Passamonti S. (Editor)
The Partners and the Objectives of Trans2Care, an Italy-Slovenia cross-border network of science and healthcare institutions
pp. 27-32 / ISBN 978-88-8303-512-8

trans2care
Transregional Network for Innovation and Technology Transfer to Improve Health Care

Chemometrics: a versatile tool to explore large dataset

Špela Župerl, Katja Stopar, Marjana Novič

PP1-National Institute of Chemistry, Ljubljana

Abstract — Chemometrics is the field of science covering the development and application of mathematical and statistical methods to identify important chemical information. It is indispensable in the evaluation of experimental results and suitable for exploration of large data sets. Within the Trans2Care project we intend to apply chemometrics methods in several areas related to the problems explored by the Project partners. In particular we shall investigate transmembrane protein transport mechanism with data-driven modelling approach and also applying biomolecular simulations. We'll combine our theoretical approach with experimental data provided by the Project partners, which will contribute to a better exploration of the available information in biomolecular systems studied, in the research of neurodegenerative diseases, in cardiovascular and pathohistological research. It will also intensify the collaboration, mobility of researchers and exchange of knowledge between partners.

Index Terms — chemometrics, data mining, predictive modelling, transmembrane proteins

1 THE NATIONAL INSTITUTE OF CHEMISTRY

The National Institute of Chemistry (NIC) was established in 1946 by the Slovenian Academy of Arts and Sciences (SAZU) as the Chemistry Laboratory of the SAZU, later it was renamed to the Chemistry Institute of the SAZU. Following Slovenian independence in 1992 the National Institute of Chemistry (NIC) became a public research institution. The NIC has 290 employees and they carry out research work in 15 laboratories and two infrastructure centers. More than 25% of the Institute's staff members represent young researchers, making NIC one of the leading Slovenian organizations for education of graduate students. Research activities of the Institute are oriented towards the development of new expertise, technologies and products,

which will help to ensure the long-term development of Slovenia. The Institute offers high-level research equipment including a Karl Zeiss Supra 35 VP Electronic Microscope with EDX analysis, a high resolution powder x-ray diffractometer, and an 800 MHz NMR spectrometer, allowing researchers to engage in advanced research challenges at the world level. Industry is also an important partner to the Institute; several Slovenian as well as foreign companies has established a close long-term cooperation with the Institute.

2 LABORATORY OF CHEMOMETRICS

In the early seventies the beginning of the chemometrics research started in Slovenia which was completely new field also in the world. At the National Institute of Chemistry prof. Dušan Hadži implemented chemometrics in his group in 1973 and in 1993 the group became an independent Laboratory of Chemometrics, first dealing with systems for identification of compounds based on infrared spectra, the study of algorithms, expert systems and modelling.

Today the Laboratory of Chemometrics has 14 associates, seven researchers, three PhD students, two post-doctoral students, one visiting professor and one professor emeritus. We are developing and applying chemometrics and statistical methods for solving problems in chemistry and related sciences, from visualization of many-dimensional data to analysis of biologically relevant information in proteomics and genomics. The research work is financed through national research programme schemes and several European projects. In 2010, the Laboratory of Chemometrics was involved in three EU projects (TRACE, CAESAR, IBAAC), bilateral projects with Turkey, Argentina and USA and made a strong collaboration with industry of asphalts (IMS-ADITOL). The laboratory has already established a strong long-term collaboration with University of Trieste and from 2010 we are partners in an international strategic project Trans2Care.

2.1 Research Areas

The researchers of the Laboratory of Chemometrics are engaged in various research activities, such as (i) introduction of chemometrics to the research and development, (ii) modelling of chemical properties and processes - QSAR and mechanistic models, (iii) development and application of artificial neural network methods in chemistry, (iv) application of discrete mathematics in structural chemistry, QSAR studies, proteomics, and in genomics, (v) development of methodologies and programme packages for mechanistic and empirical models, and (vi) development of 3D representations of chemicals structures for applications in QSAR.
We have successfully applied chemometrics methods in the field of traceability of food [1], in the prediction of toxic properties of various toxic data (developmental toxicity, mutagenicity, carcinogenicity, bio-concentration factor and skin sensitation) [2], in the optimization of pigment dyeing of polybenzimidazol fibres [3], in the optimization of bio-organic catalysts for stereo-selective reduction of prochiral ketones [4], in the research

of anti-tuberculosis drugs [5,6], in the optimization of gradient profiles in ion-exchange chromatography [7], in the investigation of a transport mechanism of a membrane protein, bilitranslocase [8,9], and in the research of structural characterization of transmembrane proteins [10]. In the field of proteomics we have recently published a review paper of Graphical representation of proteins [11].

