Implementation of multilinear regression in inhibition of thyroid hormone receptor-coactivator

A S Devi¹, D Sarwinda¹ and G F Hertono¹
¹Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia

Abstract. Thyroid Hormone Receptors (TRs) play a big role in our body. Upon its nature TRs need coactivator to bind. MSNBAs had claimed to be the inhibitor of thyroid hormone receptor-coactivator interaction. Ninety-five targeted has been obtained by in vivo which classified to active and unspecified compounds. In this study, the active compounds will be tested ex vivo using multilinear regression method to explain the relationship of the physicochemical properties and biological activity – that is Quantitative Structure-Activity Relationship (QSAR) model and to show that the model is an analogue to the synthesis of MSNBAs. It appears that MSNBAs is fit to be an inhibitors of thyroid hormone receptor-coactivator with R-sq > 0.95.

1. Introduction

TRs belong to nuclear hormone receptor superfamily, which regulate development, growth, and metabolism. [¹] TR’s activity is stimulated in vivo by thyroid hormone (T3). [²] TRs contain three functional domains which are an amino terminal transcription activation domain (AF-1), a central DNA binding domain (DBD), and a carboxyl terminal ligand binding domain (LBD) that includes a T3-inducible coactivator binding domain, AF-2. [³] TRs have been classified to two classes, TRα (TRα1 and TRβ1 and TRβ2), based on chromosomal localization and amino acid homology. [⁴] TRα1 is more expressed in bone, gastrointestinal tract, cardiac and skeletal muscle, and the central nervous system. [⁵] The faulty of TRα makes the thyroid gland can’t make enough thyroid hormone to keep the body running normally, causing hypothyroidism. [⁶] On [⁷], two scaffolds, β-aminoketones and methylsulfonylnitrobenzoates (MSNB) act as antagonists of coactivator binding to TRs by competing with NR boxes for binding to the receptor. Though the liabilities make the compounds unstable and give side effect on cardiac ion channel activity. It leads to another experiment on MSNBs with replacement the ester with an amide linkage. The result show that 5 compounds (<50µm) were active to inhibit VDR-SRC2-3, 13 compounds (<50µm) were likely to inhibit the interaction between TRα-SRC2-2, and 5 compounds (<20µm) were active to be an inhibition of TRβ-SRC2-2. [⁷]

However, in vivo experiments are very expensive and time consuming. At this thought, applied statistic in regression analysis has received attention in various science disciplines. Regression Analysis is performed to determine the correlation between two or more variables having cause-effect relation, to make predictions for the topic by using the relation. It answers some questions such as “is there any relation between dependent and independent variables”, “if there are any, what is the power of the relation”, “is it possible to make future-oriented predictions regarding the dependent variable”, and so on. [⁸] Regression using one independent variable is called univariate...
regression analysis while using more than one independent variable is called multivariate regression analysis. The regression model with one dependent variable and more than one independent variable is called multilinear regression or multiple linear regression. 

Chemistry, for example, in early and late drug discovery we can use the interaction of drugs with their biological counterparts are determined by intermolecular forces, i.e. by hydrophobic, polar, electrostatic, and steric interactions. QSAR derives models which describe the structural dependence of biological activities either by physicochemical parameters (Hansch analysis), by indicator variables encoding different structural features (Free Wilson analysis), or by three-dimensional molecular property profiles of the compounds (comparative molecular field analysis, CoMFA). QSAR, based on Hansch Analysis is the same as multilinear regression. The physicochemical parameters are group to hydrophobic, polarizability, electrostatic, and steric interactions an analogue to predictors in regression. \( \log \frac{1}{IC_{50}} \) is the logarithm of the inverse molar dose that produces or prevents a certain biological response. In this case, \( \log \frac{1}{IC_{50}} \) acts as a dependent parameter or response. The independent parameters are: Log P as the logarithm of the n-octanol/water partition coefficient P, presenting hydrophobic; Molar Refractivity (MR), a polarizability parameter which is used to bind of the substituents to polar surface; Electronic properties describe the influence of a certain group or substituent on electron density distribution, one of them is Hammett Constant (\( \sigma \)); and finally, steric parameter (\( E_s \)) mostly use the van der Waals, occurs from geometric consideration or linear free energy relationship \(^{11,12,13}\).

