BIOCHEMISTRY VERSUS BIOMATHEMATICS IN MODELLING OF BIOLOGICAL ACTIVE COMPOUNDS

Sorana D. BOLBOACĂ[1] and Lorentz JÄNTSCHI[2]

[1]„Iuliu Haţieganu“ University of Medicine and Pharmacy Cluj-Napoca, 6 Louis Pasteur, 400349 Cluj-Napoca, Romania, http://sorana.academicdirect.ro
[2]Technical University of Cluj-Napoca, 103-105 Muncii Bvd, 400641 Cluj-Napoca, http://lori.academicdirect.org

ABSTRACT

A new mathematical approach that works at the level of molecular topology is proposed for characterization of structure-activity relationship of biological active compounds. A family of molecular descriptors is generated for a set of biologic active compounds and a genetic algorithm is used for identification of the best performing multivariate regression model. A series of statistical approaches are considered for model assessment (Bolboacă and Jäntschi, 2008 [8]). In order to validate the new method the performances of the obtained model will be compared through a correlated correlation analysis with other qSAR models.

INTRODUCTION

Development of information and computer technologies induces changes into research concept, leading to the development of many in silico analytical and experimental methods [1,2] used in determination and prediction of drug metabolism [3]. These methods have some advantages, from which the most important are: allows determination of metabolic profile in early stages of drug design; experiments are done into a shorter time and with fewer expenses [2].
Mathematical approach on structure-activity relationship (SAR) for biological active compounds (begun in nineteen century) lead to the concept of quantitative structure-activity relationship (QSAR, a mathematical approach that allows the identification of the quantitative link between structure and biologic activity of investigated compounds – [4]). SAR studies have been published since 1868, when Crum-Brown & Fraser stipulated the idea that the compounds activity is o function of structure and chemical composition [5].

METHODOLOGY

A mathematical approach developed starting with the information obtained from the 2D and 3D structure of a chemical compounds leads to introduction of Molecular Descriptors Family on the Structure-Activity Relationship method [7].
A family of molecular descriptors is generated for the set of biologic active compounds and a genetic algorithm is used for identification of the best performing multivariate regression model (see Figure 1 for the formal description of the approach). A series of statistical approaches [8] are considered for model assessment:

- Simple correlation analysis; Inter-correlation analysis; Multiple correlation analysis: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/k_browse_or_query.php?database=MDFSARs/
- Qualitative vs. quantitative analysis (correlation coefficients: Pearson; Spearman; Semi-quantiative; Kendall tau-a; Kendall tau-b; Kendall tau-c; Goodman-Kruskal; test of significance and associated p-value): http://l.academicdirect.org/Statistics/linear_dependence/
- Leave-one-out cross-validation analysis: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/loo/
- Training vs. test experiment: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/qsar_qspr_s/
- Correlated correlations analysis (Steiger's test): http://l.academicdirect.org/Statistics/tests/Steiger/

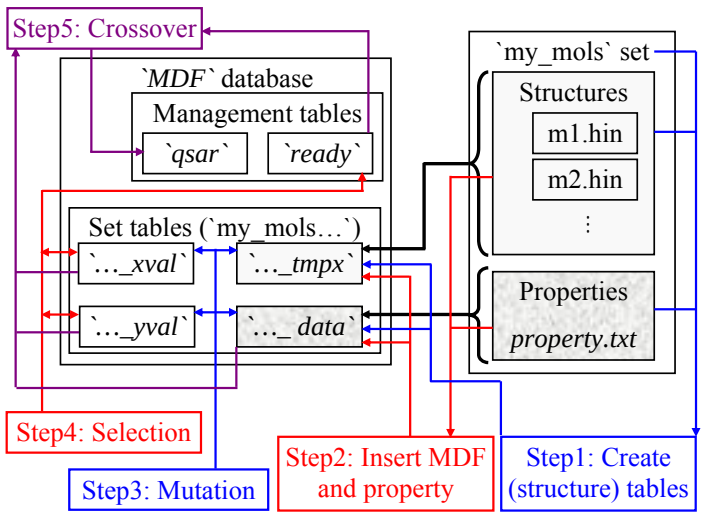


Figure 1. *Formal MDF approach*

EXAMPLES: BIOMATHEMATICS IN MODELLING BIOLOGICAL ACTIVE COMPOUNDS

AMINO ACIDS MODELLING [9]

- Analysis bulletins [9]

