行政院國家科學委員會專題研究計畫　期中進度報告

**電子郵件與虛擬藥物篩選之研究**

**計劃類別：**個別型計畫  
**計劃編號：**

**執行期間：**　年 月 日至 年 月 日  
**執行單位：**國立交通大學生物科技學系

**計劃主持人：**楊進木

**報告類型：**精簡報告  
**報告附件：**出席國際會議研究心得報告及發表論文  
**處理方式：**本計畫可公開查詢

中華民國 年 月 日
In first year, we developed a QSAR methodology associating molecular docking and feature selection with PLS. The feature of our model generates from the interaction energies of docked results, named as protein-ligand interaction profile and was extracted as atom-based, group-based and residue-based energetic terms. The concept of our QSAR analysis came from COMBINE analysis. First, GEMDOCK predicted the binding conformation for protein-ligand and generated atom paired energies. The atom paired energetic terms included hydrogen bonding, electrostatic and van der Waals interactions. These atom paired energetic terms would serve as atom-based features. The sum of atom paired energies of each residue was residue-based term and each residue could divide into two parts of main chain and side chain which called group-based terms. GEMPLS served as feature selection and model building in QSAR analysis. Potential features for contributing inhibition would be selected by evolutionary strategy and built regression by PLS.

We applied our QSAR methodology to build inhibitory QSAR models of neuraminidase, glycogen phosphorylase b and cyclooxygenase-2. We compared our preliminary results to published references and our performances showed more prediction power than published models. In the recently future, we will make more efforts to improving our methodology and
combine virtual screen to create a high through-put prediction environment.

**Keywords (keywords)**

QSAR, Virtual screening, GEMOSAR, COMBINE, GEMDOCK, GEMPLS, Neuraminidase, Glycogen Phosphorylase b, GPB, Cyclooxygenase-2, COX-2

**前言與研究目的:**

QSAR techniques are commonly regarded as a key role to computational molecular design. The major goal of QSAR is to formulate mathematical relationships between physicochemical properties of compounds and their experimentally determined *in vitro* biological activities. Thus the derived QSAR model can be subsequently used to predict the biological activities of new derivatives. A good QSAR model both enhances our understanding of the specifics of drug action and provides a theoretical foundation for lead optimization.

The QSAR methodology, COMparative BINding Energy (COMBINE), develops a linear relationship between binding free energy of the ligand-receptor complexes. In general, residue is the basic unit utilized in the feature of binding energy. But there are just a few differences between the side chains of compounds of QSAR model. For instance, the activity was changed by just one atom altered in the side chain. However, it could be difficult to detect by residue-based method. Using residue as the unit is insensitive and easy to generate outliers by comparison. In order to make QSAR model more sensitivity, we use group and atom as the new feature.

The partial least square (PLS) analysis is able to deal with strongly collinear input data and make no restriction on the number of variables used. Unfortunately, the predictive performance of PLS model drops and the PLS model becomes complicated when the number of features increases. Several feature selection methods for PLS have been proposed, in which genetic algorithm (GA) combined with PLS approach (GAPLS) has demonstrated the improvement on the prediction and interpretation of model. GEMPLS is general able to evolve the relationship between biological activities and compound features generated by COMBINE.

In the COMBINE method, the binding energy of the receptor-ligand complex is correlated to the interaction energy components. In this study, “ELE” is electrostatic interactions; “VDW” is van der Waals interactions; “HYD” is hydrogen bond interaction. Each selected energy component $u_i$ contributes to the binding free energy according to its weight $w_i$ and PLS analysis is applied to obtain the weights $w_i$: 
The purpose of GEMQSAR is to develop a novel *in silico* drug screening system combining 3D QSAR and virtual screening process. To archive the objective, we first developed the methodology of QSAR (GEMPLS\textsuperscript{6}) and applied GEMPLS to evolve the QSAR models. This model was generated by the COMBINE method\textsuperscript{2; 3} according to the three different units of interaction energy features. The 3D QSAR methodology were applied on three public data sets including 38 influenza neuraminidase inhibitor complexes\textsuperscript{7}, 76 glycogen phosphorylase b inhibitors\textsuperscript{8} and 31 cyclooxygenase-2 inhibitors\textsuperscript{9}. Experiments showed that the reduced units were able to improve the predictability and efficiency, and at the same time, the selected features in the yielded QSAR model were consistent with some experimental evidences.

