Molecules **2013**, *18*, 10789-10801; doi:10.3390/molecules180910789

OPEN ACCESS

molecules

ISSN 1420-3049

www.mdpi.com/journal/molecules

Article

Enhanced QSAR Model Performance by Integrating Structural and Gene Expression Information

Qian Chen [†], Leihong Wu [†], Wei Liu, Li Xing and Xiaohui Fan *

Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China; E-Mails: 21119019@zju.edu.cn (Q.C.); lan.seldas@gmail.com (L.W.); liuw@zju.edu.cn (W.L.); xinglizjuer@gmail.com (L.X.)

* Author to whom correspondence should be addressed; E-Mail: fanxh@zju.edu.cn; Tel.: +86-571-8820-8596; Fax: +86-571-8820-8426.

Received: 14 June 2013; in revised form: 20 July 2013 / Accepted: 26 July 2013 / Published: 4 September 2013

Abstract: Despite decades of intensive research and a number of demonstrable successes, quantitative structure-activity relationship (QSAR) models still fail to yield predictions with reasonable accuracy in some circumstances, especially when the QSAR paradox occurs. In this study, to avoid the QSAR paradox, we proposed a novel integrated approach to improve the model performance through using both structural and biological information from compounds. As a proof-of-concept, the integrated models were built on a toxicological dataset to predict non-genotoxic carcinogenicity of compounds, using not only the conventional molecular descriptors but also expression profiles of significant genes selected from microarray data. For test set data, our results demonstrated that the prediction accuracy of QSAR model was dramatically increased from 0.57 to 0.67 with incorporation of expression data of just one selected signature gene. Our successful integration of biological information into classic QSAR model provided a new insight and methodology for building predictive models especially when QSAR paradox occurred.

Keywords: quantitative structure-activity relationships (QSAR); SAR paradox; molecular modeling; gene expression; integrative analysis

1. Introduction

With the assumption that structurally similar molecules have similar biological properties, quantitative structure-activity relationship (QSAR)-based *in silico* approaches have played an essential role in risk assessment [1,2], drug discovery and development [3], and classification of active compounds in many biological research projects [4–6]. One of the primary examples would be separation of chemicals into subgroups with different toxic responses using global QSAR models [7].

To enhance the predictability of structure-activity relationship, researchers have mainly been focusing on two aspects in model generation: *i.e.*, calculating new informative descriptors and developing more powerful machine learning methods. On the one hand, the number and type of new descriptors available for QSAR studies has increased tremendously. These new descriptors contain more information about compound geometry, connectivity, physical, chemical properties and so on. For instance, the newest version of DRAGON could generate over 3,300 descriptors. On the other hand, more and more sophisticated machine learning methods, such as support vector machine (SVM) [8–10], bayesian network [11,12] and ensemble approaches [13], have been introduced into QSAR to construct predictive models in past decades.

Despite these efforts devoted to improve the predictability of models, the QSAR practice has been facing great challenges, even with the implementation of diverse machine learning approaches [14–16]. Criticism of the constructed models was largely based on the grounds of their poor predictability when they are applied to independent external datasets. Several possible reasons for this, such as over-fitting [14,17,18], improper applicability domain [19–21], activity cliffs [22], insufficient prediction reliability [23] have been intensively investigated by the research community. Our recent study [24] also suggested that the selection of molecular descriptors and the insufficient validation strategy were primarily reasons attributing to the failure of many QSAR models and demonstrated that models likely yield more promising performances when using the most-frequently-selected descriptors from a series of the equivalent models. Moreover, the central premise of QSAR that structurally similar molecules have similar biological properties has also been seriously questioned, *i.e.*, the so-called QSAR paradox [25]. Indeed, recent advances in network pharmacology have revealed that the mode of drug action was more complex than expected [26–29]. Not only could a target interact with diverse drugs, but it is far more common for a drug to be acting on multiple targets rather than a single one. Furthermore, small changes on chemical structures of drugs could lead to dramatic fluctuations in their binding affinities to protein targets [29]. This violated the traditional understanding of QSAR that two similar molecules would likely possess similar biological properties through binding to the same protein target. As a result, molecular descriptors and sophisticated computational approaches may not be enough to address these problems related with QSAR modeling.

To avoid the QSAR paradox and to improve the model performance, we proposed a novel approach to integrate both structural information of compounds and their corresponding biological effects into QSAR modeling. The successful application of this approach was demonstrated by the improved prediction results of non-genotoxic carcinogenicity of compounds from a toxicological dataset, in which we only utilized the minimum number of conventional molecular descriptors along with the gene expression profiles of significant genes from microarray data. In comparison with classical QSAR model, our integrated model clearly showed better performance in prediction accuracy and reliability. Our work provided a new insight and methodology for building predictive toxicological models, especially when the QSAR paradox occurred. Although this study was implemented on a toxicological dataset, we believed this concept could be easily generalized to other biological-relevant phenotype predictions.