Amino acid property	Parker et all., 1986	Kyte-Doolittle, 1982
MDF SPR Equation	$\hat{y} = 11.05 + x \cdot 1.85$	$\hat{y} = -7.60 + x \cdot 19.17$
SPR Determination (%)	86	87
MDF Descriptor (x)	IfPROQg	iGPdLQg
Dominant Atomic Property	Charge (Q)	Charge (Q)
Interaction via	Space (geometry)	Space (geometry)
Interaction Model	Q	$d \cdot \sqrt{Q}$
Structure on Property Scale	Logarithmic	Inversed

n = 20; r = 0.9259; F = 108*; s = 2.46; r_{loo} = 0.8935; F_{loo} = 69*; s_{loo} = 2.97

Amino acid property	Black et all., 1991	Monera et all., 1995
MDF SPR Equation	$\hat{y} = 0.86 + x \cdot (-0.96)$	$\hat{y} = 86.05 + x \cdot 843.88$
SPR Determination (%)	88	90
MDF Descriptor (x)	IAmrLQg	inMrpQg
Dominant Atomic Property	Charge (Q)	Charge (Q)
Interaction via	Space (geometry)	Space (geometry)
Interaction Model	$d \cdot \sqrt{Q}$	Q^2
Structure on Property Scale	Proportional	Inversed

n = 20; r = 0.9327; F = 120*; s = 1.11; r_{loo} = 0.9226; F_{loo} = 103*; s_{loo} = 1.18

n = 20; r = 0.9376; F = 131*; s = 0.12; r_{loo} = 0.9263; F_{loo} = 109*; s_{loo} = 0.13

n = 19 (- Prolina); r = 0.9504; F = 159*; s = 16.49; r_{loo} = 0.9382; F_{loo} = 125*; s_{loo} = 18.37

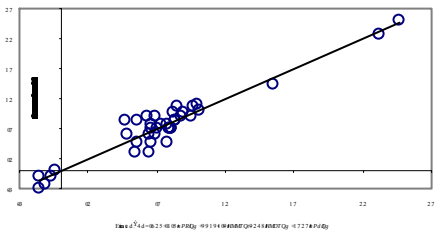
- Models assessment [10]

