MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **65** (2011) 365-382

ISSN 0340 - 6253

Chor Coefficient – Measuring Correlation in Chemistry

Damir Vukičević

Department of Mathematics, University of Split, Nikole Tesle 12, HR-21000 Split, Croatia

(Received June 21, 2010)

Abstract

QSAR and QSPR researchers try to create models that are able to predict chemical properties or activities of chemical compounds. Very often Pearson product-moment correlation coefficient is used as the measure of fitting ability of such models. We argue that in many cases this is not realistic measure of model quality and propose an alternative measure: *chor (chemical correlation) coefficient* r_c. We illustrate this on the examples of models described in published papers and on the data sets proposed by International Academy of Mathematical Chemistry. Moreover, it is shown that all algorithms for optimization of r^2 are applicable for optimization of r_c^2 with minimal programming interventions.

Introduction

QSAR and QSPR researchers are interested in predicting properties and/or activities of chemical compounds from their structure. They are often using models based on molecular descriptors [1]. Model fitting ability is often measured by Pearson product-moment correlation coefficient. For the sake of brevity Pearson product-moment correlation coefficient is often called just Pearson correlation coefficient or even simpler – correlation coefficient.

Loosely speaking, correlation is a measure of association between variables [2]. Let (X_i, Y_i), $i = 1,...,N$ be the observed pairs of values such that neither X nor Y are constants. Correlation coefficient is defined by:

$$r(X,Y)=\frac{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2\cdot\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2}}.$$

It can be easily proved using Cauchy-Schwartz inequality that $-1\le r(X,Y)\le 1$. Values $r(X,Y)=\pm 1$ imply that there is linear function that transforms variable X to variable Y. Value $r=0$ implies that there is no correlation between X and Y whatsoever. Intermediate cases are covered by $r\in(-1,0)\cup(0,1)$.

In this paper, we shall demonstrate that correlation coefficient can be quite misleading measure of the quality of the model and present another measure *chor (chemical correlation) coefficient* r_c. Let us observe one-parameter linear model for prediction of molar heat capacity C_p by distance-reduced path-code-based molecular descriptor $D=\sum_i p_i^{1/2}/i$ on the set of 131 alkanes for which all data are provided in the Table 1 of paper [3]. The results are presented by the following figure:

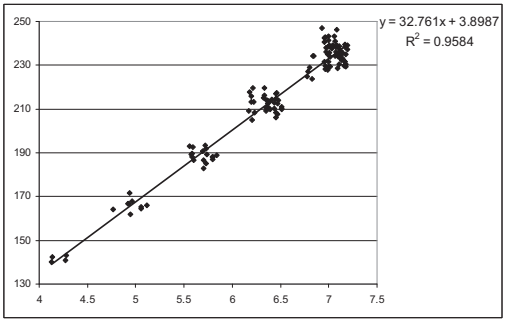

Figure 1. Estimation of molar heat capacity C_p by one parmeter linear model based on D.

At first site, this seems as great model ($r^2=0.9584$). However, as already commented in the paper [4], large correlation coefficient is primarily caused here by the fact that molar heat capacity C_p strongly depends on the number of vertices. Let $\overline{C_p}$ be obtained by calculating

the average for molecules with the same number of atoms. The following figure illustrates that $\overline{C_p}$ is better predictor of C_p then D:

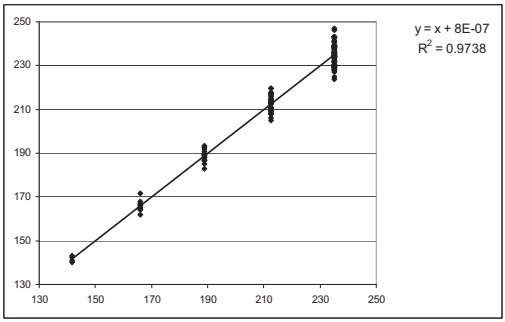

Figure 2. Estimation of molar heat capacity C_p by one parmeter linear model based on $\overline{C_p}$.

Hence, one gets better estimation just by taking the average of known isomers then by calculating molecular descriptor and using obtained linear model. This implies that model based on D-descriptor is not particularly good despite very high correlation coefficient. Therefore, we argue that correlation coefficient is not the optimal measure of the validity of models in chemistry and we propose the new measure *chor (chemical correlation) coefficient* that would amend this.

