Activity Characterization of Triazines Analogues: Statistical Parameters for Models Assessment

Sorana Daniela BOLBOACĂ & Lorentz JÄNTSCHI
"Iuliu Hatieganu" & Technical Universities, Cluj-Napoca, Romania
To analyzed three previously reported Molecular Descriptors Family (MDF) on Structure-Activity Relationships (SAR) models by the use of correlation coefficients:

- Pearson ($r_{Prs}$)
- Spearman ($\rho_{Spm}$)
- Semi-quantitative ($r_{SQ}$)
- Kendall’s tau ($\tau_{Ken,a}$, $\tau_{Ken,b}$, $\tau_{Ken,c}$)
- Gamma ($\Gamma$)
1,3,5-substituted-triazines: thirty compounds
- inhibition activity on Chorella: $pI_{50} =$ concentration required for 50% inhibition of Hill reaction

Previously reported QSAR

$\text{Est } pI_{50} = 9.614 - 0.153 \cdot X_5 - 58.888 \cdot 1/V_5 - 2.430 \cdot 1/N_3$

[Diudea at all., 2002]

- $n = 30$
- $r^2 = 0.9694$
- $F = 274.3$
- $r_{loo} = 0.9778$

where

$X_5 =$ topological descriptor for substituent number 5

$V_5 =$ fragmental volumes of the substituent in the position 5 (cm$^3$/mol)

$N_3 =$ total number of hydrogen’s in the substituent 3
MDF SAR Methodology

- Preparing chemical compounds for molecular modeling
- Generating the molecular descriptors family
- Finding
- Validating
- Comparing

\{ \text{the MDF SAR models} \}
Methods

- Correlation coefficients, Squared correlation coefficient
- Statistical test and associated significance

- Pearson
- Spearman
- Semi-quantitative

\[
\begin{align*}
\text{Correlation coefficients} & : \\
r_{Prs} &= \frac{\sum (Y_{m-i} - \bar{Y}_m)(Y_{est-i} - \bar{Y}_{est})}{\sqrt{\left(\sum (Y_{m-i} - \bar{Y}_m)^2\right)\left(\sum (Y_{est-i} - \bar{Y}_{est})^2\right)}} \\
r_{Spm} &= \frac{\sum (R_{Y_{m-i}} - \bar{R}_m)(R_{Y_{est-i}} - \bar{R}_{est})}{\sqrt{\left(\sum (R_{Y_{m-i}} - \bar{R}_m)^2\right)\left(\sum (R_{Y_{est-i}} - \bar{R}_{est})^2\right)}} \\
r_{sq} &= \sqrt{\frac{\sum (Y_{m-i} - \bar{Y}_m)(Y_{est-i} - \bar{Y}_{est})}{\sqrt{\left(\sum (Y_{m-i} - \bar{Y}_m)^2\right)\left(\sum (Y_{est-i} - \bar{Y}_{est})^2\right)}}} \cdot \sqrt{\frac{\sum (R_{Y_{m-i}} - \bar{R}_m)(R_{Y_{est-i}} - \bar{R}_{est})}{\sqrt{\left(\sum (R_{Y_{m-i}} - \bar{R}_m)^2\right)\left(\sum (R_{Y_{est-i}} - \bar{R}_{est})^2\right)}}}
\end{align*}
\]
Methods

- More correlation coefficients (!):
  \[ n(n-1)/2 \text{ pairs}, \ t = \text{number of tied } Y_m \text{ values and} \]
  \[ u = \text{number of tied } Y_{est} \text{ values}, \ C = \text{concordances}, \ D = \text{discordances} \]

  - **Kendall:**
    - \( \tau_{Ken,a} \): \( \tau_{Ken,a} = \frac{(C-D)}{[n(n-1)/2]} \)
    - \( \tau_{Ken,b} \): \( \tau_{Ken,b} = \frac{(C-D)}{\sqrt{[n(n-1)/2-t][n(n-1)/2-u]]}} \)
    - \( \tau_{Ken,c} \): \( \tau_{Ken,c} = \frac{2(C-D)}{n^2} \)