2. Results

2.1. Molecular Descriptors and Gene Expression Features Selection

As a consequence of the pretreatment process, a reduced set of 96 probes remained as the candidates for signature genes in the feature selection process, and the number of molecular descriptors was shrunk down from 929 to 108.

RFFS was applied to select best features for model construction. Since features would have appropriate predictive capability when selected in the classifiers, the frequencies of features were calculated for all the classifiers selected by RFFS. As a result, five molecular descriptors were found with a frequency higher than 0.1 in traditional QSAR models (Table 1). Similarly, gene feature classifiers were built for training set with 96 genes to select signature gene set, and information for one selected probe was also listed in Table 1. From Table 1, the frequency of appearance for JnJRn0195 (metallothionein) was 0.72 from the equivalents models, which undoubtedly exceeded the frequencies of other probes in the microarray. This result suggested that metallothionein would probably be the potential biomarker sufficient for identifying non-genotoxic carcinogens and non-carcinogens under this experimental setting. The frequencies of all the genetic probes and molecular descriptors were shown in Figure 1. The top five molecular descriptors and one genetic descriptor (JnJRn0195) were finally selected to construct the integrated model.

Table 1. Selected molecular descriptors and gene expression feature.

Feature Classes	Descriptors	Descriptions	Frequency
molecular descriptors	nN	number of nitrogen atoms	0.58
	CIC2	complementary information content (neighborhood symmetry of 2-order)	0.39
	C-005	CH_3X	0.18
	nS	number of Sulfur atoms	0.13
	C-008	CHR_2X	0.12
probes	JnJRn0195	Mt1a	0.72

2.2. Internal Validation of QSAR & Integrated Model

We constructed the final QSAR predictive model based on the five selected molecular descriptors. And integrated model was built using the same molecular descriptors with the addition of one signature gene, metallothionein. The performance of internal validation was assessed by five common metrics, which were the mean values and standard deviation for Acc., Sens., Spec., AUC, MCC derived from the equivalent models. As shown in Figure 2, the Acc., Sens., Spec., AUC, MCC of integrated model were significantly better than those of QSAR predictive model (** $p < 0.01$).

Figure 1. The frequency of features selected by equivalent models.

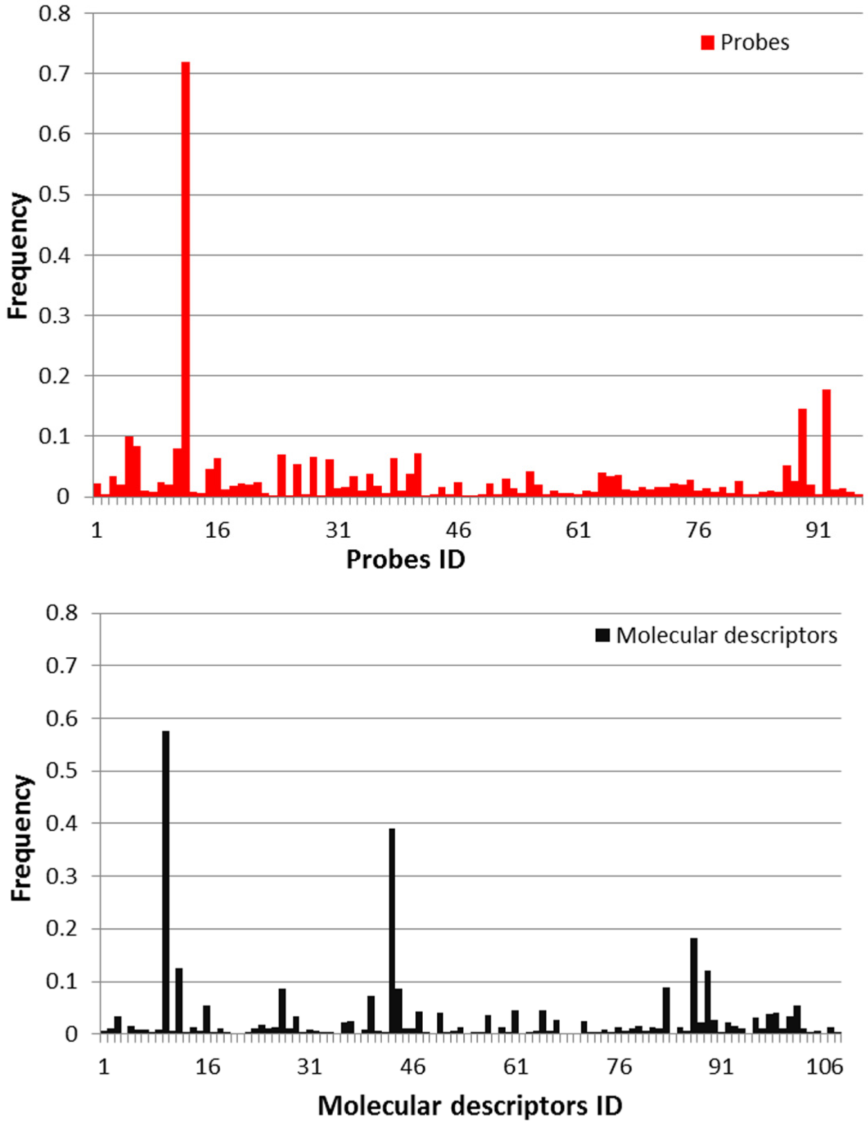
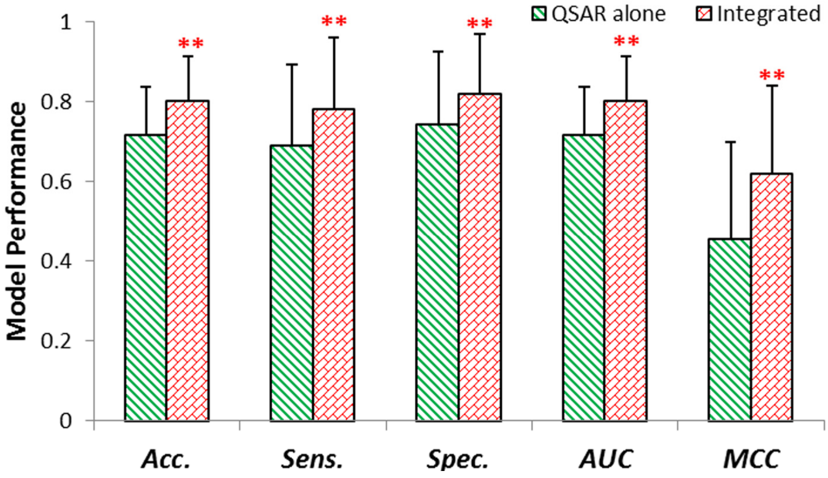


