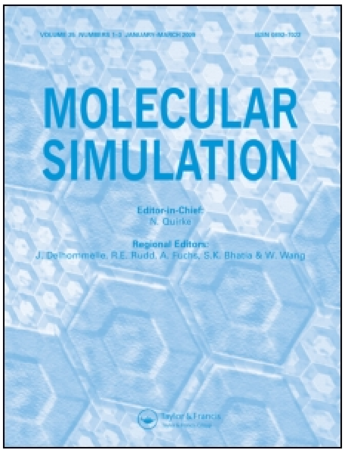

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Comparison of neural network and regression models in molecular simulation of relationships between chemical structure and biological activity

Peter Mager; Robert Reinhardt

Online publication date: 03 June 2010

To cite this Article Mager, Peter and Reinhardt, Robert(2002) 'Comparison of neural network and regression models in molecular simulation of relationships between chemical structure and biological activity', Molecular Simulation, 28: 3, 287 — 294

To link to this Article: DOI: 10.1080/08927020290014385
URL: http://dx.doi.org/10.1080/08927020290014385

Molecular Simulation, 2002, Vol. 28(3), pp. 287 – 294

COMPARISON OF NEURAL NETWORK AND REGRESSION MODELS IN MOLECULAR SIMULATION OF RELATIONSHIPS BETWEEN CHEMICAL STRUCTURE AND BIOLOGICAL ACTIVITY

PETER P. MAGER* and ROBERT REINHARDT

Research Group of Pharmacochemistry, Institute of Pharmacology and Toxicology of the University, D-04107 Leipzig, Härtelstr. 16-18, Saxony, Germany

(Received December 2000; accepted January 2001)

The effect of substitution on the phenyl ring of orally active β-lactam inhibitors (3,3'-diethyl-2-azetidinones) of human leucoctyte elastase was examined using quantitative structure-activity relationship (QSAR) models. Tradionallly used QSARs are ordinary least-squares multiple regression analysis (Hansch), nonleast-squares (NLS) and partial-least squares (PLS) regression models. The results were compared with that of backpropagation (BP) and generalized-regression genetic-neural network (GRNN) approaches. It was found that a Hansch analysis can only be used with reservation due to the multicollinearity of the physicochemical descriptors. The problem of multicollinearity using Hansch analysis is circumvented by PLS and NLS regression. BP did not improve the goodness of fit. GRNN has the ability to approximate structure-activity functions with satisfactory accuracy.

Keywords: Quantitative-structure activity relationships; QSAR; Ordinary least-squares multiple regression analysis; Hansch approach; Nonleast-squares regression; Partial-least squares regression; Backpropagation network; Generalized-regression genetic-neural network; 3,3'-diethyl-2-azetidinones; Leucoctyte elastase

1. INTRODUCTION

Inhibitors of serine human leukocyte elastase might be useful in therapy of certain systemic diseases, for example, rheumatoid arthritis. As lead structure, orally active β-lactam inhibitors (3,3'-diethyl-2-azetidinones) were used

*Corresponding author. e-mail: magp@server3.medizin.uni-leipzig.de

ISSN: 0892-7022 © 2002 Taylor & Francis Ltd DOI: 10.1080/08927020290014385

[1]. To optimize the lead structure, quantitative structure-activity relationships (QSARs) are a help. Traditionally used approaches are multiple regression or Hansch-Fujita analysis [2, 3], partial-least squares (PLS) regression [4, 5], and nonleast-squares (NLS) regression of the MASCA model [6]. Alternatively, neural network analyses [7] are applicable. Among them, backpropagation (BP) and generalized-regression genetic-neural network (GRNN) are useful in medicinal chemistry [7 – 9]. The goal of this study is a comparison of the various techniques. As decision criterion, the goodness-of-fit (squared multiple regression coefficient between biological activity and physicochemical descriptors) is used.

2. METHOD

2.1. Synthesis

The synthetic route of the compounds (Fig. 1, Tab. I) was described previously [1], together with all data that verify the structures.

2.2. Pharmacological Data

The method of an inhibition in human leukocyte elastase-1 was described elsewhere [10]. The data were taken from literature [1]. To get normalized and variance-stabilized data, natural logarithms of the scores were used.

2.3. Statistical Methods

To overcome the danger of overfitting in neural network analysis, the MASCA model [6] was used to select the important physicochemical descriptors from a large pool of physicochemical descriptors [2, 3]. The

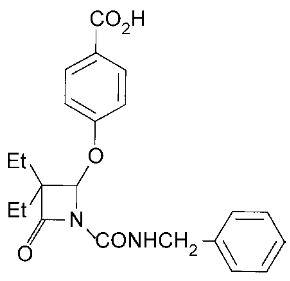

FIGURE 1 Lead structure of phenyl ring substituted 3,3’-diethyl-2-azetidinones.

