11

Atom Coloring for Chemical Interpretation and *De Novo* Design for Molecular Design

Kiyoshi Hasegawa[1], Keiya Migita[2] and Kimito Funatsu[2]
[1]*Chugai Pharmaceutical Company, Kamakura Research Laboratories,*
[2]*The University of Tokyo, Department of Chemical System Engineering,*
Japan

1. Introduction

Prediction of biological activities is valuable for finding active compounds in an effective manner, and a considerable amount of attentions has been devoted to *in silico* predictions in drug discovery process. For *in silico* predictions, quantitative structure-activity relationship (QSAR) has been widely known to be useful [1, 2]. The basic purpose of QSAR is to construct a statistical model to reveal the relationship between chemical structures and their biological activities. For the statistical analysis, chemical structures are usually represented by several kinds of chemical descriptors. The QSAR model successfully trained and scientifically validated is used for predicting the biological activities of any molecules. In addition, a physicochemical and/or mechanistic interpretation can be expected from the selected chemical descriptors in the QSAR model.

As a multivariate statistical method, partial least square (PLS) is of particular interest in QSAR study [3]. PLS can analyze data with strongly collinear, noisy and numerous descriptors, and also simultaneously model several biological activities. It can also provide us several application domains and diagnostic plots as the statistical measures. We can extract the complex patterns embedded in the data set. Recently, PLS has evolved or changed for copying with sever demands from the complex data structure [4, 5].

PLS has its major restriction that only linear relationship can be extracted from data [3]. Since many structure-activity data sets are inherently nonlinear in nature, it is desirable to have a flexible method, which can model any nonlinear relationships. Recently, there has been a considerable interest in machine learning methods (ML) such as Bayesian approach [6, 7] and support vector regression (SVR) [8, 9] for nonlinear modeling. In general, since ML employs a sort of mathematical transformations of chemical descriptors, they have drawback that any correlations between the biological activity and the original descriptors should be lost. This means that a direct interpretation of the model is not easy task. A lot of papers studying ML have reported their high performances for classification and regression rates, but unfortunately they have not referred to the aspect of chemical interpretation [10].

For chemical interpretation, we employed the extended connectivity fingerprint (ECFP) as the chemical descriptor for a statistical model. ECFP can facilitate to understand what substructures are correlated with a specific biological activity. An atom score was calculated from the degree of contribution of each substructure to the model. By visualizing the atom scores with the graded-colors, an atom color mapping onto each compound was performed.

The atom coloring is helpful as a starting point for further molecular design with the guidence of atom colors. We described herein two representative examples for application of atom colors. (classifiaction of cytochrome P450 substrates and non-substrates, and visualization of molecular selectivity in dopamine family)

After establishing a solid model, *de novo* design is available for exploring new chemical structures in computer-aided molecular design. Structure generator intended for *de novo* design generates any chemical structures that are expected to have desired biological activities. This study has been known as inverse QSAR [11, 12].

In our approach, EA-Inventor (Evolutional Algorithm-Inventor) was used as structure generator. In EA-Inventor, initial structures represented by SMILES string are modified using several mutation operations in the framework of evolutionary algorithm (EA). Biological activities of compounds are predicted by the prepared QSAR model, and their values are used as the score values in EA. After the EA cycles, i.e. prediction of biological activities and generation of new chemical structures, chemical structures that have the highest scores are obtained. We applied our *de novo* design method to two molecular design projects and demonstrated its utility. (*de novo* design using ligand-based descriptors, and protein-ligand interaction descriptors)

2. Examples of atom coloring

2.1 Classification of CYP 3A4 substrates and non-substrates

Besides an optimization of biological activity, we have to consider to avoid poor ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties of molecules. ADMET processes often involve interaction with the associated proteins [13]. For example, the cytochrome P450 (CYP) isoenzymes, such as CYP 3A4, 2D6, and 2C9, are responsible proteins for the metabolism of most drug molecules. Among them, CYP 3A4 is the most important metabolism protein in human because it metabolizes the majority of commercially available drug molecules. Therefore, prediction of CYP 3A4 substrates is a key task to be solved in molecular design [14].

We employed the Bayesian approach to classify CYP 3A4 substrates and non-substrates. Large public data set comprising of 600 data points was used to develop the Bayesian model. The extended connectivity fingerprint (ECFP) was used as chemical descriptor. ECFP can facilitate to understand what substructures are correlated with substrates or non-substrates. An atom score in molecule was calculated from the Bayesian score of each ECFP descriptor. By coloring the atom scores with the five graded-colors, an atom color mapping onto each molecule was performed. The atom coloring is an effective tool for chemical interpretaion why a specific compound becomes a CYP 3A4 substrate and what chemical parts are responsible causing CYP 3A4 metabolism.

2.1.1 Bayesian model

The data set of CYP 3A4 substrates and non-substrates was taken from a public literature [15]. According to the literature, we divided whole data set into the training and test sets. The separation was performed taking considerations of their distribution in chemical space. The total number of substrates and non-substrates are 311 and 289 in the training set, respectively. The total number of substrates and non-substrates are 56 and 44 in the test set, respectively. Also, we made further external validation experiment to our prediction model of CYP 3A4 oxidation sites [16]. All of these molecules are known to be CYP 3A4 substrates. The number of molecules in the external validation set is 61

Bayesian model was constructed through the Pipeline Pilot module in Accelrys [17]. Bayesian approach compares the frequency of occurrences of chemical descriptors that are found in two groups that discriminate best between these groups (CYP 3A4 substrates versus non-substrates) [6, 7]. ECFP_6 (ECFP with path-lengths of six) was calculated in the Pipeline Pilot module and it was used as chemical descriptor. ECFP is a novel class of topological fingerprints for molecular characterization [18]. The Bayesian model for CYP3A4 classification provided us 12595 unique bins in ECFP_6. The base-line value separating between substrates and non-substrates is -0.819. Therefore, if the score of a compound is greater than this value, the compound is predicted to be substrate. Otherwise, the compound is predicted to be non-substrate. The classification rates of substrates and non-substrates in the training set are 285/311 (92%) and 280/289 (97%), respectively. The prediction rates of substrates and non-substrates in the test set are 47/56 (84%) and 37/44 (84%), respectively. The performance of the Bayesian model is well tolerated for further prediction. The established Bayesian model was applied to the CYP 3A4 substrates in the external validation set from our prediction model of CYP 3A4 oxidation sites [16]. The prediction rate is 53/61 (87%). The prediction rates from two external validation experiments are high enough to use as a prediction filter in early stage of molecular design.

2.1.2 Atom colorings

An atom scoring is a method to calculate a score value of each atom in molecule based on the Bayesian score [19]. First, a Bayesian model is built and the corresponding ECFP_6 descriptors are scored and categorized as substrate or non-substrate features. The score value of each ECFP substructure is divided by the number of heavy atoms consisting of the substructure and the calculated score value is assigned on each atom. Then, substructures of test molecule are identified, and the respective scores of atoms in the substructure are summed up and divided by the number of frequency occurrence to calculate their averaged atom scores. The schematic illustration for calculating atom scores is shown in Figure 1. An example in Figure 1 is 'Dutasteride' containing the amide substructure. The amide substructure is mapped on 'Dutasteride' and the corresponding atom scores (C4, C8, N13, and O14) are summed up.