Abb.	Regression model						Leave-one-out		
	n	r	F (p)	s	$_{95\%}CI_{Intercept}$ ($p_{t,Stat}$)	$_{95\%}CI_{Slop}$ ($p_{t,Stat}$)	r_{loo}	F_{loo}	s_{loo}
Hyd_01	20	0.9376	131 ($1.09 \cdot 10^{-9}$)	0.12	[0.77 – 0.94]*	[-1.14 – -0.78]*	0.9263	109 ($4.73 \cdot 10^{-8}$)*	0.13
Hyd_02	20	0.9327	120 ($2.10 \cdot 10^{-9}$)	1.11	[-9.05 – -6.14]*	[15.50 - 22.84]*	0.9226	103 ($7.25 \cdot 10^{-8}$)	1.18
Hyd_03	20	0.8434	44 ($3.00 \cdot 10^{-6}$)	0.48	[-4.42 – -2.32]*	[5.03 – 9.67]*	0.8009	32 ($2.25 \cdot 10^{-5}$)	0.54
Hyd_04	20	0.9238	105 ($6.24 \cdot 10^{-9}$)	0.52	[-0.79 – -0.02]*	[5.70 – 8.65]*	0.9018	78 ($6.01 \cdot 10^{-8}$)	0.58
Hyd_05	20	0.9232	104 ($6.69 \cdot 10^{-9}$)	20.73	[66.20 – 97.23]*	[649.29-986.61]*	0.9082	85 ($3.16 \cdot 10^{-8}$)	22.58
Hyd_06	20	0.8608	52 ($1.11 \cdot 10^{-6}$)	1.01	[-2.70 – -1.29]*	[7.52 – 13.75]*	0.8288	39 ($6.49 \cdot 10^{-6}$)	1.11
Hyd_07	20	0.8309	40 ($5.70 \cdot 10^{-6}$)	1.70	[-4.30 – -1.39]*	[-2.30 – -1.15]*	0.7936	30 ($3.34 \cdot 10^{-5}$)	1.87
Hyd_08	20	0.9128	90 ($2.02 \cdot 10^{-8}$)	0.42	[1.26 – 2.10]*	[-1.12 – -0.72]*	0.8935	70 ($1.31 \cdot 10^{-7}$)	0.46
Hyd_09	20	0.8974	74 ($8.21 \cdot 10^{-8}$)	0.05	[0.82 – 0.90]*	[1.32 – 2.17]*	0.8744	58 ($4.73 \cdot 10^{-7}$)	0.06
Hyd_10	20	0.8997	76 ($6.76 \cdot 10^{-8}$)	0.32	[0.29 – 0.70]*	[-172.49 – -105.66]*	0.8599	56 ($6.37 \cdot 10^{-7}$)	0.36
Hyd_11	20	0.9116	89 ($2.26 \cdot 10^{-8}$)	2.07	[0.64 – 3.06]*	[-921.24 – -584.95]*	0.8731	51 ($1.13 \cdot 10^{-6}$)	2.56
Hyd_12	20	0.8986	75 ($7.42 \cdot 10^{-8}$)	0.45	[-4.22 – -2.50]*	[2.85 – 4.67]*	0.8812	62 ($2.93 \cdot 10^{-7}$)	0.48
Hyd_13	20	0.9252	107 ($5.30 \cdot 10^{-9}$)	0.36	[1.02 – 1.70]*	[-0.25 – -0.16]*	0.9003	75 ($8.02 \cdot 10^{-8}$)	0.42
Hyd_14	20	0.9208	100 ($8.69 \cdot 10^{-9}$)	0.80	[4.07 – 6.54]	[-4.58 – -2.99]*	0.9073	84 ($3.48 \cdot 10^{-8}$)	0.86
Hyd_15	20	0.6649	14 ($1.38 \cdot 10^{-3}$)	1.21	[-1.99 – -0.48]*	[0.17 – 0.61]*	0.5961	7 ($1.44 \cdot 10^{-2}$)	1.37
Hyd_16	20	0.9259	108 ($4.88 \cdot 10^{-9}$)	2.46	[8.71 – 13.39]*	[1.48 – 2.22]*	0.8935	69 ($4.91 \cdot 10^{-8}$)	2.97
Hyd_17	20	0.9182	97 ($1.15 \cdot 10^{-8}$)	0.52	[3.63 – 5.65]*	[-2.62 – -1.70]*	0.8984	75 ($7.94 \cdot 10^{-8}$)	0.58
Hyd_18	20	0.8814	63 ($2.84 \cdot 10^{-7}$)	0.76	[13.98 – 15.13]*	[17.22 – 29.65]*	0.8546	49 ($1.65 \cdot 10^{-6}$)	0.84
Hyd_19	20	0.8832	65 ($2.50 \cdot 10^{-7}$)	0.50	[-5.65 – -3.06]*	[4.38 – 7.50]*	0.8611	51 ($1.13 \cdot 10^{-6}$)	0.54
Hyd_20	20	0.8901	69 ($1.48 \cdot 10^{-7}$)	0.24	[1.25 – 1.61]*	[-3.42 – -2.04]*	0.8545	48 ($1.78 \cdot 10^{-6}$)	0.28
Hyd_21	20	0.8163	36 ($1.14 \cdot 10^{-5}$)	2.19	[4.66 – 8.44]*	[-37.53 – -18.06]*	0.7740	27 ($6.50 \cdot 10^{-5}$)	2.41
Hyd_22	20	0.8661	54 ($7.99 \cdot 10^{-7}$)	0.66	[0.97 – 1.96]*	[-8.45 – -4.69]*	0.8344	41 ($4.89 \cdot 10^{-6}$)	0.73
Hyd_23	20	0.9046	81 ($4.40 \cdot 10^{-8}$)	1.07	[-36.23 – -23.23]*	[-14.76 – -9.17]*	0.8819	63 ($2.85 \cdot 10^{-7}$)	1.18
Hyd_24	19	0.9504	159 ($4.77 \cdot 10^{-10}$)	16.49	[73.60 – 98.50]*	[702.55 – 985.21]*	0.9382	125 ($3.00 \cdot 10^{-8}$)	18.37

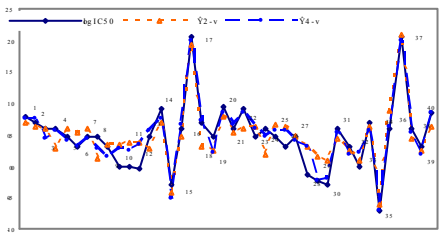
Abb. = abbreviation of hydrophobicity scale; n = sample size; r = correlation coefficient;
F = Fisher parameter and associated type I error values (p); s = standard error of estimated;
$_{95\%}CI$ = 95% confidence interval; Intercept = the intercept for the regression model;
$p_{t,Stat}$ = the type I error for the intercept and slop on regression model (Student t test);
r_{loo} = correlation coefficient obtained in leave-one-out analysis;
F_{loo} = Fisher parameter obtained in leave one out analysis;
s_{loo} = standard error of estimated in leave-one-out analysis.
* $p < 0.05$; * $p > 0.05$