Definition of chor coefficient

The squared value of correlation coefficient r^2 is sometimes called coefficient of determination and it may be interpreted as proportion of variance in one variable accounted for by differences in the other [2]. Coefficient of alienation A is defined as $A = 1 - r^2$ [5,6]. Denote by y_i estimation of y_i using one-parameter linear model. Then, it can be shown that:

$$A = \frac{\sum_{i=1}^{N} \left(y_i - y_i \right)^2}{\sum_{i=1}^{N} \left(y_i - \overline{y} \right)^2} = \frac{RSS}{TSS},$$

where *RSS* stands for residual sum of squares and TSS for total sum of squares. Namely, coefficient of alienation is the ratio of the sum of squares of errors of estimate by y_i and sum of squares of errors of estimate by $\overline{y}$. It holds:

$$r^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - y_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}$$

and

$$r = \pm\sqrt{1 - \frac{\sum_{i=1}^{N}\left(y_i - y_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}}$$

where plus implies positive and minus implies negative correlation.

Both values $|r|$ and r^2 are high when A is low. Models with good fitting ability are those models that have value of A low, i.e. models in which y_is give much better estimation than $\overline{y}$.

In general, this has been shown as very useful way of measuring the quality of the model. However, in chemistry all molecules are very strongly characterized by two values: number of atoms and number of chemical bonds. Hence, if the set of molecules consists of molecules of different numbers of atoms and different numbers of chemical bonds, it may not be the optimal strategy to compare $\sum_{i=1}^{N}\left(y_i - y_i\right)^2$ with $\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2$. Namely, the squared-error $\left(y_i - \overline{y}\right)^2$ may be much larger than the squared error of the educated guess based on observing only molecules with the same number of atoms and bonds as the observed molecule. It is even more convenient to observe the difference of the number of bonds and atoms plus one (in order to have 0 for trees) then just the number of bonds. This value is called cyclomatic number or circuit rank and it is denoted by c [7,8]. Note that this two concepts are very similar, because two molecules coincide in b (number of bonds) and n (number of atoms) if and only if they coincide in c and n.

Note that, for instance, cycloalkanes $(c>0)$ are much different then alkanes $(c=0)$ with the same number of vertices. Hence, we shall characterize molecules with the number of atoms n and cyclomatic number c and write $n(i)$ and $c(i)$ for the number of atoms of molecule i and for the cyclomatic number of molecule i respectively.

Let us illustrate this by observing the set of polyaromatic hydrocarbons (this set is proposed by International Academy of Mathematical Chemistry [9] as one of the benchmark sets [10] for testing molecular descriptors). This set consists of 82 polyaromatic hydrocarbons which number of atoms goes from 9 to 40 and which cyclomatic number goes from 2 to 11. Let us observe the difference of estimating boiling point by the average of all these molecules and only of molecules with the same cyclomatic number and number of vertices (without taking in the account the observed molecule and of course without taking into the account molecules for which boiling point is not given in [10]). For the illustration purposes, the results for three of these molecules are presented by the following table:

	BP	average *ABP* of all molecules	average *ASBP* of molecules with the same value of c and n	$(BP-ABP)^2$	$(BP-ASBP)^2$
1,5-dymetilnaphtalene	269.00	347.17	262.50	6111.03	42.25
3-methylphenantrene	352.00	345.58	358.00	41.25	36.00
1-methylfluorene	318.00	346.23	328.00	796.97	100.00

Table 1. Estimations of boiling point by average of all molecules and by average of the similar molecules (molecules with the same number of vertices and the same cyclomatic number).

One can immediately see that there is a huge difference in the quality of estimation based just on averaging in the contrast to educated guess based simply on the observing molecules with the same cyclomatic number and number of atoms.

