ORIGINAL SCIENTIFIC PAPER

CROATICA CHEMICA ACTA

Croat. Chem. Acta **2016**, *89*(4), 527–534
Published online: May 15, 2017
DOI: 10.5562/cca3117

HKD 1926

The Difference Between the Accuracy of Real and the Corresponding Random Model is a Useful Parameter for Validation of Two-State Classification Model Quality

Jadranko Batista,[1,*] Dražen Vikić-Topić,[2] Bono Lučić[2,#]

[1] University of Mostar, Faculty of Science and Education, Mostar, Bosnia and Herzegovina
[2] NMR Center, Ruđer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia
* Corresponding author's e-mail address: jadranko.batista@sve-mo.ba
# Corresponding author's e-mail address: lucic@irb.hr

RECEIVED: March 10, 2017 ⋆ REVISED: April 26, 2017 ⋆ ACCEPTED: April 26, 2017

PROCEEDING OF THE 28TH MATH/CHEM/COMP CONFERENCE, JUNE 20–25, 2016, DUBROVNIK, CROATIA
THIS PAPER IS DEDICATED TO PROF. NENAD TRINAJSTIĆ ON THE OCCASION OF HIS 80TH BIRTHDAY

Abstract: The simplest and the most commonly used measure for assess the classification model quality is parameter $Q_2 = 100\,(p + n)/N$ (%) named the classification accuracy, p, n and N are the total numbers of correctly predicted compounds in the first and in the second class, and the total number of elements of classes (compounds) in data set, respectively. Moreover, the most probable accuracy that can be obtained by a random model is calculated for two-state model by the formulae $Q_{2,\mathrm{rnd}} = 100\,[(p + u)\,(p + o) + (n + u)\,(n + o)]/N^2$ (%), where u and o are the total number of under-predictions (when class 1 is predicted by the model as class 2) and over-predictions (when class 2 is predicted by the model as class 1) in data set, respectively. Finally, the difference between these two parameter $\Delta Q_2 = Q_2 - Q_{2,rnd}$ is introduced, and it is suggested to compute and give ΔQ_2 for each two-state classification model to assess its contribution over the accuracy of the corresponding random model. When data set is ideally balanced having the same numbers of elements in both classes, the two-state classification problem is the most difficult with maximal $Q_2 = 100$ % and $Q_{2,\mathrm{rnd}} = 50$ %, giving the maximal $\Delta Q_2 = 50$ %. The usefulness of ΔQ_2 parameter is illustrated in comparative analysis on two-class classification models from literature for prediction of secondary structure of membrane proteins and on several quantitative structure-property models. Real contributions of these models over the random level of accuracy is calculated, and their ΔQ_2 values are compared mutually and with the value of ΔQ_2 (= 50 %) for the most difficult two-state classification model.

Keywords: classification model, Q_2 accuracy, overall classification accuracy, random classification accuracy, classification accuracy difference, correct class estimation, under-prediction, over-prediction, class imbalance, membrane structure modeling, QSAR classification modeling.

INTRODUCTION

In a two-state classification modeling one wants to develop a model for selected molecular property or activity (*y*-variable) using one or more input molecular attributes (descriptors, *i.e. x*-variables) which, for a molecule and to a certain accuracy, correctly estimates or predicts its property or activity class.

In estimating quality of two-state models the parameter Q_2 can be used, which is named as the classification accuracy (in %),[1] or as the percentage of all correct predictions.[2] The parameter Q_2 is the percent of correctly classified elements of the first (p) and of the second class (n) in the set having N elements belonging to one of two classes. If one reports Q_2 value for a two-state classification model of 90 % (or 95 %), it seems that the model is impressively accurate. However, the real level of model accuracy can be estimated if that Q_2 value is compared with the accuracy that can be obtained by a random model ($Q_{2,\mathrm{rnd}}$). It is evident that, in above mentioned case (*i.e.* $Q_2 = 90$ %) the real model contribution is significantly different if the most probable random accuracy is $Q_{2,\mathrm{rnd}} = 50$ %, or if it is $Q_{2,\mathrm{rnd}} = 70$ %. For each model, and also for structure-property models related to small molecules or proteins, it is possible to calculate (or to estimate by simulations) the level of accuracy which can be obtained by a random model which uses

