J Mol Model (2009) 15:343–348
DOI 10.1007/s00894-008-0361-5

SHORT COMMENTS

Three-dimensional quantitative structure-activity relationship (3D-QSAR) studies of various benzodiazepine analogues of γ-secretase inhibitors

Tarnvir Sammi · Om Silakari · Muttineni Ravikumar

Received: 28 November 2007 / Accepted: 22 September 2008 / Published online: 6 December 2008

Abstract A 3D QSAR analysis has been performed on a series of 67 benzodiazepine analogues reported as γ-secretase inhibitors using molecular field analysis (MFA), with G/PLS to predict steric and electrostatic molecular field interaction for the activity. The MFA study was carried out using a training set of 54 compounds. The predictive ability of model developed was assessed using a test set of 13 compounds (r^2_{pred} as high as 0.729). The analyzed MFA model has demonstrated a good fit, having r^2 value of 0.858 and cross validated coefficient, r^2_{cv} value as 0.790. The analysis of the best MFA model provided insight into possible modification of the molecules for better activity.

Keywords Alzheimer disease · γ-secretase · Inhibitor · Molecular Field Analysis · 3D-QSAR

Introduction

Alzheimer disease (AD) is a progressive neurodegenerative disorder affecting the elderly population. It is characterized by the presence of lesions both at an intracellular and extracellular level, identified as the neurofibrillary tangles and the amyloid plaques, respectively. The origin and the role of the neurofibrillary tangles and amyloid plaques are quite different, although both lead to neurodegeneration. Neurofibrillary tangles are composed of paired helical filaments, aggregates of phosphorylated protein tau that form, when levels of phosphorylated tau are elevated in the cell [1]. Amyloid plaques are deposits composed primarily of β-amyloid insoluble peptides of approximately 4 kDa generated from the precursor amyloid precursor protein (APP), a type Ia transmembrane protein, characterized by a large NH_2-terminal extracellular/ cytosolic domain and a small intracellular/luminal COOH terminal domain [2, 3]. In AD pathology, the production of β-amyloid peptide is the result of APP amyloidogenic processing. This involves the activity first of β-then of γ-secretase, and requires the internalization of APP from the plasma membrane to the endosomes and the lysosomes [4–6]. As γ-secretase catalyses the final step in Aβ production and because it determines the length of Aβ variants, this protease has been a prime target for the development of potential therapeutic agents for AD.

γ-Secretase enzyme is less tractable with respect to appreciating details of its active site topology. This is a complex of integral membrane proteins and crystal structure for this enzyme has not been solved to date. Therefore, in spite of great efforts, the molecular design and development of inhibitors is still dependent on random screening in the absence of structural and mechanistic information of this enzyme at atomic level.

In drug discovery, it is common to have measured activity data for a set of compounds acting on particular protein but not to have knowledge of three-dimensional structure of the protein active site. In the absence of such

Electronic supplementary material The online version of this article (doi:10.1007/s00894-008-0361-5) contains supplementary material, which is available to authorized users.

T. Sammi · O. Silakari (✉)
Department of Pharmaceutical Science and Drug Research,
Punjabi University,
Patiala, Punjab 147002, India
e-mail: omsilakari@rediffmail.com

M. Ravikumar
GVK Biosciences Pvt. Ltd.,
#210 'My Home Tycon', 6-3-1192 Begumpet,
Hyderabad 500 016, India

three-dimensional information, one can attempt to build a hypothetical model of receptor site that can provide perceptiveness about the receptor site characteristics. Molecular field analysis (MFA) is one of several such approaches, which provides compact and quantitative descriptors that capture three-dimensional information about a putative receptor site. This methodology assumes that a suitable sampling of steric and electrostatic field around a set of aligned molecules provide all the information necessary for the understanding of their biological activity.

In the present study we have generated 3D QSAR model using MFA. The generated model may guide the rational synthesis of novel compounds. The deduced model may also give perceptivity to the impact of various interactive fields on the activity and thus assists in forecasting the γ-secretase inhibitory activity of new molecules.