Hence, comparing the sum of the squared errors of some model to $TSS=\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2$ may not be the optimal strategy, since the value of $TSS=\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2$ may be deceivably high. Let

p be the observed property and i observed molecules. The obvious idea would be to use all molecules in $M_{n(i),c(i)} \setminus \{i\}$ and use the average of p on this set to estimate $y_i = p(i)$. The problem is if there is set $M_{n,c}$ consisting of only one molecule i, i.e. if $M_{n(i),c(i)} = \{i\}$. Then, $M_{n(i),c(i)} \setminus \{i\} = \varnothing$. Problem is how to determine the most similar molecules in this case. We want to find (one or more) non-empty set(s) $M_{n',c'}$ such that (n',c') is similar to (n,c). We need to define how to determine the similarity between two ordered pairs. We would like to make $(|n-n'|,|c-c'|)$ as small as possible. Of course, there are different ways to compare ordered pairs. We propose three types of orderings[1]:

1) ordering based on the sum of coordinates:

$$(u_1,u_2) \le (v_1,v_2) \text{ if and only if } u_1+u_2 \le v_1+v_2;$$

2) lexicographical order:

$$(u_1,u_2) \le (v_1,v_2) \text{ if and only if } (u_1 < v_1 \text{ or } (u_1 = v_1 \text{ and } u_2 \le v_2));$$

3) right-to-left lexicographical order:

$$(u_1,u_2) \le (v_1,v_2) \text{ if and only if } (u_2 < v_2 \text{ or } (u_2 = v_2 \text{ and } u_1 \le v_1)).$$

We use the first ordering to get the sets of the most similar molecules to eliminated molecule i, calculate their average and find the estimate $e_1(i)$. After performing this procedure for every molecule i, it can be calculated $DX_1 = \sum_{i \in M} (y_i - e_1(i))^2$. Similarly, we get e_2 and DX_2 using the second ordering and we get e_3 and DX_3 using the third ordering. Let us illustrate calculation of DX_1, DX_2 and DX_3 using the set of polyaromatic hydrocarbons observed above. In the following table cardinalities of the sets $M_{n,c}$ are given in the Table 2.

[1] Here the word ordering does not correspond to partial ordering, since ordering 1) does not satisfy antisymmetry.

Suppose that property p is known for every molecule in this set. Let $i \in M_{22,5}$. Then, $e_1(i), e_2(i)$ and $e_3(i)$ coincide and their value is the average of the values p of the remaining 6 molecules in $M_{22,5}$. On the other hand, let i be the only molecule in $M_{15,4}$. Then $e_1(i)$ is the average of the values of p of the molecules in $M_{15,3}$ and $M_{16,4}$; $e_2(i)$ is the average of the values of p of the molecules in $M_{15,3}$; and $e_3(i)$ is the average of the values of p of the molecules in $M_{16,4}$. Further, let $i \in M_{18,5}$. Then $e_1(i) = e_2(i)$ is the average of the values of p of the molecules in $M_{18,4}$; and $e_3(i)$ is the average of the values of p of the molecules in $M_{20,5}$.

n	c									
	2	3	4	5	6	7	8	9	10	11
9	2	0	0	0	0	0	0	0	0	0
10	2	0	0	0	0	0	0	0	0	0
11	2	0	0	0	0	0	0	0	0	0
12	7	2	0	0	0	0	0	0	0	0
13	3	2	0	0	0	0	0	0	0	0
14	0	7	0	0	0	0	0	0	0	0
15	0	7	1	0	0	0	0	0	0	0
16	0	7	2	0	0	0	0	0	0	0
17	0	0	6	0	0	0	0	0	0	0
18	0	0	5	1	0	0	0	0	0	0
19	0	0	2	0	0	0	0	0	0	0
20	0	0	0	6	0	0	0	0	0	0
21	0	0	0	1	0	0	0	0	0	0
22	0	0	0	7	3	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	1	1	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	1	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	1	0
33	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	1

Table 2. Cardinalities of the sets $M_{n,c}$.

Let us further analyze values $TSS, DX_1, ..., DX_3$. From $\sum_{i=1}^{N}\left(y_i - y_i\right)^2 \le TSS$, it follows that

$$A = \frac{\min\left\{\sum_{i=1}^{N}\left(y_i - y_i\right)^2, TSS\right\}}{TSS},$$

i.e. coefficient of alienation A measures the portion of the error of estimation of DX_0 that still remains after model y_i is produced. Here, we may assume that TSS, DX_1, DX_2 and DX_3 are measures of the errors of estimates that can be obtained without modeling, hence it is reasonable to define *chemical coefficient of alienation* A_c by

$$A_c = \frac{\min\left\{\sum_{i=1}^{N}\left(y_i - y_i\right)^2, TSS, DX_1, DX_2, DX_3\right\}}{\min\left\{TSS, DX_1, DX_2, DX_3\right\}}.$$