randomized original data (variables), or purely random data (variables). When the model real accuracy (estimated by a statistical parameter) has been reported, another important value that has to be given is the value of the same statistical parameter for the corresponding random model which provides information on the level of chance (random) accuracy. In such a case two classical works that addressed this topic in analysis of correlations between variables are those by Topliss *et al*. published forty years ago for multivariate linear regression models.[3,4] It is obvious that some level of random correlation is present between each pair of variables and it is also demonstrated in that papers on the randomly generated variables.[3,4] Additionally, the analysis of several real models were performed and one recommendation (later often used in chemical structure-property modeling) is given about the maximal acceptable number of variables in the Multivariate Linear Regression (MLR) models. Namely, the authors estimated that the number of variables/descriptors involved in MLR models should not exceed 1/5 of the number of cases (molecules) used in data set.[4]

Random correlation (or accuracy) is higher for real than for random pairs of variables, because real variables have (typically) more monotonous distribution of values than the random ones. In addition, real variables have, as a rule, a real common background relation to some basic properties of constituting elements of data set. In case of data sets of chemical compounds or proteins used in modeling of activities, properties, or structural characteristics of proteins (like secondary structure or topology of membrane proteins), molecular descriptors derived from chemical structure are commonly related to basic properties of compounds (*e.g*. molecular weight, size, shape, the number of specific atoms, the number of bonds) or proteins (*e.g*. sequence length, the total number of some specific amino acid types, percentage of a secondary structure). Thus, to access the real level of random accuracy (or correlation) of a model, one must ensure that generated random data used in simulations have structure and distribution similar to those of real input data.

We present here the analysis and estimate of random accuracy for two-state classification problems, and compare real and random accuracies on several data sets related to (1) the modeling and prediction of secondary structure of membrane proteins, based on their primary structure, and (b) two-class properties of small molecules from the field of Quantitative-Structure-Activity Relationships (QSAR). On examples of real data sets we will analyse the influence of balance of numbers of elements in two classes (in experimental input data and in estimated / predicted data) on the random accuracy expressed by the Q_2 parameter.

THEORY

Definition of Secondary Structure of Membrane Proteins

In most simple classification problem only two classes of experimental properties or activities are defined. Secondary structure of membrane proteins is mostly determined by the parts of sequence interacting with membrane, that are in the secondary structure alpha (forming alpha-helix) or beta (forming beta-barrel), and the rest of sequence is usually considered (taken) to be in irregular secondary structure. In this study, we will validate two-state classification models on data sets of alpha-type (*i.e*. helical) integral membrane proteins, the largest class of membrane proteins. Namely, it is assumed that 20–30 % of sequenced genomes code for helical membrane proteins, but there are less than 1.2 % (~ 1370 proteins) of solved structures of helical or beta membrane proteins among ~ 120000 known (experimentally solved) protein structures deposited in Protein Data Bank.[5] For secondary structure of alfa-type membrane proteins it is commonly to define two classes of secondary structure of amino acids in protein sequence: (1) alpha secondary structure, containing one or more transmembrane segment(s) each consisting of (mostly) 19–21 neighbouring amino acids that form integral membrane alpha helix denoted by M, and (2) extra-membrane parts having secondary structure that is named as undefined, denoted by U.

Simplified scheme of experimental secondary structure of a membrane protein having 100 amino acid residues (amino acid in primary structure is designated by '–') and one transmembrane segment of 20 amino acids in primary structure is given in Scheme 1.

Contingency Table Definition

Comparing real (experimental) and predicted structures from Scheme 1 we can define the following parameters:

- p = positive correct prediction (real M predicted as M) = 15, (underlined both in real and predicted structures)
- u = underprediction (real M predicted as U) = 5
- n = negative correct prediction (real U predicted as U) = 75
- o = overprediction (real U predicted as M) = 5.

It is evident that $p + u + n + o = N = 100$ amino acids, and that $p + u = n(\mathrm{M}) = 20$, and $n + o = n(\mathrm{U}) = 80$. In this case we say that the prediction done by the model is balanced, because the model predicts the same numbers of M and U states (classes) as it is in experimental sequence. The prediction quality of two-class model can be also described by 2 × 2 contingency table given in Table 1.