\begin{equation}
pIC_{50} = \sum_i W_i^{ELE} u_i^{ELE} + \sum_i W_i^{HYD} u_i^{HYD} + \sum_i W_i^{VDW} u_i^{VDW} + C \tag{1}
\end{equation}

The final goal of GEMQSAR is to develop a novel *in silico* drug screening system. The relationships of QSAR and docking are shown in Figure 1. The molecular docking technique was well established in our past researches\textsuperscript{10}. In the first year, we developed the core 3D QSAR methodology of GEMQSAR, named GEMPLS\textsuperscript{11}. Figure 2 shows the main steps of applying GEMPLS\textsuperscript{11} in the COMBINE analysis\textsuperscript{2; 3}: 1) prepare the inhibitor set and model protein-inhibitor complexes; 2) refine protein-inhibitor complexes and calculate features (i.e., energy interactions); 3) select important features by Mahalanobis distance; 4) select features and evolve QSAR models. 5) Performance evaluation. Each step is described in the following subsections.

The COMBINE analysis is the use of structural information about ligand-receptor complexes\textsuperscript{2; 3}. When the three-dimensional structure of macromolecule is available, ligand-receptor interaction energies could be calculated as features, which are subjected to statistical analysis in COMBINE. A subset of these features will be account for the ligand affinity. The critical interaction patterns between ligands and the receptor could be identified and be used to derive the correlation of binding affinities.
GEMDOCK
High Throughput System

Virtual Screening for Leads

Induced Ligand
3D Conformations

Protein-Ligand Interactions

Compound Structure/Physical
Chemistry Property Profile

Protein-Ligand Pharmacophore
Interaction Profile

GA-PLS/GA-κNN

Building QSAR Model

Critical Interaction and
Property Information

Predicted Biological Activities
and Discovering Potential Leads

3D QSAR Process

Figure 1. The relationships of 3D QSAR and docking process in GEMQSAR

Feature extraction

COMBINE

Step 1: Prepare data set and model protein-inhibitor complex

QSAR model evolution

GEMPLS

Step 3: GEM selects a feature set (X)

Step 4: PLS builds the relationship between feature set (X) and activity

Figure 2. The framework and steps of GEMPLS applied in the COMBINE analysis
Data preparation and feature extraction

1) Prepare Data Sets and Model Protein-Inhibitor Complexes

We have used three data sets, 38 influenza neuraminidase (NA) inhibitor complexes\(^7\), glycogen phosphorylase b (GPB)\(^{12}\) and cyclooxygenase-2 (COX2)\(^9\) as our validation and test. The inhibitory activity data and complex structures were mainly taken from the references\(^7;9;12\). Each cavity was centered the ligand in the complex with a cutoff 7.5 Å, and cavities with ligands docked by GEMDOCK. Then aligned all cavities by superposition, and the interactions of these gap residues were not considered for the COMBINE analysis.

2) Refine protein-inhibitor complexes and calculate features

The calculated ligand-receptor interaction energies were partitioned on three basis, residue, group, and atom. : 1) Residue-based method, the total binding energy of a residue is the basis unit. : 2) Group-based method, divide a residue to main chain and side chain, and the group is the basis unit. : 3) Atom-base method, the binding energy of an atom is the basis unit. The interaction profiles were outputted for QSAR model. The ligand-receptor interaction energies included van der Waals interaction (\(E_{vdW}\)) and hydrogen bonding interaction (\(E_{h-bond}\)) as below.

\[
E_{vdW} = \sum_{i=1}^{\text{lig}} F_{vdW}(r_{ij})
\]

\[
E_{h-bond} = \sum_{i=1}^{\text{lig}} F_{hb}(r_{ij})
\]

\[
F(r_{ij}) = \begin{cases} 
V_6 - \frac{V_6 r_{ij}^{V_6}}{V_1} & \text{if } r_{ij}^{V_6} \leq V_1 \\
V_2 (r_{ij}^{V_2} - V_1) & \text{if } V_1 < r_{ij}^{V_2} \leq V_2 \\
\frac{V_2 - V_1}{V_2 - V_1} & \text{if } V_2 < r_{ij}^{V_2} \leq V_3 \\
V_5 - \frac{V_5 (r_{ij}^{V_5} - V_3)}{V_4 - V_3} & \text{if } V_3 < r_{ij}^{V_5} \leq V_4 \\
0 & \text{if } r_{ij}^{V_5} > V_4
\end{cases}
\]