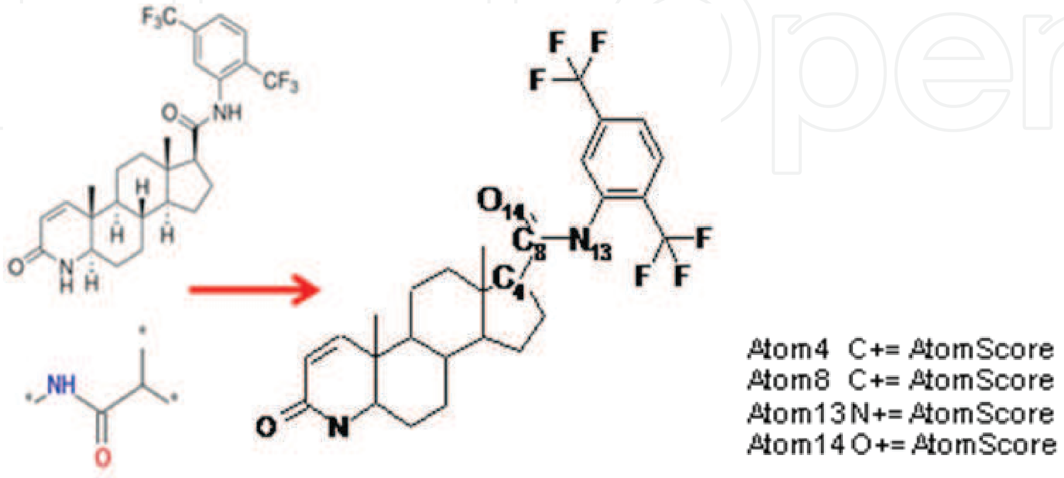
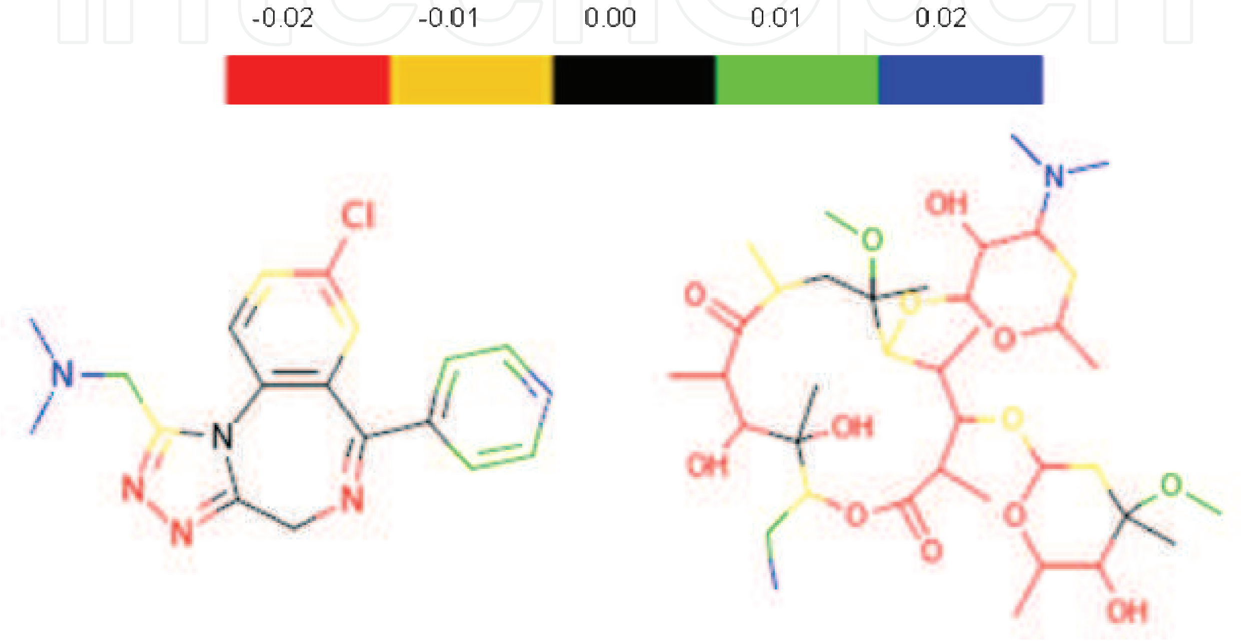


Fig. 1. Schematic illustration for calculating atom scores with example 'Dutasteride'

Four thresholds were used to color atoms in the five graded-colors (0.02, 0.01, -0.01, and -0.02). Because a higher score means a highly likelihood of CYP 3A4 substrate, the atom coloring is blue for higher, and red for lower susceptible to the CYP 3A4 metabolism. The representative examples of atom coloring are shown in Figure 2. In this figure, blue colors mean the liable atoms against CYP 3A4 metabolism. As for 'Adinazolam', tertiary N atom is identified to be liable atom. As for 'Clarithromycin', both of tertiary N atom and ethyl substituent on the macrocyclic ring are identified to be liable atoms. These predictions exactly match with the experimental results. Medicinal chemists can think any new ideas how to avoid the risk factors with the guidance of the atom coloring map.



Adinazolam

Clarithromycin

Fig. 2. The representative examples of atom coloring ('Adinazolam' and 'Clarithromycin')

Our approach is quite simple and it can be extended to other ADMET issues than CYP metabolism. The classification category between good and bad ADMET properties is prepared in advance and the Bayesian approach and the associated atom colorings are subsequently applied. From the atom colorings, we can easily identify unfavorable atoms in chemical structure for each ADMET property.

3. Visualization of molecular selectivity in dopamine family

In molecular design, computational methods that are capable of analyzing and predicting ligand selectivity profiles within target families are highly attractive [20]. This is the reason why molecules having activities against multi-target proteins cause many unfavorable side effects and toxicities. From the point of safety, data mining such as visualization of molecular selectivity against multi-target proteins cannot be ignored in molecular design [20].

Molecular selectivity was visualized using the combination of activity landscapes and atom colorings. The multiple inhibitory activites in dopamine family were selected to derive its molecular selectivity. At first, all of 390 molecular structures were mapped on 2D chemical space by preserving distance between any pairs of molecules using multi-dimensional scaling (MDS). The inhibitory activity values against each dopamine isoenzyme (D2, D3, and D4) were added independently to the data points in 2D chemical space. Activity landscapes were generated after making the color-graded interpolation between the data points. We can easily identify three specific active regions and the corresponding representative molecules for D2, D3, and D4 isoenzymes. Next, three support vector regression (SVR) models were separately built up using the above-mentioned ECFP descriptors and each inhibitory data set. By applying an atom coloring method to the representative molecules, the molecular selectivity differentiating each dopamine isoenzyme can be visually understood. Finally, the obtained molecular selectivity was validated from 3D homology structures.

3.1 Activity landscapes

Activity landscapes are defined by distance between any pairs of molecules and their biological activities [21]. If we envision a 2D projection of chemical space with the graded coloring of biological activity, this representation becomes reminiscent of geographical map that can readily be interpreted. The distance of two molecules is calculated as the Euclidean distance between their ECFP_6 descriptors. The Euclidean distance is defined according to the following equation [21]:

$$\delta_{ij} = \sqrt{N_i + N_j - 2N_{ij}} \tag{1}$$

where N_i and N_j denote the number of ECFP_6 binary bins present in molecules i and j, respectively. N_{ij} denotes the number of binary bins shared by both molecules. To map multi-dimensional data into 2D chemical space, MDS is employed. MDS aims at preserving relative similarity relationships between input data points by minimizing the deviation from the ideal relationships [22].

Biological activity values are then added to the data points in 2D chemical space for creating activity landscape. In general, however, the data points are sparse and unevenly distributed and must be interpolated to obtain coherent chemical space. For this purpose, a geostatistical technique termed Kringing is applied. Based on the expected value and a covariance function that describe the spatial dependence of the given data points, the Kringing method calculates the best linear unbiased estimator by minimizing the variance of the prediction error. We utilize the Kringing function as implemented in the 'fields' package of R [23]. After finishing the interpolation, 2D map is colored according to the predicted activity values. Areas with a value below a lower threshold are colored in blue, and areas with a value above an upper threshold are colored in red. Intermediate values are colored using a continuous gradient from blue via yellow to red.

