Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 13 (2005) 1681–1689
Bioorganic & Medicinal Chemistry

Comparative molecular field analysis and comparative molecular similarity indices analysis of human thymidine kinase 1 substrates

Achintya K. Bandyopadhyaya,[a] Jayaseharan Johnsamuel,[a] Ashraf S. Al-Madhoun,[b,†] Staffan Eriksson[b] and Werner Tjarks[a,*]

[a]*Division of Medicinal Chemistry, College of Pharmacy, The Ohio State University, Columbus, OH 43210, USA*
[b]*Department of Molecular Biosciences, Section of Veterinary Medicinal Chemistry, Swedish University of Agricultural Sciences, The Biomedical Center, SE-75123 Uppsala, Sweden*

Received 19 October 2004; revised 6 December 2004; accepted 6 December 2004
Available online 7 January 2005

Abstract—Thymidine kinase 1 (TK1) is a key target for antiviral and anticancer chemotherapy. Three-dimensional quantitative structure–activity relationship (3D-QSAR) using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) techniques was applied to analyze the phosphorylation capacity of a series of 31 TK1 substrates. The optimal predictive CoMFA model with 26 molecules provided the following values: cross-validated r^2 (q^2) = 0.651, non-cross-validated r^2 = 0.980, standard error of estimate (s) = 0.207, F = 129.3. For the optimal CoMSIA model the following values were found: q^2 = 0.619, r^2 = 0.994, s = 0.104, F = 372.2. The CoMSIA model includes steric, electrostatic, and hydrogen bond donor fields. The predictive capacity of both models was successfully validated by calculating known phosphorylation rates of five TK1 substrates that were not included in the training set. Contour maps obtained from CoMFA and CoMSIA models correlated with the experimentally developed SAR.

1. Introduction

Human thymidine kinase 1 (TK1) belongs to the family of the deoxynucleoside salvage enzymes.[1] It catalyzes the 5′-monophosphorylation of the endogenous deoxyribonucleosides thymidine (Thd) and 2′-deoxyuridine (dUrd)[1] as well as the 5′-monophosphorylation of the anti-human immunodeficiency virus (HIV) prodrugs such as 3′-azido-2′,3′-dideoxythymidine (AZT) and 2′,3′-didehydro-2′,3′-dideoxy-thymidine (d4T).[2] In the case of latter two agents, this phosphorylation is the crucial initial step in their activation as HIV reverse transcriptase inhibitors.[2]

TK1 activity is absent in resting cells, appears in late G1 cells, increases in S-phase, coinciding with the increase in DNA synthesis, and disappears during mitosis.[1,3–5] High TK1 activity in proliferating cells has proven to be of prognostic value in breast[6,7] and colon cancer[8] and has triggered the production of TK1 directed antibodies for the detection of serum TK1.[9] Early on, TK1 has also been identified as a potential target for therapeutic anticancer drugs.[10,11] However,

Abbreviations: 3D-QSAR, three-dimensional quantitative structure–activity relationship; AZT, 3′-azido-2′,3′-dideoxythymidine; BNCT, boron neutron capture therapy; CoMFA, comparative molecular field analysis; CoMSIA, comparative molecular similarity indices analysis; dUrd, 2′-deoxyuridine; D4T, 2′,3′-dideoxy-2′,3′-didehydrothymidine; dCK, deoxycytidine kinase; dGK, deoxyguanosine kinase; Dm-dNK, deoxynucleoside kinase from *Drosophila melanogaster*; GH, Gasteiger–Hückel; HIV, human immunodeficiency virus; HSV1 TK, herpes simplex virus type-1 thymidine kinase; MM+, molecular mechanics force field; Thd, thymidine; PLS, partial least square analysis; PR, phosphorylation rate; SD, standard deviation; SEE, standard error of estimation; TK1, human thymidine kinase 1; VZV TK, Varicella Zoster virus thymidine kinase

Keywords: 3D-QSAR; CoMFA; CoMSIA; Thymidine kinase 1 (TK1); TK1 substrates.

* Corresponding author. Tel.: +1 614 292 7624; fax: +1 614 292 2435; e-mail: tjarks.1@osu.edu
† Present address: Division of Cardiology, Vascular Biology Laboratory, University of Ottawa Heart Institute, Ottawa, Canada K1Y 4W7.

0968-0896/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2004.12.008

deoxycytidine kinase (dCK) is currently the only deoxynucleoside salvage enzyme utilized as a therapeutic target phosphorylating, and thus activating, anticancer prodrugs such gemcitabine, cytarabine, cladribine, and fludarabine.[12–14] This is probably due to the fact that TK1 has the most stringent substrate specificity among all nucleoside kinases allowing only phosphorylation of native Thd/dUrd and, to a limited extent, analogues with minor modifications either at the 5-position (Cl, Br, I) or at the 3′-position (N_3, F).[15] Only recently it was discovered that TK1 also tolerates bulky substitutents at the N-3 position, which has been exploited successfully in the design and synthesis of boronated Thd analogues for boron neutron capture therapy (BNCT), an experimental binary radiochemotherapeutic methods for cancer treatment.[3,16–19]

In recent years crystal structures of several human, viral, and insectoid nucleoside kinases have become available. These include dCK,[20] deoxyguanosine kinase (dGK),[4,21] herpes simplex virus type-1 thymidine kinase (HSV1 TK),[22] Varicella Zoster virus thymidine kinase (VZV TK),[23] and the deoxynucleoside kinase from *Drosophila melanogaster* (Dm-dNK).[4,21] These 3D structures may aid in the structure-based in silico design of novel anticancer and antiviral drugs. Unfortunately, there are presently neither NMR nor X-ray data of the 3D structure of human TK1 available that would allow the structure-based design of TK1 inhibitors or substrates for anticancer and antiviral applications. Also, the amino acid sequence of human TK1 has relatively high similarity with those of poxvirus and bacterial TKs but not with those of dCK, dGK, HSV1 TK, VZV TK, and Dm-dNK,[1,15,24] which prevents the development of homology models by computational methods using the 3D structures of the latter kinases as templates.[25]

In view of the problems associated with drug design of TK1 targeting drugs, including stringent TK1 substrate specificity and lack of 3D structures, three-dimensional quantitative structure–activity relationship (3D-QSAR) based techniques such as comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) could facilitate the design of novel TK1 inhibitors/substrates for therapeutic applications. Both 3D-QSAR techniques, which were successfully used in the design of variety of potential drugs,[26–32] require an existing set of molecules and the corresponding biological activities to predict the biological activities of non-synthesized compounds that are structurally related to the set of existing compounds. CoMFA is able to predict the biological activity of novel molecules based on the relationship of steric/electrostatic properties and biological activities while CoMSIA also involves hydrophobic and hydrogen bond (H-bond) donor/acceptor fields in the correlation with biological data. CoMSIA is also less sensitive to molecular alignments/conformations than CoMFA.