Material and methods

Data set

Structure and inhibitory activity (IC_{50}) data of set of 67 benzodiazepine inhibitors of γ-secretase were collected from literature [7–9]. IC_{50} value represents the dose in molar concentration that causes 50% inhibition of γ-secretase enzyme. The biological activities were converted into the corresponding pIC_{50} values (−log IC_{50}). All the IC_{50} values were obtained using the same assay method [10]. Around 13 compounds were included in the test set, based on the suggestions given by Oprea et al. [11] and remaining in training set. The biological activity data and the structures of test and training set molecules are described in Table L (supporting information).

Molecular modeling

All experiments were conducted using Cerius^2 4.10 (Accelrys; San Diego, CA) molecular modeling software running on silicon graphic workstation [12]. Molecular sketcher facilities provided in the modeling environment of Cerius^2, were used to build structures of compounds. All the molecules were initially energy minimized with smart minimizer. Geometric optimization was carried out using DREIDING force field [13]. Partial atomic charges were calculated using Gasteiger method [14]. Multiple conformations of each molecule were generated with the Boltzmann jump as a conformational search method. Further geometric optimization of each molecule was carried out with MOPAC 6 package using semi-empirical AM1 (Austin mModel) Hamiltonian [15].

Molecular alignment

To obtain effective 3D-QSAR models, proper alignment of structures is decisive step. The method used for performing the alignment was the maximum common subgroup (MCSG) method [12]. This method looks at molecules as points and lines, and uses the techniques of graph theory to identify patterns. It finds the largest subset of atoms in the shape reference compound that is shared by all the structures in the data set and uses this subset for alignment. A rigid fit of atom pairing was performed to superimpose each structure so that it overlays the shape reference compound. The bold-faced portion of the most active molecule **1**, used as the template for the superposition is shown in Fig. 1. Sterioview of aligned molecules is shown in Fig. 2.

Molecular field analysis

Molecular field analysis (MFA) calculates probe interaction energies on a rectangular grid around a bundle of aligned molecules. MFA preferences were set: rectangular grid with 1 A° step size; charges by Gasteiger algorithm; H^+ and CH_3 as probe atoms for electrostatic and steric fields respectively. The atomic coordinates of contributing model are used to compute field values on each point of the 3D grid. Steric and electrostatic interaction energies on each grid point were used as 3D QSAR field descriptors (independent variables) in regression analysis. Only 10% of total descriptors whose variance was higher were considered for further analysis. Thus the major steps in molecular field analysis were a) conformer generation and energy minimization; b) fitting atoms using MCS search and aligning the molecules; c) setting MFA preferences-rectangular grid with 1A° (default 2.00A°) step sizes, charges by Gasteiger algorithm, H^+ and CH_3 as probes; d) field creation; e) regression by G/PLS algorithm.

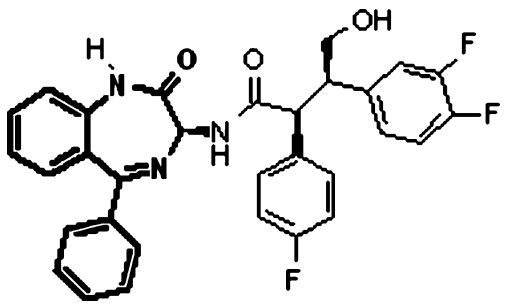

Fig. 1 The bold-faced portion of most active molecule 1 was used as template for the superposition of the rest of the molecules

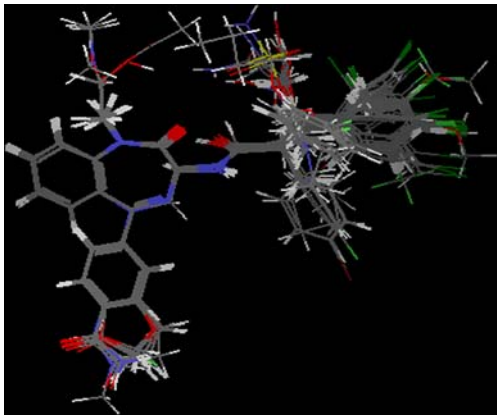

Fig. 2 Stereoview of all the aligned molecules

Genetic partial least sSquares (G/PLS)