Naturally, *chemical coefficient of determination* is defined by

$$r_c^2 = 1 - A_c = 1 - \frac{\min\left\{\sum_{i=1}^{N}\left(y_i - y_i\right)^2, TSS, DX_1, DX_2, DX_3\right\}}{\min\left\{TSS, DX_1, DX_2, DX_3\right\}}.$$

Further *chemical correlation (chor) coefficient* for one parametrical linear models is defined by

$$r_c = \pm\sqrt{1 - \frac{\min\left\{\sum_{i=1}^{N}\left(y_i - y_i\right)^2, TSS, DX_1, DX_2, DX_3\right\}}{\min\left\{TSS, DX_1, DX_2, DX_3\right\}}},$$

where r_c has the same sign as r. More precisely,

$$r_c = \operatorname{sgn}(r)\cdot\sqrt{1 - \frac{\min\left\{\sum_{i=1}^{N}\left(y_i - y_i\right)^2, TSS, DX_1, DX_2, DX_3\right\}}{\min\left\{TSS, DX_1, DX_2, DX_3\right\}}}$$

Calculation of chor coefficient

One of the reasons of the great popularity of the correlation coefficient is the simplicity of the determination of the coefficients in multilinear models for estimation of properties. Namely, if one has predictors $t_1, t_2, ..., t_k$ and wants to find linear function that estimates property p $\left(p \approx a_0 + a_1 t_1 + a_2 t_2 + ... + a_k t_k\right)$, it is sufficient to solve linear system of $k+1$ equations in

$k+1$ variables that can be done quite efficiently. This is very important, because sometimes we want to check large number of models. Suppose that we want to find (using exhaustive search) the best linear model using 4 predictors out of 100 proposed predictors. We should check $\binom{100}{4} = 3912250$ models. It the number of descriptors is larger and we are interested in models with greater number of parameters, then the exhaustive search becomes intractable. Suppose that we want to find 6-parameter model out of 1000 proposed descriptors. Then, there are $\binom{1000}{6} \approx 1.36 \cdot 10^{15}$ possible combinations. In this case various algorithms (variable neighborhood search, genetic algorithms, greedy algorithms, and so on) are used. Having in mind large theory and lot of existing algorithms that are made to optimize correlation coefficient r, coefficient r_c is developed in such way that it satisfies two important properties:

1) If μ_1 and μ_2 are two different models for estimating property p on the same set of molecules M. Then $r^2(\mu_1) \geq r^2(\mu_2)$ implies $r_c^2(\mu_1) \geq r_c^2(\mu_2)$.
2) r_c can be calculated from M, p and r.

Hence, we can apply all the algorithms for finding maximal r^2 value and then obtain r_c from r, M and p by simple and very quick calculation. Let us explain how to calculate a_c and r_c from M, p and r. It holds:

$$A_c = \frac{\min\left\{\left(y_i - y_i\right)^2, TSS, DX_1, DX_2, DX_3\right\}}{\min\left\{TSS, DX_1, DX_2, DX_3\right\}} == \frac{\min\left\{\frac{\left(y_i - y_i\right)^2}{TSS}, 1, \frac{DX_1}{TSS}, \frac{DX_2}{TSS}, \frac{DX_3}{TSS}\right\}}{\min\left\{1, \frac{DX_1}{TSS}, \frac{DX_2}{TSS}, \frac{DX_3}{TSS}\right\}}$$

$$= \frac{\min\left\{1-r^2, 1, \frac{DX_1}{TSS}, \frac{DX_2}{TSS}, \frac{DX_3}{TSS}\right\}}{\min\left\{1, \frac{DX_1}{TSS}, \frac{DX_2}{TSS}, \frac{DX_3}{TSS}\right\}} = \frac{\min\left\{1-r^2, \frac{DX_1}{TSS}, \frac{DX_2}{TSS}, \frac{DX_3}{TSS}\right\}}{\min\left\{1, \frac{DX_1}{TSS}, \frac{DX_2}{TSS}, \frac{DX_3}{TSS}\right\}};$$

and for the one parametric linear models

$$r_c = \operatorname{sgn}(r) \cdot \sqrt{1 - \frac{\min\left\{1 - r^2, \frac{DX_1}{TSS}, \frac{DX_2}{TSS}, \frac{DX_3}{TSS}\right\}}{\min\left\{1, \frac{DX_1}{TSS}, \frac{DX_2}{TSS}, \frac{DX_3}{TSS}\right\}}}.$$