In the present study, we have used a training set of structurally similar TK1 substrates and the corresponding TK1 phosphorylation rates (PRs) to predict the PRs of several TK1 substrates that were not included in the training set.

2. Materials and methods

2.1. Compound sources and phosphorylation rates

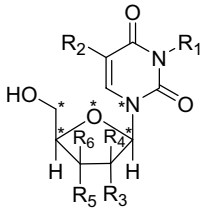
Compound **1** was obtained from Aldrich Chemicals and compounds **2**, **4**, **5**, **6**, and **27** were synthesized as described previously.[33,34] The synthesis of compound **3** will be described elsewhere (Youngjoo Byun, unpublished results).

The PRs of substrates **1**–**31** (Table 1) were obtained using the same protocol, which is described below for the determination of the PRs of compounds **1**–**6** and **27**. The PR values are expressed relative to the PR value of the endogenous TK1 substrate Thd. The PRs of compounds **7**–**26** and **28**–**31** were reported previously by Eriksson and co-workers.[3,35] The values are the mean of at least three experiments with SDs $\leqslant$ 15% (Staffan Eriksson, personal communication). All compounds were derivatives of Thd/dUrd in β-D-configuration with the base in *anti* conformation. The PRs of compounds **1**–**6** and **27** were determined in phosphoryl transfer assays using recombinant human TK1 as described previously.[16,18] Briefly, Thd and other TK1 substrates were dissolved in DMSO to produce stock solutions. The assay mixtures contained 10 μM compound, 100 μM ATP including a small fraction of 0.0325 μM [γ-^{32}P] ATP (Amersham, IL, USA), 50 mM Tris–HCl (pH 7.6), 5 mM $MgCl_2$, 125 mM KCl, 10 mM DTT, and 0.5 mg/mL bovine serum albumin (BSA). The final concentration of DMSO was set to 1% for all the reactions. In previous experiments,[3,35] TK1 substrate concentration of 100 μM were also applied. The reaction mixtures were incubated at 37 °C for 20 min in presence of 50 ng of enzyme. Following the incubation period, the enzyme was heat inactivated for 2 min at 95 °C. The reaction mixtures were centrifuged and 1 μL portions were spotted on PEI-cellulose TLC plates (Merck). The TLC plates were developed using isobutyric acid, ammonium hydroxide, and water in a ratio of 66:1:33. The radiolabeled spots were visualized by phosphoimager (Fuji Film, Science Lab., Image Gauge V3.3) and the values of each TK1 substrate is expressed relative to that of Thd. The PR values for compounds **1**–**6** and **27** are the mean of three experiments with SDs $\leqslant$ 4%.

2.2. Molecular modeling

All molecular modeling including 3D-QSAR studies were carried out on a SGI O_2 workstation using SYBYL 6.9 molecular modeling software (Tripos Inc., St. Louis, MO). Molecule **15** (Thd) was built using the 'sketch molecule' function and minimized with Gasteiger–Hückel (GH) charges and the Tripos force field, using the Powell method and an energy gradient of 0.005 kcal mol^{-1} Å^{-1}. Structures **1**–**14** and **16**–**31** were then built from Thd and minimized in a similar manner.

Table 1. The training and the test set molecules for CoMFA and CoMSIA modeling



Compounds	R_1	R_2	R_3	R_4	R_5	R_6	PR[e]	Log(PR)
Training set								
1	Me	Me	H	H	OH	H	44.3	1.65
2	Et	Me	H	H	OH	H	61.4	1.79
3	$CH_2C{\equiv}CH$	Me	H	H	OH	H	102.1	2.01
4	*n*-Bu	Me	H	H	OH	H	84.9	1.93
5	CH_2Ph	Me	H	H	OH	H	82.3	1.92
6	$CH_2CHOHCH_2OH$	Me	H	H	OH	H	27.5	1.44
7	H	H	H	H	N_3	H	70.0	1.85
8	H	Me	H	H	N_3	H	52.0	1.72
9	H	Me	H	H	H	H	40.0	1.60
10	H	Me	H	H	CH_2N_3	H	15.0	1.18
11[a]	H	Me	H	—	H	—	7.0	0.85
12[b]	H	H	H	H	OH	H	77.0	1.89
13	H	Me	H	F	OH	H	45.0	1.65
14	H	Et	H	H	OH	H	80.0	1.90
15[c]	H	Me	H	H	OH	H	100.0	2.00
16	H	I	H	F	OH	H	42.0	1.62
17	H	Br	H	H	OH	H	80.0	1.90
18	H	H	H	OH	OH	H	1.0	0.00
19	H	CH=CHBr	H	H	OH	H	1.0	0.00
20	H	CH_2CH_2Cl	H	H	OH	H	5.0	0.70
21	H	NH_2	H	H	OH	H	3.0	0.48
22	H	H	OH	H	OH	H	0.1	−1.00
23	H	H	H	H	OMe	H	0.1	−1.00
24	H	H	H	H	OEt	H	0.1	−1.00
25	H	H	F	F	OH	H	0.1	−1.00
26[d]	H	Me	$=CH_2$		OH	H	0.1	−1.00
Test set								
27	*i*-Pr	Me	H	H	OH	H	69.6	1.84
28	H	Me	H	H	F	H	30.0	1.48
29	H	F	H	H	OH	H	95.0	1.98
30	H	Me	H	OH	OH	H	1.0	0.00
31	H	Me	OH	H	OH	H	2.0	0.30

Compounds are arranged in rows with maximum structural similarity to the neighboring compounds according to the hierarchical analysis.
*These atoms were used for the alignment of the molecules.
[a] 2′,3′-Didehydro-2′,3′-dideoxythymidine (d4T).
[b] dUrd.
[c] Thd.
[d] 2′-Methylene-2′-deoxythymidine.
[e] The values are given in % relative to that of Thd, which is set at 100%.

2.3. Alignment rules and 3D molecular database

Structural alignment is one of the most sensitive parameters in 3D-QSAR analyses. The accuracy of the prediction of a CoMFA model and the reliability of the contour models strongly depend on the structural alignment of the molecules. The compound database should contain highly active compounds with various functional groups that have activities ranging from very high to very low. Generally, the low energy conformation of the most active (or toxic) compound in a given set is chosen as the template molecule.[36] In the present study, a database of the energy-minimized structures of **1**–**31** was aligned using the 'align database' option of SYBYL 6.9 taking Thd, the endogenous TK1 substrate with high PR, as the template. The atoms used for the alignments are indicated by asterisks in the drawing of the generalized molecule in Table 1 and the final alignments are shown in Figure 1. All molecules were aligned based on the assumption that they bind to TK1 in the same manner.

