Bioorganic & Medicinal Chemistry 20 (2012) 5410–5415

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

New description of protein–ligand interactions using a spherical self-organizing map

Kiyoshi Hasegawa [a], Kimito Funatsu [b,*]

[a] Chugai Pharmaceutical Company, Kamakura Research Laboratories, Kajiwara 200, Kamakura, Kanagawa 247-8530, Japan
[b] The University of Tokyo, Department of Chemical System Engineering, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

ARTICLE INFO

Article history:
Available online 24 March 2012

Keywords:
Protein–ligand interactions
Spherical self-organizing map
Molecular electrostatic potential
Caspase-3 inhibitors
Quantitative structure–activity relationship

ABSTRACT

In a previous report, we studied the mapping ability of the spherical self-organizing map (SSOM). The original 3D structure of the active site of the β2 protein structure was well reproduced by the SSOM. To validate the geometrical transformation and the resulting molecular electrostatic potential (MEP) distribution, the molecular surfaces of 20 β2 ligands were mapped onto the protein SSOM sphere. The MEP values of the two spheres derived from the ligand and the β2 receptor protein were compared. In most cases involving potent ligands, the two spheres had a moderate negative correlation. This indicates that the SSOM approach has excellent potential to represent a complex protein surface as a simple spherical structure.

In this study, we perform a quantitative structure–activity relationship (QSAR) study of caspase-3 inhibitors based on the SSOM technique. Initially, the active site of the protein structure 'caspase-3' was characterized by the SSOM using the MEP values. Each inhibitor was then projected onto the protein SSOM sphere and the chemical descriptors were derived from the ligand SSOM sphere. The correlation of the chemical descriptors and the inhibitory activities was investigated using the support vector regression (SVR) method. Finally, the important MEP descriptors from the final SVR model were examined. The structural requirements of caspase-3 inhibitors are discussed from the perspectives of both the ligand and protein structures.

1. Introduction

Prediction of biological activities is crucial for the effective identification of active compounds, and considerable effort has been devoted to in silico drug design. For in silico predictions, the quantitative structure–activity relationship (QSAR) approach has been widely used.[1,2] In QSAR, compounds are represented by chemical descriptors and statistical models are built to predict biological activities of candidate compounds. When a molecule is represented by several kinds of descriptors derived from the 3D structure, the method is called 3D-QSAR.

Polanski et al. proposed the so-called comparative molecular surface analysis as a new 3D-QSAR method.[3] They used a self-organizing map (SOM) to transform the 3D molecular surface into a 2D map.[4] They defined the chemical descriptors by unfolding all the nodes with molecular electrostatic potential (MEP) values in SOM. The relationship between the activity and the chemical descriptors was analyzed by partial least squares (PLS).[5] They successfully constructed a 3D-QSAR model; however, important neighboring relationship information between the nodes in SOM was lost. Hasegawa et al. proposed a new surface-based 3D-QSAR method using SOM and a three-way PLS.[6] In the same manner as the Polanski work, a 2D SOM map with MEP values was generated. The three-way array was constructed by collecting all 2D SOM maps. The correlation between the biological activity and the three-way array was investigated by the three-way PLS.[7] They extended their approach to a more general case where both the electrostatic and lipophilic potentials on the molecular surface change simultaneously.[8] In addition to these two studies, Hasegawa et al. also attempted to apply the SOM method to protein structure determination. However, due to the border effect and the strain caused by the forced 2D embedding the generated SOM map could not reproduce the complicated surface structure of a protein.

Recently, Erdas et al. reported the prediction of the binding affinities of phencyclidine-like compounds.[9] In their study, the molecular surface was mapped onto a tessellated sphere using the spherical SOM (SSOM) algorithm. The SSOM approach was chosen because of its ability to preserve the geometry of the molecular surface. In addition, the SSOM does not have topological defects present in the 2D map. As a result, Erdas et al. could extract important features that were useful in predicting binding affinities.

* Corresponding author.
E-mail address: funatsu@chemsys.t.u-tokyo.ac.jp (K. Funatsu).

