http://dx.doi.org/10.5012/bkcs.2013.34.9.2680

Descriptor-Based Profile Analysis of Kinase Inhibitors to Predict Inhibitory Activity and to Grasp Kinase Selectivity

Hyejin Park,[†,‡] Kyeung Kyu Kim,[‡,*] ChangHoon Kim,[#] Jae-Min Shin,[†] and Kyoung Tai No[§,*]

[†]Bioinformatics & Molecular Design Research Center, Seoul 120-749, Korea
[‡]Department of Molecular Cell Biology, Samsung Biomedical Research Institute, Sungkyunkwan University School of Medicine, Suwon 440-746, Korea. [*]E-mail: kyeongkyu@skku.edu
[§]MACROGEN, Seoul 153-781, Korea. [*]E-mail: ktno@yonsei.ac.kr
[#]Department of Biotechnology, Yonsei University, Seoul 120-749, Korea

Received June 3, 2013, Accepted June 12, 2013

Protein kinases (PKs) are an important source of drug targets, especially in oncology. With 500 or more kinases in the human genome and only few kinase inhibitors approved, kinase inhibitor discovery is becoming more and more valuable. Because the discovery of kinase inhibitors with an increased selectivity is an important therapeutic concept, many researchers have been trying to address this issue with various methodologies. Although many attempts to predict the activity and selectivity of kinase inhibitors have been made, the issue of selectivity has not yet been resolved. Here, we studied kinase selectivity by generating predictive models and analyzing their descriptors by using kinase-profiling data. The 5-fold cross-validation accuracies for the 51 models were between 72.4% and 93.7% and the ROC values for all the 51 models were over 0.7. The phylogenetic tree based on the descriptor distance is quite different from that generated on the basis of sequence alignment.

Key Words : Descriptor, Profile, Kinase, SVM

Introduction

Protein kinases (PKs) are a family of enzymes that transfer a phosphate group to a substrate protein. A total of 518 kinases have been identified in the human genome. They consist of about 1.7% of all human genes, and the set of protein kinases is called kinome, a term that was coined by Gerard Manning and colleagues in 2002.[1] With the advancements in genomics, the systematical and evolutional analysis of the kinome was attempted. Representatively, the 518 human PKs and the evolution of PKs throughout eukaryotes were analyzed.[1,2]

Many kinases are common players in the cellular signaling and are involved in cellular processes such as cell differentiation, proliferation, and apoptosis. Since kinases represent a major control point of cell behavior, they have been implicated not only in oncological, but also in a number of non-oncological conditions, including central nervous system disorders,[3,4] autoimmune diseases,[5] osteoporosis,[6] and metabolic disorders.[7] Because of its biological importance, the kinase family has also been the target of drug discovery efforts. Although many kinase inhibitors have been developed, only 22 small molecules (including 3 indirect mTOR inhibitors) have been approved by the FDA. A reason for the low approval rate of kinase inhibitors is that their cross-reactivity with unintended targets can cause undesired side effects. Since the structural variation in the kinase family is very low and the catalytic domains of many kinases are highly conserved, many kinase inhibitors share the problem of broad selectivity.

Recent advances in high throughput techniques have facilitated large-scale bioactivity profiling experiments that are useful in not only predicting the compound activity against various types of kinases but also in providing information on kinase selectivity. However, the systematic screening of large numbers of compounds against a large number of kinases is a difficult, expensive, and time-consuming process. Therefore, computational kinase profiling is a promising approach in the effort to discover novel kinase inhibitors.

Previously, many researchers and companies have used computational kinase profiling.[8-10] This strategy is routinely applied to predict the kinase enzymatic and cellular activity. However, those studies have focused on activity prediction and have a limited ability to predict the selectivity of a kinase inhibitor.

To date, studies examining kinase inhibitor selectivity have been carried out using chemogenomic approaches.[11-17] Because these approaches depend on the variation of specific kinase sequences, they were limited to the elucidation of the kinase inhibitor selectivity for weak inhibitor-sequence relation.

Here, we studied the inhibition profiling data for 51 kinases to predict the kinase inhibitory activity of the corresponding inhibitors and to understand their selectivity using a ligand-related descriptor-based approach. Our approach enables a kinomic view of kinase inhibitor selectivity.