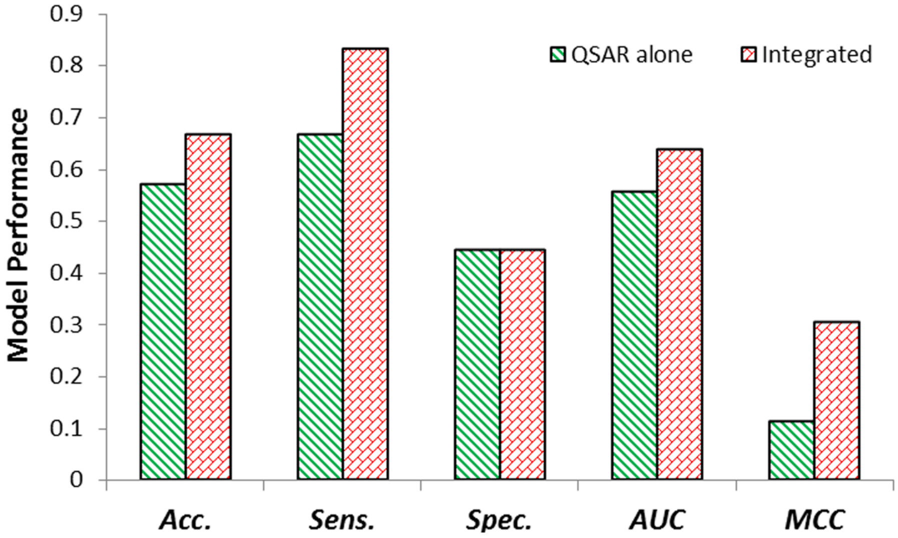
Figure 2. The performance of models using internal validation. QSAR alone model and Integrated model were evaluated by averaged Acc., Sens., Spec., AUC, MCC. ** p < 0.01 (compared to QSAR alone model).



2.3. External Validation of QSAR & Integrated Model

We tested our algorithms on the test set with 21 samples using the best QSAR model and integrated model from internal validation process. Again the integrated model outperformed the QSAR model, which was consistent with modeling results on training set (Figure 3).

Figure 3. The performance of models on test set.