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmc.2012.03.041

In a previous study, we studied the mapping ability of the SSOM.[10] The original 3D structure of the active site of the β2 protein structure was correctly reproduced by the SSOM. To validate the geometrical transformation and the resulting MEP distribution, the molecular surfaces of 20 β2 ligands were mapped onto the established protein SSOM sphere. The MEP values of the two spheres derived from the ligand and the β2 receptor protein were compared. In most cases involving potent ligands, the two spheres had a moderate negative correlation. This indicates that the SSOM approach has excellent potential to represent a complex protein surface as a simple spherical structure.

In this study, we performed a QSAR study of caspase-3 inhibitors based on the SSOM technique. Initially, the active site of the protein structure 'caspase-3' was characterized by the SSOM using MEP values. Each inhibitor was subsequently projected onto the protein SSOM sphere and the chemical descriptors were derived from the ligand SSOM sphere. The correlation of the chemical descriptors and the inhibitory activities was investigated using the support vector regression (SVR) method. Finally, the important MEP descriptors from the final SVR model were examined. The structural requirements of caspase-3 inhibitors are discussed from the perspectives of both the ligand and protein structures.

2. Material and methods

2.1. Data set

We collected the caspase-3 inhibitory data set from the literature.[11] The inhibitory activity was represented by the logarithm of the reciprocal value of the IC_{50} in the unit of nanomole (pIC_{50}), where IC_{50} represents the nano-molar concentration at which 50% inhibition of caspase-3 is achieved. We removed the cyano compounds and used 35 molecular data points for QSAR analysis, because the cyano compounds are known to show biological action against caspase-3 through a different mechanism. Table 1 shows the chemical structures and inhibitory activities of the caspase-3 inhibitors.

2.2. Spherical self-organizing map

SOM is a type of neural network that is applied in many fields.[12] SOM is based on the idea that human brains tend to spontaneously compress and organize sensory data. SOM can be used to generate a projection of objects from a higher-dimensional space onto a 2D space. In other words, this method enables a reduction in dimensions while conserving the topology of the information as much as possible. In many applications, there are borders surrounding the node arrangement. This causes a problem called the 'border effect' because the number of neighborhood nodes for a node near a border differs from that of a node near the center. This problem can be solved by arranging the nodes uniformly on the surface of a sphere.

Sangole et al. developed a spherical SOM (SSOM) approach. SSOM is a particular type of SOM in which nodes are arranged on a tessellated unit sphere with uniform triangular elements.[13] A tessellated unit sphere with uniform triangular elements is called an icosahedron. In the architecture of SSOM, each vertex corresponds to a node. The neighborhood of SSOM is similar to the hexagonal neighborhood of the 2D SOM because a node has six nearest neighbors. Moreover, the distance between a node and its nearest neighbors is uniform. The architecture of the SSOM used in this study is shown in Figure 1.

The applied SSOM algorithm is briefly described. If readers would be interested in the SSOM algorithm, please refer to our previous report.[10]

1. Initialize the weight vectors w^i of all nodes on the SSOM map. At each epoch of training, the following four steps are taken:

   a. An input point, x^p is chosen at random. Apply the steps b-d to each input point.
   b. Calculate the distance between the input point and the weight vectors of all nodes.
   c. Choose the node with minimum distance as the winner node.
   d. Update the weight vectors of the winner node and its neighbors.

2. Stop if the number of epoch exceeds the predefined training rounds. Otherwise, go back to step b.

The SSOM map is trained using the above algorithm. At the end, each point on the molecular surface is assigned to a node with specific values of weights.

2.3. Receptor mapping and molecular modeling

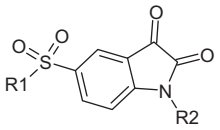
The X-ray crystal structure of caspase-3 deposited in the Protein Data Bank (PDB) was used to map the protein surface (PDB code: 1GFW).[14] The receptor surface within a 4.5 Å resolution from the ligand was extracted from the whole protein structure. The distance between the sampling points was set at 0.25 Å. The total number of sampling points was 12,171. On each sampling point, the value of the MEP was calculated using the electrostatic potential (ESP) suiting method in the MOE software package.[15] The particle mesh Ewald approach (PME)[16] was used as the ESP method in MOE. The sampling points on the receptor surface were mapped onto the tessellated nodes of the SSOM sphere. The SSOM sphere was trained by sampling all points on the van der Waals surface of the receptor protein using the learning algorithm of SSOM. According to the final established weight vectors determined by the SSOM training, each sampling point was placed on a specific node in the SSOM sphere. Subsequently, each node was coded by the associated MEP value of the original sampling point. An average value was taken when two or more sampling points were assigned to the same specific node. If an empty node was detected the corresponding MEP value was set to zero.