Methods

Dataset. In order to generate the dataset, we extracted the kinase bioactivity data from the Kinase SARfari database (https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari), which is the largest public database for protein kinases.[18] This database provides curated data sets comprising the ChEMBL SAR data derived from published literature. The data set used in this study contains 51 human kinase domains that have over 100 compounds for which the IC_{50} values are available. For our study, we divided the kinaseinhibitor pairs into 2 classes (active and inactive) on the basis of the IC_{50} value. The class with IC_{50} value smaller than 10 μM was defined as active, while the class with IC_{50} value higher than 10 μM was defined as inactive.

Then, 1058 2D molecular descriptors for each compound were calculated using the preADMET program.[19] The number of molecular descriptors was filtered based on their deviation and correlation. Finally, the molecular descriptors were ranked by their signal-to-noise ratio. The signal-to-noise ratio is defined as:

$$\text{Signal to noise ratio} = \frac{\text{AVG(Active)} - \text{AVG(Inactive)}}{\text{STD(Active)} + \text{STD(Inactive)}}$$

In the above equation, AVG (class) means the average of descriptor value on each class, STD (class) means standard deviation of each class.

Model Building. Support vector machine (SVM) is a statistical method for classification, wherein the samples are projected into a feature space and a hyperplane is constructed in that space. The decision function can be used to make predictions for test samples. Both the training and test samples can be efficiently projected into the feature space *via* kernel functions.

The E1071 package of the R program (http://cran.r-project.org/web/packages/e1071/) was used to build the kinase models for radial basis SVM. The gamma, cost, and the number of descriptors were optimized by a 5-fold cross validation. The performance of a 5-fold cross validation was measured by accuracy = (TP + TN)/(TP + TN+ FP + FN), where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. To reliably estimate the performance, receiver operating characteristic curve (ROC) analysis was performed. In this analysis, the area under the ROC curve (AUC) corresponding to random classification is 0.5, while that at an optimal state is 1.

Construction of Descriptor-Based Phylogenetic Tree for 51 Kinase Models. To compare the 51 kinase models, the descriptor similarity matrix was generated. In this matrix, each descriptor was encoded by Boolean values according to the usage of each kinase model. To generate the distance matrix for the 51 kinase models, the distance between each kinase model was measured using the Tanimoto distance. Using this distance matrix, a hierarchical clustering by the Ward method was performed to generate the phylogenetic tree. The sequence similarity matrix was generated using aligned sequence in Kinase SARfari database. The Figtree program (http://tree.bio.ed.ac.uk/software/figtree/) was used to draw the phylogenetic tree.

Results and Discussion

To understand the kinase selectivity that represents the major hurdle in kinase drug discovery studies, we tried to profile the kinase inhibitory activity. In order to profile the kinase inhibitor activity, we constructed protein kinase inhibitor classification models based on SVM. The schematic diagram of our method is shown in Figure 1. As a first step, we chose 51 human protein kinases from the dataset obtained from ChEMBL kinase SARfari and constructed classification models for them using the SVM method with an IC_{50} cutoff of 10 mM. The 5-fold cross-validation method was used to validate our models. As shown in Figure 2 and Table 1, the 5-fold cross-validation accuracies for the 51 models were between 72.4% and 93.7% and the ROC values for all the 51 models were over 0.7. These measures indicate the robustness of the constructed kinase models.

To predict the selectivity of the kinase inhibitors in our model, we analyzed the distribution of the descriptors for all

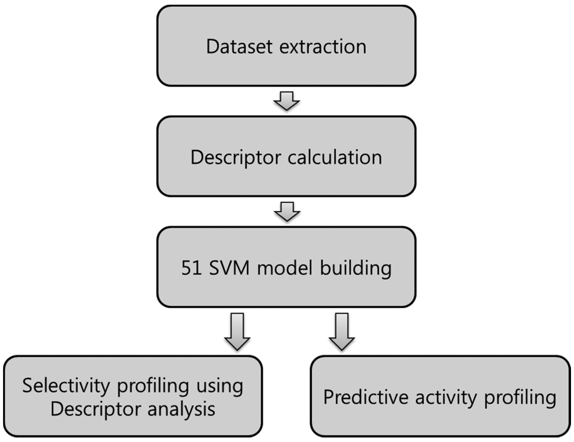
Dataset extraction
Descriptor calculation
51 SVM model building
Selectivity profiling using Descriptor analysis
Predictive activity profiling

Figure 1. Schematic diagram of the descriptor-based profile analysis.