G/PLS technique available in QSAR^{+} environment of Cerius^{2} software was used to perform regression analysis of data. As there were large numbers of points used as independent variables, genetic partial least squares (G/PLS) were used to derive QSAR models. G/PLS is derived from two QSAR calculation methods: Genetic function approximation (GFA) and partial least squares (PLS). The GFA algorithm approach builds multiple models rather than single model; it automatically selects which features are to be used in model. Further it is better at discovering combinations of features that take advantage of correlations between multiple features. In PLS, variables might be overlooked during interpretation or in designing the next experiment even though cumulatively they are important. It gives a reduced solution, which is statistically more robust than multiple linear regression (MLA). The linear PLS model finds "new variables" (latent variables or X scores) which are linear combinations of original variables. To avoid over fitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. Cross validation is a practical and reliable way of testing this significance [16]. G/PLS combines the best features of GFA and PLS [17]. In GFA; equation models have a randomly chosen proper subset of independent variables. As a result of multiple linear regressions (MLA) on each model, the best ones become the next generation and two of them produce an offspring. This was repeated 50,000 (default, 5000 times). For other settings, all defaults were used. Application of G/PLS thus allows the construction of large QSAR equations while still avoiding over fitting and eliminating most variables. The best model was selected on statistical measures such as data points (n), square correlation coefficient (r^2), cross-validated correlation coefficient (r^2_{cv}), predicted correlation coefficient (r^2_{pred}), predicted sum of squares (PRESS), bootstrap correlation coefficient (r^2_{bs}). Values are given in Table 1.

Result and discussion

The process of the QSAR model developed can be generally divided into three stages: data preparation, data analysis, and model validation. The first stage includes selection of a molecular dataset for QSAR studies, calculation of molecular descriptors and selection of a QSAR (statistical analysis and correlation) method [18]. MFA is the most widely used computational tool which considers the steric and electrostatic influences [19]. The best model (Eq. 1) for 54 training set molecules was developed. Regression analysis was performed using G/PLS.

Table 1 Various statistical parameters along with their numerical value obtained for the best model

S.No.	Parameters	Value
1.	Data points (n)	54
2.	Square of correlation coefficient (r^2) for training set	0.858
3.	Leave one out cross validated correlation coefficient (r^2_{cv})	0.790
4.	Predicted sum of squares (PRESS)	16.086
5.	Number of PLS components (C)	5
6.	Simple correlation coefficient (r^2_{pred}) for test set	0.729
7.	Predicted correlation coefficient (R^2_{pred})[a]	0.685
8.	Bootstrap correlation coefficient (r^2_{bs})	0.843
9.	Lest Square error (LSE)	0.208
10.	Predicted root mean square error (RMSE $_{pred}$)	0.579
11.	Slope of regression line of observed vs predicted activity passing through origin(k)	1.004
12.	Slope of regression line of predicted vs observed activity passing through origin(k')	0.987
13.	Correlation coefficient for regression line of observed vs predicted activity passing through origin(R^2_0)	0.999
14.	Correlation coefficient for regression line of predicted vs observed activity passing through origin(R'^2_0)	0.998

[a] Correlation coefficient calculated using Eq. 3

Table 2 The structure and γ-secretase inhibitory activity data for training and test set molecules