A) Experimental protein secondary structure scheme

1----------50

UUUUUUUUUUMMMMMMMMMMMMMMMMMMMMMUUUUUUUUUUUUUUUUUUUUUU

51----------100

UUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUU

The total numbers of amino acids experimentally determined to be in states U and M are n(U) = 80 and n(M) = 20, respectively. Sequence length of protein: N = 100 amino acids, $N = n(U) + n(M)$.

B) Estimated protein secondary structure scheme - estimation done by a method (algorithm)

1----------50

UUUUUMMMMMMMMMMMMMMMMMMMMUUUUUUUUUUUUUUUUUUUUUU

51----------100

UUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUU

The total numbers of amino acids predicted by the model to be in states U and M are 80 and 20, respectively.

Scheme 1. Simplified view of experimental membrane protein secondary structure and secondary structure estimated or predicted by a balanced model.

Table 1. Contingency table for experimental and estimated (by a model) membrane protein secondary structures.

		(est/pred) M	(est/pred) U	Σ rows (experimental)
(A) Elements of a general contingency table				
(exp)	M	p	u	$p + u$
(exp)	U	o	n	$n + o$
Σ columns (estimated or predicted)		$p + o$	$n + u$	
(B) Contingency table obtained from experimental and estimated membrane protein structures from Scheme 1				
(exp)	M	15	5	20
(exp)	U	5	75	80
Σ columns (estimated or predicted)		20	80	

An ideal model would be the one with $u = 0$ and $o = 0$ ($N = p + n$), when all elements in both classes are correctly predicted.

Balanced Data Sets and Balanced Models

Real two-class data sets usually have different numbers of elements in both classes. However, in some cases it is possible to create an ideally *balanced experimental data set* with the same number of elements of classes ($p + u = n + o$). If possible, it is desirable to use balanced experimental data set in model development and optimization, because in that case both classes are equivalently treated during the model training, and one expects that characteristics of both classes will be evenly memorized by the model, *i.e*. evenly represented by the model's parameters.

Another case is *balanced set in estimation (or prediction)*, *i.e*. when the same numbers of classes are estimated by the model ($p + o = n + u$), and, in that case, it is not necessary that a totally balanced experimental set was used for model training.

Model balanced in estimation or prediction is the third concept introduced and used for models that conserve in estimation the total numbers of classes from experimental set which is used for model training. In that case we have both $p + u = p + o$ and $n + o = n + u$, what gives $u = o$. However, u and o do not need to be equal to zero, and it can be $p + u \neq n + o$ (for experimental set) or $p + o \neq n + u$ (for estimation of classes). Thus, a well performed modeling will normally end after the model achieves the balance between u and o in estimation on the training (or validation) set, and only in the case when $u = o$ it is possible to reach (in an ideal case) the maximal possible classification accuracy Q_2 of 100 %.

Real and Random Accuracies of a Model

Starting from contingency table, different statistical parameters are defined, used (and also cited) in scientific literature in estimating the model accuracy.[2,6] Parameter Q_2 [Eq. (1)], related to classification accuracy of a real model, is the simplest one that can be calculated from the contingency table:

$$Q_2 = 100\frac{p+n}{p+n+u+o}\ (\%) \qquad (1)$$

If a random model predicts ($p + o$) amino acids to be in class M for ($p + u$) experimentally determined amino acids in class M, and ($n + u$) = N – ($p + o$) amino acids to be in class U for ($n + o$) experimentally determined amino acids in class U, then the random accuracy $Q_{2,rnd}$ can be estimated as:

$$Q_{2,\text{random}} = Q_{2,\text{rnd}} = 100\left[\frac{(p+u)}{N}\frac{(p+o)}{N} + \frac{(n+o)}{N}\frac{(n+u)}{N}\right] (\%) \quad (2)$$

or shortly

$$Q_{2,\text{rnd}} = 100\frac{(p+u)(p+o)+(n+o)(n+u)}{N^2} (\%) \quad (3)$$

Also, this is the most probable value of Q_2 parameter that can be obtained by any random model.