3. Discussion

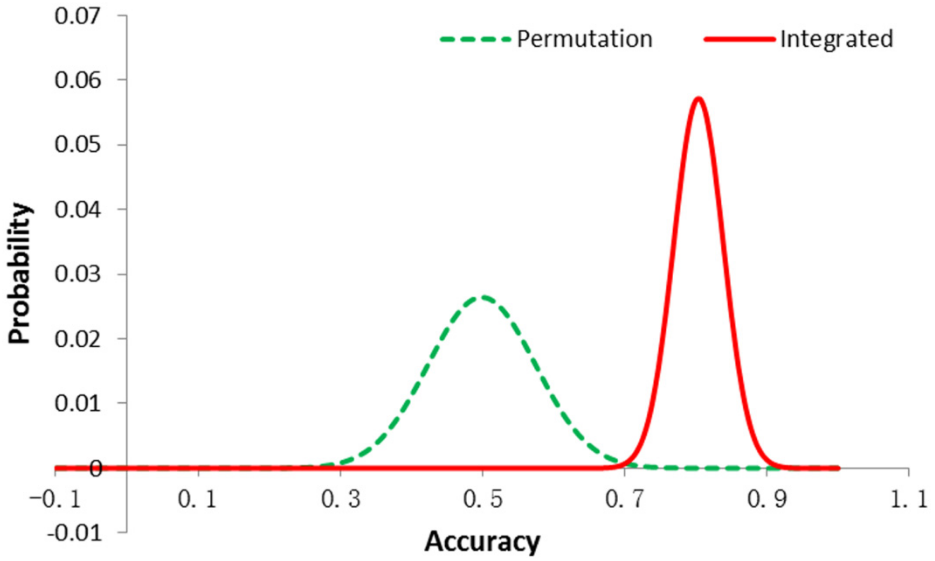
3.1. Permutation Test

Y-randomization test was applied in our study to evaluate the reliability of model training process. Fifty-seven samples with random labels were used to select features and build predictive models. The total number of active and inactive samples was kept same as training set accuracy of prediction from models was used as the determining index to compare the performance of the integrated model and Y-randomization model with top five molecular descriptors and one signature gene. It is assumed that the accuracy would fit to normal distribution for random models and integrated models. Based on the mean value and its standard deviation of equivalent models from training process, normal distribution curves were generated for integrated method and Y-randomization, separately. Figure 4 shows that ACC. of Y-randomization model was near 0.5 and the ACC. of integrated model was significantly higher than Y-randomization in training process, which confirmed the reliability of our algorithm.

3.2. Selection of Descriptors

Based on LOO-SVM modified by RFFS, the molecular descriptors with the frequencies above the predefined threshold were selected for the equivalent models as a near optimal descriptor set to develop a best predictive model. The reliability of this ranking approach has been confirmed on several toxicological and active data sets [24,30]. As shown in Figure 1, we set 0.1 as the cutoff value and considered the top five descriptors as the near optimal descriptor set for model construction. With the addition of the next descriptor in the frequency ranking list, the five statistical metrics of resulting QSAR model were all poorer than the current model with five descriptors. We also showed that the predicting model with top 4 molecular descriptors showed a better performance than top five descriptors in the QSAR model (See Supporting Information Table S2). Analysis on the exact values for the frequency of descriptor 4 and descriptor 5 in the equivalent models indicated these two values were extremely close, 0.126 *versus* 0.120, respectively. This result implied that these two descriptors may possess similar level of significance to the final model construction. Based on our feature selection criteria, we still selected five descriptors for model construction despite the fact that model with four descriptors gave better performance on internal validation. In addition, we have tried to incorporate two gene expression probes in the integrated model and the model showed worse predicting performance compared to the model with only one gene expression probe (See Supporting Information Figure S1). Moreover, single biological marker detection would be more rapid and much cheaper than the whole gene expression profile experiment, and it worked well with molecular descriptors as demonstrated in our study. Therefore, in this study, we only incorporated one gene expression probe in the integrated model.

Figure 4. Prediction accuracy of integrated model and permutated model.



3.3. Potential Action of the Signature Gene in the Carcinogenesis of Compounds

Based on our work, the integration of molecular descriptors and biological information can generate more reliable models, which indicated that information extracted from QSAR and genetic features cooperated each other in certain manner to provide deeper understanding of underlying mechanism of biological processes. We also have performed the model construction by only one significant gene (metallothionein) without traditional QSAR descriptors. The performance of resulting model was not as good as the integrated model (See Supporting Information Figure S2). According to our method, metallothionein was identified as the significant gene involved in non-genotoxic carcinogenesis, which is a metal-binding protein from a family of cysteine-rich, low molecular weight protein family. And it has been thought to play an important role in the toxicology of several kinds of metals [31–33].

Metallothionein has also been considered to play a crucial role in carcinogenesis [34,35] but the exact mechanisms are not very clear. Many studies suggested that metallothionein could have a protective effect on some carcinomas induced by either physical factor or chemical factor, such as ultraviolet radiation [36], the azo-dye *p*-dimethylaminoazobenzene [37] and copper gluconate [38]. More importantly, some researchers [39] found that the inactivation of metallothionein by gene knock-out experiments could increase the risk of carcinogenesis in the test animals. All of these studies supported that misfunction or abnormal expression of MT may be closely related to the carcinogenesis of compounds and our results in this work validated that inclusion of this biological information into QSAR model enhanced its performance on carcinogenesis propensity prediction for small molecule compounds.