Thirty-five casapase-3 inhibitors were modeled according to the 3D structure of the X-ray crystal structure. That is, the R1 and R2 substituents in Table 1 were replaced by the fragment library deposited in the MOE software package.[15] The new constructed molecule was then docked into the active site of the casapase-3 protein using the docking software Glide.[17] The default setting was used for the docking calculation. The pose structure with the lowest docking energy was supposed to be the active conformation in the casapase-3 protein. For each docked inhibitor, the molecular surface was generated at the same resolution as for the protein structure. Each sampling point of the ligand was then mapped onto the established SSOM sphere using the weight vectors used in the receptor protein. Subsequently, the ligand SSOM with the MEP values was constructed.

2.4. Support vector regression

SVR was used to develop the predictive model for each caspase-3 inhibitor using the SSOM node with the MEP value. SVR is a regression-type support vector machine (SVM).[18] The general principle of SVM is to perform a classification by constructing an n-dimensional hyperplane that optimally separates the data set into two categories. The SVM minimizes the empirical classification error and maximizes the geometric margin. The margin is defined as the distance from the separating hyperplane to its nearest sample. Original data is initially nonlinearly mapped into a high

Table 1
Chemical structures and inhibitory activities of caspase-3 inhibitors



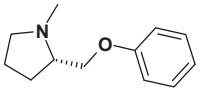
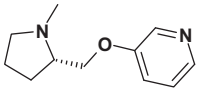
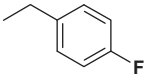
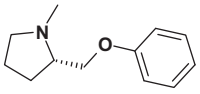
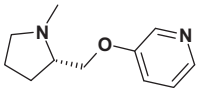
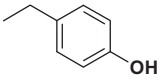
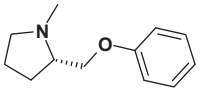
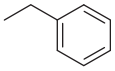
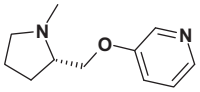
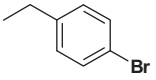
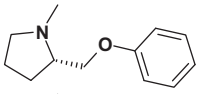
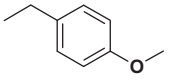
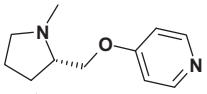
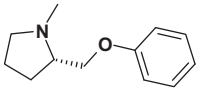
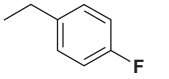
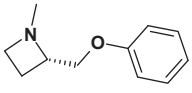
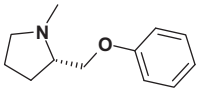
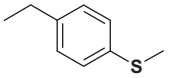
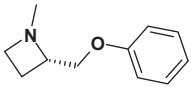
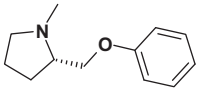
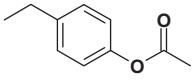
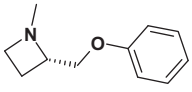
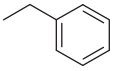
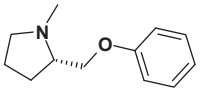
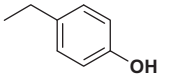
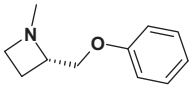
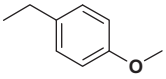
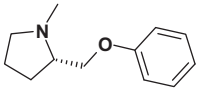
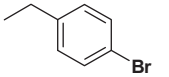
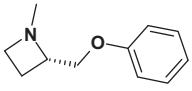
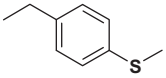
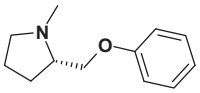
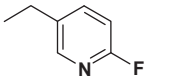
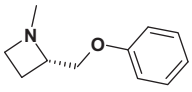
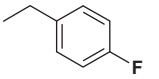
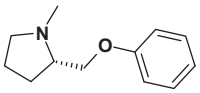
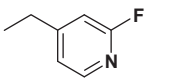
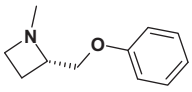
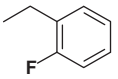
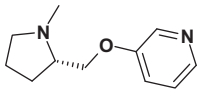
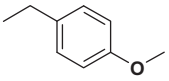
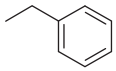
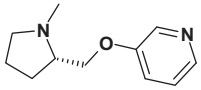
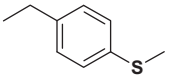
No.	R1	R2	pIC_{50}	No.	R1	R2	pIC_{50}
1		$-CH_3$	6.92	19			8.08
2		–H	6..62	20			8.41
3			7.91	21			8.44
4			7.84	22		–H	7.69
5			7.92	23		–H	6.54
6			7.91	24		$-CH_3$	7.04
7			7.92	25			8.01
8			7.87	26			8.08
9			8.01	27			7.95
10			7.99	28			8.06
11			7.67	29			8.03
12			8.04	30			7.96
13			8.01	31			7.53
14		–H	7.23	32			8.24
15		$-CH_3$	7.63	33			5.84
16			8.28	34			5.99