2.2 Structural Elucidation of Transmembrane Protein, Bilitranslocase

In the collaboration with University of Trieste, Department of Life Science we have started the research project in which we address the problem of structural elucidation of transmembrane proteins. The experimental work performed at the Department of Life Science was the basis for the application of computational methods in the Laboratory of Chemometrics. In-silico methods are strongly dependent on experimental data available and a combination of experimental and theoretical approach can lead to a successful resolution of the specific problem. Experimental data are treated with computer algorithms, whose output are theoretical predictions of biological properties, and the predictions can be directly validated in additional experimental work. Below is a schematic presentation of a QSAR model for prediction of biological properties from structural data (Fig.1).

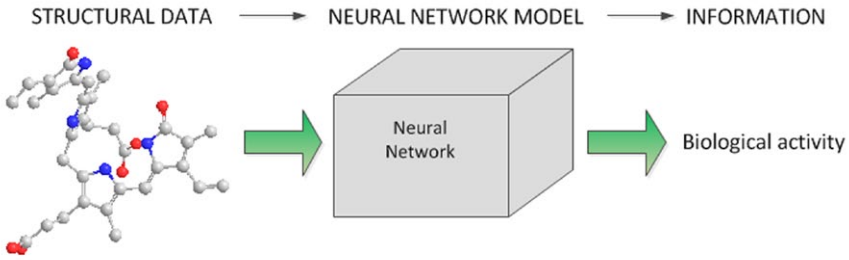

Fig.1. Schematic presentation of a QSAR model for prediction of biological property from structural data.

In the latest publication [12] we have presented an approach towards structure elucidation of bilitranslocase, the membrane protein which transports bilirubin from blood to liver cells. In this research we combined two approaches: first, the prediction of transmembrane segments of the protein based on the mathematical descriptors obtained from the information of membrane proteins of known 3D structure, and second, the information about the transport mechanism from experimentally tested set of small molecules for their competitive inhibition of bilitranslocase.
In the first approach the sequence and secondary structure information of transmembrane segments of proteins with known 3D structure available from public databases (PDB and PDBTP) was exploited to build a model for prediction of transmembrane segments of structurally unresolved target protein. The prediction error of the model for prediction of alpha transmembrane segments for the external

validation set was below 10%. The model was challenged with bilitranslocase and it proposed four transmembrane alpha helices, each containing around 20 amino acids, which is partially confirmed with experimental studies using particular antibodies corresponding to parts of amino acid sequences of bilitranslocase.
In the second approach we have tested a set of 89 non-congeneric compounds for their competitive inhibition constants in the investigated protein-substrate system. The information about 3D chemical structure of small molecules (represented by molecular descriptors) and the experimental data assessed by evaluating the kinetics of inhibition of bilitranslocase transport activity was used for development of a data-driven model using artificial neural networks.
QSAR models showed good predictive ability for bilitranslocase binding affinity. From the mechanistic interpretation of selected molecular descriptors, obtained with genetic algorithm, it was found that interactions between bilitranslocase and small molecules rely on the ability to establish hydrogen bonds, diminishing the involvement of charge interactions. The results of this work show that, contrary to dietary anthocyanins, most of dietary flavonols do not interact with bilitranslocase, whereas, some flavonol aglycones act as poor ligands of that carrier. In case of nucleobases and their derivatives the phosphate group in principle improved the transport ability by bilitranslocase.

3 CHEMOMETRICS IN TRANS2CARE

3.1 Competences or what can we offer to T2C

The long term collaboration with the University of Trieste in the field of transmembrane proteins is the basis for the research work within the T2C project. As described in the paragraph above, we have recently studied transmembrane protein, bilitranslocase, its transport activity and transmembrane segments. We will continue with the development of the prediction models, either for inhibition constants or for transmembrane alpha helices of beta sheets. We will also study the 3D structural model and the molecular dynamics simulations of bilitranslocase. A complementary study and characterization of the sequence of bilitranslocase isolated from plants will be another interesting research topic and will offer an interesting comparison with the bilitranslocase homologue obtained from rat liver. Furthermore, structure elucidation studies will be extended to other transmembrane transporters, such as SbmA transporter (protein present in the membrane of bacteria).