Clearly, our method would not be suitable for the virtual screening experiments or molecular library optimization. With the increasing amount of genome information related to small molecule compounds [40–42], this work has opened a new window for toxicological genomics. QSAR and gene expression profile are two important tools in toxicity prediction realm, but few integrated studies of them have been done to take advantages from both methods. QSAR is excellent with its low cost and quick prediction, while gene expression profile focuses on its high-throughput property to generate amounts of genome information with relatively high expense. Integrated study of QSAR and gene expression profile might not only solve model performance problems on QSAR but also decrease the number of expensive gene expression profile experiments to be taken for toxicity assessments. The detection of single biological marker or a few biological markers would be much cheaper than the whole gene expression profile experiment, and it works well with molecular descriptors in our study. Therefore, it possesses considerable merit to include biological information into QSAR model to improve model performance when QSAR model alone is not ideal to predict some toxicity endpoints.

4. Materials and Methods

4.1. Toxicogenomics Dataset

The toxicogenomics dataset was taken from the literature [43], in which the non-genotoxic carcinogens (NGTCs) were labeled as ‘active’ and the non-carcinogens (NCs) were labeled as ‘inactive’. In this study, 78 compounds with expression data of 1,471 probes were identified, including their gene expression data at high dosage. Table 2 summarizes the information about the dataset used in our research. Briefly, these 78 compounds were divided into two sets as described in the reference [43]: the training set with 57 compounds (27 actives and 30 inactives) and the test set with 21 compounds (12 actives and nine inactives). More details were listed in the Supporting Information Table S1, such as chemical structures, molecular weight and their experimental classifications. In the pre-treatment process of 57 compounds, probes with standard deviation of expression values less than 0.5 or missing values were filtered out.

4.2. Molecular Descriptors

The 929 2-Dimensional (2D) molecular descriptors were calculated by a commercial software, DRAGON [44]. Similar to the pre-treatment process of gene expression data, molecular descriptors with 80% zeros, standard deviation less than 0.5 or correlation coefficient over 0.9 were removed for 57 compounds in the training set. The remaining descriptors represented a variety of structural information, such as number of rings, number of bonds, number of atoms, *etc.* Since NGTC compounds belonged to diverse chemical classes and may not share a common structural scaffold, a global QSAR strategy was employed to identify the key signatures of these compounds in common. All data were scaled to normal distribution to avoid predominant features.

Table 2. Summary of dataset.

	Training set	**Test set**
Samples	57	21
Actives	27	12
Inactives	30	9
Descriptors	929 → 108 → 5	
Probes	1471 → 96 → 1	

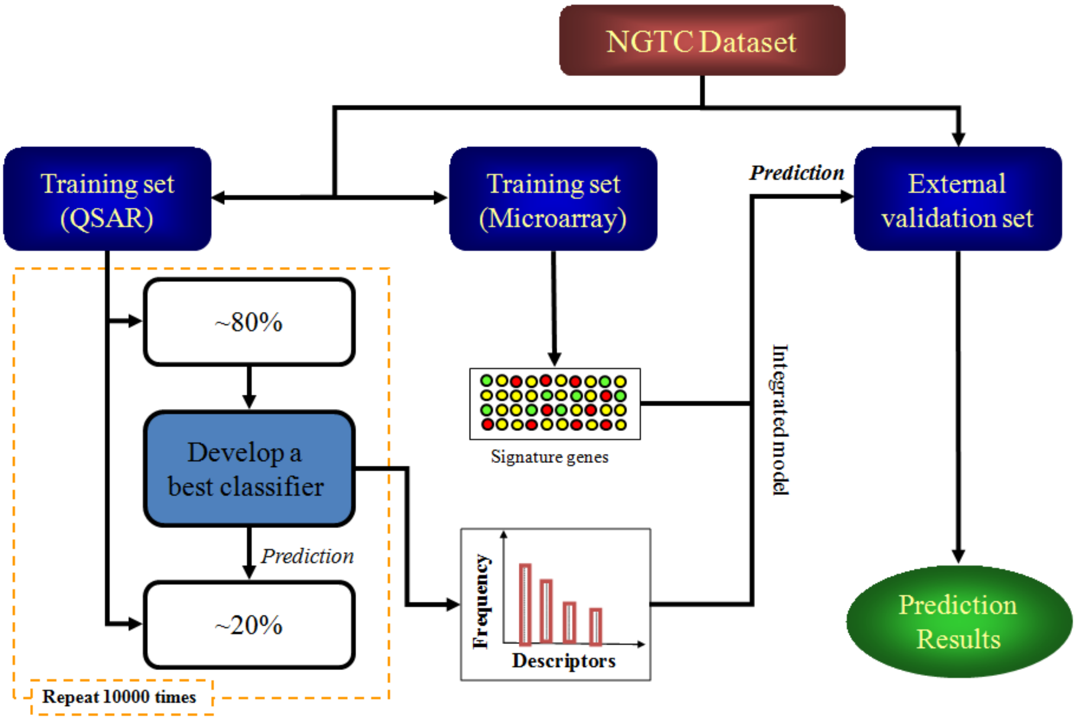
4.3. Software

The SVM used in this study has been implemented by libsvm-3.12 package, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. All the calculations were carried out on Windows7 operating system.