We collected three inhibitory data sets of D2, D3, and D4 from GVK data base [24]. The logarithm of reciprocal value of Ki in the micro molar unit (log(-Ki)) was used as the inhibitory activity. We selected 390 molecular data points for filling all elements in three inhibitory data matrices. Three activity landscapes of D2, D3, and D4 are shown in Figure 3. In Figure 3, the region showing both high inhibitory activity and high specificity toward each dopamine isoenzyme was highlighted by the dashed circle. The representative molecules are shown beside each activity landscape. Three molecules have the common tertiary N atoms in their chemical structures to interact with the common Asp in their target dopamine proteins [25]. However, chemical fragments at two ends in molecule significantly differ to each other. They are reflected on the molecular specificity in dopamine family.

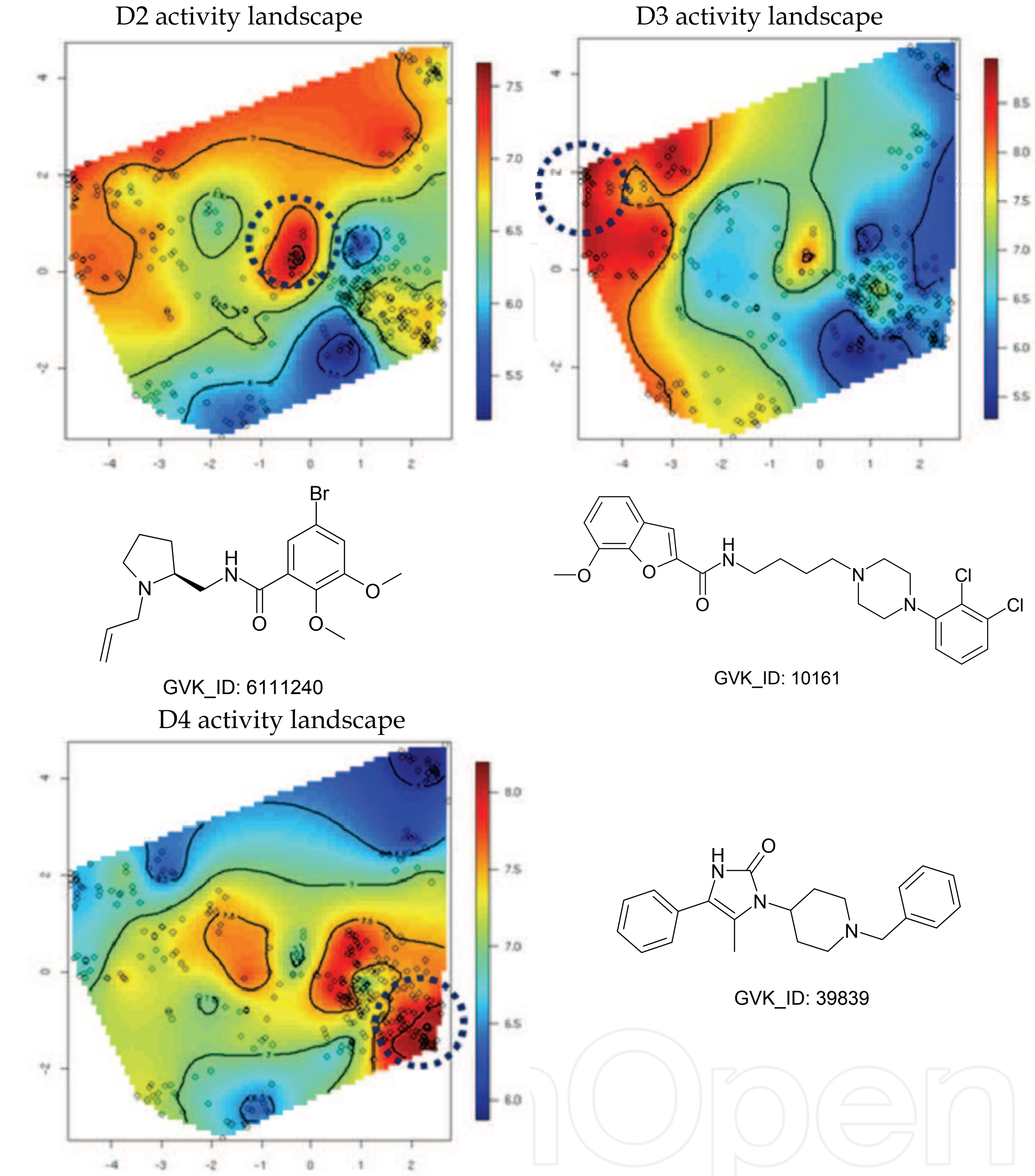


Fig. 3. Three activity landscapes and the representative molecules

3.2 SVR models

SVR was used to develop the predictive model for each dopamine isoenzyme using ECFP_6 descriptors. SVR is a regression type of support vector machine (SVM) by the introduction of a loss function comprising of the squared deviations [8-10]. The general principle of SVM is to perform a classification by constructing an n-dimensional hyper plane that optimally separates the data set into two categories.

Before SVR analysis, 3859 ECFP_6 descriptors were reduced to 82 by the filter of variance cut off of 0.1 After that, a systematic grid search was used to determine the best parameter values for D2 data set based on 10-fold cross-validation (C=2, ν=0.30, and γ=0.03125). The procedure of backward-elimination was applied to produce the final D2 model with 44 descriptors. The R^2 and Q^2 values of the final D2 model are 0.782 and 0.615, respectively. The same procedures were applied to D3 and D4 data sets. As for D3, the final model has 45 descriptor with C=2, ν=0.35, and γ=0.03125. The R^2 and Q^2 values of the final D3 model are 0.900 and 0.783, respectively. As for D4, the final model has 40 descriptor with C=1, ν=0.55, and γ=0.03125. The R^2 and Q^2 values of the final D4 model are 0.793 and 0.637, respectively. All statistical procedures were performed using the 'kernlab' package of R [26] and some written scripts in our laboratory.

3.3 Atom colorings

After building each SVR model, a local gradient was calculated according to the following formula:

$$w_{ij} = \left(\frac{\partial}{\partial x} f_j(x) \,|_{x=x_i} \right) \tag{2}$$

where i means index of compounds. J means the index of ECFP_6 descriptors. The value of local gradient indicates the descriptor importance in a specific compound. For a specific compound i, the value of descriptor j is slightly changed keeping other descriptor values to be fixed. Then, the differences of the predicted values from the SVR model are calculated. The local gradient is derived from the differences divided by the shifted values. The original idea of local gradient came from the financial study [27]. One of authors (K. Migita) has made further invention for improving the calculation efficacy and precision [28]. The value of local gradient of each substructure was used as the Bayesian score value for next atom coloring. The calculation of local gradient was performed using the written scripts in our laboratory under the R environment.

The atom colorings of each specific dopamine inhibitor are shown in Figure 4. Four thresholds were used to color atoms in the five graded-colors (0.008, 0.004, -0.004, and -0.008). The atom coloring scheme is blue for the most contributed atom to the inhibitory activity, and red for least. As shown in Figure 4, compound 6111240 is a specific inhibitor toward D2 protein and this is emphasized in the remarkable blue and green colors. Compound 6111240 has unfavorable amide part toward D4 protein, which are highlighted in red color. Other specific compounds (10161 and 39839) also show clear molecular specificity owing to the atom colorings.