\(lig\) presents the number of atom and \(r_{ij}\) is the distance of atom \(i\) and protein \(j\). \(F_{hb}\) and \(F_{vdW}\) are hydrogen bonding and van der Waals interaction, respectively. \(F(r_{ij})\) is the linear function of six parameters. The six parameters of \(V_1-\)\(V_6\) are shown in figure 3.
Figure 3. The linear functions and six parameters of hydrogen bonding interaction and van der Waals interaction in GEMDOCK

**QSAR Model Evolution and building: GEMPLS**

PLS has played a critical role in the derivation of QSAR in CoMFA or COMBINE studies. Recently, more and more people recognize the benefits of feature selection before PLS regression. GAPLS has been shown as a practical solution. But when the number of features becomes large, GAPLS still has difficulty in driving out noises. And scanning for best \( l_v \) is too inefficient and time consuming. Here, we introduce a number of successive enhancements, which are described in the following paragraphs, to construct our model GEMPLS to overcome the drawbacks of GAPLS.

The general idea of PLS is to try to extract these latent variables, accounting for as much of the manifest feature variation as possible while modeling the inhibitory activities well. To decide both the optimum number of latent variables and prediction error of a QSAR model, we defined the weighted standard deviation error of the predictions (WSDEP) as the scoring function of our GEMPLS:

\[
WSDEP = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N - l_v - 1}} \left( \frac{100}{95} \right)^{l_v}
\]

where \( y_i \) and \( \hat{y}_{pred,i} \) are the observed and predicted inhibitory activities belong to inhibitor \( i \), \( N \) is the total number of samples, and \( l_v \) is the number of latent variables in the current model. In order to improve on the efficiency, we append an extra bit \( l_v \), representing the number of latent
variables, to the original chromosome and expect GEMPLS model to efficiently solve the problem of the optimum number of latent variables through evolutionary process.

3) Select Features by Mahalanobis Distance

Mahalanobis distance is able be used to measure the deviation of a sample from the mean of the distribution in multivariable calculus. Therefore, the Mahalanobis distance is adopted to identify significant features from all of those.

\[ M^2 = (v - u)' \sum^{-1} (v - u) \]  

(6)

\( M \) is the Mahalanobis distance from the feature vector \( v \) (column vector of data matrix here) to the mean vector \( \mu \), where \( \Sigma \) is the covariance matrix of the features.

4) Feature Selections and QSAR Models Evolution

The inhibitory activity usually correlates with few important interaction energy features, that is, most of interaction energy features are meaningless or not apparently distinct from each other. GEM was applied to find out the significant interaction energy features and PLS was used to build the QSAR models based on these selected features. \( WSDEP \) was used as the objective function to provide a measure of how the internal predictability with respect to the selected features. The fittest individual will have the lowest \( WSDEP \).

GEM, modified and enhanced from our previous works\(^6\), consists of five steps briefly described in the following:

1) **Initiation and evaluation of the initial population** \((G_{t=0})\). Each chromosome is composed by an array of feature set and an \( lv \) value. For example, a chromosome has \( n+3 \) bits if the number of candidate feature is \( n \) and three bits for \( lv \) value. The initial population \((G_{t=1})\) of population size \((N_p)\) is created by setting feature bits (0 denote the absence of corresponding feature, and 1 denote its presence) and an \( lv \) value (denote the number of latent variables and range in [1~5]) of each chromosome to random values and one, respectively. Then PLS is used to build a quasi-QSAR model, and evaluated by the scoring function \( WSDEP \), for each chromosome.

2) **Selection of the reproductive population**. The chromosomes of reproductive population \((P,G_t)\) are selected from the population \((G_t)\) with a fixed proportion \((P_s)\) according to the stochastic universal sampling.

3) **Crossover and mutate the reproductive population** \((P,G_t)\). The offspring population \((G_{off})\) is generated by uniform crossover with a probability (crossover rate: \( P_c \)) and mutation operators, including uniform and biased mutation operators, with a probability (mutation rate: \( P_m \)).

4) **Evaluation of the offspring population** \((G_{off})\). PLS is then used to build a quasi-QSAR model, and evaluated by the scoring function \( WSDEP \), for each chromosome.
model, evaluated by WSDEP, for each chromosome in the offspring population.