If the experimental secondary structure of a protein (or a data set of proteins as a whole) contains the same number of M and U states/classes, then $p + u = n + o = N/2$ and we can factorize ($p + u$) in the numerator of [Eq. (3)]. Thus, for equally populated two states in experimental structure (*i.e.* for balanced experimental data set), one has [Eq. (4)]:

$$Q_{2,\text{rnd}} = 100\frac{(p+u)(p+o+n+u)}{N^2} = = 100\frac{N/2}{N} = 100\frac{1}{2} = 50\ (\%) \quad (4)$$

This means that the random value of Q_2 parameter is $Q_{2,rnd}$ = 50 %, for data with ideal balance of numbers of M and U states in experimental structure, regardless how big or small is the ratio between the numbers of M and U classes estimated (or predicted) by a model. Note that this also holds ($Q_{2,rnd}$ = 50 %) for ideally balanced estimation (or prediction) by the model, *i.e.* when $p + o = n + u = N/2$, regardless how large or small is the ratio of numbers of M and U classes in experimental structure.

If one obtains a balanced model which estimates (or predicts) in secondary structure of proteins the same numbers of states/classes M and U as in the experimental structure ($p + u = p + o$ and $n + o = n + u$), then $Q_{2,rnd}$ from [Eq. (3)] becomes $Q_{2,rnd–bal}$ in [Eq. (5)]:

$$Q_{2,\text{rnd}} = Q_{2,\text{rnd–bal}} = 100\frac{(p+u)^2+(n+o)^2}{N^2} (\%) \quad (5)$$

Equation (5) enable us to estimate the most probable random accuracy for balanced model that one plans to develop, and in that case $Q_{2,rnd}$ can be calculated only using experimental numbers of classes, *i.e.* here $p + u$ for state M (class 1) and $n + o$ for state U (class 2). It follows from [Eq. (5)] that for balanced model the minimal value of $Q_{2,rnd}$ is 50 %, when both classes are equally represented in experimental data set ($p + u = n + o$).

The Difference Between Real and Random Accuracies of a Model

Finally, the difference (in %) between the real model accuracy Q_2 and the corresponding random accuracy $Q_{2,rnd}$ is calculated by [Eq. (6)]:

$$\Delta Q_2 = Q_2 - Q_{2,\text{rnd}} (\%) \quad (6)$$

This value has its maximum of $(\Delta Q_2)_{max}$ = 50 % for balanced model estimation or prediction ($u = o$) when:

a) the maximal value of Q_2 = 100 %, and
b) the experimental data set is balanced having the same numbers of both classes (M and U for proteins, or class 1 and class 2, for general two-state QSAR model) when $Q_{2,rnd}$ = 50 %.

At the same time, the balanced model developed on such an experimental set of data is the most difficult problem for modeling (and analogous to the coin-tossing problem repeated N times, where $N = p + n + u + o$). Thus, the maximal range (Q_2 difference) for development and optimization of a model (*i.e.* our 'algorithm' is guessing) from the random level to the maximal level is 50 %. Any real two state classification model developed on the imbalanced experimental (training) data set with different total numbers of elements of two classes will have the difference ΔQ_2 between the real and random Q_2 accuracies smaller than 50 %.

RESULTS

Assume that for a sequence like the one from Scheme 1 (N = 100) with the 2 × 2 contingency table given in Table 1, containing 20 % amino acids in state/class M ($p + u$ = 20) and 80 % in state/class U ($n + o$ = 80) in experimentally determined structures, an optimized balanced model ($u = o$ = 5) estimates that 20 % amino acids are in state M ($p + o$ = 20) and 80 % in state U ($n + u$ = 80). For this two-class problem one can calculate random accuracy using [Eq. (3)] or [Eq. (5)], and the result is $Q_{2,rnd}$ = 68 %. For such a balanced model p = 15 and n = 75, and [Eq. (1)] gives that the value of the classification accuracy Q_2 is 90 %. Thus, the maximal model contribution above the most probable random estimate is ΔQ_2 = 90 % – 68 % = 22 %.

Knowing that the most difficult two-state classification problem having equal number of both classes in experimental set has $Q_{2,rnd}$ = 50 %, and that an ideally balanced model has the maximal classification accuracy Q_2 of 100 %, the maximal possible contribution of such a model ΔQ_2 is 50 % (Eq. (6)). Based on it, one can see that the maximal possible contribution of the model from Scheme 1 is significantly lower than it is for the most difficult two-state problem for which ΔQ_2 = 50 %. The real model contribution is from the random level of 68 % (which is primarily defined by the class imbalance in the experimental set used for model development because class M has 20 % and class U has 80 % elements), to the classification accuracy of Q_2.= 90 % obtained by the model estimation.