2. Method

We used multilinear regression to interpret the interaction between biological activity and its physicochemical parameters. It has been successfully found that lipophilicity makes a significant contribution to medicinal chemistry. It offered more comprehensive framework to optimize general absorption, distribution, metabolism and excretion properties, toxicology profiles and ultimately pharmacological response. Refer to QSAR predictors, Ligand Lipophilicity Efficiency (LLE) combines Log P and electrostatic which makes it hard to solve specific problems with \(^{14,15,16}\). Ligand Efficiency (LE) assesses the binding affinity in relation to the number of heavy atoms in a molecule \(^{17}\). It is like steric parameter on Hansch analysis. Binding Efficiency Index (BEI) and Surface Efficiency Index (SEI) are introduced after the Ligand Efficiency discoveries become hits. These two indices reduce the number of predictors by combining potency with molecular weight and polar surface area – as replacement to polarizability. It is suggested that these indices, either individually or in combination, are useful for effective and efficient drug discovery \(^{18}\). We used n (numbers of compound), R-sq (Coefficient of Determination), S (Standard Error of the Regression), F-value (Fischer Statistic) and P-value to evaluate the model.

3. Result

All data is acquired from PubChem AID 738019, AID 738020, AID 738021 (https://pubchem.ncbi.nlm.nih.gov/) and previous paper. Four parameters are used in this model; SEI (Surface Efficiency Index), BEI (Binding Efficiency Index), LE (Ligand Efficiency) and LLE (Ligand Lipophilic Efficiency). The response is \( \log \frac{1}{IC_{50}} \) represent the biological activity and four parameters as the independent variables or predictors. We use QSAR model by Hansch to find interaction between
\[ \log \frac{1}{IC_{50}} \] and four predictors; SEI, BEI, LE, and LLE. We execute this in statistical software, Minitab 17.

**Table 1. Biochemical assay data of Inhibition of VDR**

| SID       | \( \log \frac{1}{IC_{50}} \) | SEI   | BEI   | LE    | LLE  |
|-----------|-------------------------------|-------|-------|-------|------|
| 164135898 | -1.57978                      | 3.21  | 8.92  | 0.17  | 1.28 |
| 164138757 | -1.60206                      | 3.19  | 10.15 | 0.20  | 1.98 |
| 164138759 | -0.84510                      | 3.66  | 9.50  | 0.19  | 2.28 |
| 164147066 | -1.56820                      | 3.22  | 9.95  | 0.20  | 2.50 |
| 164147067 | -1.64345                      | 2.97  | 9.69  | 0.19  | 2.45 |

The equation of Inhibition of VDR:

\[ \log \frac{1}{IC_{50}} = -9.114 + 1793 \text{SEI} + 1744 \text{BEI} - 85.81 \text{LE} + 0.6324 \text{LLE} \quad (1) \]

\( n = 5, R^{2} = 100\% \)

**Table 2. Biochemical assay data of Inhibition of TR\( \alpha \)**

| SID       | \( \log \frac{1}{IC_{50}} \) | SEI   | BEI   | LE    | LLE  |
|-----------|-------------------------------|-------|-------|-------|------|
| 164130247 | -0.27875                      | 0.00415 | 0.01364 | 0.27 | 3.55 |
| 164130253 | -1.4624                       | 0.00329 | 0.0089 | 0.17 | 1.08 |
| 164133096 | -0.23045                      | 0.00419 | 0.01198 | 0.23 | 2.88 |
| 164135896 | -1.64345                      | 0.00316 | 0.01179 | 0.24 | 3.72 |
| 164135897 | -1.51851                      | 0.00325 | 0.01169 | 0.24 | 3.59 |
| 164135898 | -1.51851                      | 0.00325 | 0.00904 | 0.17 | 1.34 |
| 164138756 | -1.5682                       | 0.00322 | 0.01093 | 0.22 | 2.88 |
| 164138757 | -1.50515                      | 0.00326 | 0.01037 | 0.2  | 2.07 |
| 164141530 | -0.62325                      | 0.0039  | 0.01374 | 0.27 | 4.08 |
| 164147065 | -0.98227                      | 0.00342 | 0.01152 | 0.23 | 3.36 |
| 164147066 | -1.64345                      | 0.00316 | 0.00978 | 0.19 | 2.43 |
| 164147067 | -1.39794                      | 0.00314 | 0.01024 | 0.2  | 2.69 |
| 164147068 | -1.4624                       | 0.00329 | 0.01037 | 0.21 | 2.78 |