Table 1 (*continued*)

No.	R1	R2	pIC_{50}	No.	R1	R2	pIC_{50}
17			8.41	35			6.94
18			8.36				

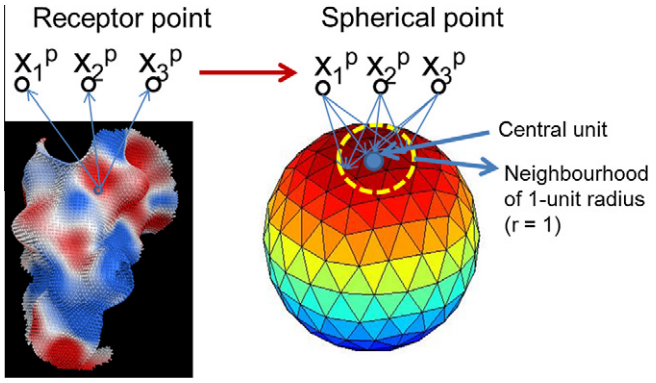


Figure 1. Architecture of the SSOM.

dimensional feature space, and then a linear function is fitted to approximate the decision function between **x** and **y**.

The regression problem can be transformed into the binary classification problem. For each sample x_i in the training data, the corresponding y_i is added by a positive number d to produce one new sample (x_i, y_i) belonging to class 1. Similarly, the y_i can also be subtracted by the same d to produce another new sample (x_i, y_i) belonging to class −1. Repeating this procedure, the N samples for regression are doubled and classified into two classes. The regression function can then be calculated using the same algorithm in SVM. The predictive performance of the SVR model is evaluated based on the Q^2 value derived from leave-one-out cross-validation (LOOCV).

After building the SVR model, local gradients were calculated to obtain the correlation coefficient of each descriptor as is the case for multiple-linear regression.[19] This method is called as the sensitivity analysis. A set of values of all descriptors is randomly generated and the value of activity is predicted. Then, the value of each descriptor is slightly changed, and differences of the predicted value are calculated for all descriptors. Sensitivity of each descriptor is defined as the mean of differences of predicted activities. It is thought that the magnitude of the differences of predicted value reflects the importance of each variable. Therefore, the absolute values of these sensitivities can be used as the descriptor importance. The local gradient was calculated using the scripts written in our laboratory in the R environment.