Analysis of Real and Random Accuracies of Models for Prediction of Membrane Proteins' Secondary Structure

We analysed random accuracies in several sets of membrane proteins from literature. In case when it was not possible to find *p*, *n*, *o* and *u* values for model predictions in literature (or calculate them from data given in published paper), but we had information on experimentally determined secondary structure, *e.g.* $p + u$ and N (the total number of amino acids), we calculate $Q_{2,rnd}$ by [Eq. (5)] assuming an ideal case, *i.e.* that the model is balanced ($p + u = p + o$ and, consequently, $n + o = n + u$). In cases of balanced estimation/prediction by a model, we denoted $Q_{2,rnd}$ as $Q_{2,rnd–bal}$ and used [Eq. (5)].

From Table 2 one can see that the most probable random classification accuracy for selected real data sets varies from 54 % to (even) 64 %, indicating a remarkable imbalance of the numbers of elements/states M and U of classes in experimental set. Values of $Q_{2,rnd–bal}$ (= $Q_{2,rnd}$) from Table 2 can be reduced to some extent by balancing

Table 2. Analysis of the most probably random level of Q_2 accuracy for data sets of membrane proteins based on the class-distribution of experimental data.[a]

Method	*N*	*p* + *u*	*n* + *o*	$Q_{2,rnd–bal}$ / %
Zhou and Zhou 2003, 73 proteins[7]	18399	6240	12159	55.17
Zhou and Zhou 2003, 79 proteins[7]	18471	6518	11953	54.33
Zhou and Zhou 2003, 147 proteins[7]	72598	23392	49206	56.32
Bernsel *et al.* 2008, low-res. set, 147 proteins[8],[b]	64074	15098	48976	63.98
Bernsel *et al*. 2008, high-res. set, 123 proteins[8],[b]	26971	9732	17239	53.87
Rost et al. 1995, 131 proteins[9]	32615	10130	22485	57.17

[a] $Q_{2,rnd-bal} = Q_{2,rnd}$ calculated by Eq. (5) for balanced model in estimation (under the assumption $u = o$, *i.e.* $p + u = p + o$ and $n + o = n + u$); N is the total number of amino acids in all sequences in data set of proteins; *p*, *n*, *u*, *o* are (respectively) the numbers of positive and negative correct predictions, and under-predictions and over-predictions, as explained in the manuscript.

[b] 'low-res' and 'high-res' are acronyms for two data sets of protein structures analysed in [8].

data set of membrane proteins. Because the numbers of positive (*p*) and negative (*n*) correct predictions are not separately reported in analysed manuscripts,[7–9] it was not possible to calculate neither Q_2 nor ΔQ_2 parameters. In any case, the contributions of models do not need to be counted as they starts from $Q_{2,rnd}$ = 50 % but from $Q_{2,rnd–bal}$, which is higher than 50 % for each of models presented in Table 2.

Analyses given in Table 3 include seven data sets from different versions of two methods (SPLIT[10,11] and TopPred_ΔG[13]) developed for prediction of secondary structure of membrane proteins.

The differences between real and random accuracies ΔQ_2 for methods and data sets from Table 3 are in the range between 24.5 % and 35 %, and are significantly lower than

Table 3. Real and random classification accuracies and their differences (all in %) of data sets used for development of methods for prediction of secondary structure of membrane proteins.[a]

Method	*N*	*p*	*n*	*u*	*o*	Q_2	$Q_{2,rnd}$	ΔQ_2	$Q_{2,rnd–bal}$
SPLIT 4.0, 52 proteins[10,11],[b]	11037	3809	5737	1066	425	86.49	51.36	35.13	50.68
SPLIT 2.0, 71 proteins[12]	28487	4502	22246	1060	679	93.90	69.39	24.51	68.57
SPLIT 2.0, 95 proteins[12]	43336	7747	33404	1078	1107	94.96	67.53	27.43	67.57
TopPred_ΔG(m-hr), 123 proteins[13],[c]	26971	8203	15687	1529	1552	88.58	53.85	34.73	53.87
TopPred_ΔG(s-hr), 123 proteins[13],[c]	26971	7906	15662	1826	1577	87.38	54.13	33.25	53.87
TopPred_ΔG(m-lr), 146 proteins[13],[c]	64074	12816	46294	2282	2682	92.25	63.65	28.60	63.98
TopPred_ΔG(s-lr), 146 proteins[13],[c]	64074	12342	46367	2756	2609	91.63	64.10	27.53	63.98

[a] *N, p, n, u, o,* are defined in the footnote (a) of Table 2; Q_2, $Q_{2,rnd}$, ΔQ_2, $Q_{2,rnd–bal}$ are calculated (as %) by [Eq. (1)], [Eq. (4)], [Eq. (6)] and [Eq. (5)], respectively.