The equation of MSNBA as inhibitor of TR\( \alpha \):

\[ \log \frac{1}{IC_{50}} = -5.491 + 1364SEI + 144BEI - 11.7LE + 0.195LLE \quad (2) \]

\( n = 13, S = 0.116716, R^{2} = 96.56\%, R^{2}(\text{adj}) = 94.84\%, \text{PRESS} = 0.239842, R^{2}(\text{pred}) = 92.43\% \)

**Table 3. Biochemical assay data of Inhibition of TR\( \beta \)**

| SID       | \( \log \frac{1}{IC_{50}} \) | SEI   | BEI   | LE    | LLE  |
|-----------|-------------------------------|-------|-------|-------|------|
| 164130247 | -0.81954                      | 0.00321 | 0.00892 | 0.24 | 3.01 |
| 164133096 | -0.5563                       | 0.00319 | 0.01015 | 0.22 | 2.55 |
| 164133099 | -1.02531                      | 0.00366 | 0.0095 | 0.21  | 2.74 |
| 164141530 | -0.76343                      | 0.00322 | 0.00995 | 0.26 | 3.94 |
| 164147065 | -1.20412                      | 0.00297 | 0.00969 | 0.22 | 3.14 |

The equation of Inhibition of TR\( \beta \):

\[ \log \frac{1}{IC_{50}} = -3.301 + 1016SEI + 223.1BEI - 19.41LE + 0.1869LLE \quad (3) \]
n = 5, R-sq = 100%

### 3.1 The goodness of fit

The goodness of fit gives the expression on how well a model fits given set of data, or how well it will predict a future set of observations. There is various way to test the goodness of the model. R-sq measures the proportion of the total variation in \[ \log \frac{1}{IC_{50}} \] explained by the model. R-sq is said to be “good” if R-sq > 60%. Therefore, eq. 1, eq. 2 and eq. 3 show the model with R-sq 100%, 96.56% and 100% are good and can tell the relation between predictors and response. For eq. 2, it gives us more than just R-sq. such as R-sq(adj). R-sq(pred). and S. R-sq(adj) shows the contribution of predictors in the model. It explains that the four predictors in eq. 2 contribute 94.84% of total variance, which is good. R-sq(pred) expresses how well the model will adjust in the future with new observations. It gives 92.43% to overall model, S (Standard Error of the Regression), the smaller error we get, the better the model. The error is 0.116716 which means the average distance of the data from fitted line is 12%.

Table 4. Analysis of Variance of the Model (Eq. 2)

|       | F-Value | P-Value |
|-------|---------|---------|
| Constant | 56.11   | 0.000   |
| SEI    | 33.84   | 0.000   |
| BEI    | 0.33    | 0.582   |
| LE     | 1.01    | 0.344   |
| LLE    | 1.67    | 0.232   |

### 3.2 The Significance of the Model

The significance of the eq. 2 presents us the F-Value and P-Value. It uses to see which parameters in particular are significant predictors of the response. The F-Value from the ANOVA should be smaller than F-value on F-table with the judgement of \( \alpha = 0.01 \). From the table. \( F_{\alpha,\alpha} = 7.01 \), which make BEI, LE and LLE are the significant predictors. We should consider removing SEI although the P-Value < \( \alpha \). However, the overall performance of model is shown by F-Value (56.11%) is greater than the P-Value. So, we can state that the model is fit better than removing the SEI.