INHIBITORY ACTIVITY ON CARBONIC ANHYDRASE (SUBSTITUTED 1,3,4-THIADIAZOLE- AND 1,3,4-THIADIAZOLINE-DISULFONAMIDES)

- **CA IV** [11]
  n = 40
  v = 4
  r = 0.9593
  s_{est} = 0.1599
  F (p) = 101 (< 0.001)
  $r^2_{cv\text{-}loo}$ = 0.9034
  $r^2 - r^2_{cv\text{-}loo}$ = 0.0168



- **CA II [12]**: $\hat{Y}_{4\text{-}v} = -9.9859 + 4.5643 \cdot imDdSCg + 2.945 \cdot 10^{-3} \cdot isDrqQg + 5.2036 \cdot IIMDQQg + 1.4832 \cdot lmMrsGg$
  n = 40
  v = 4
  r^2 = 0.9037
  s_{est} = 0.1706
  F (p) = 82 ($2.7 \cdot 10^{-15}$)
  $r^2_{cv\text{-}loo}$ = 0.8804



measured versus activity estimated by models

- **CA I** [13]

QSAR vs. MDF-SAR	Steiger's Z parameter	p-value
Model no. 1* - QSAR vs. Eq.(1) - MDF-SAR model	0.582	0.2803
Model no. 2* - QSAR vs. Eq.(1) - MDF-SAR model	1.041	0.1489
Model no. 1* - QSAR vs. Eq.(2) - MDF-SAR model	2.563	0.0052
Model no. 2* - QSAR vs. Eq.(2) - MDF-SAR model	2.965	0.0015

* table 1 and 2

CONCLUSION

The proposed mathematical model proved to have abilities in prediction and estimation of property and activity of chemical compounds in terms of estimation as well prediction.

REFERENCES

1. Plant N. Strategies for using in vitro screens in drug metabolism. Drug Discov Today 2004;9:328-336.
2. Baranczewski P, Stanczak A, Sundberg K, Svensson R, Wallin A, Jansson J, Garberg P, Postlind H. Introduction to in vitro estimation of metabolic stability and drug interactions of new chemical entities in drug discovery and development. Pharmacol Rep. 2006;58(4):453-72.
3. Schuster D, Steindl TM, Langer T. Predicting drug metabolism induction in silico. Curr Top Med Chem 2006;6(15):1627-40.
4. Hammett LP. Some Relations between Reaction Rates and Equilibrium Constants. Chemical Reviews 1935;17:125-136.
5. Crum-Brown A, Fraser TR. On the Connectionbetween Chemical Constitution and Physiological Ac-tion. Part I. On the Physiological Action of the Salts ofthe Ammonium Bases, derived from Strychnia, Brucia,Thebaia, Codeia, Morphia, and Nicotia. Philosophical Transactions of the Royal Society of London 1868;25:151-203.
6. Hansch C, Leo A. Substituent Constants for Correlation Analysis in Chemistry and Biology. John Wiley & Sons: New York, 1979.
7. Jäntschi L. Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology. Leonardo Electronic Journal of Practices and Technologies 2005;4(6):76-98.
8. Bolboacă SD, Jäntschi L. Modelling the property of compounds from structure: statistical methods for models validation. Environmental Chemistry Letters. 2008;6:175–181.
9. Bolboacă SD, Jäntschi L. A Structural Informatics Study on Collagen. Chemical Biology & Drug Design 2008;71(2):173-179.
10. Bolboacă SD, Jäntschi L. Modelling Analysis of Amino Acids Hydrophobicity. MATCH Communications in Mathematical and in Computer Chemistry 2008;60(3):1021-1032.
11. Jäntschi L, Bolboacă SD. Modelling the Inhibitory Activity on Carbonic Anhydrase IV of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information. El J Biomed 2006;2:22-33.
12. Jäntschi L, Ungureşan ML, Bolboacă SD. Complex Structural Information Integration: Inhibitor Activity on Carbonic Anhydrase II of Substituted Disulfonamides. Appl Med Inform 2005;17:12-21.
13. Bolboacă SD, Jäntschi L. Modelling the Inhibitory Activity on Carbonic Anhydrase I of Some Substituted Thiadiazoleand Thiadiazoline-Disulfonamides: Integration of Structure Information. Computer-Aided Chemical Engineering, Elsevier Netherlands & UK. 2007;24:965-970.

The research was partly supported by UEFISCSU Romania through project ID_458/2007.