[b] Data set of 52 proteins available on the SPLIT 4.0 server (http://splitbioinf.pmfst.hr/split/4/) was used.

[c] In the TopPred_ΔG method, the meanings of letters 'm', 's', 'hr' and 'lr' in acronyms m-hr, s-hr, m-lr, and s-lr are (respectively); 'm' -'multiple sequence information used in the method'; 's' - 'single sequence information used in the method'; 'hr' - 'high-resolution set of sequences', and 'lr' - 'low-resolution set of protein sequences'.

the maximal possible (*i.e.* ΔQ_2 = 50 %) which can be obtained only for the most difficult two-state classification model. Real accuracies ranged from 86.5 % to 94 %. Relatively small differences between $Q_{2,rnd}$ and $Q_{2,rnd-bal}$ suggest like the estimations by the models are balanced (*i.e.* $u = o$), but it is not entirely correct. Namely, the numbers u and o are much smaller then p and n, causing that the contributions of u and o to classification accuracies in [Eq. (3)] and [Eq. (5)] are overmatched by the contributions of p and n.

Analysis of Real and Random Accuracy of Classification Quantitative Structure-Activity Models

Quantitative structure-property/activity classification models have been often developed and applied in different subfields of chemistry like drug-design, environmental, physical or material chemistry. In Table 4 we give a set of two-state classification models from drug-design[14–18] and one from environmental science.[1]

One can see from Table 4 that the average contribution of developed models (i.e. computational methods) over the level of the most probable random accuracy ($Q_{2,rnd}$) measured by the ΔQ_2 parameter is higher than for previous models related to structure of membrane proteins (Table 3, the average of ΔQ_2 values is 30.2 %), and range from 26.9 % for model no. 6 to 47.2 % for model no. 12, with an average of 35.6 %. The main reason for higher ΔQ_2 values could be ascribed to possibility of selection of more balanced data sets with much closer numbers of elements of classes (high or low activities) in the field of QSAR modeling comparing with membrane protein data sets, in which the total number of elements of structure (class) U is considerably larger than for class M (compare $p + u$ and $n + o$ values in Tables 2 and 3). This disbalance in the numbers of M and U secondary structure states is defined by the length and by the nature of membrane protein sequences which contain (usually) more U than M secondary structure states, and in creating data set only complete sequences have to be selected (*i.e.* we cannot take only a part of a sequence into the data set, but the sequence as a whole).

For more balanced data sets $Q_{2,rnd}$ decreases, and from [Eq. (6)] it follows that ΔQ_2 will increase. This is also confirmed by the values of $Q_{2,rnd-bal}$ from Table 4, a parameter actually calculated by [Eq. (5)] from squares of frequencies of class 1 and class 2, which will be the lowest and equal to 50 % if frequencies of both classes are equal. Finally, the average of $Q_{2,rnd-bal}$ values from Table 4 is 53 %, and is lower that $Q_{2,rnd-bal}$ averages from Table 2 (56.8 %) and Table 3 (59.8 %).

It should be stressed that this comparative analysis of magnitudes of ΔQ_2 parameters for different models does not suggest anything about the level of significance of these models (*per se*). Namely, ΔQ_2 parameter calculated as the difference of two parameters ($Q_{2,rnd}$) and Q_2 will be more significant if each of two quality parameters used for its

Table 4. Real and random classification accuracy and their differences (all in %) of data sets of compounds used for development of quantitative structure-activity two-state classification models. [a]