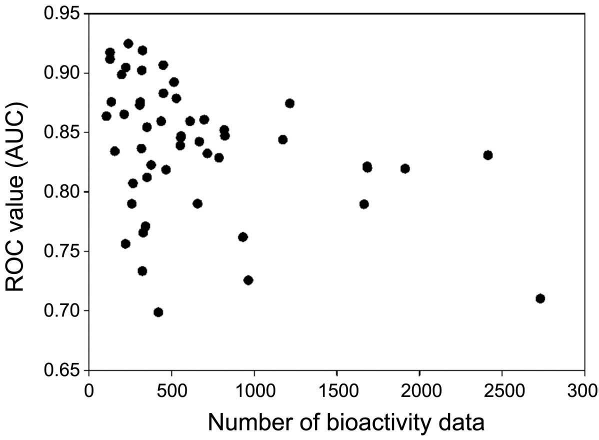

Figure 2. The bioactivity data-derived ROC values for the 51 kinase models.

Table 1. Performance of the 51 kinase models based on a 5-fold cross-validation

Kinase	No. of IC_{50}	No. of Descriptors	Accuracy	ROC
hABL1	350	30	0.871	0.854
hABL	135	30	0.874	0.876
hAKT1	611	20	0.882	0.859
hAURa	376	30	0.824	0.823
hBRAF	238	30	0.924	0.925
hCAMK2A	127	30	0.937	0.917
hCAMK2B	127	30	0.929	0.912
hCDK1	1664	30	0.792	0.790
hCDK2	1913	30	0.848	0.820
hCDK4	1172	30	0.85	0.844
hCDK5	667	30	0.843	0.842
hChk1	931	30	0.843	0.762
hCSNK2A1	156	30	0.885	0.834
hEGFR	2414	30	0.85	0.831
hFGFR1	716	30	0.832	0.832
hFLT3	266	40	0.82	0.807
hGSK3a	436	30	0.856	0.859
hGSK3b	786	30	0.837	0.829
hHER2	818	40	0.856	0.852
hIGF1R	557	30	0.86	0.847
hIKBKB	323	20	0.817	0.733
hINSR	221	30	0.792	0.756
HJNK1	342	30	0.769	0.771
hJNK2	212	30	0.849	0.865
hJNK3	197	30	0.898	0.899
hMEK1	327	40	0.801	0.766
hMET	259	20	0.826	0.790
hMK2	317	20	0.836	0.836
hp38a	1685	30	0.896	0.820
hp38b	351	30	0.815	0.812
hp38d	307	20	0.879	0.873
hp38g	310	30	0.881	0.876
hPDGFRa	420	30	0.724	0.699
hPDPK1	105	30	0.867	0.864
hPKAa	552	30	0.837	0.839
hPKCa	822	30	0.848	0.847
hPKCb1	697	30	0.864	0.861
hPKCe	555	30	0.865	0.846
hPKCg	528	30	0.886	0.879
hPKCh	449	30	0.906	0.907
hPKCt	514	30	0.893	0.892
hPKCz	450	30	0.918	0.883
hPKD1	324	30	0.932	0.919
hPKD3	320	30	0.922	0.902
hPLK1	223	20	0.924	0.905
hSRC	1683	30	0.844	0.821
hTIE2	466	30	0.843	0.819
hVEGFR1	657	30	0.796	0.790
hLCK	1215	30	0.88	0.874
hVEGFR2	2731	30	0.785	0.710
hPDGFRb	964	30	0.789	0.726

the kinase models. A total of 466 descriptors were used to generate the models and 40 highly correlated descriptors were collected for each model. As shown in Figure 3, the maximum number of kinase models that was described by at least one descriptor was 13, and 146 descriptors were uniquely explained by a single kinase model. Only 10 descriptors were repeated in over 10 kinase models and three quarters of the descriptors were repeated in fewer than 3 models. These results indicate that the descriptors that were used to generate the kinase models could clearly elucidate the kinase selectivity.

We analyzed the distribution of the 146 descriptors that were not repeated in any kinase models (Figure 4). For example, while the 13 descriptors of the MEK1 model have no redundancies in other kinase models, the descriptors of several other kinases (*e.g.*, ABL1, an isoform of p38 and protein kinase C (PKC), and an isoform of protein kinase D (PKD)) are repeated in other models more than once. This indicates that these kinases may be expected to be less selective than the others in our model.