These molecular specificities were validated by 3D structures of D2 and D4 proteins. The 3D structures were built by the homology modeling procedure based on the template X-ray crystal structure of β2 (PDB code: 2RH1). The 3D coordinates of homology structures of D2 and D4 proteins were cited from a recent public study [25]. In Figure 5, 3D structures of D2 and D4 proteins are shown with a specific compound 6111240. Cyan and white colors represent D2 and D4 protein structures, respectively. Compound 6111240 is shown in yellow bold color. The compound has nice polar interactions with Asp 114 and Cys 118 residues in D2 protein. However, in the case of D4 protein, due to the bulky residues (Phe 91 and Arg 186), compound cannot maintain the polar interactions in D4 protein. This could be one of reasons why compound 6111240 has high molecular specificity only toward D2 protein. Different coloring patters on amide part are nicely accordance with the differences of active site pockets between D2 and D4 proteins.

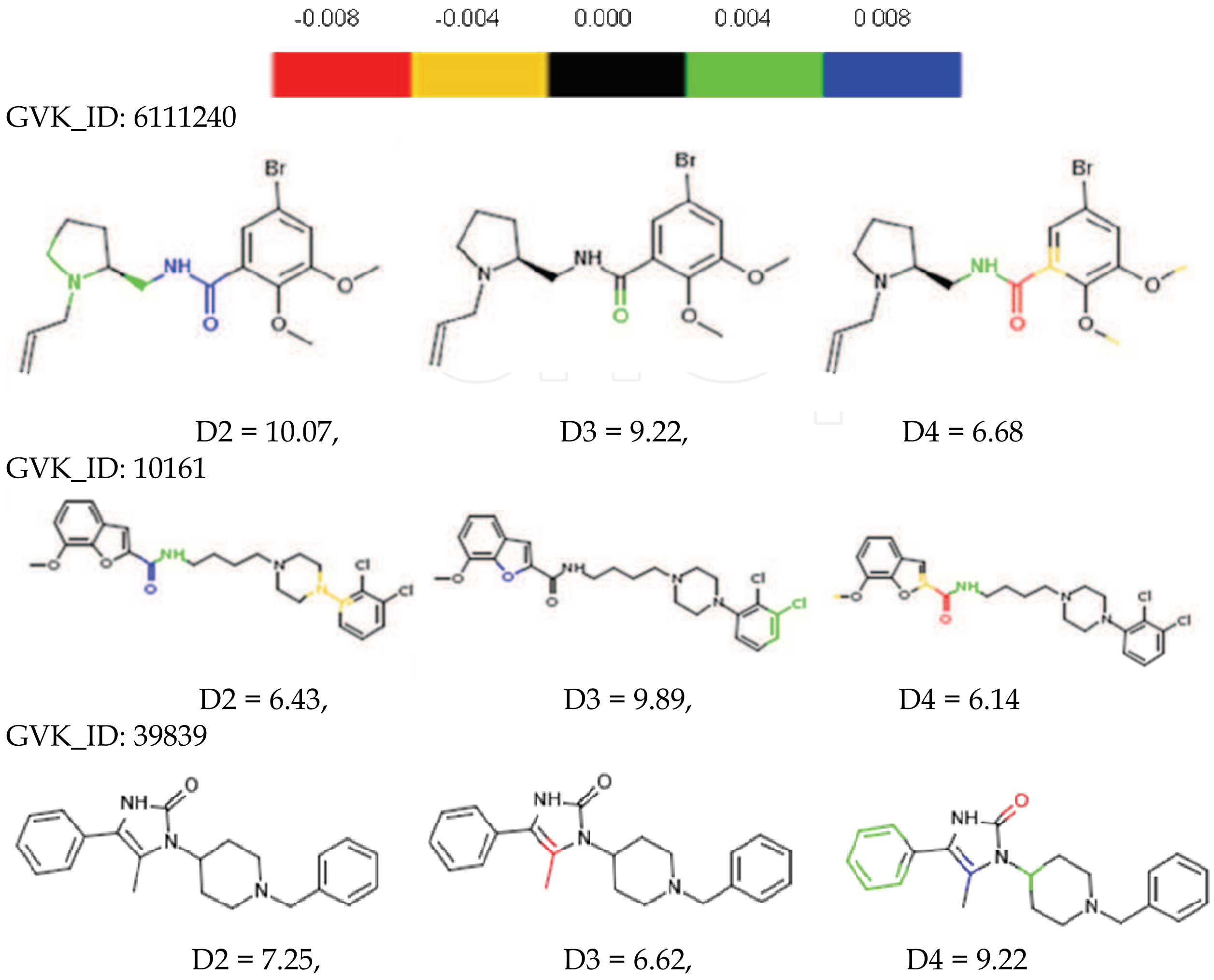


Fig. 4. Atom colorings of each specific dopamine inhibitor

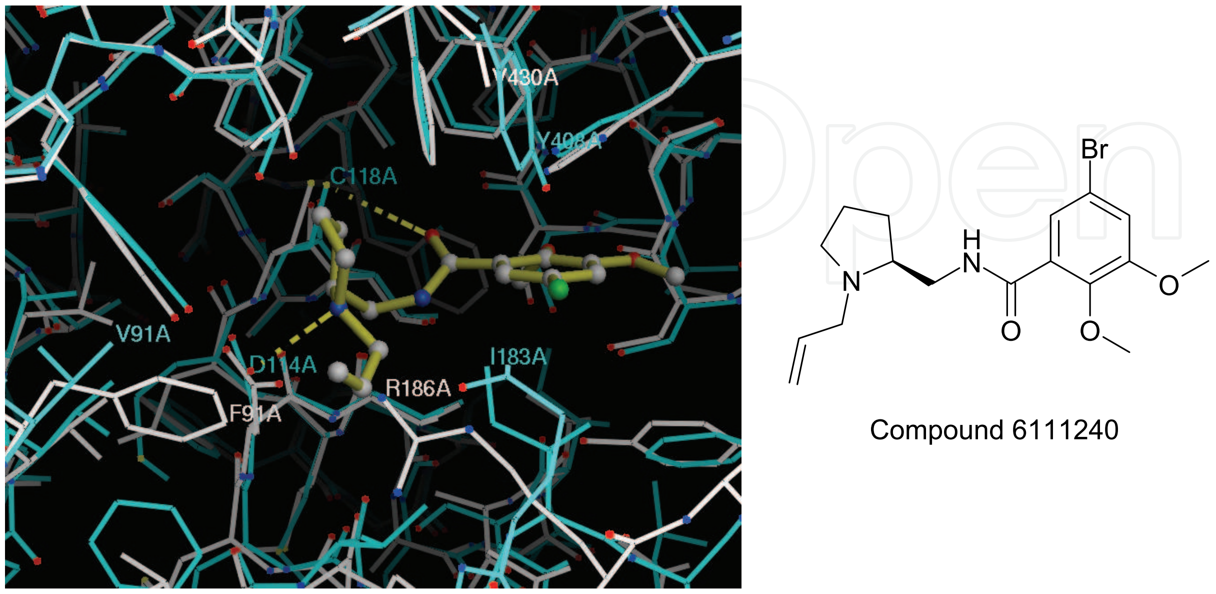


Fig. 5. Two homology models and their interactions with compound 6111240

Including biological activity, the landscape technique can be applied to multiple ADMET endpoints. This makes help us identify some interesting chemical space where compounds show high biological activity and also satisfy good ADMET properties. We are planning to construct graphical user interface (GUI) handling many landscape panels [29]. When medicinal chemists touch a data point in a specific landscape, the associated data is shown immediately with a chemical structure. This system might be a good compass to know us where synthetic compounds are in chemical space or what extent of chemical efforts is necessary to achieve chemical optimization.

4. Examples of *de novo* design

4.1 *De novo* design using ligand-based descriptors

QSAR has been successfully applied in molecular design [30]. However, when a precise QSAR model is created, chemical interpretation generally becomes difficult. In particular, in the case of a non-linear model, the relationship between chemical descriptors and biological activity cannot be described explicitly, and chemical interpretation becomes challenging [31]. In this case, an inverse QSAR approach is attractive to design practical chemical structures. Inverse QSAR is a relatively new concept that chemical structures having high biological activities are computationally generated using a structure generator [11, 12].