No	Method	*N*	*p*	*n*	*u*	*o*	Q_2	$Q_{2,rnd}$	ΔQ_2	$Q_{2,rnd-bal}$
1	ANN anti-alergic model, training set[14],[b]	251	114	103	13	21	86.4	50.0	36.4	50.0
2	ANN anti-alergic model, test set[14],[b]	84	44	32	0	8	90.5	50.6	39.9	50.1
3	Anti-alergic LDA model, training set[14],[c]	330	155	136	14	25	88.2	50.0	38.2	50.1
4	Anti-alergic LDA model, test set[14],[c]	91	38	37	8	8	82.4	50.0	32.4	50.0
5	Trypanosomicidal activity model 1, train. set[15],[d]	346	101	203	19	23	87.9	54.3	33.5	54.7
6	Trypanosomicidal activity model 1, test set[15],[d]	94	20	61	3	20	86.2	59.2	26.9	63.0
7	ACE inhibition, model 1 (Eq. (17)), train. set[16],[e]	23	5	17	1	0	95.6	63.5	32.1	61.4
8	ACE inhibition, model 1 (Eq. (17)), test set[16],[e]	9	3	6	0	0	100.0	55.6	44.4	55.6
9	Drug-induced anorexia LDA model, train. set[17],[f]	122	36	61	9	16	79.5	51.9	27.6	53.4
10	Drug-induced anorexia LDA model, test set[17],[f]	55	16	29	4	6	81.8	52.7	29.1	53.7
11	Anthelmintic activity model, training set[18],[g]	273	137	109	7	20	90.1	50.4	39.7	50.2
12	Model 1 C.lyt Retention. training set[1],[h]	21	13	8	0	0	100.0	52.8	47.2	52.8

[a] *N, p, n, u, o,* are defined in the footnote (a) of Table 2; and Q_2, $Q_{2,rnd}$, ΔQ_2, $Q_{2,rnd-bal}$ in the footnote (a) of Table 3. The acronym 'train.' is for 'training'.
[b] 'ANN' is the acronym of 'Artificial Neural Network'.
[c] LDA is the acronym of 'Linear Discriminant Analysis'.
[d] Trypanosomicidal activity model 1 based on Linear Discriminant Analysis, given by [Eq. (1)] in [16] for the training and test sets.
[e] ACE is the acronym of Angiotesin-Converting Enzyme (ACE) inhibition. Model given by [Eq. (17)] is based on the Linear Discriminant Analysis.
[f] LDA is described in footnote (c). Model is given by [Eq. (2)] in [17].
[g] Anthelmintic activity LDA classification model given by [Eq. (10)] in [18].
[h] Model 1 is developed for modeling of adhesion of marine bacteria *Cellulophaga lytica* to polymer coating and mentioned in Table 3 in [1].

calculation will be more significant. The significance of a model (and also significance of model quality parameters) is primarily defined by the relation between the size of data sets (*i.e*. by the number of elements) used for training and by the number of optimized model parameters. Taking this into account one can conclude that models for prediction of structure of membrane proteins from Tables 2 and 3, which are based on much larger data sets, seem to be more significant than QSAR models from Table 4.[1]

CONCLUSION

Presented results show that the accuracy that can be obtained by a random model, is determined, to a large extent, (1) by the ratio of numbers of elements belonging to each of two classes in experimental input data (*i.e*. $(p + u)/(n + o)$), and (2) by the corresponding ratio of numbers of elements in two classes (*i.e*. $(p + o)/(n + u)$) estimated or predicted by the model. In both cases, optimal value is equal to 1, *i.e*. when both classes are equally populated (balanced). Finally, the balanced model for prediction of classes is the model which estimates or predicts the total numbers of elements in classes to be (almost) the same as those in experimental data, and, only such a model can reach (ideally) a maximal accuracy of 100 %.

For analysed models ΔQ_2 values were mostly between 25 % and 45 %, which is lower than the value of 50 % for the most difficult (coin-tossing) model for which the maximal Q_2 is 100 % and $Q_{2,\mathrm{rnd}}$ = 50 %, giving the maximal ΔQ_2 of 50 %. This is a useful parameter for estimation of the 'space' for improvement of models, what can be realized either by improvement of modeling procedures by increasing the Q_2 value, or by selection of more informative (and balanced) data sets by decreasing of $Q_{2,\mathrm{rnd}}$.