(5) **Reinsertion of the child population.** To form the population of the next generation ($G_{\text{next}}$), the chromosomes of the current population ($G_t$) with lower objectives in the preceding ($1-\text{P}_s$) proportion are protected to the next generation, while the others are replaced with better ones from the offspring population ($G_{\text{off}}$). Let $t = t+1$ and $G_t = G_{\text{next}}$.

(6) The cycle of above four steps (from step 2 to 5) is repeated until the number of generation reaches to the maximum number of generations ($N_{\text{max}}$). The values of empirical parameters are defined as follows: $N_p = 100$, $N_{\text{max}} = 200$, $P_s = 0.9$, $P_e = 0.6$, and $P_m = 0.05$.

Genetic operator: Biased Mutation. The uniform mutation may incur a risk of local convergence and slow evolution because plenty of features will raise the combinatorial complexity of feature space. To reduce the phenomena, the uniform mutation was cooperated with biased mutation to lead the evolution of GA toward significant feature set and to reduce the interference of noise features.

$$F(x_i) = \text{MIN} + (\text{MAX} - \text{MIN}) \times \left( \frac{N_f - X_i}{N_f - 1} \right)$$ (7)

where $F(x_i)$ is the probability of selection of feature $i$; $x_i$ is the rank of feature $i$ in the descending order of Mahalanobis distance of all features, MIN and MAX are the lower and upper bounds, respectively, of probability of biased mutation; $N_f$ is the number of significant features. The value of $F(x_i)$ is derived from $x_i$ only when $x_i$ is ahead of $N_f$, otherwise $F(x_i)$ is set to MIN. The meaning of $F(x_i)$ is that the more significant feature, the more higher probability of selection. In this study, $\text{MAX}=0.8$, $\text{MIN}=0.2$ and $N_f=39$.

5) **Performance Evaluation**

The predictability of QSAR model was assessed by the conventional correlation coefficient ($r^2$), the cross-validated correlation coefficient ($q^2$), the cross-validated SDEP ($SDEP_{cv}$), and external SDEP ($SDEP_{ex}$):

$$q^2 = 1 - \frac{\sum (y_i - y_{\text{pred},i})^2}{\sum (y_i - \bar{y})^2}$$ (8)

$$SDEP = \sqrt{\frac{\sum (y_i - y_{\text{pred},i})^2}{N}}$$ (9)

where $y_i$ and $y_{\text{pred},i}$ are the observed and predicted activity of inhibitor $i$, $y_{\text{pred},i}$, respectively, $\bar{y}$ is the average activity value of the inhibitor set, and $N$ is the total number of inhibitors. The model with more remarkable predictability can provide the higher correlation coefficient ($r^2$, $q^2$) and
the lower SDEP between the observed and predicted inhibitory activities.

結果與討論

The ligands were divided to training set and testing set according to the reference$^7;^9;^{12}$. The three methods of feature exaction (residue-based, group-based, and atom-based) were used to train three QSAR models by Leave-One-Out method to optimize WSDEP. Our results of neuraminidase were shown in Table 1. The $q^2$ of the reference in training set was better than the results of the three GEMQSAR models, but the $q^2$ of the GEMQSAR models in testing set were better than the $q^2$ of the reference. Our model although had sight lower training correlation to Wade et al but we showed more superior prediction power than Wade et al. The energy basis unit was reduced (residue->group->atom), and the $q^2$ in training set improved with feature unit reduced. However the $q^2$ in testing set didn’t become better like in training set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Original features$^a$</th>
<th>Selected features$^b$</th>
<th>$Lv^c$</th>
<th>Train $r^2^d$</th>
<th>Train $q^2^e$</th>
<th>Test $r^2$</th>
<th>Test $q^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residue-based</td>
<td>153</td>
<td>13</td>
<td>3</td>
<td>0.753</td>
<td>0.611</td>
<td>0.85</td>
<td>0.754</td>
</tr>
<tr>
<td>Group-based</td>
<td>306</td>
<td>14</td>
<td>3</td>
<td>0.737</td>
<td>0.621</td>
<td>0.876</td>
<td>0.798</td>
</tr>
<tr>
<td>Atom-based</td>
<td>1233</td>
<td>33</td>
<td>3</td>
<td>0.794</td>
<td>0.688</td>
<td>0.830</td>
<td>0.769</td>
</tr>
<tr>
<td>Wang et al$^7$</td>
<td>770</td>
<td>330</td>
<td>3</td>
<td>0.877</td>
<td>0.875</td>
<td>0.582</td>
<td>0.566</td>
</tr>
</tbody>
</table>

$^a$ The number of features extracted from original data by different methods

$^b$ The number of features selected by GEMPLS

$^c$ Latent variable

$^d$ The conventional correlation coefficient.