To prove the results of our analysis, we tried to perform a classification of the kinases using 2 principles: 1) a classifi-

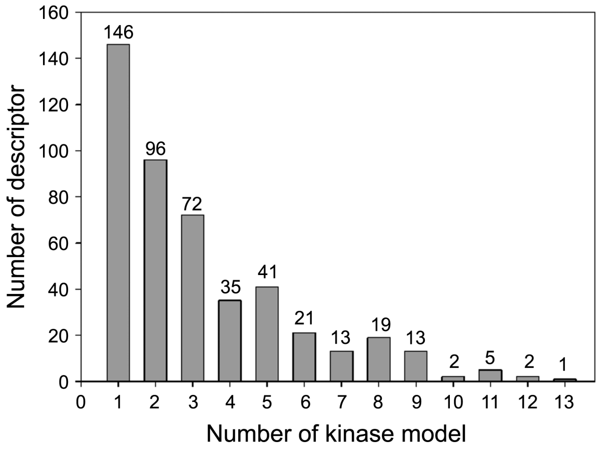

Figure 3. The distribution of the descriptor of the 51 kinase models.

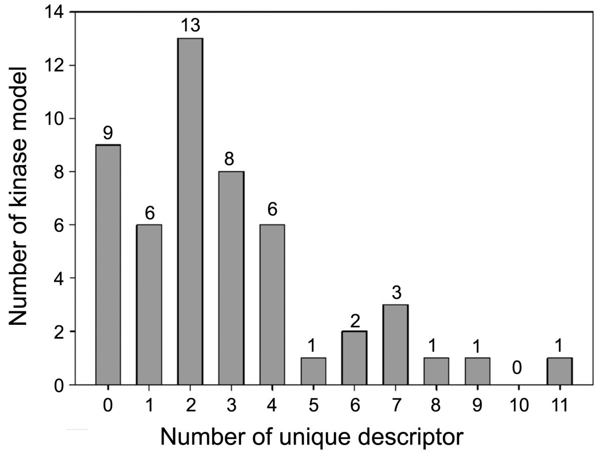

Figure 4. The distribution of the nonrepeated descriptors of the 51 kinase models.

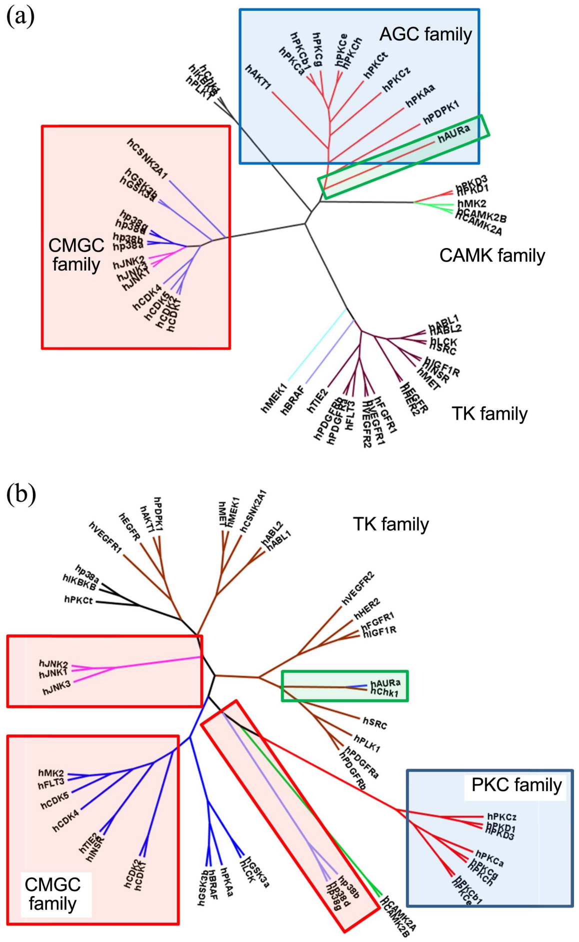

Figure 5. A phylogenetic tree of the 51 kinases using (a) the kinase sequence-based clustering and (b) the descriptor distance-based clustering methods. Major differences between 2 trees are JNK kinase (Red box), PKC family (Blue box), and Aurora A kinase (Green box).