Note that TSS, DX_1, DX_2, DX_3 depend solely on M and p and they do not depend on the observed model. Further, let

$$\phi = \min\left\{1, \frac{DX_1}{DX_0}, \frac{DX_2}{DX_0}, \frac{DX_3}{DX_0}\right\},$$

Then last two formulas can be rewritten as:

$$A_c = \frac{\min\left\{1 - r^2, \phi\right\}}{\phi};$$

$$r_c = \operatorname{sgn}(r) \cdot \sqrt{1 - \frac{\min\left\{1 - r^2, \phi\right\}}{\phi}}.$$

Definition of chor coefficient for trees

In many cases all molecules in the observed set are just trees (molecules with $c = 0$). One such set is set of 18 octane isomers proposed in [10]. Here, the procedure outlined above is not optimal, because $c = 0$ for all molecules and it is not a discriminatory parameter. Hence, in this situation, we propose to use number of leaves l (pendant vertices or vertices of degree 1) instead of c and following completely analogous procedure to define DX_1, DX_2 and DX_3. Let us define by $T_{n,l}$ set of trees with n vertices and l leaves and let us illustrate the calculation of e_1, e_2 and e_3 on the set of alkanes observed in the paper [11]. Cardinalities of $T_{n,l}$ are presented in the following table:

n	l				
	2	3	4	5	6
3	1	0	0	0	0
4	1	1	0	0	0
5	1	1	1	0	0
6	1	2	2	0	0
7	1	3	4	1	0
8	1	4	8	4	1

Table 3. Cardinalities of the sets $T_{n,l}$.

Suppose that property p is known for every molecule in this set. Let $i \in T_{7,3}$. Then, $e_1(i), e_2(i)$ and $e_3(i)$ coincide and their value is the average of the values p of the remaining 2 molecules in $T_{7,3}$. On the other hand, let i be the only molecule in $T_{5,3}$. Then $e_1(i)$ is the average of the values of p of the molecules in $T_{4,3}, T_{6,3}, T_{5,2}$ and $T_{5,4}$; $e_2(i)$ is the average of the values of p of the molecules in $T_{5,2}$ and $T_{5,4}$; and $e_3(i)$ is the average of the values of p of the molecules in $T_{4,3}$ and $T_{6,3}$.

Then, $TSS, ..., DX_3$ are calculated completely analogously as above and similarly values a_c and r_c are obtained. As above, all algorithms for maximizing standard correlation can be easily adopted to this situation.

Applications of chor coefficient

Let us start with the very simple question: Is it simpler to predict boiling point of octane isomers or to predict boiling point of alkanes in general? One should expect that correct answer would be that it is easier to make predictions in more specific setting, i.e. that it is easier to predict properties of octane isomers. It is well known that Randić index is very useful for prediction of the boiling point of alkanes. Here, we analyze data from paper [11]. In the following table, we present the values of r^2 and r_c^2 when different families of alkanes are taken under consideration:

alkanes	number of molecules	r^2	ϕ	r_c^2
octane isomers	18	0.67177	0.79025	0.58465
heptane and octane isomers	27	0.89426	0.14019	0.24574
hexane, heptane and octane isomers	32	0.90514	0.06569	0.00000
propane, butane, pentane, hexane, heptane and octane isomers	38	0.95577	0.05239	0.15571

Table 4. Estimation of boiling points of alkanes by Randić number.

It can be readily seen that correlation coefficient r^2 is highest when the largest family of alkanes is observed. On the other hand, the correlation coefficient is the lowest in the specific case of the octane isomers. This is very counter-intuitive.

On the other hand, r_c^2 behaves much more according to our intuition. It has the highest value for octane isomers. It has reasonably high value for heptane and octane isomers. It has value 0 for hexane, heptane and octane isomers detecting malfunctioning of linear correlation when molecules are clustered in three clusters (linear function usually can not optimize three cluster centers). Finally, very low value is for all 38 observed alkanes.