4.4. Study Design

The workflow of this study included a training process and an external validation process, which was depicted in Figure 5.

Figure 5. Work flow for developing the integrated model.



In order to select the best features (both molecular descriptors and significant genes) in the model training process, a random forward feature selection algorithm (RFFS) combined with a leave-one-out support vector machine (LOO-SVM) was used to achieve this aim [45–47]. RFFS was based on straight forward feature selection algorithm (SFFS) by overcoming the influence of feature order during the selection, a known weakness associated with SFFS. In other word, RFFS randomly re-ordered the features and samples for model construction then SFFS was taken to select features. LOO-SVM was the classifier to assess the "goodness" of the selected features. In brief, with the ratio of active to inactive samples fixed at 7:8, 45 samples were randomly selected from the training set to pick on the best features in the context of LOO-SVM modified by RFFS. Then, the models that were built on 45 samples with the selected features were employed to predict carcinogenicity of the remaining 12 samples. The process was repeated 10,000 times and the features would be kept for analysis if the accuracy of corresponding model was over 0.55. The selected features for these equivalent models were counted and their frequencies of presence were calculated to find best features [24,30]. Molecular descriptors and gene expression data were treated with the same method respectively to find out the most essential molecular descriptors and signature gene set with the best cross-validation performance in training set.

Leave-many-out support vector machine (LMO-SVM) was employed to build the final model, which was finally validated by using the test set. For SVM, radial kernel was used, and the parameters were optimized by grid search strategy. The best QSAR training model was built with the molecular descriptors obtained aforementioned, which were chosen according to their ranking on frequency of all the molecular descriptors for the equivalent models. The final predictive integrated model was built with top molecular descriptors and expression values of the signature genes from microarray, and the performance of QSAR model and integrated model were tested and compared on 21 compounds in the test set that were not included in the feature selection and model construction.

4.5. Model Construction and Validation

Leave-many-out strategy was applied in training process. In detail, 57 compounds in the training set were randomly separated into two parts, 45 samples for model training and 12 samples for internal validation. The ratio of active and inactive samples were fixed in the 45 samples, *i.e.*, 21 actives *versus* 24 inactives. During the model training, the best features from feature selection process were taken to build classifiers, which were then applied to the remaining 12 compounds to evaluate model performance. This process was repeated for 1,000 times to generate enough classifiers. Five common metrics, accuracy (Acc.), sensitivity (Sens.), specificity (Spec.), area under curve (AUC) and Matthew's coefficient correlation (MCC) [48,49], were calculated to evaluate the prediction model performance for all 1,000 models. Models with Acc. over 0.55 were considered as equivalent models. Mean value and standard deviation was calculated for these five metrics based on these equivalent models on training set.

A permutation test, Y-randomization, was also employed to compare classifier accuracy of integrated model *versus* chance alone [50,51]. Specifically, permutation test was carried out on training set compounds with the best features, five molecular descriptors and the expression level of one single signature gene. In each permutation, the 57 compounds' class labels in the training set were randomly scrambled. A model was developed using the randomly selected 21 actives and 24 inactives to predict the remaining 12 compounds. These processes were repeated for 10,000 times to construct enough classifiers to prove robustness of our integrated model. Again, accuracies of 10,000 models were calculated, as well as the mean value and standard deviation.

5. Conclusions

The novel modeling method in our work showed a significant enhancement on model performance by integrating molecular descriptors and biological information. The performance of final predictive model on test set was increased from 0.57 to 0.67 in term of accuracy with only one biological marker incorporated into QSAR model. The successful application of our integrated method suggested its great potential in predictive toxicological research. Although this study was implemented on a toxicological data, we believed the concept could be easily generalized to other biological-relevant phenotype predictions.

Supplementary Materials

Supplementary materials can be accessed at: http://www.mdpi.com/1420-3049/18/9/10789/s1.