2.4. CoMFA analysis

The aligned molecules (**1**–**31**) were placed in a three-dimensional grid (2 Å spacing) extending at least 2 Å

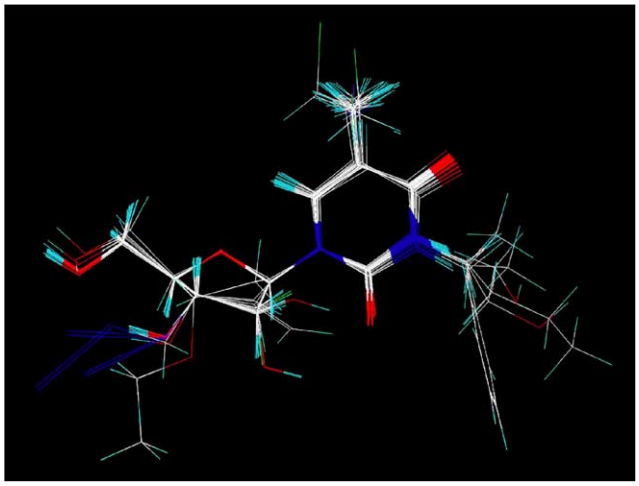


Figure 1. Alignment of all 31 structures.

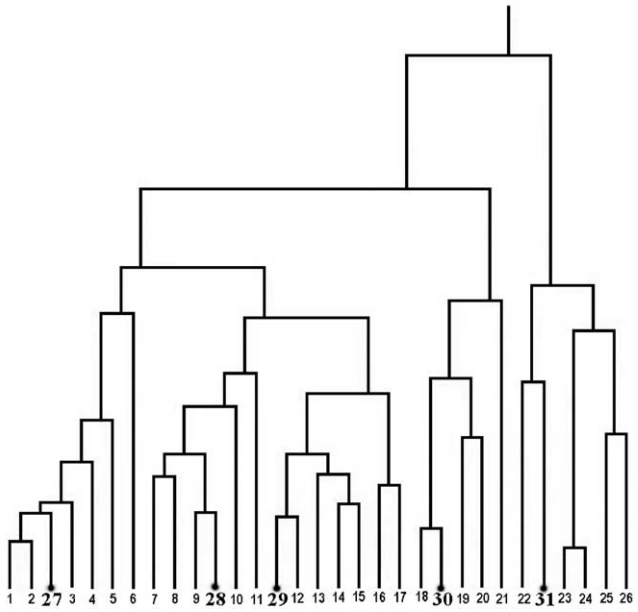


Figure 2. Dendrogram generated from log(PR)s and all descriptors of the CoMSIA fields. For clarity, only the test set compounds (**27**–**31**) are indicated by bold numbers in the diagram. Compounds **1**–**26** (from left) constitute the training set.

beyond the volumes of all investigated molecules on all axes. In the CoMFA analysis, Lennard-Jones potential and Coulomb potentials were employed to calculate the CoMFA steric and electrostatic interaction fields, respectively. At each grid point the steric energy and electrostatic energy were measured for each molecule using the default probe atom, sp^3 carbon atom with a charge of +1. The cutoff values for both the steric and electrostatic energies were set to 30 kcal/mol with a distance dependent dielectric field. CoMFA uses partial least squares (PLS) to predict the activity of the molecules from the energy values at the grid points. Smaller cutoff values (20 and 25 kcal/mol) for the final model resulted in comparable q^2 and r^2 values with larger standard error of estimation (SEE), and thus, decreased predictability.

2.5. CoMSIA analysis

The molecular alignment was placed in a three-dimensional grid (2.0 Å spacing) similar to that of CoMFA analysis. CoMSIA differs from CoMFA in the implementation of the fields. It calculates steric and electrostatic fields, in addition to hydrophobic, H-bond donor, and H-bond acceptor fields, and it uses Gaussian equations for field calculation that do not require cutoff values. In the CoMSIA analysis a radius of 1 Å, a charge of +1, hydrophobicity value of +1, and H-bond donor or acceptor properties of +1 were used as standard parameters for the probe atom. A default value of 0.3 was used as the attenuation factor, α. Five columns were created with steric, electrostatic, hydrophobic as well as H-bond donor and acceptor descriptors. PLS analyses were performed and the best model was used to predict the biological properties. CoMSIA analyses with grid spacings of 1.0 and 1.5 Å resulted in decreased predictability.

2.6. Design of the training and the test sets

The application of statistical methods depends on a suitable design of the training set and the test set, for which biological data will be predicted.[37] The structures and all relevant properties of the test set compounds should represent those of the training set. A hierarchical analysis (as implemented in SYBYL 6.9) was performed to cluster structurally similar molecules in groups. A dendrogram (Fig. 2), generated by using all the descriptors of the CoMSIA fields, was utilized to divide the molecules into the training (**1**–**26**) and the test set (**27**–**31**), thereby ensuring compatibility of the relevant properties between the training and the test sets. Based on this dendrogram, 3-*i*-PrThd (**27**) and FdUrd (**28**) were selected as the representatives of all N-3 substituted Thds and 5-substituted dUrds, respectively, for the test set. Since training set compounds have various substituents at the 3′-position, 3′-FThd (**29**) was included in the test set. The training set also contains a number of compounds with substituents at the 2′-positions both in ribose and arabinose configuration. Therefore, AraT (**30**) and 2′-hydroxylthymidine (**31**) were included in the test set.

2.7. PLS analysis and validations

CoMFA and CoMSIA descriptors were used as independent variables and log(PR) as the dependent variables in the PLS regression analyses for the development of 3D-QSAR models. SAMPLS (SAMple distance PLS)[38] analysis was carried out using the cross-validated 'leave-one-out' option to determine the optimum number of components to be used in the final non-cross-validated analysis. The number of components used was not greater than 1/3 of the number of rows (26) in the training set. The optimum number of components produces the smallest root mean predictive sum of squared errors, which corresponds to the highest cross-validated coefficient (q^2). The non-cross-validated

analyses were then performed using the optimum number of components, with the column filtration set to 1.0.

