Linear Regression Modeling and Validation Strategies for Structure-Activity Relationships

Sorana D. Bolboaca[1], Lorentz Jantshi[2]
[1] Department of Medical Informatics and Biostatistics, 'Iuliu Hatieganu' University of Medicine and Pharmacy Cluj-Napoca, 400349 Cluj-Napoca, Cluj, Romania.
sbolboaca@umfcluj.ro
[2] Department of Physics and Chemistry, Technical University of Cluj-Napoca, 101-103 Muncii Bvd., 400641 Cluj-Napoca, Cluj, Romania.
lorentz.jantschi@gmail.com

Keywords: quantitative structure-activity relationships (QSARs), linear regression analysis, model design, validation and diagnosis.

Identification and development of a new active compounds is an extremely expensive (reflected in time - between 10 and 15 years [1] and costs) and difficult process without a guaranteed result [2] ($\sim$ 90% of the initial candidates fail to be produces due to their toxicological properties [3]). Traditional strategies based on experiments (animal models [4]) are not anymore able to meet the actual needs in identification of new active compounds while in silico approaches such as computer-aided drug design [5], structure-based drug design [6], or virtual screening [7], are used nowadays.

Quantitative structure-activity relationships (QSARs) are mathematical relationships linking chemical structure and pharmacological activity/property in a quantitative manner for a series of compounds [8]. The approaches are based on the assumption that the structure of chemical compounds (such as geometric, topologic, steric, electronic properties, etc.) contains features responsible for its physical, chemical and biological properties [9]. The linear regression analysis is the statistical method frequently used in QSAR analysis since the main aim of the modeling is to identify a model able to predict the activity of new compounds [10].

Problems solving strategies in linear regression modeling include approaches for dealing with effective assessment of assumptions (linearity, independence of the errors, homoscedasticity, normality [11]), which seems to be broken in QSARs analyses [12,13]; effective methods for model selection [11,14,15]; efficient methods for model diagnosis [16,17]; and adequate approaches for assessment of predictive power of a QSAR model [18,19].

Here we emphasize problem solving strategies that address the main issues that arise when developing multivariate linear regression models using

real data.

Other problems not addressed here include the dealing with not normal distributed errors [20,21] and additional methods for estimation of regression parameters [22].

References

[1] Congressional Budget Office, Research and Development in the Pharmaceutical Industry, CBO, Washington DC, 2006.

[2] X.-P. Chen, and G.-H. Du, *Target validation: A door to drug discovery*, Drug Discov Ther **1**(1), 2007, 23–29.

[3] H. van de Waterbeemd, and E. Gifford, *Admet in silico modelling: towards prediction paradise?*, Nat Rev Drug Discov **2**(3), 2003, 192–204.

[4] R. McArthur, and F. Borsini, *Animal models of depression in drug discovery: A historical perspective*, Pharmacol Biochem Behav **84**(3), 2006, 436–452.

[5] P. Vanhee, A.M. van der Sloot, E. Verschueren, L. Serrano, F. Rousseau, and J. Schymkowitz, *Computational design of peptide ligands*, Trends Biotechnol **29**(5), 2011, 231–239.

[6] S.S. Phatak, H.T. Tran, and S. Zhang, *Novel computational biology methods and their applications to drug discovery*, Front Biol **6**(4), 2011, 289–299.

[7] D.L. Ma, D.S.H. Chan, P. Lee, M.H.T. Kwan, and C.H. Leung, *Molecular modeling of drug-DNA interactions: Virtual screening to structure-based design*, Biochimie **93**(8), 2011, 1252–1266.

[8] L.P. Hammett, *The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives*, J Am Chem Soc **59**(1), 1937, 96–103.

[9] A.M. Johnson, and G.M. Maggiora, Concepts and Applications of Molecular Similarity, John Willey & Sons, New York, 1990.

[10] M. Goodarzi, B. Dejaegher, and Y.V. Heyden, *Feature selection methods in QSAR studies*, Journal of AOAC International **95**(3), 2012, 636–651.

[11] S.D. Bolboaca, and L. Jantschi, *Modelling the property of compounds from structure: statistical methods for models validation*, Environ Chem Lett **6**, 2008, 175–181.

[12] S.D. Bolboaca, and L. Jantschi, *Distribution Fitting 3. Analysis under Normality Asumptions*, Bulletin UASVM Horticulture **66**(2), 2009, 698-705.

[13] L. Jantschi, and S.D. Bolboaca, *Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics*, Bulletin UASVM Horticulture **66**(2), 2009, 691–697.

[14] S.D. Bolboaca, *Assessment of Random Assignment in Training and Test Sets using Generalized Cluster Analysis Technique*, Appl Med Inform **28**(2), 2010, 9–14.

[15] S.D. Bolboaca, and L. Jantschi, *Dependence between determination coefficient and number of regressors: a case study on retention times of mycotoxins*, Stud U Babes-Bol Che **LVI**(1), 2011, 157–166.

[16] R.E. Sestras, L. Jantschi, and S.D. Bolboaca, *Poisson Parameters of Antimicrobial Activity: A Quantitative Structure-Activity Approach*, Int J Mol Sci **13**(4), 2012, 5207–5229.

[17] S.D. Bolboaca, and L. Jantschi, *Predictivity Approach for Quantitative Structure-Property Models. Application for Blood-Brain Barrier Permeation of Diverse Drug-Like Compounds*, Int J Mol Sci **12**(7), 2011, 4348–4364.

[18] S.D. Bolboaca, and L. Jantschi, *Comparison of QSAR Performances on Carboquinone Derivatives*, TheScientificWorldJOURNAL **9**(10), 2009, 1148–1166.

[19] S.D. Bolboaca, and L. Jantschi, *The Effect of Leverage and/or Influential on Structure-Activity Relationships*, Comb Chem High Throughput Screen **16**(4), 2013, 288–297.

[20] L. Jantschi, and S.D. Bolboaca, *Observation vs. Observable: Maximum Likelihood Estimations according to the Assumption of Generalized Gauss and Laplace Distributions*, Leonardo El J Pract Technol **8**(15), 2009, 81–104.

[21] L. Jantschi, and S.D. Bolboaca, *The Jungle of Linear Regression Revisited*, Leonardo El J Pract Technol **6**(10), 2007, 169-187.

[22] L. Jantschi, *Distribution Fitting 1. Parameters Estimation under Assumption of Agreement between Observation and Model*, Bulletin UASVM Horticulture **66**(2), 2009, 684-690.