  - **Gamma:**
    - \( \Gamma = \frac{(C-D)}{(C+D)} \)
Results for MDF SARs for $5.52 - 8112.2 \cdot iSMMWHg + 194.35 \cdot iSMmEQt$ (2v structure vs. $pI_{50}$, from [Bolboacă & Jäntschi, LEJPT (DOAJ), 2006])

- Gamma: 0.7243
- Kendall tau c: 0.6889
- Kendall tau b: 0.7192
- Kendall tau a: 0.7127
- Semi-Quantitative: 0.9234
- Spearman: 0.8652
- Pearson: 0.9855
## Results for MDF SARs

<table>
<thead>
<tr>
<th>Method</th>
<th>95% CI</th>
<th>( r_{P_{rs}}^2 = 0.9712 )</th>
<th>( t_{P_{rs,1}} = 30.71^{+} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>( r_{Prs} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spearman</td>
<td>( \rho_{Sp_{m}} )</td>
<td>( \rho_{Sp_{m}}^2 = 0.7485 )</td>
<td>( t_{Sp_{m,1}} = 9.13^{+} )</td>
</tr>
<tr>
<td>semi-Quantitative</td>
<td>( r_{sQ} )</td>
<td>( r_{sQ}^2 = 0.8526 )</td>
<td>( t_{sQ} = 12.73^{+} )</td>
</tr>
<tr>
<td>Kendall a</td>
<td>( \tau_{Ken,a} )</td>
<td>( \tau_{Ken,a}^2 = 0.5079 )</td>
<td>( Z_{Ken,ta} = 5.53^{+} )</td>
</tr>
<tr>
<td>Kendall b</td>
<td>( \tau_{Ken,b} )</td>
<td>( \tau_{Ken,b}^2 = 0.5173 )</td>
<td>( Z_{Ken,tb} = 5.54^{+} )</td>
</tr>
<tr>
<td>Kendall c</td>
<td>( \tau_{Ken,c} )</td>
<td>( \tau_{Ken,c}^2 = 0.4746 )</td>
<td>( Z_{Ken,tc} = 5.35^{+} )</td>
</tr>
<tr>
<td>Gamma</td>
<td>( \Gamma )</td>
<td>( \Gamma^2 = 0.5246 )</td>
<td>( Z_{\Gamma} = 4.07^{+} )</td>
</tr>
</tbody>
</table>

\( ^{+} \) p-value less than 0.0001

\[
\text{min}(Y) \leq \text{HM}(Y) \leq \text{GM}(Y) \leq \text{AGM}(Y) \leq \text{AM}(Y) \leq \text{EM}(Y) \leq \text{max}(Y)
\]
Results for MDF SARs

for $1.74 - 9261 \cdot iSMMWHg + 10.34 \cdot iAMdEHg + 3.89 \cdot INDRLQg$

(3v structure vs. $pl_{50}$, from [idem])
## Results for MDF SARs

<table>
<thead>
<tr>
<th>Method</th>
<th>95%CI r_{Prs} [0.9752-0.9944]</th>
<th>r_{Prs} = 0.9768</th>
<th>t_{Prs,1} = 34.35$^\dagger$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman</td>
<td>95%CI $p_{Spm} = 0.8185-0.9568$</td>
<td>$p_{Spm} = 0.8288$</td>
<td>$t_{Spm,1} = 11.64$ $^\dagger$</td>
</tr>
<tr>
<td>semi-Qntitative</td>
<td>95%CI $r_{sQ} = 0.8937-0.9754$</td>
<td>$r_{sQ} = 0.8998$</td>
<td>$t_{sQ} = 15.85$ $^\dagger$</td>
</tr>
<tr>
<td>Kendall a</td>
<td>95%CI $\tau_{Ken,a} = 0.5854-0.8906$</td>
<td>$\tau_{Ken,a} = 0.6002$</td>
<td>$Z_{Ken,ta} = 6.01$ $^\dagger$</td>
</tr>
<tr>
<td>Kendall b</td>
<td>95%CI $\tau_{Ken,b} = 0.5854-0.8906$</td>
<td>$\tau_{Ken,b} = 0.6099$</td>
<td>$Z_{Ken,tb} = 6.02$ $^\dagger$</td>
</tr>
<tr>
<td>Kendall c</td>
<td>95%CI $\tau_{Ken,c} = 0.5321-0.8734$</td>
<td>$\tau_{Ken,c} = 0.5608$</td>
<td>$Z_{Ken,tc} = 5.82$ $^\dagger$</td>
</tr>
<tr>
<td>Gamma</td>
<td>95%CI $\Gamma = 0.5931-0.8930$</td>
<td>$\Gamma = 0.6171$</td>
<td>$Z_{\Gamma} = 4.79$ $^\dagger$</td>
</tr>
</tbody>
</table>