Acknowledgments

This work was supported by the National S&T Major Project (No. 2012ZX09505001-001), the National Basic Research Program of China (No. 2012CB518405), the National Science Foundation of China (No. 81173465) and program for New Century Excellent Talents in University (NCET-12-0488).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Bradbury, S.P. Quantitative structure-activity relationships and ecological risk assessment: An overview of predictive aquatic toxicology research. *Toxicol. Lett.* **1995**, *79*, 229–237.
2. Loizou, G.; Spendiff, M.; Barton, H.A.; Bessems, J.; Bois, F.Y.; d'Yvoire, M.B.; Buist, H.; Clewell, H.J., III; Meek, B.; Gundert-Remy, U. Development of good modelling practice for physiologically based pharmacokinetic models for use in risk assessment: The first steps. *Regul. Toxicol. Pharmacol.* **2008**, *50*, 400–411.
3. Lill, M.A. Multi-dimensional QSAR in drug discovery. *Drug Discov. Today* **2007**, *12*, 1013–1017.
4. Frecer, V.; Miertus, S. Design, structure-based focusing and in silico screening of combinatorial library of peptidomimetic inhibitors of Dengue virus NS2B-NS3 protease. *J. Comput. Aided Mol. Des.* **2010**, *24*, 195–212.
5. González-Díaz, H.; Dea-Ayuela, M.A.; Pérez-Montoto, L.G.; Prado-Prado, F.J.; Agüero-Chapín, G.; Bolas-Fernández, F.; Vazquez-Padrón, R.I.; Ubeira, F.M. QSAR for RNases and theoretic–experimental study of Mol. Divers. on peptide mass fingerprints of a new Leishmania infantum protein. *Mol. Divers.* **2010**, *14*, 349–369.

6. Mercader, A.G.; Pomilio, A.B. QSAR study of flavonoids and biflavonoids as influenza H1N1 virus neuraminidase inhibitors. *Eur. J. Med. Chem.* **2010**, *45*, 1724–1730.
7. Enoch, S.; Cronin, M.; Schultz, T.W.; Madden, J. An evaluation of global QSAR models for the prediction of the toxicity of phenols to *Tetrahymena pyriformis*. *Chemosphere* **2008**, *71*, 1225–1232.
8. Czermiński, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct. Act. Relat.* **2001**, *20*, 227–240.
9. Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P.A.; Igglessi-Markopoulou, O.; Kollias, G. A combined LS-SVM & MLR QSAR workflow for predicting the inhibition of CXCR3 receptor by quinazolinone analogs. *Mol. Divers.* **2010**, *14*, 225–235.
10. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM TIST* **2011**, *2*, 27.
11. Burden, F.R.; Winkler, D.A. Predictive Bayesian neural network models of MHC class II peptide binding. *J. Mol. Graph. Model.* **2005**, *23*, 481–489.
12. Burden, F.R.; Winkler, D.A. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **1999**, *42*, 3183–3187.
13. Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R.P.; Song, Q. Boosting: An ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786–799.
14. Cronin, M.T.; Schultz, T.W. Pitfalls in QSAR. *J. Mol. Struct.* **2003**, *622*, 39–51.
15. Doweyko, A.M. QSAR: dead or alive? *J. Comput. Aided Mol. Des.* **2008**, *22*, 81–89.
16. Tetko, I.V.; Bruneau, P.; Mewes, H.-W.; Rohrer, D.C.; Poda, G.I. Can we estimate the accuracy of ADME–Tox predictions? *Drug Discov. Today* **2006**, *11*, 700–707.
17. Kolossov, E.; Stanforth, R. The quality of QSAR models: Problems and solutions†. *SAR QSAR Environ. Res.* **2007**, *18*, 89–100.
18. Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694–701.
19. Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.
20. Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **2004**, *112*, 1249.
21. Sushko, I.; Novotarskyi, S.; Ko□rner, R.; Pandey, A.K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Mu□ller, K.-R. Applicability domains for classification problems: Benchmarking of distance to models for AMES mutagenicity set. *J. Chem. Inf. Model.* **2010**, *50*, 2094–2111.
22. Maggiora, G.M. On Outliers and Activity Cliffs Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
23. Huang, J.; Fan, X. Reliably assessing prediction reliability for high dimensional QSAR data. *Mol. Divers.* **2013**, *17*, 63–73.
24. Huang, J.; Fan, X. Why QSAR fails: An empirical evaluation using conventional computational approach. *Mol. Pharm.* **2011**, *8*, 600–608.