TABLE I Substituents of the lead structure (Fig. 1), physicochemical descriptors (X_1 = scaled molar refraction 0.1 MR), X_2 = STERIMOL parameter B5, X_3 = lipophilic substituent constant π, octanol/water), and transformed inhibition constants (Y) of human leukocyte elastase

		Activity	*Chemical Descriptors*		
Compd	*Substituents R*	*Y*	X_1	X_2	X_3
1	H	7.31322	0.10	1.00	0.00
2	4-Cl	8.13153	0.59	1.80	0.71
3	4-F	7.74066	0.09	1.35	0.15
4	4-Me	8.29405	0.57	2.04	0.50
5	3-Me	8.85367	0.57	2.04	0.50
6	4-OMe	8.55641	0.79	3.07	0.15
7	4-SMe	9.14846	1.38	3.26	0.42
8	4-Et	9.01189	1.03	3.17	1.00
9	4-COMe	8.68271	1.12	3.13	−0.55
10	4-NO_2	8.53700	0.74	2.44	−0.28
11	4-NH_2	7.49554	0.54	1.97	−1.23
12	4-NMe_2	9.21034	1.56	3.08	0.18
13	4-OH	8.00637	0.29	1.93	−0.67
14	2-NH_2	7.09008	0.54	1.97	−1.23
15	2-OH	7.31322	0.29	1.93	−0.67
16	4-CO_2H	7.69621	0.69	2.66	−0.55
17	4-$CONH_2$	8.36637	0.98	3.07	−1.49
18	4-Ph	10.05191	2.54	3.11	2.13
19	3,4-diMe	9.95228	1.14	4.08	1.00
20	3,5-diMe	9.59560	1.14	4.08	1.00
21	4-OMe, 3-Me	9.48037	1.36	5.11	0.65
22	3-OMe, 4-Me	8.77956	1.36	5.11	0.65

squared multiple correlation coefficient (R^2) was statistically tested using the maximum-likelihood criterion [6]. Multiple regression analysis [2, 3], partial-least squares (PLS) regression [4, 5], and nonleast-squares (NLS) regression of the MASCA model [6] were applied.

2.4. Physicochemical Data

The scaled molar refraction (0.1 MR), the STERIMOL parameter B5, and the lipophilic substituent constant π (octanol/water) were used [2, 3] as substituent constants.

2.5. Neural Networks

The algorithm of the generalized backpropagation neural network was applied [7–9]. The learning rate momentum was equal to 0.8, the learning rate minimum and maximum were equal to 0.001 and 0.3. Optimization was achieved by (i) estimating the global error vector prior to adjusting weights, and (ii) updating successively the weights until convergence was reached.

The neural network architectures, learning algorithms, and genetic selection procedures of the generalized-regression genetic-neural network [9] were based on the proposed default parameters of the NeuroGenetic Optimizer program (BioComp Systems, Redmond, WA).

3. RESULTS AND DISCUSSION

The lead structure and substituents of the compounds are given in Figure 1 and Table I, together with the design matrix of the chemical parameters and the biological activity (Y; 5 digits to avoid rounding errors). The molar refraction (MR) and the STERIMOL parameter B5 describe steric substituent effects. The molar refraction correlates with the lipophilicity for apolar groups. Therefore, a certain degree of multicollinearity [6] might be expected due to physical reasons.

The traditional ordinary least squares (OLS) regression resp. the **Hansch analysis** led to:

$$Y = 7.293 + 0.543\,X_1 + 0.250\,X_2 + 0.408\,X_3 \tag{1}$$

The squared multiple correlation coefficient $R^2 = 0.833$ is significant (the critical quantile, Roy's largest root criterion, is $\theta_\alpha = 0.345$ at the conventional significance level $\alpha = 0.05$). The test statistics $TS_1 = 2.37$, $TS_2 = 2.39$, $TS_3 = 3.53$ are used for examining the regression coefficients (the critical quantile $c_0 = 2.101$ at the 5% level). The theoretically calculated biological activity data are given in Table II.

Diagnostic statistics [6] shows that there is no influential point. Unfortunately, there is a high leverage point (compound 18, $TS = 0.76 > c_0 = 0.375$ at the 5% significance level) and an outlier (compound 22, externally Studentized residual, $TS = 2.98 > c_0 = 2.101$ at the 5% significance level). The major problem is the significant multicollinearity, however: the internal determination coefficients are $D_1 = 0.585$, $D_2 = 0.475$, $D_3 = 0.344$ (the critical quantile is $c_0 = 0.270$; maximum likelihood criterion). There is also a high collinearity (Table III) among the physicochemical descriptors (Studentized maximum modulus test, $c_0 = 0.557$).

In summary: Eq. (1) can only be used with reservation to predict the activity of novel compounds not included into the training set (Tab. I).