Contrary to the low value of r_c^2 for 38 alkanes, the value $r^2 = 0.95577$ is misleadingly high and may lead us to conclusion that one-parameter model is enough. However, value of r_c^2 provides the argumentation to try to use bi-parametric linear model. In the next table, we present the results for bi-parametric linear model using Randić number and number of (carbon) atoms:

alkanes	number of molecules	r^2	ϕ	r_c^2
octane isomers	18	0.67177	0.79025	0.58465
heptane and octane isomers	27	0.94470	0.14019	0.60554
hexane, heptane and octane isomers	32	0.94876	0.06569	0.22002
propane, butane, pentane, hexane, heptane and octane izomers	38	0.97924	0.05239	0.60365

Table 5. Estimation of boiling points of alkanes by Randić number and number of atoms.

We can immediately see that values of r_c significantly increased and that they are relatively stable. Hence, new model is much better then model that used only Randić number. Further, we can see that r_c^2 provides important information by observing values for the second and the third family of alkanes. Note that r^2 values are almost the same. However, it is much easier to make educated guess about the third family of alkanes and hence there is a significant difference between r_c^2 in these two families of molecules.

Now, let us analyze the data given in paper [3]. Beside already mentioned D descriptor, the following descriptors have been analyzed in that paper:

- W, Wiener index, half-sum of all entries d_i in the graph distance matrix, $W = \frac{1}{2} \sum_{u,v \in V(G)} d_{uv}$;
- J, the average distance-based molecular connectivity, $J = \frac{|E(G)|}{c+1} \cdot \sum_{uv \in E(G)} (D_u \cdot D_v)^{1/2}$;
- Q, the quadratic path-code-based molecular descriptor, $Q = \sum_i p_i^2$;
- S, the **s**quare-root path-code-based molecular descriptor, $S = \sum_i p_i^{1/2}$;
- A, the distance-attenuated path-code-based molecular descriptor, $A = \sum_i p_i / i$;
- P, the path count molecular descriptor, $P = \sum_i p_i^{1/2} / i^{1/2}$,

where $V(G)$ is the set of vertices (atoms) of G, $E(G)$ is the set of edges (bonds) of G, d_{uv} is distance between vertices u and v, p_i is the number of paths of length i, and $D_u = \sum_{v \in V(G)} d_{uv}$.

In the same paper, beside C_p the following properties have been analyzed:

- density at 25ºC- $\rho(kg/m^3)$,
- refractive index at 25ºC - $n_{D^{25}}$,
- Gibbs energy of formation in gaseous state- $\Delta_f G^0_{300}$, in kJ / mol
- Vaporization enthalpy - ΔH^{300}_{vap}, in kJ / mol
- normal boiling point - $NBP(^0C)$

Coefficients r^2 of one-parameter linear models are given by the following table:

descriptor	property					
	$C_p\left(J/K\cdot mol\right)$	$\rho\left(kg/m^3\right)$	$n_{D^{25}}$	$\Delta_f G^0_{300}$	ΔH^{300}_{vap}	$NBP\left({}^0C\right)$
W	0.887	0.483	0.528	0.410	0.926	0.844
J	0.272	0.635	0.602	0.679	0.069	0.285
Q	0.785	0.854	0.870	0.878	0.540	0.786
S	0.802	0.381	0.423	0.287	0.915	0.770
D	0.958	0.648	0.689	0.549	0.905	0.925
A	0.936	0.812	0.842	0.763	0.752	0.913
P	0.900	0.509	0.552	0.415	0.939	0.865

Table 6. Coefficients r^2 of one-parameter linear models

As discussed in paper [4] some of this coefficients are misleadingly high. The quality of the models will be much better expressed by r_c^2 which is presented in the following table:

descriptor	property					
	$C_p\left(J/K\cdot mol\right)$	$\rho\left(kg/m^3\right)$	$n_{D^{25}}$	$\Delta_f G^0_{300}$	ΔH^{300}_{vap}	$NBP\left({}^0C\right)$
W	0.000	0.000	0.000	0.000	0.000	0.000
J	0.000	0.000	0.000	0.000	0.000	0.000
Q	0.000	0.446	0.406	0.563	0.000	0.000
S	0.000	0.000	0.000	0.000	0.000	0.000
D	0.000	0.000	0.000	0.000	0.000	0.000
A	0.000	0.290	0.279	0.150	0.000	0.000
P	0.000	0.000	0.000	0.000	0.000	0.000

Table 7. Coefficients r_c^2 of one-parameter linear models.