25. Martin, Y.C.; Kofron, J.L.; Traphagen, L.M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.
26. Hopkins, A.L. Network pharmacology: The next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4*, 682–690.
27. Taylor, I.W.; Linding, R.; Warde-Farley, D.; Liu, Y.; Pesquita, C.; Faria, D.; Bull, S.; Pawson, T.; Morris, Q.; Wrana, J.L. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **2009**, *27*, 199–204.
28. Yildirim, M.A.; Goh, K.-I.; Cusick, M.E.; Barabási, A.-L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25*, 1119.
29. Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68.
30. Shao, L.; Wu, L.; Fan, X.; Cheng, Y. Consensus ranking approach to understanding the underlying mechanism with QSAR. *J. Chem. Inf. Model.* **2010**, *50*, 1941–1948.
31. Klaassen, C.D.; Liu, J.; Diwan, B.A. Metallothionein protection of cadmium toxicity. *Toxicol. Appl. Pharmacol.* **2009**, *238*, 215–220.
32. Park, J.D.; Liu, Y.; Klaassen, C.D. Protective effect of metallothionein against the toxicity of cadmium and other metals. *Toxicology* **2001**, *163*, 93–100.
33. Cobbett, C.; Goldsbrough, P. Phytochelatins and metallothioneins: Roles in heavy metal detoxification and homeostasis. *Annu. Rev. Plant Biol.* **2002**, *53*, 159–182.
34. Krizkova, S.; Ryvolova, M.; Hrabeta, J.; Adam, V.; Stiborova, M.; Eckschlager, T.; Kizek, R. Metallothioneins and zinc in cancer diagnosis and therapy. *Drug Metab. Rev.* **2012**, *44*, 287–301.
35. Cherian, M.G.; Jayasurya, A.; Bay, B.-H. Metallothioneins in human tumors and potential roles in carcinogenesis. *Mutat. Res-Fund. Mol. M.* **2003**, *533*, 201–209.
36. McGee, H.M.; Woods, G.M.; Bennett, B.; Chung, R.S. The two faces of metallothionein in carcinogenesis: Photoprotection against UVR-induced cancer and promotion of tumour survival. *Photoch. Photobio. Sci.* **2010**, *9*, 586–596.
37. Mishra, D.; Sudarshan, M.; Chakraborty, A. Elemental alteration, iron overloading and metallothionein induction in experimental hepatocarcinogenesis: A free radical-mediated process? *Toxicol. Lett.* **2011**, *203*, 40–47.
38. Abe, M.; Usuda, K.; Hayashi, S.; Ogawa, I.; Furukawa, S.; Igarashi, M.; Nakae, D. Carcinogenic risk of copper gluconate evaluated by a rat medium-term liver carcinogenicity bioassay protocol. *Arch. Toxicol.* **2008**, *82*, 563–571.
39. Waalkes, M.P.; Liu, J.; Goyer, R.A.; Diwan, B.A. Metallothionein-I/II Double Knockout Mice Are Hypersensitive to Lead-Induced Kidney Carcinogenesis Role of Inclusion Body Formation. *Cancer Res.* **2004**, *64*, 7766–7772.
40. Mattingly, C.; Rosenstein, M.; Colby, G.; Forrest, J., Jr.; Boyer, J. The Comparative Toxicogenomics Database (CTD): A resource for comparative toxicological studies. *J. Exp. Zool. Comp. Exp. Biol.* **2006**, 305, 689–692.
41. Uehara, T.; Ono, A.; Maruyama, T.; Kato, I.; Yamada, H.; Ohno, Y.; Urushidani, T. The Japanese toxicogenomics project: Application of toxicogenomics. *Mol. Nutr. Food. Res.* **2010**, *54*, 218–227.

42. Ganter, B.; Tugendreich, S.; Pearson, C.I.; Ayanoglu, E.; Baumhueter, S.; Bostian, K.A.; Brady, L.; Browne, L.J.; Calvin, J.T.; Day, G.-J. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.* **2005**, *119*, 219–244.
43. Nie, A.Y.; McMillian, M.; Brandon Parker, J.; Leone, A.; Bryant, S.; Yieh, L.; Bittner, A.; Nelson, J.; Carmen, A.; Wan, J. Predictive toxicogenomics approaches reveal underlying molecular mechanisms of nongenotoxic carcinogenicity. *Mol. Carcinog.* **2006**, *45*, 914–933.
44. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. DRAGON software: An easy approach to molecular descriptor calculations. *Match* **2006**, *56*, 237–248.
45. Chen, Y.-W.; Lin, C.-J. Combining SVMs with various feature selection strategies. *Feature Extraction* **2006**, *207*, 315–324.
46. Liu, Y.; Zheng, Y.F. FS_SFS: A novel feature selection method for support vector machines. *Pattern. Recogn.* **2006**, *39*, 1333–1345.
47. Maldonado, S.; Weber, R. A wrapper method for feature selection using Support Vector Machines. *Inform. Sci.* **2009**, *179*, 2208–2217.
48. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.
49. Shao, L.; Wu, L.; Fang, H.; Tong, W.; Fan, X. Does applicability domain exist in microarray-based genomic research? *PLoS One* **2010**, *5*, e11055.
50. Radmacher, M.D.; McShane, L.M.; Simon, R. A paradigm for class prediction using gene expression profiles. *J. Comput. Biol.* **2002**, *9*, 505–511.
51. Fan, X.; Shi, L.; Fang, H.; Cheng, Y.; Perkins, R.; Tong, W. DNA microarrays are predictive of cancer prognosis: A re-evaluation. *Clin. Cancer Res.* **2010**, *16*, 629–636.

Sample Availability: Not available.