EA-Inventor (Evolutionary Algorithm-Inventor) was used as structure generator in our study. In EA-Inventor, chemical structures are updated using several mutation operations in an iterative manner to achieve high inhibitory activities. We applied EA-Inventor to the data set of 33 matrix metallopeptidase 2 (MMP-2) inhibitors. The scoring function for input to EA-Inventor is the prediction value derived from the SVR model. In order to prevent generated structures deviating from chemical space, the leverage value was added to the scoring function as the penalty value. The generated chemical structures are reasonable judging from 3D homology structure of MMP-2 protein as their counterparts.

4.2 EA-Inventor

EA-Inventor has been successfully applied along with the docking algorithm [32]. In our study, EA-Inventor was used in combination with prediction value derived from a specific QSAR model. Initial chemical structures are set up as a SMILE string. As mutation operations to chemical structures, 33 different transformations are available in EA-Inventor. Some typical transformations are shown in Figure 6. Transformations (i) and (ii) involve changing the atoms and bonds, respectively. Transformations (iii) and (iv) involve breaking the ring structure and changing the ring size, respectively. Transformation (v) involves the addition of the prepared chemical fragments in advance. Around 1400 libraries of fragments have been compiled by breaking down the chemical structures of drug-like molecules. Therefore, new chemical structures, including those of drug-like compounds, can be formed by combining these fragments [33]. The algorithm of EA-Inventor is schematically shown in Figure 7. The experiment of EA-Inventor was performed on the Sybyl environment [34].

4.3 SVR model

The data set of 33 MMP-2 inhibitors were collected from a literature [35] and the SVR model was developed with 496 ligand-based descriptors. These descriptors were calculated by the descriptor module in MOE [36]. After pre-processing of descriptors based on variance cut and p-values (0.1 and 0.01), 26 descriptors were selected. Then, a grid search based on leave-one-out cross-validation was performed to determine the best parameter values (C=64, ν=0.15 and γ=0.008). The procedure of backward-elimination was applied to produce the final SVR model with 15 descriptors. The R^2 and Q^2 values of the final SVR model are 0.867 and 0.795, respectively. The selected 15 descriptors have the ambiguous physicochemical meanings and their chemical interpretations are far to be ease. In this situation, an inverse QSAR becomes a powerful approach [11, 12].

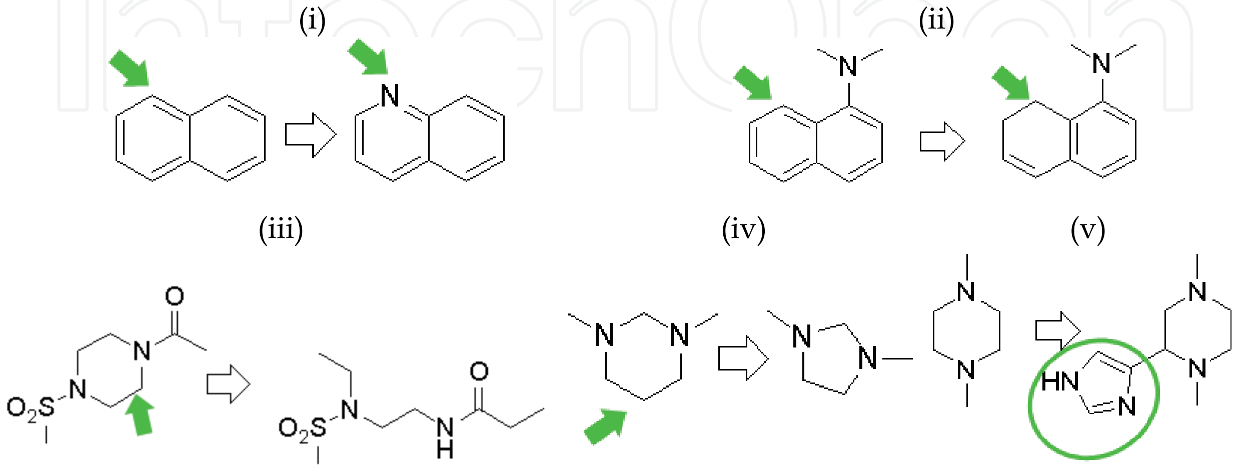


Fig. 6. Typical transformations in EA-Inventor: (i) modification of atom type, (ii) modification of bond type, (iii) breaking of ring, (iv) modification of ring size, (v) addition of fragment.

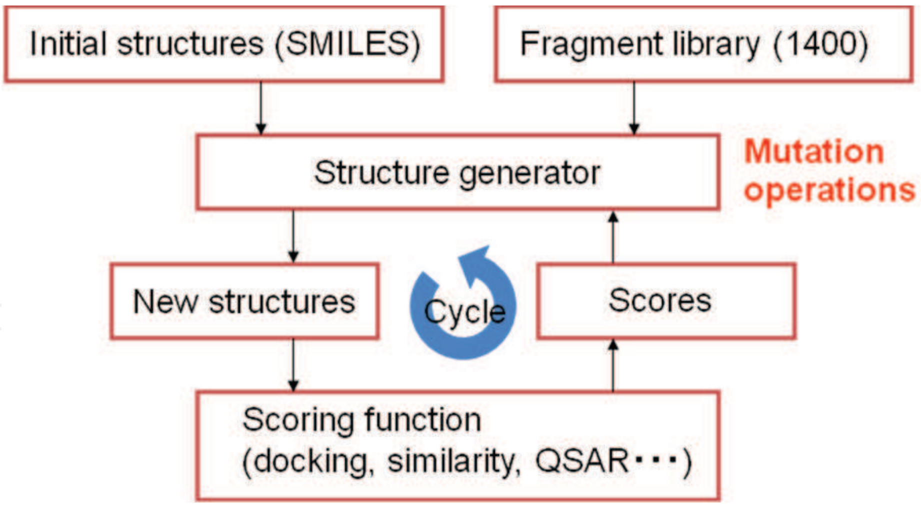
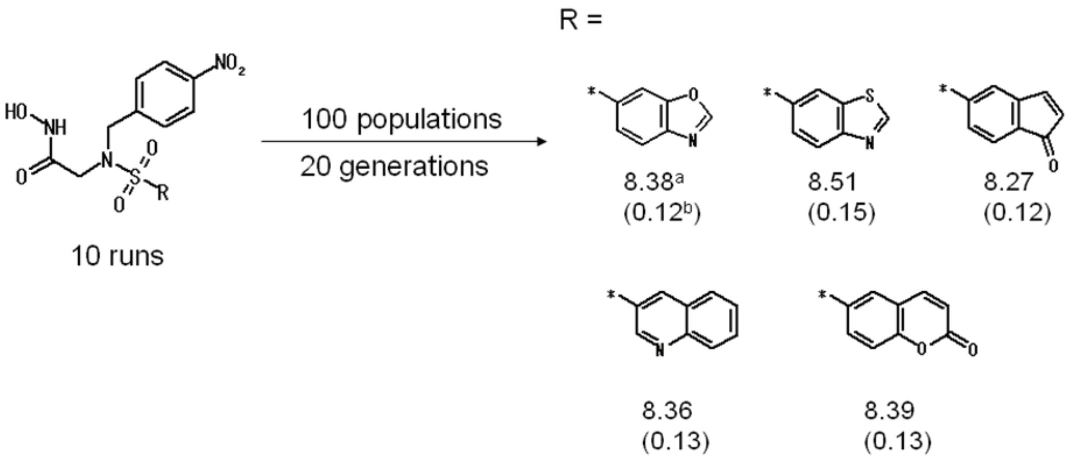


Fig. 7. Algorithm of EA-Inventor as structure generator.