One can see that in fact there are only several models with the significant fitting ability. Now, let us analyze bi-parameter models that use number of atoms and one of these descriptors. Coefficients r^2 are given by the following table:

descriptor	property					
	$C_p\left(J/K\cdot mol\right)$	$\rho\left(kg/m^3\right)$	$n_{D^{25}}$	$\Delta_f G^0_{300}$	ΔH^{300}_{vap}	$NBP\left({}^0C\right)$
W	0.974	0.888	0.895	0.838	0.928	0.944
J	0.974	0.892	0.899	0.868	0.941	0.943
Q	0.974	0.855	0.875	0.888	0.929	0.943
S	0.974	0.862	0.875	0.861	0.938	0.943
D	0.974	0.821	0.839	0.840	0.913	0.943
A	0.974	0.843	0.864	0.844	0.934	0.943
P	0.974	0.873	0.883	0.865	0.939	0.943

Table 8. Coefficients r^2 of two-parameter linear models (one parameter being number of atoms)

One can see that correlations are now (as expected) somewhat higher. Let us observe coeffitients r_c^2:

descriptor	property					
	$C_p\left(J/K\cdot mol\right)$	$\rho\left(kg/m^3\right)$	$n_{D^{25}}$	$\Delta_f G^0_{300}$	ΔH^{300}_{vap}	$NBP\left({}^0C\right)$
W	0.089	0.576	0.521	0.417	0.000	0.034
J	0.090	0.593	0.541	0.526	0.000	0.028
Q	0.096	0.453	0.430	0.597	0.000	0.028
S	0.090	0.479	0.429	0.500	0.000	0.021
D	0.092	0.323	0.267	0.425	0.000	0.021
A	0.095	0.407	0.380	0.440	0.000	0.022
P	0.090	0.519	0.466	0.517	0.000	0.022

Table 9. Coefficients r_c^2 of two-parameter linear models (one parameter being number of atoms)

From the last table, it can be seen that models for $\rho\left(kg/m^3\right)$, $n_{D^{25}}$ and $\Delta_f G^0_{300}$ are pretty good, while models for $C_p\left(J/K\cdot mol\right)$, ΔH^{300}_{vap} and $NBP\left({}^0C\right)$ are not so good. Note that estimation of C_p by all 7 bi-parameter models is assumed not to be too good although r^2 values are above 0.97. This shows that r_c^2 is indeed new measure substantially different from r^2.

In order to verify the importance of the results given in Table 9, we perform the following test. For each of these models we use leave-one-out method and calculate the coefficient of determination r_1^2. Further, the analogous calculation is made using just the number of vertices and coefficient of determination r_2^2 is obtained. In the following table we present $\left(1-r_1^2\right)/\left(1-r_2^2\right)$ in order to compare RSSs of estimates:

descriptor	property					
	$C_p\left(J/K\cdot mol\right)$	$\rho\left(kg/m^3\right)$	$n_{D^{25}}$	$\Delta_f G^0_{300}$	ΔH^{300}_{vap}	$NBP\left({}^0C\right)$
W	1.017	0.421	0.456	0.480	0.554	1.002
J	1.017	0.409	0.441	0.394	0.445	1.011
Q	1.015	0.550	0.551	0.333	0.539	1.011
S	1.020	0.519	0.546	0.412	0.470	1.018
D	1.023	0.676	0.702	0.474	0.654	1.020
A	1.016	0.595	0.598	0.460	0.504	1.018
P	1.019	0.481	0.513	0.399	0.460	1.019

Table 10. Ratios $\left(1-r_1^2\right)/\left(1-r_2^2\right)$.