As consequence, **PLS regression** was firstly used (one component of the descriptor matrix was significant). The retransformed biased PLS regression

TABLE II Comparison between experimentally obtained (obtd) and theoretically calculated (calcd) biological activity*

		Statistical methods calcd			*Neural networks calcd*	
Compd	*Obtd*	*OLS*	*PLS*	*NLS*	*BP*	*GRNN*
1	7.31	7.60	7.53	7.51	7.58	7.37
2	8.13	8.35	8.28	8.26	7.82	8.27
3	7.74	7.74	7.67	7.65	7.63	7.69
4	8.29	8.32	8.26	8.24	7.88	8.53
5	8.85	8.32	8.26	8.24	7.88	8.53
6	8.56	8.55	8.55	8.55	8.53	8.56
7	9.15	9.03	9.04	9.05	9.18	9.15
8	9.01	9.05	9.01	9.00	8.79	9.01
9	8.68	8.46	8.52	8.54	8.83	8.68
10	8.54	8.19	8.20	8.21	8.11	8.43
11	7.50	7.58	7.63	7.65	7.85	7.29
12	9.21	8.98	9.02	8.03	9.17	9.21
13	8.01	7.66	7.66	7.67	7.76	7.66
14	7.09	7.76	7.63	7.65	7.85	7.29
15	7.31	7.66	7.66	7.66	7.76	7.66
16	7.70	8.11	8.14	8.15	8.20	7.80
17	8.37	7.98	8.09	8.14	8.67	8.37
18	10.05	10.36	10.28	10.25	10.07	10.05
19	9.95	9.34	9.32	9.32	9.71	9.77
20	9.60	9.34	9.32	9.32	9.71	9.77
21	9.48	9.57	9.61	9.63	9.13	9.13
22	8.78	9.57	9.61	9.63	9.13	9.13

*OLS, Hansch analysis; PLS, partial least squares regression; NLS, nonleast-squares regression; BP, backpropagation neural network; GRNN, generalized-regression genetic-neural network.

TABLE III Simple correlation matrix of the data of Table I

Y	X_1	X_2	X_3
1	0.815	0.728	0.748
	1	0.689	0.578
		1	0.396
			1

equation is:

$$Y = 7.201 + 0.590\,X_1 + 0.270\,X_2 + 0.347\,X_3 \tag{2}$$

The squared multiple correlation coefficient is $R^2 = 0.829$ (critical quantile of R^2 see above). Simultaneous test statistics for examining the regression coefficients are not available, however. The theoretically calculated values are listed in Table II. In strict analogy, **NLS regression** was applied. The complete ranked, unbiased principal-component regression analysis (PCRA)

$$Y = 8.514 + 0.783\,PC_1 + 0.066\,PC_2 - 0.014\,PC_3 \tag{3a}$$

shows that the absolute PCRA coefficients decrease monotonically (*PC* denotes the principal components of the descriptors). This finding and subsequent diagnostic statistics [6] are an evidence that there is no mixed or even a complete multicollinearity without predictive model power. The squared multiple correlation coefficient and fitted scores are equivalent to those of Eq. (1). The test statistics $TS_1 = 9.429$, $TS_2 = 0.793$, $TS_3 = 0.172$ and subsequent tests on significance of the eigenvalues of the descriptor matrix indicate clearly that one the first component is statistically significant, like the result obtained by using the rules of PLS regression. Elimination of the insignificant components of Eq. (3a), maintaining the first component ($TS = 9.762$) and rectification leads to the biased NLS regression equation:

$$Y = 7.165 + 0.599\,X_1 + 0.281\,X_2 + 0.323\,X_2 \qquad (3b)$$

The squared multiple correlation coefficient is $R^2 = 0.827$, and Table II collects the theoretically calculated data. As the principal components are orthogonal, there is no high-leverage point. However, an outlier can again be found (compound 22). The apparent reason is that position-dependent chemical constants were not included to analysis because the size of compounds having different positions differ markedly.

In summary: the problem of multicollinearity is circumvented by Eqs. (2) and (3b), the two QSAR equations can be used for predicting novel drugs. However, there are no qualitative differences compared with a Hansch analysis.

In **backpropagation neural network analysis**, the following layers are used: one input layer with linear transfer functions and 3 nodes, one hidden layer with sigmoid transfer function and 3 nodes, one output layer with sigmoid transfer function and one node. Full connections between the nodes were assumed. The sigmoidal backpropagation functions were solved by the nonlinear Levenberg-Marquardt algorithm. The analysis gives $R^2 = 0.795$, the weight coefficients are collected in Table IV. The theoretically calculated values are listed in Table II.