4.4 *De novo* design

With fixing the core scaffold structure, EA-Inventor was performed with 100 populations and 20 generations in 10 runs. The output of generated chemical structures is shown in Figure 8. In Figure 8, symbol *a* means the predicted inhibitory activity. Symbol *b* means the leverage value. All of the generate structures have both of high predicted inhibitory activities and low values of leverage.

There are two ways for making validation about the output structure from *de novo* design: One is that actually synthesizing compounds and measuring their inhibitory activities. Other is that comparing the output structure with the target structure. Because we are not available on experimental way, we employed the latter approach for validation of our results. In Figure 9, the binding mode of one of examples (benzothiazole molecule) in the pocket of MMP-2 homology model is shown. Homology model of MMP-2 was built from a complex X-ray structure between MMP-12 and hydroxamate ligand (PDB: 1RMZ). White and orange colors represent MMP-2 protein and benzothiazole molecule, respectively. Benzothiazole molecule has the good interactions with Val 400, His 403, and Tyr 425 highlighted in red colors. Then, it is highly expected from the protein-ligand interactions that the benzothiazole molecule could show the high inhibitory activity. All of the fused 6-5 or 6-6 ring systems exemplified in Figure 8 are favorable to occupy the cavity pocket in MMP-2 protein. These results indicate that EA-Inventor can actually generate new chemical structures with the high inhibitory activities.



a: Predicted activity; b: Leverage value

Fig. 8. Output of generated chemical structures from EA-Inventor

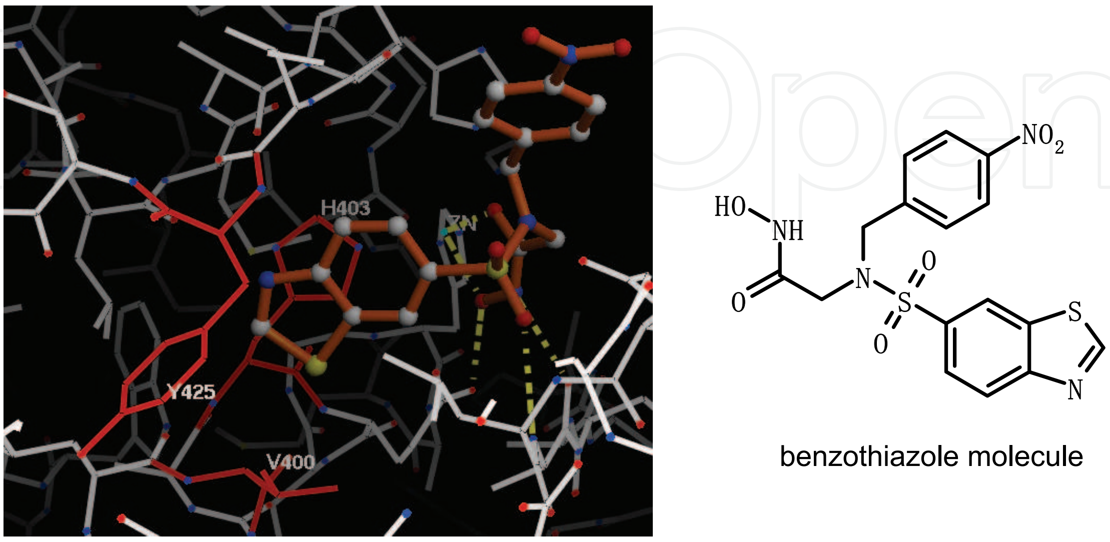


Fig. 9. Docking mode of benzothiazole molecule in MMP-2

We can consider the chemical structures cited in Figure 8 as the starting molecules. For example, ‘benzothiazole’ and ‘indole’ can be transformed to any aromatic rings such as benzofuran. It is rationale transformation from five member ring ‘pyrolidine’ to six member ring ‘piperidine’. Therefore, the proposed chemical structures from EA-Inventor are good hints for further molecular design.

4.5 De novo design using protein-ligand interaction descriptors

Structure-based drug design (SBDD) is a technique for designing novel compounds by use of physico-chemical interactions in a complex X-ray structure between a protein and a ligand molecule [37]. The most important thing for SBDD is an accurate estimation of binding affinity of the ligand molecule against the target protein. This issue has not been solved accompanying with prediction of docking pose of molecule in the protein structure. Therefore, for a specific target protein, the scenario that the docking scoring function is customized in the framework of QSAR is highly attractive [38]. The integration of the customized docking scoring function into *de novo* design engine is interesting for designing potent molecules more accurately than those of traditional SBDD.

We customized the docking scoring function using comparative molecular binding energy (COMBINE) descriptors and SVR. COMBINE descriptors are energy terms between the ligand molecule and each amino acid residue of the target protein. The data set of 35 human caspase-3 inhibitors was used in our study. The SVR model can successfully identify important amino acid residues for explaining inhibitory activities against human caspase-3. Then, we integrated the docking scoring function into EA-Inventor. A number of molecules were virtually generated by EA-Inventor, and they were evaluated using the customized docking scoring function. EA-Inventor produced some interesting compounds and the rationale was validated using X-ray crystal structure of human caspase-3.

4.6 SVR model

The data set of 35 human caspase-3 inhibitors focusing Isatin sulfonamide analogues was taken from a literature [39]. COMBINE descriptors expressing the protein-ligand interactions were used as chemical descriptors [40, 41]. From the X-ray crystal structure of human caspase-3 (PDB code: 1GFW), total 47 amino acids were detected as the nearest amino acids forming the binding pocket. Since van der Waals, Coulomb, and hydrogen bonding interaction energies are considered, and total number of COMBINE descriptors becomes 141. All of 35 inhibitors were docked into human caspase-3 structure and COMBINE descriptors were calculated by the GLIDE docking module [42]. After pre-processing of descriptors based on variance cut of 0.05, 29 significant descriptors were selected. A grid search based on leave-one-out cross validation was used to determine the best parameter values ($C=4$, $\nu=0.95$ and $\gamma=0.03125$). The procedure of backward-elimination was applied to produce the final SVR model with 13 COMBINE descriptors. The R^2 and Q^2 values of the final SVR model are 0.977 and 0.873, respectively.

The selected 13 COMBINE descriptors are as follows in the descending order of the above-mentioned local gradient values: His 121 vdw, Met 61 vdw, Ser 251 hbonds, Glu 123 coul, Phe 250 vdw, Ser 251 vdw, Asp 253 coul, Arg 207 vdw, Phe 256 vdw, Asn 208 coul, Tyr 204 coul, Tyr 204 vdw, and Lys 259 coul. Three abbreviations (vdw, coul, and hbonds) mean van der Waals, Coulombic, and hydrogen bonding interactions, respectively. Many COMBINE descriptors are derived from the van der Waals interaction. The presumed binding mode of compound 21 in human caspase-3 is shown in Figure 10. White and green present the caspase-3 protein and compound 21, respectively. The selected amino acid residues in the final SVR model are highlighted in yellow bold colors in Figure 10. They are located close to two ends of chemical structure in compound 21.