S. No	Observed Activity[a]	Predicted Activity[b]
1	10.222	9.824
2	10.155	9.797
3[c]	9.699	8.849
4	9.523	8.777
5	9.097	9.121
6	9.046	8.600
7	8.921	8.934
8[c]	8.921	8.845
9	8.920	8.740
10	8.824	9.017
11	8.745	8.753
12	8.745	8.619
13[c]	8.658	8.383
14	8.444	8.555
15	8.432	8.694
16	8.420	8.238
17	8.328	7.764
18	8.260	8.773
19[c]	8.174	9.305
20	8.161	8.185
21	8.097	7.149
22	8.018	7.583
23	7.920	7.232
24[c]	7.896	8.701
25	7.879	8.661
26	7.824	7.992
27	7.824	7.030
28[c]	7.721	7.215
29	7.658	6.939
30	7.553	7.621
31	7.538	7.480
32	7.456	7.226
33	7.398	6.724
34	7.229	7.219
35[c]	7.222	6.596
36	7.174	7.641
37	7.143	7.171
38	7.032	6.942
39	7.000	7.929
40	6.997	7.623
41	6.959	6.715
42	6.959	7.160
43[c]	6.745	6.558
44	6.744	6.735
45	6.674	7.244
46	6.658	6.684
47[c]	6.516	6.326
48	6.495	6.408
49	6.495	6.441
50	6.469	6.676
51[c]	6.456	5.566
52	6.420	6.374
53	6.398	7.408
54	6.398	6.722
55[c]	6.393	5.633
56	6.347	6.109
57	6.125	6.415
58	6.086	6.232
59	6.056	6.135
60[c]	6.032	6.632
61	5.873	7.312
62	5.873	6.143
63	5.796	5.980
64[c]	5.495	6.362
65	5.328	5.012
66	5.323	4.874
67	5.301	6.011

[a]= -log $IC_{50,}$ where IC_{50} is molar dose required to produce 50% inhibition of γ-secretase
[b]= $_{pIC50}$ predicted from the best model.
[c]= test set molecules

In G/PLS, pIC_{50} of compounds were considered as dependent variables whereas molecular field descriptors, i.e., steric (CH_3) and electrostatic (H^+) as independent variables. G/PLS was carried out over 50, 000 generations with a population size of 100.

$$\begin{aligned} pIC_{50} = {} & 5.09937 - 0.04833\,H^+/553 + 0.036404\,CH_3/667 \\ & + 0.083552\,H^+/1018 - 0.10192\,H^+/464 \\ & + 0.078734\,CH_3/1046 - 0.03743\,H^+/1019 \\ & + 0.064932\,CH_3/428 \end{aligned} \quad (1)$$

The second part of QSAR model development consists of analysis of regression output. The various statistical parameters calculated in G/PLS for Eq. 1 are shown in Table 1. The steric (CH_3) and electrostatic (H^+) descriptors, in the equation specify the regions where variations in structural features (steric or electrostatic) of different compounds in the training set, lead to an increase or decrease in activities. The number accompanying descriptors represents its position in the three dimensional MFA grid. An energy cutoff of −30 to +30 kcal/mol was set for both steric and electrostatic contributions. The optimal number of components was set to 5. The smoothing parameter *d,* was set to 1.0 to control the bias in the scoring factors between equations with different numbers of terms. The length of final equation was fixed to nine terms. Cross-validation was performed with leave-one-out procedure. PLS analysis was scaled, with all variables normalized to a variance of 1.0. Equation 1 explains the 85.8% variance in the activity with respect to steric and electrostatic fields. The activities of training set molecules were predicted from this equation and are given in Table 2. Graph between observed and predicted activity is shown in Fig. 3a.