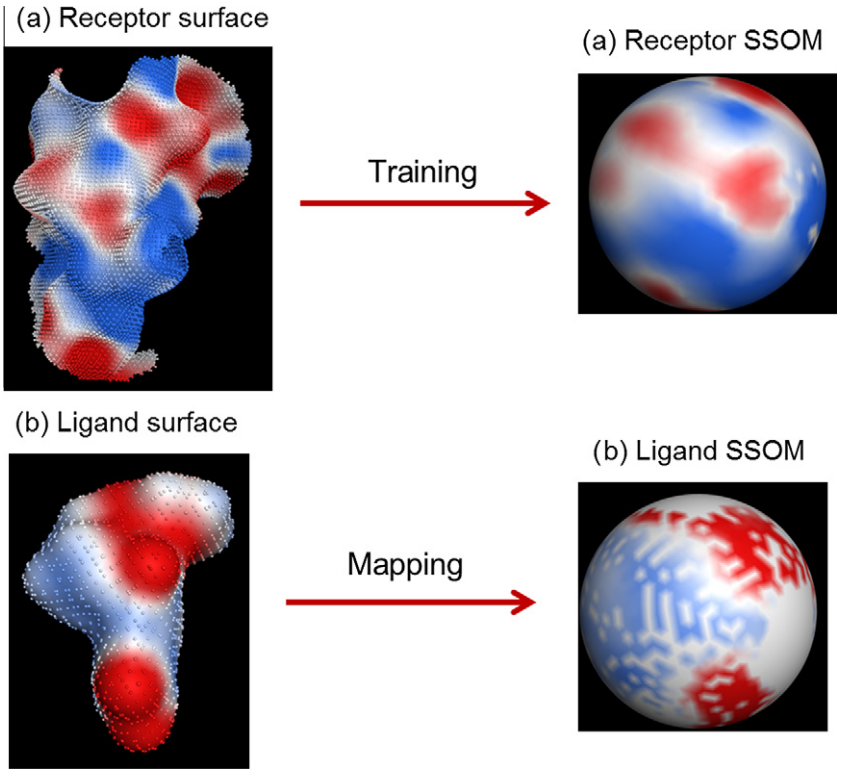


Figure 2. (a) Result of the training from the receptor surface. (b) Result of the mapping from the ligand surface. Red and blue colors represent negative and positive MEP values, respectively.

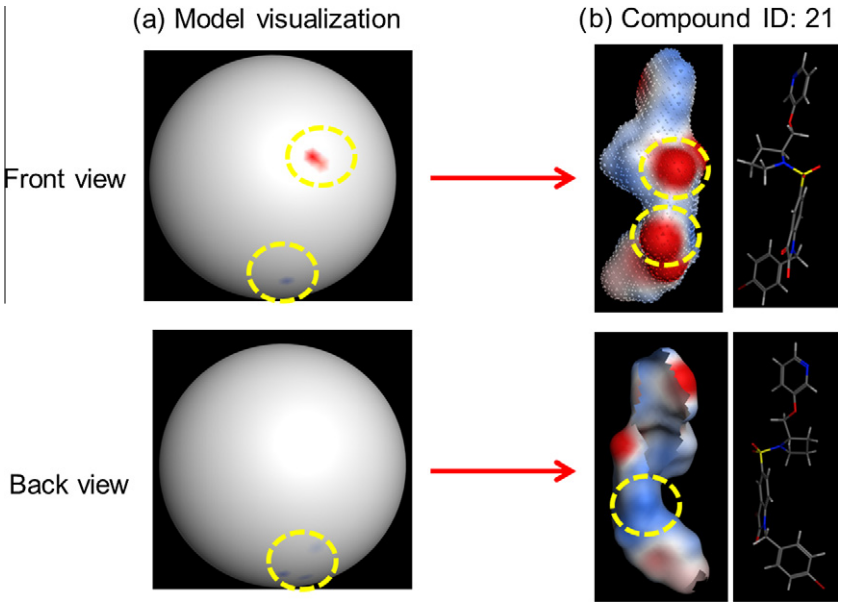


Figure 3. (a) Visualization of the correlation coefficient on the ligand SSOM. Red and blue colors indicate negative and positive values, respectively. (b) The original molecular surface with MEP values and the 3D structure. Red and blue colors represent negative and positive MEP values, respectively.