The equation with three predictors which are BEI, LE and LLE :

\[
\log \frac{1}{IC_{50}} = -5.251 + 1173BEI - 38.1LE - 0.234LLE \
\]

\( n = 13, S = 0.251641, R-sq = 82.00\%, R-sq(adj) = 76.00\%, \text{PRESS} = 1.10066, R-sq(pred) = 65.24\% \)

The idea of removing SEI is seen to be bad decision. R-sq to these three predictors is smaller than the previous model. Neither the R-sq(adj), which decrease to 18.84% below, it means the loss of the model is facing caused by removing SEI. It applies also in R-sq(pred) and S. The model barely fits in the future with new observations which is affected by the decrease of R-sq(pred) and the error we get is greater than before. Therefore, we can be sure from previous statement that the model is fit better with four predictors which are SEI, BEI, LE and LLE.

From eq. 2 we can interpret the model to:

- For every 1 \( \mu \)m of SEI increase, it produces/prevents 1364 \( \mu \)m of biological responses.
- For every 1 \( \mu \)m of BEI increase, it decreases 144 \( \mu \)m of molar dose that produces/prevents biological responses.
• For every 10 points of LE increase, it decreases 117 points of biological responses.
• For every 1000 points of LLE increase, it increases 195 points of biological responses.

4. Conclusion
We can conclude that multilinear regression can be used to explain the relationship of physicochemical and biological activity between MSNBAs as inhibitor of thyroid hormone receptors (TRs). In this study, we present the multilinear regression model that we have are an analogue to the result of synthesis of MSNBAs. It gives us an overall model R-sq > 95%.

References
[1] Cheng S Y, Leonard JL and Davis PJ 2010 Molecular Aspects of Thyroid Hormone Actions PubMed
[2] Fraichard A et al 1997 The T3R Alpha Gene Encoding a Thyroid Hormone Receptor is Essential for Post-Natal Development and Thyroid Hormone Production EMBO Journal p 4412-20
[3] Harvey C B and Williams G R 2002 Mechanism of Thyroid Hormone Action PubMed
[4] Mangelsdorf D J et al 1995 The Nuclear Receptor Superfamily: The Second Decade PubMed
[5] Visser W E, Friesema F C and Visser T J 2011 Minireview: Thyroid Hormone Transporters: The Knowns and The Unknowns Molecular Endocrinology p 1 - 14
[6] www.thyroid.org/hypothyroidism
[7] Hwang J Y 2013 Synthesis and Evaluation of Methylsulfonylnitrobenzamides (MSNBAS) as Inhibitors of the Thyroid Hormone Receptor-Coactivator Interaction Bioorg Med Chem Lett
[8] Uyanik G K and Guler N 2013 A Study on Multiple Linear Regression Analysis Procedia p 234 - 40
[9] Tabachnick B G and Fidell S L 1996 Using Multivariate Statistics 3rd edition (Harper Collins College Publisher, New York)
[10] Kubinyi H 2008 QSAR: Hansch Analysis and Related Approaches 1 ed R Mannhold et al (Wienhem, Germany) p 1 – 47
[11] Hansen O R 1962 Hammett Series with Biological Activity Acta Chemica Scandinavica p 1593 - 1600
[12] Hansch C, Maloney P, Fujita T and Muir R M 1962 Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients Nature p 178 - 80
[13] Hansch C and Fujita T 1964 ρ-σ-π Analysis, A Method for the Correlation of Biological Activity and Chemical Structure Journal of the American Chemical Society p 1616-26
[14] Waring M J 2010 Lipophilicity in Drug Discovery Expert Opinion on Drug Discovery p 235 – 48
[15] van de Waterbeemd H 2009 Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability Vol 40 ed van de Waterbeemd H and Testa B (Wiley-VCH Verlag GmbH and Co, Weinheim, Germany) p 71 – 99
[16] Arnott J A and Planey S L 2012 The Influence of Lipophilicity in Drug Discovery and Design Expert Opinion on Drug Discovery p 863 – 75
[17] Hopkins et al, Groom C R and Alex A 2004 Ligand Efficiency: A Useful Metric for
Lead Selection Science Direct 9 p 430 – 1

[18] Abad-Zapatero C and Metz J T 2005 *Ligand Efficiency Indices as Guideposts for Drug Discovery* Science Direct 10 p 464 - 9