cation based on sequence alignment and 2) a classification based on the descriptors. To better visualize the results, we constructed the corresponding phylogenetic trees. In the phylogenetic tree constructed on the basis of sequence alignment [Figure 5(a)], the kinome family is well defined. The general human kinome is evolutionarily divided into 7 major groups[1].The AGC family contains the kinases protein kinase A(PKA), protein kinase C (PKC), and protein kinase G (PKG). The CaMK family contains the calcium/calmodulin-dependent protein kinases. The CK1 family contains the casein kinase 1 group. The CMGC family contains cyclin dependent kinase (CDK), MAP kinase (MAPK), Glycogen synthase kinase 3 (GSK3), and CDC-like kinase (CLK). The STE family contains the homologs of yeast Sterile 7, Sterile 11, and Sterile 20 kinases. The TK family contains the tyrosine kinases, and the TKL family contains the tyrosine kinase-like group of kinases. Many of the 51 selected kinases belong to the TK, CMGC, and AGC families.

The phylogenetic tree based on the descriptor distance is quite different from that generated on the basis of sequence alignment. Distinct from the sequence-based tree, the descriptor-based tree was divided into 4 kinases families, the TK, AGC, CaMK, and CMGC families. The CMGC family comprised p38, c-Jun kinase (JNK), and other kinases. Most of the proteins that in the previous tree belonged to the AGC and CaMK families are rearranged into the TK family. According to the descriptor-based phylogenetic tree, the similarity between the inhibitors of 2 kinases is well explained by this rearrangement. In particular, while the PKC isoform family is included in the AGC family in the sequence-based tree, the PKC family constructed an independent tree in our model tree [Figure 5(b)]. This result correlates with the analysis performed using the distribution of the descriptors.

Table 2. Comparison of the overlapped descriptors among the 8 PKC family models and the 51 kinase models

Descriptor	8 PKC clusters	51 models
AlogP98 atom type 025 C	7	13
Information Content	6	12
Valence bound charge index03	8	12
BCUT highest eigenvalue 03 MPEOE charge	2	11
Kier steric descriptor	6	11
Number of aromatic bonds	6	11
SK atom type melting point	0	11
Valence charge index_03	8	11
BCUT lowest eigenvalue 02 mass	7	10
BCUT lowest eigenvalue 03 mass	5	10

As previously mentioned, the descriptors that represented the PKC isoform models are not unique, meaning that PKC kinases have a lower selectivity than the other kinases. The known PKC inhibitors have a similar structure to Staurosporine, which is a well-known non-specific kinase inhibitor. This finding indicates low kinase selectivity in the PKC family.[20]

The descriptors with “*BCUT highest eigenvalue 03 MPEOE charge*” and “*SK atom type melting point*” are highly ranked in the descriptor distribution but they are not used for the PKC kinase model (Table 2). Therefore, these descriptors are presumed to be useful to predict the PKC kinase sensitivity. As the MPEOE charge and SK atom type are a unique parameters used in the preADMET program, these descriptors can only be used by this program.[19]

In the descriptor-based tree, Aurora A kinase is rearranged into the TK family. This is a remarkable difference from the arrangement in the sequence-based tree. Certain well-known inhibitors of Aurora A kinase also have inhibitory activity against c-Src tyrosine kinase. Even though Aurora A kinase belongs to the family of mitotic serine/threonine kinases and it shares sequences homology with members of the AGC family, the structure and function of its inhibitors are similar to the inhibitors of the tyrosine kinase family.[20] Our analysis underpinned the results of previous reports.

Therefore, we suggest that the descriptor-based profile analysis for kinase inhibitors can provide new insights on how to classify a kinase family based on the selectivity of its inhibitors. Moreover, our analysis proved to be more informative than the sequence-based phylogenetic tree in the description of the kinase selectivity.

Conclusions

In this study, we built kinase models for 51 kinases and tried to profile the activity and selectivity of each model. For each model, we used a descriptor-based simulation using the SVM method and analyzed the repeated descriptors. From this approach, we could suggest a new classification of the kinase families and successfully described the selectivity of JNK and PKC. Furthermore, our model can explain why the Aurora kinase inhibitors are similar to the c-Src kinase inhibitors. Therefore, our ligand-related descriptor-based approach is useful for obtaining systematical information on kinase inhibitor activity and selectivity by suggesting virtual inhibitory profiles and can describe certain features of kinase selectivity. Future studies using this approach could further elucidate kinase selectivity.