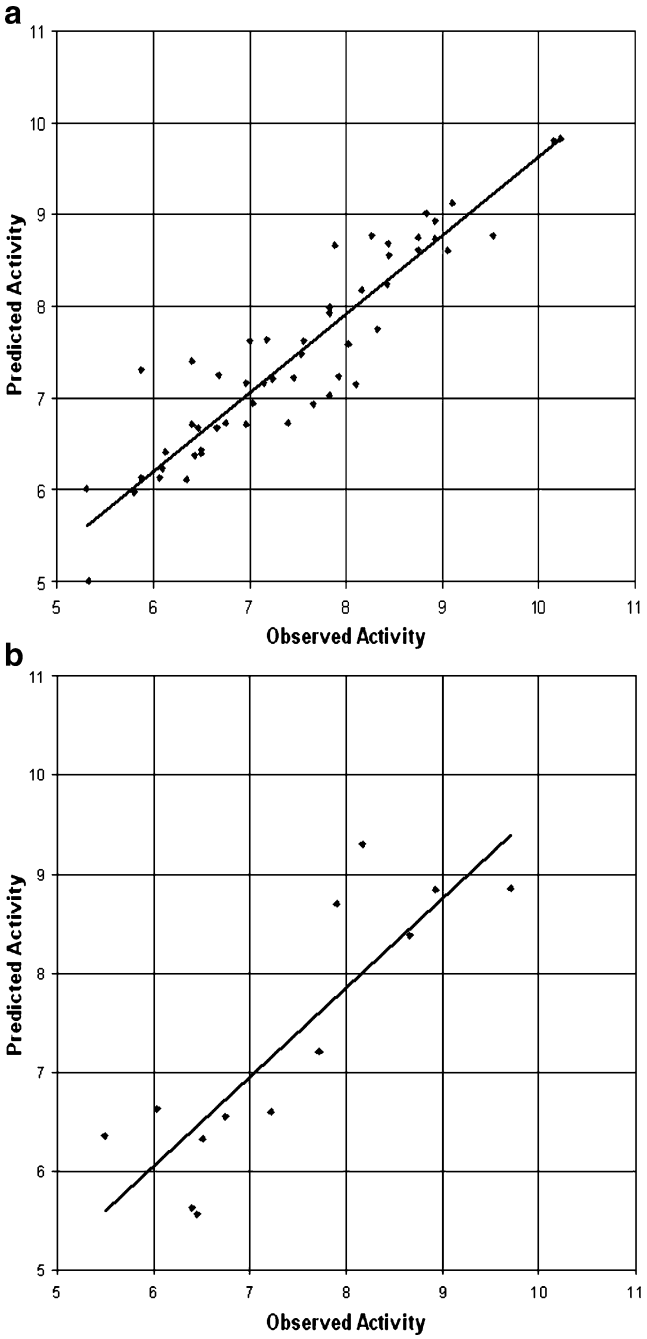

Fig. 3 Correlation graph between observed and predicted activities from the best MFA model for training set (**a**) and test set molecules (**b**)

The most important part of QSAR model development is the model validation. In present study generated model were subjected to internal as well as external validation. In the case of internal validation, r^2_{cv} is used as criterion of both robustness and predictive ability of the model [20]. r^2_{cv} is calculated according to following formula [21]:

$$r^2_{cv} = 1 - \frac{\sum \left(Y_{obs} - Y_{pred}\right)}{\sum \left(Y_{obs} - Y\right)^2}. \quad (2)$$

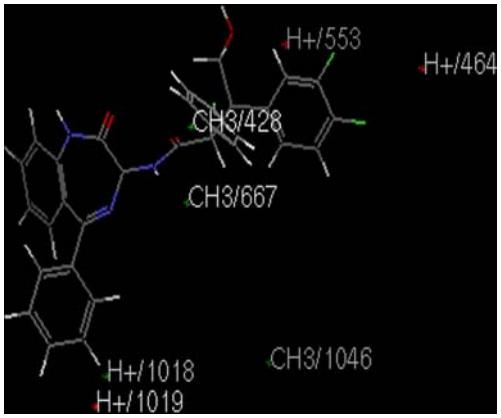

Fig. 4 Stereoview of the best MFA model corresponding to Eq. 1. The most active compound 1 is shown in the background as a reference

In the above equation, Y means average activity value of the entire dataset while Y_{obs} and Y_{pred} represent observed and predicted activity values, respectively. The analyzed model demonstrated r^2_{cv} value of 0.790. Often a high r^2_{cv} ($r^2_{cv} = 0.5$) is considered as a proof of high predictive ability of the model [20]. However, there exists no correlation between r^2_{cv} and r^2 between the predicted and observed activities for the test set [18, 22, 23]. High value of r^2_{cv} alone is an insufficient criterion for QSAR model to be robust and highly predictive. Apparently, the only way to estimate the true predictive power of a model is to test it on a sufficient large collection of compounds from an external test set [18]. The external predictability of generated model was determined by calculating r^2_{pred} values for a test set of 13 compounds according to the following equation [21]:

$$R^2_{pred} = 1 - \frac{\sum \left(Y_{pred(test)} - T_{test}\right)^2}{\sum \left(Y_{test} - Y_{training}\right)^2}. \quad (3)$$