To further access the statistical confidence of the derived models, a 100-cycle bootstrap analysis was performed using the optimized number of latent variables determined in the PLS model.[39]

3. Results and discussion

3.1. Biological data

The logarithms of the 'relative' TK1 PRs of compounds **1**–**26** (Table 1) were used as the dependent variable in the 3D-QSAR analyses. In our own studies,[16,18] the PRs of a small library of 12 boronated Thd analogues correlated with their catalytic efficiencies (K_{cat}/K_M). High PRs are indicative of good TK1 substrate characteristics while low PRs are indicative of poor TK1 substrate characteristics. They are currently the only biological data available for a large number of TK1 substrates.[3,18,35,40,41] In addition, these PRs were obtained by Eriksson and co-workers[3,35] applying the same experimental protocol.

3.2. Results of the CoMFA analysis

The statistical data obtained from the standard CoMFA model constructed with steric and electrostatic fields are shown in Table 2. The optimal number of components (7) was determined using SAMPLS analysis implemented in SYBYL 6.9 with a leave-one-out (LOO) cross-validated q^2 of 0.651, showing a good predictive capacity of the model ($q^2 > 0.5$). A high correlation coefficient (r^2) of 0.980 for the non-cross-validated final model indicates the self-consistency of the model ($r^2 > 0.9$). Predictions for the log(PR)s (residuals less than one log units) for the holdout test compounds were achieved with a predictive r^2 of 0.837. The satisfactory quality of the CoMFA model is represented in Figure 3, which shows a scatter plot of experimental versus the predicted log(PR)s for the training and the test set.

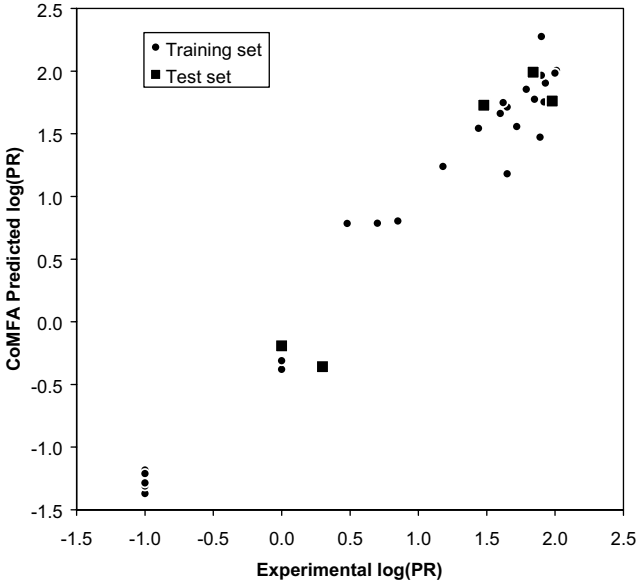


Figure 3. Experimental versus predicted log(PR)s of compounds in the training set and the test set for the CoMFA model.

3.3. Results of the CoMSIA analysis

Eight models were constructed by varying the steric, electrostatic, hydrophobic, H-bond donor and acceptor descriptor fields (Table 2). The optimal components of each of these models were determined by SAMPLS analysis implemented in SYBYL 6.9 with maximum q^2 values. The model developed by using only steric and electrostatic fields produced a cross-validated q^2 of 0.603 with five latent variables. The corresponding conventional r^2 value was 0.916. The quality of the model deteriorated drastically when the hydrophobic field or the H-bond acceptor fields were added to the steric and electrostatic field descriptors ($q^2 < 0.55$). The best eight component

Table 2. Summary of CoMFA and CoMSIA results for various models

	CoMFA	CoMSIA (SE)	CoMSIA (SEH)	CoMSIA (SEA)	CoMSIA (SEHD)	CoMSIA (ALL)	CoMSIA (SEDA)	CoMSIA (SED)	CoMSIA (SEHA)
Components	7	5	2	6	6	5	8	8	2
q^2	0.651	0.603	0.532	0.515	0.548	0.534	0.553	0.619	0.550
r^2	0.980	0.916	0.839	0.943	0.974	0.956	0.988	0.994	0.805
SEE	0.207	0.369	0.475	0.311	0.211	0.265	0.148	0.104	0.522
F	129.3	43.4	59.9	52.1	117.2	87.9	182.4	372.2	47.5
Field contributions (%)									
S	0.528	0.214	0.127	0.140	0.083	0.062	0.098	0.099	0.156
E	0.472	0.786	0.511	0.648	0.357	0.308	0.480	0.502	0.335
H			0.362		0.221	0.234			0.323
D					0.338	0.257	0.313	0.399	
A				0.211		0.138	0.109		0.187
r^2_{bs}	0.995	0.960	0.862	0.976	0.990	0.977	0.995	0.996	0.852
SD_{bs}	0.005	0.020	0.049	0.017	0.006	0.012	0.004	0.003	0.048

q^2, squared cross-validated coefficient; r^2, squared conventional coefficient; SEE, standard error of estimate; F, F-test value; S, E, H, A, D, ALL denote the steric, electrostatic, hydrophobic, hydrogen-bond acceptor and donor, and the combination of all the five fields, respectively; r^2_{bs} and SD_{bs} represent the mean value of 100-run bootstrap predicted cross-validated r^2 and its standard deviation.

model with the maximum q^2 value consisted of steric, electrostatic, and H-bond donor fields as the descriptors, which gave the highest cross-validated coefficient q^2 (0.619), thereby indicating the highest predictive capacity. The highest squared correlation coefficient r^2 (0.994) was also obtained for this model showing a strong internal consistency and therefore it was chosen for the final analysis. The standard error of estimate and the *F*-test values of this model were 0.104 and 372.2, respectively.

The scatter plot of experimental versus predicted log(PR)s for the training and the test set are presented in Figure 4 for the best CoMSIA model with steric, electrostatic, and H-bond donor fields. The predicted log(PR) residuals for both the training and test set compounds were all lower than 0.20 log units, except for compound **31** (0.85 log unit). However, the experimentally determined PR of compound **31** was 2.0 while the predicted PR was 0.3. Thus, the poor TK1 substrate characteristics of **31** were still accurately indicated (Table 3).

3.4. Comparison of CoMFA and CoMSIA analysis

Table 3 shows experimental and predicted log(PR)s for all the compounds. PRs were calculated from the log(PR)s in order to achieve an improved correlation with the actual experimental PRs. Overall, the residuals for the CoMFA model showed higher values for several compounds with higher standard deviation (SD 0.18 with high and low values of 0.47 and −0.38, respectively) for the residues compared with the CoMSIA model (SD 0.09 with high and low values of 0.12 and −0.19, respectively). The scatter plot (Fig. 4) of experimental versus predicted log(PR)s shows overall higher linearity for the CoMSIA model compared with the CoMFA model. Similarly, the plots of the test set compounds results showed overall higher linearity for the best CoMSIA model.