$^\dagger$ p-value less than 0.0001
Results for MDF SARs

for $5.75 + 199 \cdot iSMMmEQt - 9010 \cdot iSMMWHg - 0.071 \cdot LADmkQt + 2.86 \cdot \text{INPRJQg}$ (2v structure vs. $pI_{50}$, from \cite{ibidem})

- **Gamma**: 0.7809
- **Kendall tau c**: 0.7444
- **Kendall tau b**: 0.7764
- **Kendall tau a**: 0.7701
- **semi-Qantitative**: 0.9469
- **Spearman**: 0.9034
- **Pearson**: 0.9925
Results for MDF SARs

Plot:

- Cross-validation leave-one-out correlation score (more correlation coefficient !!):
  \[ r^2_{cv-loo} = 0.9849 \]

- Steiger’s Z test (vs. previously reported QSAR) = 2.828 (p-value < 0.05)
### Results for MDF SARs

<table>
<thead>
<tr>
<th>Method</th>
<th>95% CI $r_{Prs}$</th>
<th>$r_{Prs}^2$ = 0.9850</th>
<th>$t_{Prs,1} = 42.85$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>[0.9841-0.9964]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spearman</td>
<td>$\rho_{Spm}$ [0.8051-0.9533]</td>
<td>$\rho_{Spm}^2 = 0.8162$</td>
<td>$t_{Spm,1} = 11.15$</td>
</tr>
<tr>
<td>semi-Quantitative</td>
<td>$r_{sQ}$ [0.8903-0.9746]</td>
<td>$r_{sQ}^2 = 0.8967$</td>
<td>$t_{sQ} = 15.59$</td>
</tr>
<tr>
<td>Kendall a</td>
<td>$\tau_{Ken,a}$ [0.5671-0.8848]</td>
<td>$\tau_{Ken,a}^2 = 0.5931$</td>
<td>$Z_{Ken,ta} = 5.98$</td>
</tr>
<tr>
<td>Kendall b</td>
<td>$\tau_{Ken,b}$ [0.5776-0.8881]</td>
<td>$\tau_{Ken,b}^2 = 0.6028$</td>
<td>$Z_{Ken,tb} = 5.98$</td>
</tr>
<tr>
<td>Kendall c</td>
<td>$\tau_{Ken,c}$ [0.5248-0.8710]</td>
<td>$\tau_{Ken,c}^2 = 0.5542$</td>
<td>$Z_{Ken,tc} = 5.78$</td>
</tr>
<tr>
<td>Gamma</td>
<td>$\Gamma$ [0.5852-0.8905]</td>
<td>$\Gamma^2 = 0.6098$</td>
<td>$Z_\Gamma = 4.73$</td>
</tr>
</tbody>
</table>

† p-value less than 0.0001
Concluding Remarks

- All seven methods for appreciating of the correlation between measured and estimated herbicidal activity of studied triazines analogues had statistical significance (see tables) → any of can serve for biological activity estimators evaluating (!)
Concluding Remarks

- If ... is considered as being a quantitative variable, then *Pearson* correlation coefficients is the statistical parameter that can be use in evaluation of relationship between the measured activity an the activity estimated by models.
- If ... is considering as being a qualitative variable (only the relative ordering can be assumed to be reproductible) then Spearman (or Kendall ...)
- If ... is considered as being a semi-quantitative variable, then proposed semi-Quantitative r is the best SAR evaluator.
Acknowledgements:

- The research was partly supported by UEFISCSU Romania through projects ET36/2005 & ET108/2006.
- A. v. Humboldt Foundation
- Organizers of I.S.O.C.
All are online (!!):
http://vl.academicdirect.org/molecular_topology
/mdf_findings/
MDF (Demo) Calculator
MDF SAR Predictor
Leave One Out Analysis
MDF Investigator
Training vs. Test Experiment
/mdf_findings/rank/
[presented methods]

Thank you for your attention!

http://sorana.academicdirect.ro
&
http://lori.academicdirect.org
Activity Characterization of Triazines Analogues: Statistical Parameters for Models Assessment

Sorana Daniela BOLBOACĂ¹, Lorentz JÄNTSCHI²

¹ „Iuliu Hațieganu“ University of Medicine and Pharmacy, 13 Emil Isac Street, 400023 Cluj-Napoca, Romania. E-mail: sbolboaca@umfcluj.ro
² Technical University of Cluj-Napoca, Romania, 15 Constantin Daicoviciu Street, 400020 Cluj-Napoca, Romania

Abstract

Correlation coefficients and associated squared values are used as parameters in validation of structure-activity relationship models. By using of the molecular descriptors family on structure-activity relationship method [1] the herbicidal activity of a sample of triazines analogues was modelled [2]. A number of three multivariate models proved to have estimation and prediction abilities [2].

Starting from the hypothesis that the measured activity of triazines analogue is a semi-quantitative variable, the aim of the research was to analyzed the three previously reported models by using the Pearson, Spearman, Kendall’s and Gamma correlation coefficients. The structure-activity relationship models were previously reported [2]. The measured herbicidal activity of triazines analogues and the value estimated by the previously reported models were investigated by using the Pearson, Spearman, Kendall's and Gamma correlation coefficients ($r_{Prs}$, $\rho_{Spm}$, $\tau_{Ken,a}$, $\tau_{Ken,b}$, $\tau_{Ken,c}$, $\Gamma$) and associated squared correlation coefficient ($r_{Prs}^2$, $\rho_{Spm}^2$, $\tau_{Ken,a}^2$, $\tau_{Ken,b}^2$, $\tau_{Ken,c}^2$, $\Gamma^2$).

The results of investigation, express as correlation coefficients and associated 95% confidence intervals, squared correlation coefficient and Student’s t, respectively the parameter of the Z test were calculated and analyzed.

The correlation coefficients obtained with all methods were statistical significant ($p < 0.0001$). The correlation coefficients vary according with the model as follow:

- Model with two descriptors: from $r = 0.6889$ ($\tau_{Ken,c}$) - 95%CI [0.4370-0.8405] to $r = 0.9855$ ($r_{Prs}$) - 95% CI [0.9694-0.9931];
- Model with three descriptors: from 0.7489 ($\tau_{Ken,c}$) - 95%CI [0.5321-0.8734] to the value equal with 0.9883 ($r_{Prs}$) - 95% CI [0.9752-0.9944];
Model with four descriptors: from 0.7444 ($\tau_{Ken,c}$) - 95% CI [0.5248-0.8710] to the value equal with 0.9925 ($r_{Prs}$) - 95% CI [0.9841-0.9964].

If there is considered that the herbicidal activity of triazines analogues is a quantitative variable, the Pearson correlation coefficients is the statistical parameter that must be used in evaluation of relationship between the measured and estimated activity.

Considering the measured activity of triazines analogues as a semi-quantitative variable, a rank correlation coefficient is the statistical parameter able to provide more reliable estimation rather than Pearson correlation coefficient. However, which is the proper rank correlation coefficients that can be use in this circumstance? The comparisons between the rank correlation coefficients are discussed and the further plan of research is highlighted.

Keywords: Triazines Analogues, Statistical Models, Correlation Coefficients

Acknowledgements
The research was partly supported by UEFISCSU Romania through projects ET36/2005 & ET108/2006.

References