Chemometrics methods are suitable for exploration of large data sets, especially in the case when the information contained in the data is not obvious and the knowledge is not easily extracted. For this reason we intend to collaborate with the T2C partners whose role is a compilation of large amounts of data, not only in the research laboratories, but also in the hospitals, where a lot of data are collected during their everyday practise.

4 CONCLUSION

In Trans2Care project our priority will be to explore experimental data collected by several partners, from biochemical data on membrane transporters to various studies of the neurodegenerative disease on one side, and statistical analysis of different data-bases and registers on the other side. The interdisciplinary approach of our research work will offer a good opportunity for young researchers starting their research career. It will also straighten already established collaborations and connect the research institutes, universities and hospitals within the Project, having in mind a common goal, a network of participants for exchange of ideas, knowledge dissemination and technology transfer.

ACKNOWLEDGEMENT

The financial support of the Fondo europeo di sviluppo regionale (Evropski sklad za teritorialni razvoj) for the Trans2Care project is greatly appreciated. The financial support by the Slovene Research Agency through the research grant P1-017 is acknowledged.

REFERENCES

[1] N. Grošelj, G. van der Veer, M. Tušar, M. Vračko, M. Novič, "Verification of the geological origin of bottled mineral water using artificial neural networks," Food Chemistry, vol. 118, pp. 941–947, 2010.

[2] N. Fjodorova, M. Vračko, M. Novič, A. Roncaglioni, E. Benfenati, "New public QSAR models for carcinogenicity," Chemistry Central J., vol. 4, pp. 1-15, 2009.

[3] N. Fjodorova , M. Novič, T. Diankova, "Optimization of pigment dyeing process of high performance fibers using feed-forward bottleneck neural networks mapping technique," Anal. Chim. Acta, vol. 705, pp. 148 –154, 2011.

[4] S. Nandi, M. Chaumontet, F. Taran, M. Novič, "Prediction of new [3+2] dipolar cycloaddition reactions," M. VRAČKO (ed.), Abstract book, National Institute of Chemistry, CMTPI 2011 conference, pp.100.

[5] N. Minovski, T. Šolmajer, "Chemometrical exploration of combinatorially generated drug-like space of 6-fluoroquinolone analogs : a QSAR study," Acta chim. slov., vol. 57, pp. 529-540, 2010.

[6] N. Minovski, A. Perdih, T. Šolmajer, "Combinatorially-generated library of 6-fluoroquinolone analogs as potential novel antitubercular agents : a chemometric and molecular modeling assessment," J. mol. model., vol. 17, pp. 19, 2011.

[7] V. Drgan, D. Kotnik, M. Novič, "Optimization of gradient profiles in ion-exchange chromatography using computer simulation programs," Anal. chim. Acta, vol. 705, pp. 315-321, 2011.

[8] A. Karawajczyk, V. Drgan, N. Medic, G. Oboh, S. Passamonti, M. Novič, "Properties of flavonoids influencing the binding to bilitranslocase investigated by neural network modelling," Biochem. Pharm., vol. 73, pp. 308-320, 2007.

[9] Š. Župerl, S. Fornasaro, M. Novič, S. Passamonti, "Experimental determination and prediction of bilitranslocase transport activity," Anal. Chim. Acta, vol.705, pp. 322-333, 2011.

[10] A. Roy Choudhury, M. Novič, "Data-driven model for the prediction of protein transmembrane regions," SAR QSAR environ. res., vol. 20, pp. 741-754, 2009.

[11] M. Randić, J. Zupan, A.T. Balaban, D. Vikić-Topić, D. Plavsić, "Graphical representation of proteins," Chem. Rev., vol. 111, pp.790-862, 2011.

[12] A. Roy Choudhury, Š. Župerl, S. Passamonti, M. Novič, "Structure Elucidation of Transmembrane Proteins Using Public-Available Databases and Experimental Data on Competitive Inhibition," Acta chim. slov., vol. 58, pp. 385-392, 2011.

CONTACT INFO

Špela Župerl, Katja Stopar and Marjana Novič are with National Institute of Chemistry, Laboratory of Chemometrics, Hajdrihova 19 - POB 660, 1001 Ljubljana, Slovenia.