3.5. CoMFA contour maps

Figure 5 shows the steric contour maps for the CoMFA model with dUrd (**12**) as a reference. Sterically favored regions (green) around C-5 of the pyrimidine ring suggest an increase in the activity for the compounds having a small group at C-5. This is indeed the case for Thd (**15**), 5-EtdUrd (**14**), 5-BrdUrd (**17**), and 5-FdUrd (**29**) having methyl, ethyl, bromo, and fluoro substituents, respectively, at the 5-position. However, a disfavored region (yellow) slightly above the green isopleths at C-5 accounts for the inactivity of the compounds **19** and **20**, having bulky bromovinyl and chloroethyl

Table 3. Experimental versus predicted PRs of CoMFA and CoMSIA models

	Experimental data		Predicted CoMFA data				Predicted CoMSIA data			
	PR	Log(PR)	Log(PR)	Calcd PR	ΔLog(PR)	ΔPR	Log(PR)	Calcd PR	ΔLog(PR)	ΔPR
Training set										
1	44.3	1.65	1.71	51.5	−0.06	−7.2	1.71	51.5	−0.06	−7.2
2	61.4	1.79	1.86	71.6	−0.06	−10.2	1.72	52.4	0.07	9.0
3	102.1	2.01	2.01	101.2	0.00	0.9	2.07	118.6	−0.06	−16.5
4	84.9	1.93	1.90	80.2	0.03	4.7	1.82	66.7	0.11	18.2
5	82.3	1.92	1.75	56.8	0.17	25.5	1.98	94.4	−0.06	−12.1
6	27.5	1.44	1.54	34.9	−0.10	−7.4	1.42	26.1	0.02	1.4
7	70.0	1.85	1.77	59.4	0.08	10.6	1.78	60.4	0.07	9.6
8	52.0	1.72	1.56	36.1	0.16	15.9	1.79	61.8	−0.07	−9.8
9	40.0	1.60	1.66	45.8	−0.06	−5.8	1.71	51.2	−0.11	−11.2
10	15.0	1.18	1.24	17.3	−0.06	−2.3	1.10	12.4	0.09	2.6
11	7.0	0.85	0.80	6.4	0.05	0.6	0.88	7.5	−0.03	−0.5
12	77.0	1.89	1.47	29.6	0.42	47.4	1.88	75.2	0.01	1.8
13	45.0	1.65	1.18	15.1	0.47	29.9	1.63	42.6	0.02	2.4
14	80.0	1.90	1.97	92.5	−0.07	−12.5	1.81	65.0	0.09	15.0
15	100.0	2.00	1.98	96.4	0.02	3.6	2.15	140.9	−0.15	−40.9
16	42.0	1.62	1.75	56.0	−0.13	−14.0	1.50	31.6	0.12	10.4
17	80.0	1.90	2.28	188.8	−0.38	−108.8	1.79	62.2	0.11	17.8
18	1.0	0.00	−0.31	0.5	0.31	0.5	−0.05	0.9	0.05	0.1
19	1.0	0.00	−0.38	0.4	0.38	0.6	−0.06	0.9	0.06	0.1
20	5.0	0.70	0.79	6.1	−0.09	−1.1	0.89	7.8	−0.19	−2.8
21	3.0	0.48	0.78	6.1	−0.30	−3.1	0.46	2.9	0.02	0.1
22	0.1	−1.00	−1.18	0.1	0.18	0.0	−0.87	0.1	−0.13	0.0
23	0.1	−1.00	−1.31	0.0	0.31	0.1	−1.03	0.1	0.03	0.0
24	0.1	−1.00	−1.29	0.1	0.29	0.0	−1.00	0.1	0.00	0.0
25	0.1	−1.00	−1.21	0.1	0.21	0.0	−0.95	0.1	−0.05	0.0
26	0.1	−1.00	−1.37	0.0	0.37	0.1	−1.06	0.1	0.06	0.0
Test set										
27	69.6	1.84	1.99	97.9	−0.15	−28.4	1.72	52.0	0.12	17.6
28	30.0	1.48	1.73	53.3	−0.25	−23.3	1.39	24.5	0.09	5.5
29	95.0	1.98	1.76	57.5	0.22	37.5	2.07	116.9	−0.09	−21.9
30	1.0	0.00	−0.19	0.6	0.19	0.4	−0.03	0.9	0.03	0.1
31	2.0	0.30	−0.36	0.4	0.66	1.6	−0.55	0.3	0.85	1.7

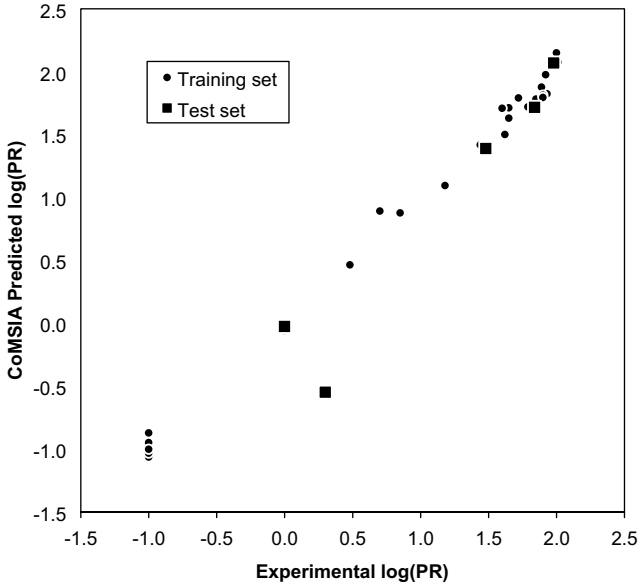


Figure 4. Experimental versus predicted log(PR)s of compounds in the training set and the test set for the CoMSIA model.

