthalpy of formation with a hybrid approach for feature selection

Ana L. Teixeira[a,b,*], João P. Leal[b,c], André O. Falcão[a]
[a] LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal; [b] Centro de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal; [c] Unidade de Ciências Químicas e Radiofarmacêuticas, IST/ITN, Instituto Superior Técnico, UTL, Estrada Nacional 10, P-2686-953 Sacavém, Portugal; [*]ateixeira@lasige.di.fc.ul.pt

Due to the high rate of new compounds discovered each day and the fact that laboratory techniques for experimental measurements are still expensive, there is a significant gap between the number of known chemical compounds and the amount of experimental thermochemical property data in the literature. Thus it is clear the great need to foster the application of prediction methods with a good predictive performance when experimental values are not available. Quantitative structure-property/activity relationship (QSPR/QSAR) methods are widely used for prediction and their goal is to relate molecular descriptors, from molecular structure, with experimental chemical, physical and/or biological properties by means of data-mining methods. The three major difficulties in the development of QSPR/QSAR models are (1) quantifying the inherently abstract molecular structure, (2) determining which structural features most influence the given property (representation problem) and (3) establishing the functional relationship that best describes the relationship between these structure descriptors and the property/activity data (mapping problem). The first difficulty can be overcome by the use of calculated molecular descriptors, developed to quantify various aspects of molecular structure. In fact, this approach is one of the causes of the second difficulty since thousands of molecular descriptors are currently extant. The problem lies then in the identification of the appropriate set of descriptors that allow the desired property of the compound to be adequately predicted.

We propose an alternative selection method, based on Random Forests to determine the variable importance in the context of QSPR regression problems, with an application to a manually curated dataset for predicting standard enthalpy of formation. The subsequent predictive models are trained with support vector machines introducing the variables sequentially from a ranked list based on the variable importance [1]. The model generalizes well even with a high dimensional dataset and in the presence of highly correlated variables. The feature selection step was shown to yield lower prediction errors with RMSE values 23% lower than without feature selection, albeit using only 6% of the total number of variables (89 from the original 1485) [1]. The proposed approach further compared favorably with other feature selection methods and dimension reduction of the feature space [1]. The methodology seemingly improves the prediction performance of standard enthalpy of formation of hydrocarbons using a limited set of molecular descriptors, providing faster and more cost-effective calculation of descriptors by reducing their numbers.

Acknowledgments
ALT gratefully acknowledges Fundação para a Ciência e Tecnologia (FCT) for a doctoral grant (SFRH/BD/64487/2009). AOF and ALT acknowledge the Multiannual Funding Programme of LaSIGE by FCT. JPL acknowledge the Strategic Project (PEst-OE/QUI/UI0612/2011) to CQB by FCT.

References
[1] A. L. Teixeira, J. P. Leal, A. O. Falcão, *J. Cheminformatics.* **2013**, *5:9*.