In the **generalized-regression genetic-neural network**, the following layers were used: one input layer with linear transfer function and 3 nodes (physicochemical descriptors); two hidden layer with sigmoidal and summation transfer functions having each 3 nodes; and one output layer with direct transfer function and 1 node (biological activity). The following net parameters were used: generations run = 10, population size = 42, minimum network training passes for each network = 20, cutoff for network

TABLE IV Weights and adjustment deltas of the backpropagation neural network analysis

Layer	*Node*	*Connection*	*Weight*	*Weight delta*
2	1	1	−2.18472	0.000002
2	1	2	−3.92458	0.000008
2	2	1	−1.58524	−0.000014
2	2	2	−19.47955	−0.000063
3	1	1	−11.64952	−0.000040
3	1	2	13.37093	0.000024

training passes = 50, input neural node influence factor = 0, hidden neural node influence factor = 0, limit on hidden neurons = 8; selection was performed by the top 50% surviving, refilling of the population was done by cloning the survivors, mating was performed by using the TailSwap technique (the system picks up a cut point and exchanges "genetic material" between the cut point and the end of the string of the "parents", essentially swapping tails). Mutations were performed using the random bit exchange technique at a rate of 25%. The analysis leads to $R^2 = 0.946$, the theoretically calculated activity data are listed in Table II.

In summary: In contrast to other QSAR examples [8, 9], BP did not improve the goodness of fit in this example. GRNN has the ability to approximate structure-activity functions with satisfactory accuracy.

APPENDIX: TECHNICAL NOTE

A proposal to test the predictive model power by sequential resampling was not realized. First, cross-validation with, *e.g.*, by using 80% of the data as training and 20% as test set, was not made because it was shown by a random-number simulation experiment that cross-validation requires approximately *equal* sample sizes of the subgroups [11]. Second, subsamples of one-leaving-out procedures must be taken *randomly* from the same finite population; there can be no difference between the subsamples other than for sampling errors, because all possible subsamples are taken from a population in such way that they have the same probability of being selected [14]. The gain in precision is illusory in most cases. Sequential resampling may be useful for diagnostic statistics which prove the assumptions of the underlying theory of hypothesis testing, however. Also, the root-mean squared error and related measures were not applied because their significance cannot be examined by exact significance tests.

References

[1] Shah, S. K., Dorn, C. P., Finke, P. E., Hale, J. J., Hagmann, W. K., Brause, K. A., Chandler, G. O., Kissinger, A. L., Ashe, B. M., Weston, H., Knight, W. B., Maycock, A. L., Dellea, P. S., Fletcher, D. S., Hand, K. M., Mumford, R. A., Underwood, D. J. and Doherty, J. B. (1992). Orally Active β-Lactam Inhibitors of Human Leukocyte Elastase-1. Activity of 3,3-Diethyl-2-azetidinones, *J. Med. Chem.*, **35**, 3745–3754.

[2] Hansch, C. and Leo, A. (1995). Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, *Am. Chem. Soc.*, Washington, DC.

[3] Fujita, T. (1990). The Extrathermodynamic Approach to Drug Design, *Compr. Med. Chem.*, **4**, 497–560.

[4] Wold, S., Ruhe, A., Wold, H. and Dunn, W. J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverse, *SIAM J. Sci. Stat. Comput.*, **5**, 735–743.

[5] Frank, E. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools, *Technometrics*, **35**, 109–135.

[6] Mager, P. P. (1991). *Design Statistics in Pharmacochemistry*, Wiley, New York, NY. The free software for NLS regression can be downloaded: http://www.uni-leipzig.de/~pharma/ppm2.htm.

[7] Schneider, G. (2000). Neural Networks are Useful Tools for Drug Design, *Neural Netw.*, **13**, 15–60.

[8] Mager, P. P. (1998). Neural Network Approaches Applied to Selective A_{2a} Adenosine Receptor Agonists, *Med. Chem. Res.*, **8**, 277–290.

[9] Mager, P. P. and Reinhardt, R. (1999). A Comparison of Backpropagation and Generalized-Regression Genetic-Neural Network Models, *Drug Design and Discovery*, **16**, 49–53.

[10] Doherty, J. B., Ashe, B. M., Parker, P. L., Blacklock, T. J., Butcher, J. W., Chandler, G. O., Dahlgren, M. E., Davies, P., Dorn, C. P., Finke, P. E., Firestone, R. A., Hagman, W. K., Halgren, T., Knight, W. B., Maycock, A. L., Navia, M. A., O'Grady, L., Pisano, J. M., Shah, S. K., Thompson, K. R., Weston, H. and Zimmerman, M. (1990). Inhibition of Human Leukocyte Elastase. 1. Inhibition by C-7-Substituted Cephalosporin *tert*-Butyl Esters, *J. Med. Chem.*, **33**, 2513–2521.

[11] Mager, P. P. (1996). A Random Numbers Experiment to Simulate Resample Model Evaluation, *J. Chemometrics*, **10**, 221–240.