In the above equation $Y_{pred(test)}$ and Y_{test} indicate predicted and observed activity values respectively, of the test set compounds and $Y_{training}$ indicates mean activity values of the training set. The prediction of model was reasonably good with an r^2_{pred} of 0.729. The graph between observed and predicted activity values of test is shown in Fig. 3b.

Generated model was further validated by bootstrapping. It is a procedure in which several times n, random selection out of n objects are performed to simulate different sampling from a large set of objects. In each run some objects are not included in PLS analysis, some others are included more than once. Confidence interval for each term can be estimated from such a procedure, giving an independent measure of stability of PLS model [24]. The generated model estimated r^2_{bs} value of 0.843.

Golbraikh and Tropsha [18] reported that the criteria discussed above may not be sufficient for a QSAR model to be truly predictive. An additional more strict conditions including $r^2_{cv}>0.5$, $r^2_{pred} > 0.6$, R^2_0 or R'^2_0 close to r^2_{pred}, i.e., $\left[\left(r^2_{pred} - R^2_0\right)/r^2_{pred}\right] < 0.1$ or $\left[\left(r^2_{pred} - R'^2_0\right)/r^2_{pred}\right] < 0.1$, and the corresponding 0.85< =k< =1.15 or 0.85< =k'<=1.15 are needed for QSAR model to have high predictive ability. The above said parameters were also calculated for our model (Table 1) and they satisfied all these recommendations.

Stereoview of the best model corresponding to Eq. 1 is shown in Fig. 4. The most active compound **1** is shown in the background as reference. The presence of three steric descriptors CH_3 /428, CH_3/667 and CH3/1046 with positive coefficient, near 4-flourophenyl ring, suggests the importance of bulky substituent in this region. (Refer to Fig. 5, in supporting information). This is also evident from the fact that compound **1, 2, 10, 16,** and **22** in which bulky group is present in this region are relatively more active than compounds **42, 50, 54, 61, 63, 65,** and **67** which do not have any substituent in this region. Presence of electrostatic descriptor H^+/553 and H^+/464 with negative coefficient around the *ortho, meta* and *para* position of m, p-diflourophenyl ring, suggest that electrostatic environment is favorable in this region. This can be observed in compounds **1, 2, 4–7, 9–12,** and **14–18** which are more active than compounds **50, 61, 63, 63,** and **67** with less electronegative substituent in this region. Further, presence of two electrostatic descriptors, H^+/1018 and H^+/1019 with opposite signs, indicates a subtle balance must exist in this region for a molecule to be effective γ-secretase inhibitor.

Conclusion

MFA model of γ-secretase inhibitory activity has been developed grounded on steric and electrostatic descriptors to investigate the substitution requirement for favorable receptor-drug interaction. This study conceded a consistant and statistically significant model with high correlation coefficient and sufficiently reliable predictability. On the basis of this study, it can be concluded that steric and electrostatic interaction energies play an important role in complimentary fit of inhibitors in the active site of γ-secretase enzyme.

Thus, the present study looks into the vital structural features, which can be exploited for modification in present analogs in order to achieve improved γ-secretase inhibitory activity. Substantial ability of the model obtained to predict the external test set molecules supports that the deduced model can be used for the designing of benzodiazepine analogues as γ-secretase inhibitors.

Acknowledgments The authors thank the head, A.K. Tiwary, Department of Pharmaceutical Science and Drug Research, for his steady advice and cooperation and Dr. J.A.R.P.Sarma, Director, Bioinformatics division, GVK Biosciences Pvt. Ltd. for providing software and giving a great chance to work there.