Acknowledgments. This study was supported by a grant of the Korea Healthcare technology R&D Project, Ministry for Health, Welfare & Family Affairs, Republic of Korea (A100096). Also, Special thanks to Dr. Jee-Young Lee (Bioinformatics & Molecular Design Center) and Dr. Ky-Youb Nam (Gachon University) for organize this manuscript.

References

1. Manning, G.; Whyte D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. *Science* **2002**, *298*, 1912.
2. Schiffer, C. A. *N. Engl. J. Med.* **2007**, *357*, 258.
3. Smith, W. W; Pei, Z.; Jiang, H.; Dawson, V. L.; Dawson, T. M.; Ross, C. A. *Nature Neurosci.* **2006**, *9*, 1231.
4. Hayashi, M. L.; Rao, B. S.; Seo, J.; Choi, H.; Dolan, B. M.; Choi, S.; Chattarji, S.; Tonegawa, S. *Proc. Natl. Acad. Sci.* USA **2007**, *104*, 11489.
5. Whartenby, K. A.; Calabresi, P. A.; McCadden, E.; Nguyen, B.; Kardian, D.; Wang, T.; Mosse, C.; Pardoll, D. M.; Small, D. *Proc. Natl. Acad. Sci.* USA **2005**, *102*, 16741.
6. Buckbinder, L.; Crawford, D. T.; Qi, H.; Ke, H. Z.; Olson, L. M.; Long, K. R.; Bonnette, P. C.; Baumann, A. P.; Hambor, J. E.; Grasser, W. A. 3rd; Pan, L. C.; Owen, T. A.; Luzzio, M. J.; Hulford, C. A.; Gebhard, D. F.; Paralkar, V. M.; Simmons, H. A.; Kath, J. C.; Roberts, W. G.; Smock, S. L.; Guzman-Perez, A.; Brown, T. A.; Li, M. *Proc. Natl Acad. Sci.* USA **2007**, *104*, 10619-10624 .
7. Solinas, G.; Vilcu, C.; Neels, J. G.; Bandyopadhyay, G. K.; Luo, J. L.; Naugler, W.; Grivennikov, S.; Wynshaw-Boris, A.; Scadeng, M.; Olefsky, J. M.; Karin, M. *Cell Metab.* **2007**, *6*, 386.
8. Martin, E. J.; Sullivan, D. C. *J. Chem. Inf. Model.* **2008**, *48*, 873.
9. Martin, E. J.; Sullivan, D. C. *J. Chem. Inf. Model.* **2008**, *48*, 861.
10. Martin, E. J.; Mukherjee, P.; Sullivan, D. C.; Jansen, J. *J. Chem. Inf. Model.* **2011**, *51*, 1942.
11. Caffrey, D.; Lunney, E.; Moshinsky, D. *BMC Bioinf.* **2008**, *9*, 491.
12. Zhang, X.; Fernaìndez, A. *Mol. Pharm.* **2008**, *5*, 728.
13. Sciabola, S.; Stanton, R. V.; Wittkopp, S.; Wildman, S.; Moshinsky, D.; Potluri, S.; Xi, H. *J. Chem. Inf. Model.* **2008**, *48*, 1851.
14. Sheridan, R. P.; Nam, K.; Maiorov, V. N.; McMasters, D. R.; Cornell, W. D. *J. Chem. Inf. Model.* **2009**, *49*, 1974.
15. Lapins, M.; Wikberg, J. *BMC Bioinf.* **2010**, *11*, 339.
16. Ma, X. H.; Wang, R.; Tan, C. Y.; Jiang, Y. Y.; Lu, T.; Rao, H. B.; Li, X. Y.; Go, M. L.; Low, B. C.; Chen, Y. Z. *Mol. Pharm.* **2010**, *7*, 1545.
17. Niijima, S.; Shiraishi, A.; Okuno, Y. *J. Chem. Inf. Model.* **2012**, *52*, 901.
18. Gaulton, A.; Bellis, L. J.; Bento A. P.; Chambers, J.; Davies M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucleic Acids Res.* **2012**, *40*, D1100.
19. Lee, S. K.; Chang, G. S.; Lee, I. H.; Chung, J. E.; Sung, K. Y.; No, K. T EuroQSAR 2004.
20. Bain, J.; Plater, L.; Elliott, M.; Shpiro, N.; Hastie, C.; McLauchlan, H.; Klevernic, I.; Arthur, J. S.; Alessi, D. R.; Cohen, P. *The Biochemical Journal* **2007**, *408*, 297.