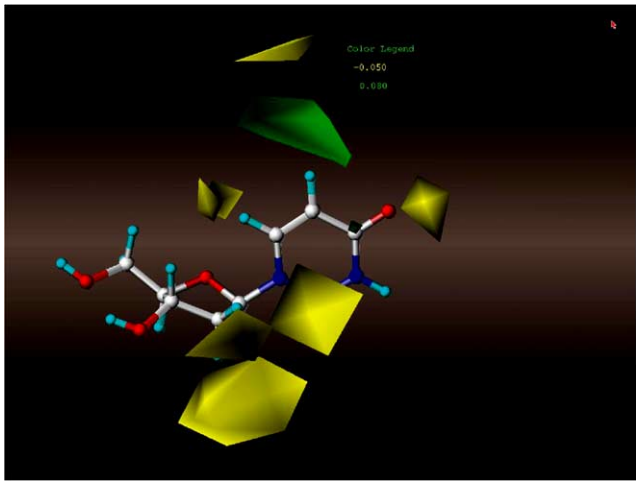


Figure 5. CoMFA SD * Coeff contour plots for steric fields with dUrd (**12**) as the reference. Sterically favored (contour level 0.080) areas are shown in green, while the yellow isopleths depict sterically disfavored areas (contour level −0.050).

substituents at this position. Two large yellow isopleths near the 2′-position of the sugar unit represent unfavorable steric conditions at the 2′-α (ribose) and the 2′-β (arabinose) positions. The presence of 2′-β hydroxyl groups in AraU (**18**) and AraT (**30**) are in agreement with their inactivity. The fluorine atoms at 2′-β are responsible for reduced activities of FMAU (**13**) and FIAU (**16**) although both compounds have favorable small groups at C-5. Both of the fluorine atoms at the 2′-position of dFdU (**25**) are located near the yellow regions at the α and β faces of C-2′, which is coherent with the inactivity of this compound. The third smaller yellow polyhedron is in proximity to the methylene group at 2′-position of 2′-methylene-2′-deoxythymidine (**26**), which accounts for the inactivity of this compound.

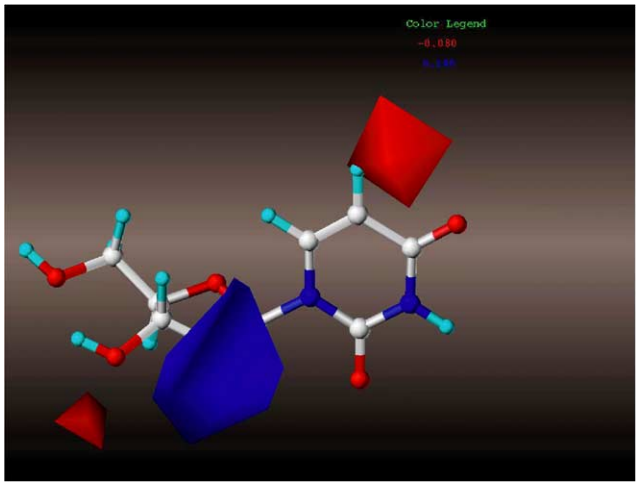


Figure 6. CoMFA SD * Coeff contour plots for electrostatic fields with dUrd (**12**) as the reference. Blue isopleths depict areas where positively charged groups increase activity (contour level 0.140) and red areas indicate increase in the activity with negatively charged groups (contour level −0.080).

Electrostatic fields based on the PLS analysis of the CoMFA model are shown in Figure 6. A large blue isopleth near the 2′-position is representing an area where a positive charge is favored. Fluoro or hydroxyl groups at the 2′-position, bearing negative GH charges on the 'O' and 'F' atoms diminish (**18**, **22**, **25**, **30**, **31**) or decrease the activity (**13**, **16**). Smaller electronegative groups (hydroxyl or azido) at 3′-position are essential for high activities, represented by the red favorable isopleth near this area. Compound **10** having a larger methylazido group at C-3′ shows a lower activity. Halogen atoms bearing negative GH charges at C-5 position, as in 5-BrdUrd (**17**) and 5-FdUrd (**29**) have high activity, whereas the amino group in compound **21**, having positive GH charges on the hydrogen atoms, diminished activity.

3.6. CoMSIA contour maps

Unlike the CoMFA contour maps that represent the 'pseudo-receptor' region where aligned molecules would interact favorably or disfavorably, the CoMSIA method provides contours that allow each field contributions to be mapped directly to regions within the structures.[42] The color schemes for steric and electrostatic fields are represented in a similar fashion as those of CoMFA contours. Both steric (Fig. 7) and electrostatic (Fig. 8) contours obtained from the CoMSIA model appear to be superior to the CoMFA contours providing excellent correlation with the predicted data. The reason for the lack of activity of the compounds **19** and **20** are clearly demonstrated in Figure 7, which shows a very large yellow isopleth near C-5 corresponding with the positions of the halogen atoms in both molecules. The green region closer to C-5 indicates increase in activities compared to dUrd (**12**) for compounds with halogen- (**17**, **29**), methyl- (**15**), and ethyl substituents (**14**). The yellow isopleth at the 2′-position covers both the α (ribose) and the β (arabinose) faces of the sugar portion indicating

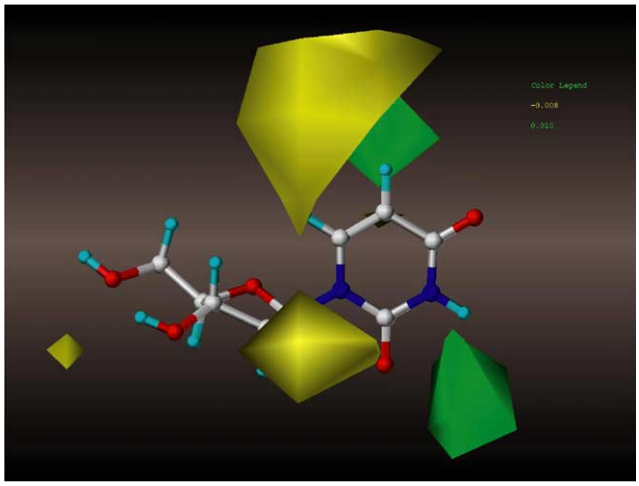


Figure 7. CoMSIA SD * Coeff contour plots for steric fields with dUrd (**12**) as the reference. Sterically favored (contour level 0.010) areas are shown in green, while the yellow isopleths depict sterically disfavored areas (contour level −0.008).