3. Results and discussion

3.1. SSOM analysis

Initially, the L-curve approach was used to select the two important parameters (accuracy and smoothness).[10] In the L-curve criterion, smoothness is defined as the sum of distances between the weight vector of each node (w^i) and the mean of its nearest neighbors (W^i). The smaller this sum, the smoother the SSOM model becomes. Likewise, accuracy is defined as the sum of the distances between each input point (x^p) and its closest weight vector (w^i). The smaller this sum, the better the original 3D structure can be reproduced by the SSOM model. The number of training rounds was chosen as 100 from monitoring the convergence plot. In this case, the values of accuracy and smoothness were determined to be 0.74×10^4 and 82, respectively. In these combinations, the molecular surface on the caspase-3 receptor was trained on the SSOM sphere with 2562 points. The MEP value on each point was then added with two colors. The result of the training from the receptor surface is shown in Figure 2(a). Red and blue colors represent negative and positive MEP values, respectively. As shown in Figure 2(a), the distributions of the two MEP values derived from the original receptor surface and the SSOM sphere are preserved.

Each caspase-3 inhibitor was mapped onto the protein SSOM sphere. The SSOM sphere derived from the most potent ligand (compound **21**) is shown in Figure 2(b). Comparing the protein and ligand spheres, the two MEP values inversely match. The value of the correlation coefficient was 0.758.

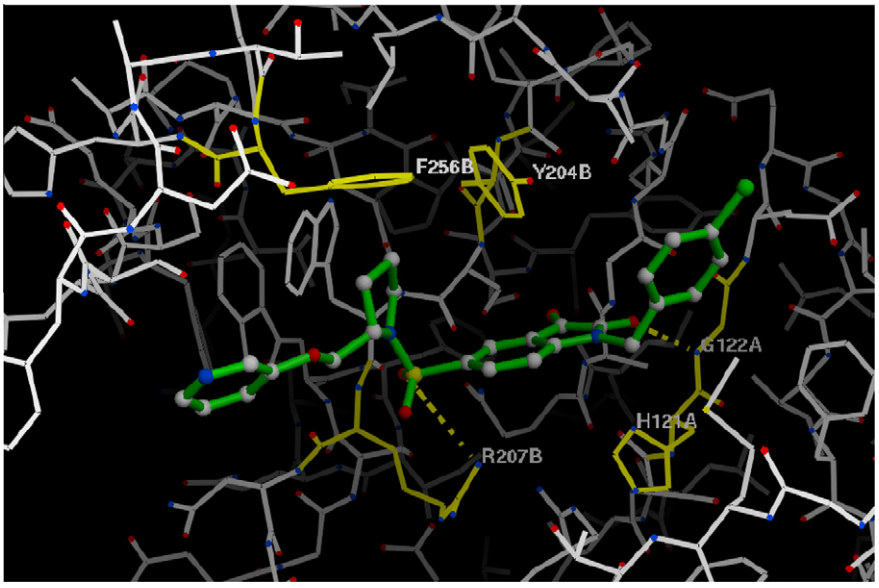


Figure 4. Molecular modeling of compound **21** in the caspase-3 protein.

3.2. SVR model

Prior to SVR analysis, the 2562 SSOM descriptors were reduced to 190 using a variance cut-off filter value of 0.5. The descriptors were further reduced to 24 using a *p*-value of 0.001. A systematic grid search[19] was then used to determine the best parameter values based on the Q^2 value derived from LOOCV ($C = 8$, $v = 0.15$ and $g = 0.03125$), where C is a trade-off parameter between the training error and the model complexity, v is the lower bound of the proportion of support vectors to the total samples and g determines the shape of the radial basis function. Finally, the procedure of backward-elimination was applied to obtain the final SVR model with 10 descriptors. The R^2 and Q^2 values of the final model were 0.983 and 0.638, respectively.

3.3. Structural requirements

The local gradient method was applied to ten descriptors in the final SVR model and the correlation coefficient value was back-projected onto the SSOM sphere. The resulting back-projection map from both front and back views is presented in Figure 3(a). In this figure, red and blue colors indicate the negative and positive correlation coefficient values, respectively.

In Figure 3(a), ten MEP descriptors have converged to three main regions indicated by the yellow dashed circles: one negative and two positive correlation coefficient regions. The original molecular surface with the MEP values and the 3D structure is shown in Figure 3(b). Red and blue colors represent negative and positive MEP values, respectively. The corresponding regions are highlighted by yellow dashed circles in order to make direct comparisons. The red region in Figure 3(a) corresponds to two oxygen atoms of the sulfone-amide moiety in Figure 3(b). The blue region in Figure 3(a) corresponds to the second oxygen atom on the oxy-indazole moiety in Figure 3(b). Another blue region in Figure 3(a) corresponds to the benzene ring of the oxy-indazole moiety in Figure 3(b). From them, we can estimate the following structural features required to provide a higher inhibitory activity:

1. Two oxygen atoms of the sulfone-amide have high electronegativity.
2. The second oxygen atom on the oxy-indazole has low electrone gativity.
3. The benzene ring of the oxy-indazole has low electronegativity.