Because of simplicity of calculation and interpretation of parameter ΔQ_2, possibility of its calculation for models only from the frequencies of classes, as well as due to its usefulness through giving the information about the contribution (to accuracy, measured by Q_2) of real models over the corresponding random models, we suggest the use of parameter ΔQ_2 in reporting the quality of models, together with standard and frequently used Q_2 parameter, or other parameters used in the field, like Matthews correlation coefficient.[19] Additionally, the parameter ΔQ_2 can be calculated for a single protein sequence, for a set of protein sequences, and also for estimating the accuracy of a method in prediction of secondary structure of each of 20 amino acids on a single protein sequence or on a set of protein sequences (*e.g*. for alanine: Q_2, $Q_{2,\mathrm{rnd}}$ and ΔQ_2 are calculated for estimation of accuracy of prediction of secondary structure of all alanines in a protein sequence or for all alanines in a set of protein sequences – and analogously, it can be done for other 19 amino acids).[20]

Acknowledgment. The authors were supported by the Croatian Ministry of Science and Education through basic grant given to the Ruđer Bošković Institute and also by The Centre of Excelence for Marine Bioprospecting (BioProCro). The work of Jadranko Batista was additionally supported by the University of Mostar and by the grant of the Foundation of the Croatian Academy of Sciences and Arts. One of authors (Bono Lučić) stated that he was not involved in editorial work and decisions related to this manuscript. We thank reviewers, editors, and editorial staff for the time they spent in evaluation of our paper, and also reviewers for their constructive remarks that helped us to improve the manuscript.

REFERENCES

[1] B. Rasulev, F. Jabeen, S. Stafslien, B. J. Chisholm, J. Bahr, M. Ossowski, P. Boudjouk, *ACS Appl. Mater. Interfaces* **2017**, *9*, 1781.
[2] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, H. Nielsen, *Bioinformatics* **2000**, *16*, 412.
[3] J. G. Topliss, R. J. Costello, *J. Med. Chem.* **1972**, *15*, 1066.
[4] J. G. Topliss, R. P. Edwards, *J. Med. Chem.* **1979**, *22*, 1238.
[5] M. Punta, L. R. Forrest, H. Bigelow, A. Kernytsky, J. Liu, B. Rost, *Methods* **2007**, *41*, 460.
[6] D. M. W. Powers, *J. Machine Learning Techn*. **2011**, *2*, 37.
[7] H. Zhou, Y. Zhou, *Protein Sci.* **2003**, *12*, 1547.
[8] A. Bernsel, H. Viklund, J. Falk, E. Lindahl, G. von Heijne, A. Elofsson, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 7177.
[9] B. Rost, R. Casadio, P. Fariselli, C. Sander, *Protein Sci.* **1995**, *4*, 521.
[10] D. Juretić, L. Zoranić, D. Zucić, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 632.
[11] D. Juretić, A. Jerončić, D. Zucić, *Croat. Chem. Acta*, **1999**, *72*, 975.
[12] B. Lučić, N. Trinajstić, D. Juretić in *Chemical Topology to Three-Dimensional Geometry* (Ed: A. T. Balaban), New York, Plenum Press, **1997**, pp. 117–158.
[13] T. Hessa, N. M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S. H. White, G. von Heijne, *Nature* **2007**, *450*, 1026.
[14] R. Garcia-Domenech, R. Zanni, M. Galvez-Llompart, J. Vicente de Julian-Ortiz, *Comb. Chem. High T. Scr.* **2013**, *16*, 628.
[15] J. A. Castillo-Garit, O. del Toro-Cortés, M. C. Vega, M. Rolón, A. Rojas de Arias, G. M. Casañola-Martin, J. A. Escario, A. Gómez-Barrio, Y. Marrero-Ponce, F. Torrens, C. Abad, *Eur. J. Med. Chem.* **2015**, *96*, 238.
[16] J. A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, R. García-Domenech, V. Romero-Zaldivar, *J. Comput. Chem.* **2008**, *29*, 2500.

[17] M. Gálvez-Llompart, J. Gálvez, R. García-Domenech, L. B. Kier, *J. Chem. Inf. Model.* **2012**, *52*, 1337.
[18] Y. Marrero-Ponce, J. A. Castillo-Garit, E. Olazabal, H. S. Serrano, A. Morales, N. Castañedo, F. Ibarra-Velarde, A. Huesca-Guillen, A. M. Sánchez, F. Torrens, E. A. Castro, *Bioorg. Med. Chem.* **2005**, *13*, 1005.
[19] B. W. Matthews, *Biochim. Biophys. Acta* **1975**, *13*, 1005.
[20] D. Juretić, B. Lučić, N. Trinajstić, *J. Mol. Struct. (THEOCHEM)* **1995**, *338*, 43.