References

1. Grundke-Iqbal I, Iqbal K, Tung YC, Quinlan M, Wisniewski HM, Binder LI (1986) Proc Natl Acad Sci 83:4913–4917 doi:10.1073/pnas.83.13.4913
2. De Strooper B, Annaert W (2000) J Cell Sci 113:1857–1870
3. Nunan J, Small DH (2002) Essays Biochem 38:37–49
4. Koo EH, Squazzo SL (1994) J Biol Chem 269:17386–17389
5. Koo EH, Squazzo SL, Selkoe DJ, Koo CH (1996) J Cell Sci 109:991–998
6. Perez RG, Soriano S, Hayes JD, Ostaszewski B, Xia W, Selkoe DJ, Chen X, Stokin GB, Koo EH (1999) J Biol Chem 274:18851–18856 doi:10.1074/jbc.274.27.18851
7. Owens AP, Nadin A, Talbot AC, Clarke EE, Harrison T, Lewis HD, Reilly M, Wrigley JDJ, Castroa JC (2003) Bioorg Med Chem Lett 13:4143–4145 doi:10.1016/j.bmcl.2003.07.031
8. Churcher I, Ashton K, Butcher JW, Clarke EE, Harrison T, Lewis HD, Owens AP, Teall MR, Williams S, Wrigley JDJ (2003) Bioorg Med Chem Lett 13:179–183 doi:10.1016/S0960-894X(02)00909-5
9. Churcher I, Susie SW, Kerrad S, Harrison T, Castro JL, Shearman MS, Lewis HD, Clarke EE, Wrigley JDJ, Beher D, Tang YS, Liu W (2003) J Med Chem 46:2275–2278 doi:10.1021/jm034058a
10. Clarke EE, Sherman MS (2000) J Neurosci Methods 102:61–68 doi:10.1016/S0165-0270(00)00280-6
11. Oprea T I, Waller CL, Marshal GR (1994) J Med Chem 37:2206–2215 doi:10.1021/jm00040a013
12. Cerius^2 Molecular Modeling Program Package, Molecular Simulations (Accelrys) Inc.; San Diego, CA 92121-3752, USA
13. Mayo SL, Olafson BD, Goddard WA (1990) J Phys Chem 94:8897–8909 doi:10.1021/j100389a010
14. Gasteiger J, Marsili M (1980) Tetrahedron 36:3219–3222 doi:10.1016/0040-4020(80)80168-2
15. Dewar MJS, Zoebish EG, Healy EF, Stewart JP (1985) J Am Chem Soc 107:3902–3909 doi:10.1021/ja00299a024
16. Fan Y, Shi LM, Kohn KW, Pommier Y, Weinstein JN (2001) J Med Chem 44:3254–3263 doi:10.1021/jm0005151
17. Molecular Simulations. Cerius2 Tutorial, version 3.5. Molecular Simulations, Inc., San Diego, CA. (http://www.msi.com/doc/cerius35/)
18. Golbraikh A, Tropsha A (2002) J Mol Graph Mod 20:269–276 doi:10.1016/S1093-3263(01)00123-1
19. Hopfinger AJ, Tokarsi JS, Charifson PS (eds) (1997) Practical applications of computer-aided drug design. Marcel Dekker, New York
20. Hawkins DM, Basak SC, Mills D (2003) J Chem Inf Comput Sci 43:579–586 doi:10.1021/ci025626i
21. Roy PP, Roy K (2007) QSAR Comb Sci 27:302–313 doi:10.1002/qsar.200710043
22. Aptula AO, Jeliazkova NG, Schultz TW, Cronin MTD (2005) QSAR Comb Sci 24:385–396 doi:10.1002/qsar.200430909
23. Roy K, Leonard JT (2006) QSAR Comb Sci 25:235–251 doi:10.1002/qsar.200510161
24. QSAR (1993) Hansch analysis and related approaches. In: Mannhold R, Kroogsgrad-Larsen P, Timmerman H (eds) Methods and principles in medicinal chemistry, vol 1. VCH, Weinheim, p 105