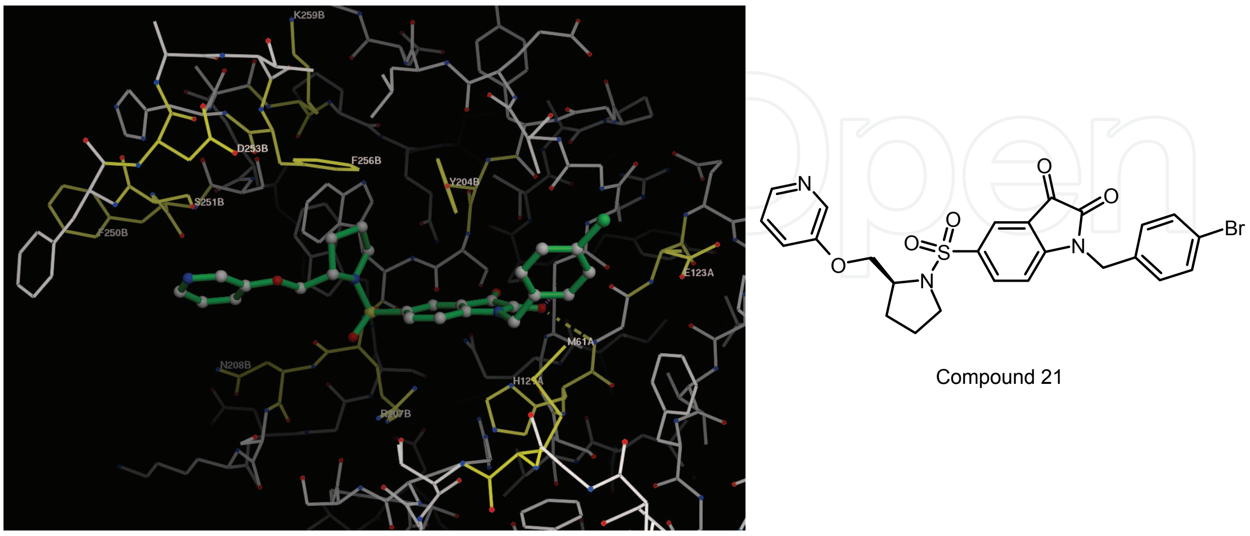


Fig. 10. Presumed binding mode of compound 21.

4.7 *De novo* design

With fixing the core scaffold structure, EA-Inventor was performed with 50 populations and 3 generations in 10 runs. The output of generated chemical structures from EA-Inventor is shown in Figure 11. The number below each chemical structure means the inhibitory activity predicted by the SVR model. The generated chemical structures are mainly 5 and 6-5 fused ring systems. In order to validate the output structures, they were docked into the X-ray crystal structure of human caspase-3. The presumed binding mode of indole molecule is shown in Figure 12. Three strong interactions with the indole part of molecule are highlighted in red colors: CH-π interactions of side chains of Glu 123 and Met 61, and π-π interaction of imidazole ring of His 121. These amino acid residues are included in the selected 13 COMBINE descriptors. Judging from these 3D molecular interactions, 6-5 fused ring analogues seems to be highly expected for high inhibitory activity.

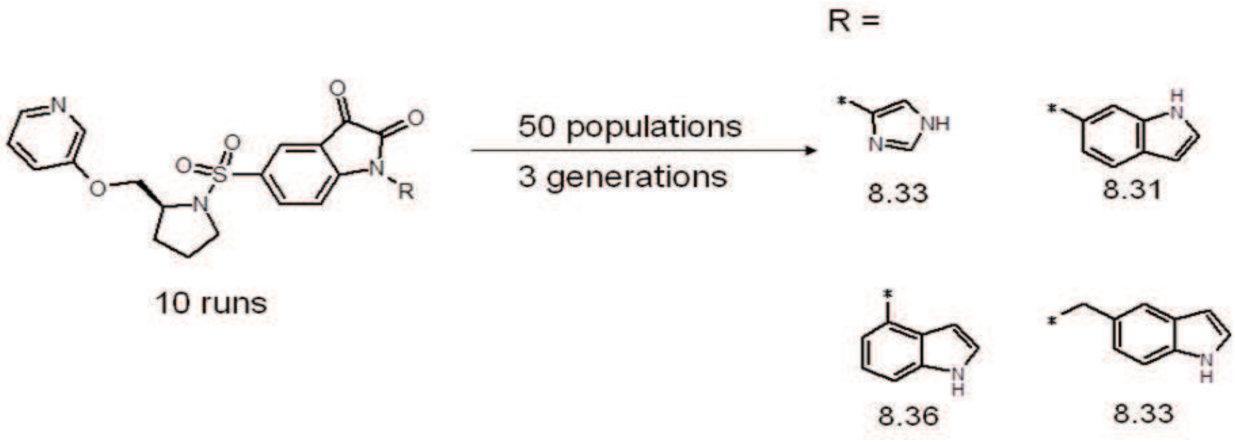


Fig. 11. EAInventor for producing chemical structures.

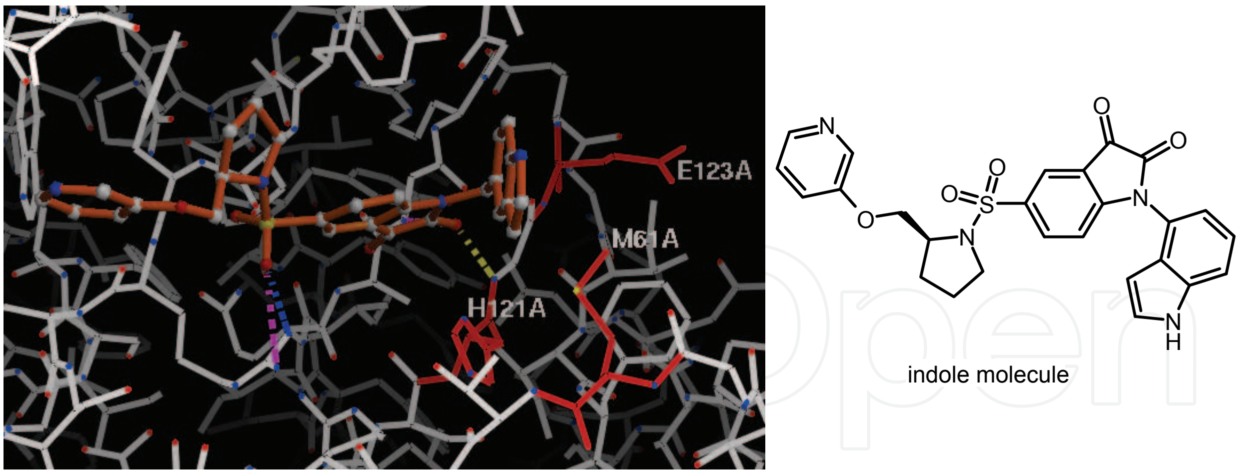


Fig. 12. Presumed binding mode of indole molecule.

Thanks to the transparent COMBINE descriptors, medicinal chemists can easily find which amino acid residues or which types of interactions are important for increasing inhibitory activity. Medicinal chemists can focus on the specific chemical parts interacting with these important amino acid residues. Their chemical parts are effectively replaced by another moiety using *de novo* design engine.

5. Conclusion

Two techniques of atom coloring and *de novo* design were introduced with their representative examples. By coloring atom scores derived from the Bayesian or SVR models with the five graded-colors, we can easily identify the key atoms in chemical structure responsible to each biological activity. The visualization encourages us to think design ideas how to modify chemical structure in next synthetic plan. *De novo* design provides us chemical structures with high inhibitory activities in the framework of QSAR. The generated chemical structures are good starting molecules for further molecular design.

As mentioned before, the landscape technique is useful to grasp rough trend in whole chemical space. The combination of landscape with *de novo* design engine is especially attractive as new strategy of molecular design. At first, we select an interesting region in chemical space. The interesting region means that it contains a chemical structure with high biological activity and good ADMET properties. Any chemical structures are virtually generated starting from the chemical structure in the region. During *de novo* simulation, the generated chemical structures are forced to stay inside the region. After that, chemical structures with same chemical profile as that of the starting molecule could be found according to this scenario. This study is in progress and the results will be disclosed in near future [43].

