Advanced data mining, data visualization and deep learning for kinase research

Peyrat, G., [1*] Carles, F., [1] Bourg, S., Meyer, C., [2] Bonnet, P. [1]
gautier.peyrat@univ-orleans.fr
[1] Institut de Chimie Organique et Analytique, UMR CNRS-Université d'Orléans 7311, Université d'Orléans BP 6759, 45067 Orléans Cedex 2, France
[2] Janssen-Cilag, Centre de Recherche Pharma, CS10615 – Chaussée du Vexin, 27106 Val-de-Reuil, France

Major progress has been made in the field of Machine Learning in the past two decades. The improved Deep Learning algorithms combined with the large scale-input data has resulted in boosting performance and quality of prediction models in a lot of research fields. Unfortunately, in the area of pharmaceutical science, only a small community of experts is currently able to get the best out of the plethora of databases and last-generation prediction tools leaving aside a large amount of dark data i.e. data that has been already collected but not fully exploited for decision-making.

Here we present two new databases and web service to make more effective use of available kinase data. The first one is a database of Protein Kinase Inhibitors called PKIDB. It contains actually more than 210 kinase inhibitors approved or in development (phase varying from 0 to 4 in clinical trials). Each compound is annotated with data gathered from public database related to structural information, drug indication or name. Moreover main physical properties are described by the means of calculated molecular descriptors and their distributions are studied too. This freely redesigned monthly updated database is accessible on the Structural Bioinformatics & Chemoinformatics (SB&C) webservice platform at http://www.icoa.fr/pkidb [1].

The second one, called KinoMine, is a web portal allowing to access all available data, chemical and biological, on protein kinase (e.g. bioactivities and 3D structures) through a new interactive visualization and data mining web-based interface. It aims to provide the most curated kinase data and is accessible on SB&C platform too at http://kinomine.icoa.fr.

Finally, we also introduce and compare performances of different proteochemometrics PCM models aiming to classify kinase-ligand complexes and differencing each other regarding two aspects:

- first, the nature of descriptors used to describe the interaction (1D, 2D, 2.5D [2], 3D and interaction fingerprints [3])

- second, the nature of the selected algorithm to establish the model (well established algorithms such as Random Forest [4] or Support Vector Machine [5] or new Deep Learning algorithm such as the outperformer Deep Neural Network [6]). Our best obtained results will be presented.

Bibliography :

[1] Carles, F., Bourg, S., Meyer, C., Bonnet , P., Molecules, 2018, 23, 908
[2] Bosc, N., Wroblowski, B, Meyer, C, Bonnet, P., J. Chem. Inf. Model., 2017, 57, 93-101.
[3] Singh, J., Deng, Z., Narale, G., Chuaqui, C., Chem. Biol. Drug. Design., 2006, 67, 5-12.
[4] Breiman, L., Mach. Learn., 2001, 45, 5-32.
[5] Boser, B. E., Guyon, I. M., Vapnik, V. N., Proceedings of the fifth annual workshop on computational learning theory,
1992, 144-152.
[6] Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W. T., Kowalczyk, W., Ijzerman, A. P., Van Westen, G. J. P., J. Cheminformatics, 2017, 9, 45