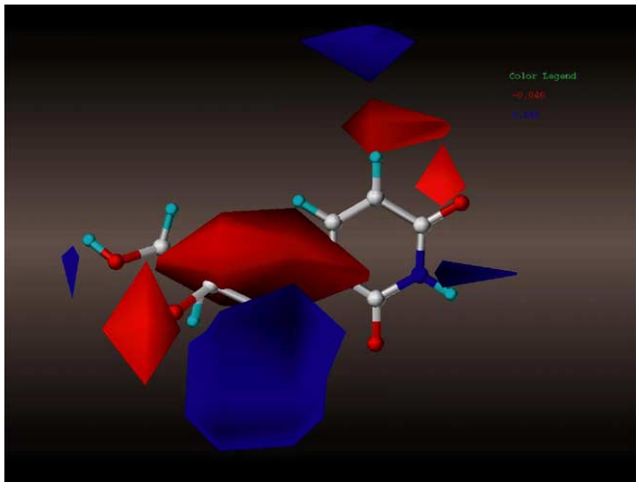


Figure 8. CoMSIA SD * Coeff contour plots for electrostatic fields with dUrd (**12**) as the reference. Blue isopleths depict areas where positively charged groups increase activity (contour level 0.040) and red areas indicate increase in the activity with negatively charged groups (contour level −0.040).

reduced (**13**, **16**) or diminished activities (**18**, **22**, **30**, **31**). Among the compounds with azido groups at the 3′-position (**7**, **8** and **10**), compound **10** has the lowest activity as it is expressed by the yellow region in some distance to C-3′. The benzyl group of 3-BnThd (**5**) and most of the bulky butyl group of 3-BuThd (**4**) are embedded in the green region near N-3 of the pyrimidine ring, which corresponds with their higher activities compared with 3-MeThd (**1**), 3-EtThd (**2**), and 3-*i*-PrThd (**27**).

The presence of a blue region beyond C-5 (Fig. 8) correlates with the position of the halogen atoms of compounds **19** and **20**, which accounts for the lack of activity of both compounds. A large red region on the β-face of the sugar ring appears to be due to negatively charged 'backbone' including the ring oxygen atom of the sugar portion and the N-1 nitrogen of the pyrimidine base. Electronegative fluoro- or hydroxyl groups at the 2′-position are located in a very large blue region, which indicates diminished (**18**, **22**, **25**, **30**, **31**) or decreased activity (**13**, **16**). The red polyhedron near C-3′ indicates high activities for compounds having hydroxyl or azido at 3′-position. The red isopleth in close proximity to the C-5 position corresponds with the high activities of 5-BrdUrd (**17**) and 5-FdUrd (**29**). It is noteworthy that the central nitrogen atoms of the azido groups in AZT (**7**), AZU (**8**) and compound **10** have negative GH charges, whereas the encompassing nitrogen atoms have positive GH charges. The central nitrogen atoms of AZT (**7**) and AZU (**8**) are favorably located within the red region whereas the central nitrogen atom of compound **10** is located outside of the red region indicating its lower activity.

The graphical interpretation of the H-bonding donor interaction in the CoMSIA model is represented in Figure 9. It highlights areas beyond the molecules where putative hydrogen acceptor groups in the enzyme can form H-bonds with molecule thereby influencing binding affinities. The cyan isopleths near the hydrogen atoms of the 3′- and 5′-OH groups represent favorable interactions with the H-bond donor surface of the molecule. This indicates the necessity of hydroxyl groups at 3′- and 5′-positions for activity. Likewise, a large cyan isopleth (favored) is present near the hydrogen bound to N-3. Therefore, Thd (**15**) is more active than its N-3 alkylated derivatives 3-MeThd (**1**), 3-EtThd (**2**), 3-BuThd (**4**), 3-BnThd (**5**), and 3-*i*-PrThd (**27**). Compound **6**, having two OH groups in close proximity to the large purple regions beyond the carbonyl oxygens at C-2 and C-4 of the pyrimidine ring, shows decreased activity compared to the other N-3 alkylated Thds. A purple polyhedron near the hydrogen atom of the amino group at C-5 of compound **21** is indicative of a disfavored

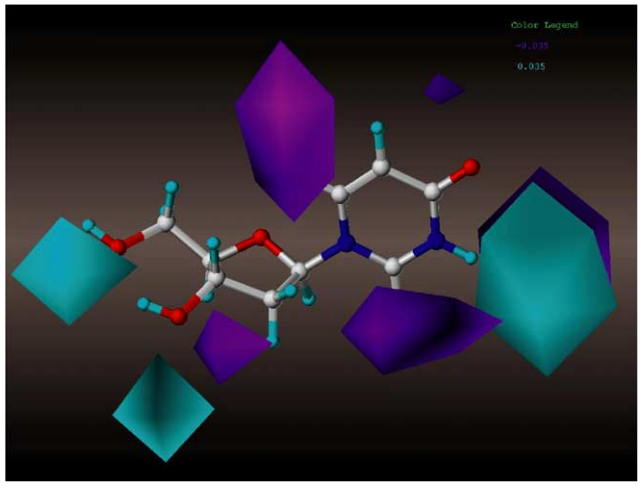


Figure 9. The contour plots of the CoMSIA H-bond donor fields (SD * Coeff) with dUrd (**12**) as the reference. Cyan isopleth contours (contour level 0.035) beyond the structures where an H-bond donor group in the compound will be favorable, while purple isopleths (contour level −0.035) represent H-bond donor in the compound unfavorable.

H-bond donor surface, as it is reflected by the low activity of this compound. The purple region near the 2′-position of the ribose portion demonstrates lack of activity associated with compound **22** and **31** having 2′-α hydroxyl groups. Both arabino nucleosides AraU (**18**) and AraT (**30**) show very low activity due to the proximity of the 2′-β hydroxyl groups to the purple regions around C-5 and the carbonyl group at C-2.

4. Conclusions

CoMFA and CoMSIA methods were successfully applied in our study to build 3D-QSAR models that can accurately predict relative TK1 phosphorylation rates (PRs) of TK1 substrates. Overall, the degree of predictivity of the CoMSIA model appeared to be superior to that of the CoMFA model. However, a combined use of both the CoMFA and the CoMSIA model may be most suitable to predict the activities of novel TK1 substrates designed in silico. Three-dimensional structural coordinates obtained from CoMFA and CoMSIA contour maps could be used in database mining for existing unknown TK1 substrates.

Acknowledgements

This work was supported by the US Department of Energy Grant DE-FG02-90ER60972 and grants from The Swedish Research Council. The authors thank Mr. Rohit Tiwari for inspiring discussions.