We can immediately see that the results in the first and the last column suggest that adding the descriptor does not contribute to the accuracy of the estimate (as expected from the results in Table 9) and the second, third and fourth column suggest that adding a descriptor contributes to the accuracy of the estimate (as expected from the results in Table 9). The results in the fifth column are surprising. However, it seems that here the value of cyclomatic number plays an important role. In order, to show that it is so, we perform the following calculation. For each of these models we add cyclomatic number as parameter, use leave-one-out method and calculate the coefficient of determination r_3^2. Further, the analogous calculation is made using just the number of vertices and cyclomatic number and coefficient of determination r_4^2 is obtained. In the following table we present $\left(1-r_3^2\right)/\left(1-r_4^2\right)$ in order to compare RSSs of estimates:

descriptor	property					
	$C_p\left(J/K\cdot mol\right)$	$\rho\left(kg/m^3\right)$	$n_{D^{25}}$	$\Delta_f G^0_{300}$	ΔH^{300}_{vap}	$NBP\left({}^0C\right)$
W	0.998	0.433	0.514	0.556	1.008	0.792
J	1.001	0.325	0.427	0.345	1.001	0.752
Q	1.029	0.627	0.652	0.336	1.003	0.844
S	1.015	0.527	0.609	0.381	1.019	0.857
D	1.031	0.795	0.841	0.564	1.027	0.940
A	1.024	0.681	0.711	0.512	0.998	0.895
P	1.016	0.460	0.553	0.359	1.013	0.838

Table 11. Ratios $\left(1-r_3^2\right)/\left(1-r_4^2\right)$.

One can see that columns that give good results in both tables are only second, third and fourth column, which is in accordance with the results given in Table 9.

Finally, let us return to the study of benchmark data set of polycyclic hydrocarbons [10]. We analyze four classes of models:

Models A: one parameter linear models consisting of one benchmark descriptor;

Models B: two parameter linear models consisting of one benchmark descriptor and number of atoms;

Models C: two parameter linear models consisting of one benchmark descriptor and cyclomatic number;

Models D: three parameter linear models consisting of one benchmark descriptor, number of atoms and cyclomatic number.

In each of these classes we find the model with the highest correlation coefficient. We present our results in the following table:

property	class of models	highest value of r^2	highest value of r_c^2
melting point	A	0.74424	0.25753
	B	0.75583	0.29119
	C	0.77945	0.35975
	D	0.77962	0.36023
boiling point	A	0.97978	0.44581
	B	0.98140	0.49021
	C	0.98322	0.54019
	D	0.98406	0.56315
octanol-water partition coefficient (LogP)	A	0.94128	0.49056
	B	0.94271	0.50297
	C	0.94951	0.56190
	D	0.96109	0.66244

Table 12. Analyes of the models A, B, C and D.

As expected, the results are the lowest for the melting point and it is well known that it can not be predicted well by this kind of descriptors.

Conclusions

In this paper, we have present new measure of the fitting ability of the model - *chemical correlation (chor)*. We compare it to Pearson correlation coefficient and illustrate its advantages. Also, we show that it is strongly connected with Pearson correlation coefficient and that all algorithms for optimization of r can be applied to optimize r_c.

Acknowledgment

The partial support of Croatian Ministry of Science, Education and Sport (grants no. 177-0000000-0884 and 037-0000000-2779) is gratefully acknowledged. The author thanks anonymous referee whose suggestions improved the quality of this paper.

References

[1] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.

[2] J. L. Rodgers, W. A. Nicewander, *Amer. Statistician* **42** (1988) 59–66.

[3] A. T. Balaban, A. Beteringhe, T. Constantinescu, P. A. Filip, O. Ivanciuc, *J. Chem. Inf. Model.* **47** (2007) 716–731.

[4] D. Vukičević, A. Beteringhe, T. Constantinsceu, M. Pompe, A. T. Balaban, *Chem. Phys. Lett*. **464** (2008) 155–159.

[5] S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*, Wiley, Hoboken, 2005.

[6] B. S. Everitt, *The Cambridge Dictionary of Statistics*, Cambridge Univ. Press, New York, 2006.

[7] http://mathworld.wolfram.com/CircuitRank.html

[8] D. Babić, D. J. Klein, I. Lukovits, S. Nikolić, N. Trinajstić, *Int. J. Quantum Chem.* **90** (2002) 166–176.

[9] http://www.iamc-online.org/

[10] http://www.moleculardescriptors.eu/dataset/dataset.htm

[11] S. Nikolić, G. Kovačević, A. Miličević, N. Trinajstić, *Croat. Chem. Acta* **76** (2003) 113–124.