6. References

[1] P. Gedeck, R.A. Lewis, Exploiting QSAR models in lead optimization, *Curr.Opin.Drug Dis.Dev.*, 2008, 11, 569-575.

[2] C.W. Yap, H. Li, Z.L. Ji, Y.Z. Chen, Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic and toxicological properties, *Mini-Rev.Med.Chem.*, 2007, 7, 1097-1107.

[3] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemom.Intell.Lab.Syst.*, 2001, 58, 109-130.

[4] K. Hasegawa, K. Funatsu, Advanced PLS Techniques in Chemometrics and Their Applications to Molecular Design. In *Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques;* Lodhi, H.; Yamanishi, Y.; Eds., IGI publishing, 2011.

[5] K. Hasegawa, K. Funatsu, Advanced PLS Techniques in Chemoinformatics Studies, *Curr.Comput.-Aided Drug Des.*, 2010, 6, 103-127.

[6] X. Xia, E. Maliski, P. Gallant, D. Rogers, Classification of Kinase Inhibitors Using a Bayesian Model, *J.Med.Chem.*, 2004, 47, 4463-4470.

[7] P. Prathipati, N.L. Ma, T.H. Keller, Global Bayesian Models for the Prioritization of Antitubercular Agents, *J.Chem.Inf.Model.*, 2008, 17, 2362-2370.

[8] J.-P. Doucet, F. Barbault, H. Xia, A. Panaye, B. Fan, Nonlinear SVM approaches to QSPR/QSAR studies and drug design, *Curr.Comput.-Aided Drug Des.*, 2007, *3*, 263-289.

[9] H. Li, Y. Liang, Q. Xu, Support vector machines and its applications in chemistry *Chemom.Intell.Lab.Syst.*, 2009, 95, 188-198.

[10] K. Hasegawa, K. Funatsu, Non-Linear Modeling and Chemical Interpretation with Aid of Support Vector Machine and Regression, *Curr.Comput.-Aided Drug Des.*, 2010, 6, 24-36.

[11] D. Hecht, G.B. Fogel, A Novel In Silico Approach to Drug Discovery via Computational Intelligence, *J.Chem.Inf.Model.*, 2009, 49, 1105–1121

[12] T. Miyao, M. Arakawa, K. Funatsu, Exhaustive Structure Generation for Inverse-QSPR/QSAR, *Mol.Inf.*, 2010, 29, 111-125.

[13] H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.R. Li, L.Y. Han, H.H. Lin, Y.Z. Chen, Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins, *J.Pharm.Sci.*, 2007, 96, 2838-2860.

[14] R. Arimoto, Computational models for predicting interactions with cytochrome p450 enzyme, *Curr.Top.Med.Chem.*, 2006, 6, 1609-1618.

[15] C.W. Yap, Y.Z. Chen, Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines, *J.Chem.Inf.Model.*, 2005, 45, 982-992.

[16] K. Hasegawa, M. Koyama, K. Funatsu, Quantitative Prediction of Regioselectivity Toward Cytochrome P450/3A4 Using Machine Learning Approaches, *Mol.Inf.*, 2010, 29, 243–249.

[17] Pipeline Pilot Basic Chemistry Component Collection, Accelrys, Inc. 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA.

[18] D. Rogers, M. Hahn, Extended-Connectivity Fingerprints, *J.Chem.Inf.Model.*, 2010, 50, 742–754.

[19] J.T. Metz, D.A. Stonich, D. Rogers, Visualization of Atomic Contributions to Ligand Properties, SciTegic Users Group Meeting 2007.

[20] E. Lounkine, M. Wawer, A.M. Wassermann, J. Bajorath, SARANEA – a freely available program to mine structure-activity and structure-selectivity relationship information in compound data sets, *J.Chem.Inf.Model.*, 2010, 50, 68-78.

[21] L. Peltason, P. Iyer, J. Bajorath, Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and formation of activity cliffs, *J.Chem.Inf.Model.*, 2010, 50, 1021–1033.

[22] I. Borg, P.J.F. Groenen, Modern Multidimensional Scaling. Theory and Applications, 2nd ed.; Springer: New York, NY, 2005.

[23] http://cran.r-project.org/web/packages/fields/index.html

[24] http://www.gvkbio.com/

[25] F.M. McRobb, B. Capuano, I.T. Crosby, D.K. Chalmers, E. Yuriev, Homology Modeling and Docking Evaluation of Aminergic G Protein-Coupled Receptors, *J.Chem.Inf.Model.*, 2010, 50, 626-637.

[26] http://cran.r-project.org/web/packages/kernlab/index.html

[27] L.J. CAO, F.E.H. TAY, Feature Selection for Support Vector Machines in Financial Time Series Forecasting, In Intelligent Data Engineering and Automated Learning (Lecture Notes in Computer Science), K.S. Leung, L.-W. Chan, H. Meng (Eds.), Springer Verlag, Berlin, 2009.

[28] K. Migita, personal communications.

[29] K. Hasegawa, T. Fukami, M. Ohta, Y. Shiratori, *2nd Chemoinformatics Strasbourg Summer School*, P5: Construction of ADMET local models and development of Web GUI for chemists, 2010, Obernai, France.

[30] K. Hasegawa, K. Funatsu, Data Modeling and Chemical Interpretation of ADME Properties Using Regression and Rule Mining Techniques, In Gary, W. C. (Ed.), Frontier in

Drug Design & Discovery 4. Bentham Science Publisher, 2009.

[31] R. Guha, On the interpretation and interpretability of quantitative structure-activity relationship models, *J.Comput.-Aided Mol.Des.*, 2008, 22, 857-871.

[32] L. Qian, M. Brian, S. Karl, S. Julian, Tagged Fragment Method for Evolutionary Structure-Based De Novo Lead Generation and Optimization, *J.Med.Chem.*, 2007, 50, 5392–5402.

[33] K. Hasegawa, T. Kimura, K. Funatsu, Inverse QSAR Study Using Evolutionary Algorithm *J.Comput.-Aided Chem.* , 2009, 10, 10-15.

[34] http://tripos.com/index.php

[35] D. Kumar, S.P. Gupta, A Quantitative Structure-Activity Relationship Study on Some Matrix Metalloproteinase and Collagenase Inhibitors, *Bioorg.Med.Chem.*, 2003, 11, 421-426.

[36] http://www.chemcomp.com/software-moe2009.htm

[37] I.D. Kuntz, Structure-Based Strategies for Drug Design and Discovery, *Science*, 1992, 257, 1078-1082.

[38] M. Arakawa, K. Hasegawa, K. Funatsu, Tailored scoring function of Trypsin-benzamidine complex using COMBINE descriptors and support vector regression, *Chemom.Intell.Lab.Syst.*, 2008, 92, 145-151.

[39] Q. Wang, R.H. Mach, D.E. Reichert, Docking and 3D-QSAR Studies on Isatin Sulfonamide Analogues as Caspase-3 Inhibitors, *J.Chem.Inf.Model.*, 2009, 49, 1963–1973.

[40] A.R. Ortiz, M.T. Pisabarro, F. Gago, R.C. Wade, Prediction of Drug Binding Affinities by Comparative Binding Energy Analysis, *J.Med.Chem.*, 1995, 38, 2681–2691.

[41] K. Hasegawa, T. Kimura, K. Funatsu, GA strategy for variable selection in QSAR studies. Enhancement of comparative molecular binding energy analysis by GA-based PLS method, *QSAR*, 1999, 18, 262–272.

[42] http://www.schrodinger.com/

[43] K. Hasegawa, K. Migita, K. Funatsu, in preparation of manuscript.

KNOWLEDGE - ORIENTED APPLICATIONS IN
DATA MINING
Edited by Kimito Funatsu and Kiyoshi Hasegawa
INTECH