References and notes

1. Arner, E. S. J.; Eriksson, S. *Pharmac. Ther.* **1995**, *67*, 155.
2. Schinazi, R. F.; Mead, J. R.; Feorino, P. M. *AIDS Res. Human Retrov.* **1992**, *8*, 963.
3. Al-Madhoun, A. S.; Tjarks, W.; Eriksson, S. *Mini-Rev. Med. Chem.* **2004**, *4*, 341.
4. Eriksson, S.; Munch-Petersen, B.; Johansson, K.; Eklund, H. *Cell. Mol. Life Sci.* **2002**, *59*, 1327.
5. Eriksson, S.; Wang, L. *Recent Adv. Nucleos.* **2002**, 455.
6. Broet, P.; Romain, S.; Daver, A.; Ricolleau, G.; Quillien, V.; Rallet, A.; Asselain, B.; Martin, P. M.; Spyratos, F. *J. Clin. Oncol.* **2001**, *19*, 2778.
7. Romain, S.; Bendahl, P. O.; Guirou, O.; Malmstrom, P.; Martin, P. M.; Ferno, M. *Int. J. Cancer* **2001**, *95*, 56.
8. Tanigawa, N.; Masuda, Y.; Muraoka, R.; Tanaka, T. *J. Surg. Oncol.* **1994**, *55*, 209.
9. Wu, C.; Yang, R.; Zhou, J.; Bao, S.; Zou, L.; Zhang, P.; Mao, Y.; Wu, J.; He, Q. *J. Immunol. Methods* **2003**, *277*, 157.
10. Weber, G. *Cancer Res.* **1983**, *43*, 3466.
11. Hampton, A.; Kappler, F.; Maeda, M.; Patel, A. D. *J. Med. Chem.* **1978**, *21*, 1137.
12. Chabner, B. A. *Cytidine Analogues. Cancer Chemotherapy and Biotherapy*, 2nd ed.; Lippincott-Raven: Philadelphia, 1996.
13. Cheson, B. D.; Keating, M. J.; Plunkett, W. *Nucleoside Analogs in Cancer Therapy*; Marcel Dekker: New York, 1997.
14. Galmarini, C. M.; Mackey, J. R.; Dumontet, C. *Leukemia* **2001**, *15*, 875.
15. Eriksson, S.; Wang, L. *Nucleos. Nucleot.* **1997**, *16*, 653.
16. Al-Madhoun, A. S.; Johnsamuel, J.; Barth, R. F.; Tjarks, W.; Eriksson, S. *Cancer Res.* **2004**, *64*, 6280.
17. Barth, R. F.; Yang, W.; Al-Madhoun, A. S.; Johnsamuel, J.; Byun, Y.; Chandra, S.; Smith, D. R.; Tjarks, W.; Eriksson, S. *Cancer Res.* **2004**, *64*, 6287.
18. Al-Madhoun, A. S.; Johnsamuel, J.; Yan, J.; Ji, W.; Wang, J.; Zhuo, J.-C.; Lunato, A. J.; Woollard, J. E.; Hawk, A. E.; Cosquer, G. Y.; Blue, T. E.; Eriksson, S.; Tjarks, W. *J. Med. Chem.* **2002**, *45*, 4018.
19. Lunato, A. J.; Wang, J.; Woollard, J. E.; Anisuzzaman, A. K. M.; Ji, W.; Rong, F.-G.; Ikeda, S.; Soloway, A. H.; Eriksson, S.; Ives, D. H.; Blue, T. E.; Tjarks, W. *J. Med. Chem.* **1999**, *42*, 3378.
20. Sabini, E.; Ort, S.; Monnerjahn, C.; Konrad, M.; Lavie, A. *Nat. Struct. Biol.* **2003**, *10*, 513.
21. Johansson, K.; Ramaswamy, S.; Ljungcrantz, C.; Knecht, W.; Piskur, J.; Munch-Petersen, B.; Eriksson, S.; Eklund, H. *Nat. Struct. Biol.* **2001**, *8*, 616.
22. Champness, J. N.; Bennett, M. S.; Wien, F.; Visse, R.; Summers, W. C.; Herdewijn, P.; De Clercq, E.; Ostrowski, T.; Jarvest, R. L.; Sanderson, M. R. *Proteins: Struct. Funct. Genet.* **1998**, *32*, 350.
23. Bird, L. E.; Ren, J.; Wright, A.; Leslie, K. D.; Degreve, B.; Balzarini, J.; Stammers, D. K. *J. Biol. Chem.* **2003**, *278*, 24680.
24. Gentry, G. A. *Pharmacol. Thera.* **1992**, *54*, 319.
25. Spadola, L.; Novellino, E.; Folkers, G.; Scapozza, L. *Eur. J. Med. Chem.* **2003**, *38*, 413.
26. Bhongade, B. A.; Gadad, A. K. *Bioorg. Med. Chem.* **2004**, *12*, 2797.
27. Datar, P.; Desai, P.; Coutinho, E.; Iyer, K. *J. Mol. Model.* **2002**, *8*, 290.
28. Doytchinova, I.; Valkova, I.; Natcheva, R. *Quant. Struct.–Act. Relat.* **2001**, *20*, 124.
29. Lepper, E. R.; Ng, S. S. W.; Guetschow, M.; Weiss, M.; Hauschildt, S.; Hecker, T. K.; Luzzio, F. A.; Eger, K.; Figg, W. D. *J. Med. Chem.* **2004**, *47*, 2219.
30. Sperandio da Silva, G. M.; Sant'Anna, C. M. R.; Barreiro, E. J. *Bioorg. Med. Chem.* **2004**, *12*, 3159.
31. Ungwitayatorn, J.; Samee, W.; Pimthon, J. *J. Mol. Struct.* **2004**, *689*, 99.
32. Zhu, L.; Hou, T.; Xu, X. *J. Mol. Model.* **2001**, *7*, 223.
33. Markiw, R. T.; Canellakis, E. S. *J. Org. Chem.* **1969**, *34*, 3707.
34. Segal, A.; Solomon, J. J.; Mukai, F. *Cancer Biochem. Biophy.* **1990**, *11*, 59.
35. Johansson, N. G.; Eriksson, S. *Acta Biochim. Pol.* **1996**, *43*, 143.
36. Xu, M.; Zhang, A.; Han, S.; Wang, L. *Chemosphere* **2002**, *48*, 707.
37. Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinhem, 1994.
38. Bush, B. L.; Nachbar, R. B., Jr. *J. Comput. Aided Mol. Des.* **1993**, *7*, 587.
39. Song, M.; Breneman, C. M.; Sukumar, N. *Bioorg. Med. Chem.* **2004**, *12*, 489.
40. Johnsamuel, J.; Lakhi, N.; Al-Madhoun, A. S.; Byun, Y.; Yan, J.; Eriksson, S.; Tjarks, W. *Bioorg. Med. Chem.* **2004**, *12*, 4769.
41. Byun, Y.; Yan, J.; Al-Madhoun, A. S.; Johnsamuel, J.; Yang, W.; Barth, R. F.; Eriksson, S.; Tjarks, W. *Appl. Radiat. Isot.* **2004**, *61*, 1125.
42. Klebe, G.; Abraham, U.; Mietzner, T. *J. Med. Chem.* **1994**, *37*, 4130.