These structural requirements have rationale from the X-ray crystal structure of the caspase-3 protein (Fig. 4). The sulfone-amide interacts with the R207B side chain through hydrogen-bonding. The first oxygen atom on the oxy-indazole hydrogen-bonds with the G122A back-bone NH group. The lower electronegativity of the second oxygen atom may induce the strong hydrogen bonding interactions of first oxygen atom. The benzene ring of the oxy-indazole has three π–π interactions with the side-chains of Y204B, F256B and H121A. The lower electronegativity facilitates π–π interactions.

4. Conclusion

In this study, we perform a QSAR study of caspase-3 inhibitors based on the SSOM technique. The MEP values on the ligand SSOM sphere were used as chemical descriptors. The correlation of the chemical descriptors and the inhibitory activities was investigated by the SVR method. The important MEP descriptors were derived from the final SVR model. Based on the X-ray crystal structure of the protein, the descriptors matched the structural requirements of caspase-3 inhibitors.

This study is one aspect of the SSOM application focusing on the QSAR model. In a previous study, we demonstrated that SSOM has a high mapping ability for reproducing the original surface of the protein structure. Using two SSOM maps derived from the ligand and protein surfaces, we can create the comparable descriptors for the ligand and protein molecules. This means that SSOM has more potential extensions in other fields such as chemogenomics. The purpose of chemogenomics is to identify suitable inhibitors against an orphan target receptor whose function has not been understood.[20] The multiple molecular features such as the electrostatic, lipophilic and hydrogen-bonding potentials can be used to characterize the surfaces of two counterparts using SSOM. By combining two SSOM descriptor blocks, we are able to correlate chemical and protein structures with binding affinities in more systematic way.

References and notes

1. Gedeck, P.; Lewis, R. A. *Curr. Opin. Drug Dis. Dev.* **2008**, *11*, 569.
2. Yap, C. W.; Li, H.; Ji, Z. L.; Chen, Y. Z. *Mini-Rev. Med. Chem.* **2007**, *7*, 1097.
3. Polanski, J.; Walczak, B. *Comput. Chem.* **2000**, *24*, 615.
4. Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. *J. Comput. -Aided Mol. Des.* **1996**, *10*, 521.
5. Hasegawa, K.; Funatsu, K. 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 Global, 2011; pp 145–168.
6. Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. *Comput. Chem.* **2002**, *26*, 583.
7. Bro, R. *J. Chemom.* **1996**, *10*, 47.
8. Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. *Comput. Biol Chem.* **2003**, *27*, 381.
9. Erdas, O.; Buyukbingol, E.; Alpaslan, F. N.; Adejare, A. *J. Chemom.* **2010**, *24*, 1.
10. Hasegawa, K.; Funatsu, K. *Mol. Inf.* **2012**, *31*, 161.
11. Wang, Q.; Mach, R. H.; Reichert, D. E. *J. Chem. Inf. Model.* **2009**, *49*, 1963.
12. Digles, D.; Ecker, G. F. *Mol. Inf.* **2011**, *30*, 838.
13. Sangole, A.; Knopf, G. *Int. J. Smart Eng. Syst. Des.* **2003**, *5*, 11.
14. http://www.rcsb.org/pdb/home/home.do.
15. http://www.chemcomp.com/software.htm.
16. Darden, T.; York, D.; Pederson, L. *J. Chem. Phys.* **1993**, *98*, 10089.
17. http://www.schrodinger.com/productsguide/.
18. Hasegawa, K.; Funatsu, K. *Curr. Comput. -Aided Drug Des.* **2010**, *6*, 24.
19. Arakawa, M.; Hasegawa, K.; Funatsu, K. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 145.
20. Rognan, D. *Mol. Inf.* **2010**, *29*, 176.