$^e$ The cross-validated correlation coefficient

Generally, PLS easily trended to over-fit when the number of features increasing. Therefore, a well-developed strategy of feature extraction should consider the balance of feature number and predicting ability. We develop three QSAR model regarding three strategies of feature extraction and analysis the relationship of performance and feature numbers. Comparing with group-based and residue-based models, the residue-based model had better training quality but lower predicting power. The higher predicting ability showed that group-based method extracted more accuracy information in our QSAR analysis. In other words, one residue-based unit might contain the noises and meaningful features. After dividing feature unit form residue to backbone
and side-chain, the meaningful feature was correctly recovered in GEMPLS (Table 2.). Sum up the above, the advantages of group-based were: 1) Include residue-based information and delete noise of feature by feature selection. 2) Select fewer features than atom-based and uneasy over fitting. According to present results, it is better to use group-based unit to build a QSAR model.

Table 2. The important groups and their GEMPLS coefficient.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Group type</th>
<th>Energy type</th>
<th>GEMPLS coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARG118</td>
<td>side</td>
<td>VDW</td>
<td>-0.192</td>
</tr>
<tr>
<td>ARG118</td>
<td>side</td>
<td>ELE</td>
<td>-0.12</td>
</tr>
<tr>
<td>GLU119</td>
<td>side</td>
<td>ELE</td>
<td>-0.1</td>
</tr>
<tr>
<td>ARG152</td>
<td>main</td>
<td>VDW</td>
<td>-0.396</td>
</tr>
<tr>
<td>TRP178</td>
<td>side</td>
<td>HYD</td>
<td>-0.359</td>
</tr>
<tr>
<td>SER179</td>
<td>main</td>
<td>VDW</td>
<td>-0.135</td>
</tr>
<tr>
<td>SER179</td>
<td>side</td>
<td>VDW</td>
<td>-0.124</td>
</tr>
<tr>
<td>ILE222</td>
<td>side</td>
<td>VDW</td>
<td>-0.236</td>
</tr>
<tr>
<td>GLU227</td>
<td>main</td>
<td>HYD</td>
<td>-0.313</td>
</tr>
<tr>
<td>ALA246</td>
<td>main</td>
<td>VDW</td>
<td>-0.133</td>
</tr>
<tr>
<td>ALA250</td>
<td>side</td>
<td>VDW</td>
<td>-0.674</td>
</tr>
<tr>
<td>ALA250</td>
<td>side</td>
<td>ELE</td>
<td>-0.169</td>
</tr>
<tr>
<td>VAL275</td>
<td>side</td>
<td>ELE</td>
<td>-0.227</td>
</tr>
<tr>
<td>ARG292</td>
<td>main</td>
<td>VDW</td>
<td>-0.232</td>
</tr>
<tr>
<td>VAL349</td>
<td>side</td>
<td>VDW</td>
<td>-0.207</td>
</tr>
</tbody>
</table>

The predicted pIC50 values were plotted against experimental pIC50 values for the group model with three latent variables in Figure 4. From Figure 4, the group GEMQSAR model has a good prediction. The results of GEMPLS were shown in Figure 5 (a), which showed the coefficients of the electrostatic, hydrogen bond and van der Waals. The negative coefficients contributed to the activity, and the important groups were listed in table 2. The important structural features for a strong inhibitor and corresponding groups were shown in Figure 5 (b). Some interactions played critical role to contribute higher activity: 1) The electrostatic interactions in orange groups. 2) The hydrogen bond interactions in green groups. 3) The van der Waals interactions in gray groups. On the basis of the above COMBINE analysis, we could
develop new inhibitors which have high activity according to the suggested properties of the GEMQ SAR model.

Figure 4. Experimental pIC50 values versus predicted pIC50 values for the group model derived from 38 complexes: ●, predicted values for training set from leave-one-out cross-validation at three latent variables; △, predicted values for testing set.

Figure 5. Selected features of neuraminidase inhibitor model (a) The coefficients of the electrostatic, hydrogen bond and van der Waals. (b) The important structural features of binding site.
Table 3. Performance comparison of GEMPLS and the reference for glycogen phosphorylase b

<table>
<thead>
<tr>
<th>Model</th>
<th>Original features(^a)</th>
<th>Selected features(^b)</th>
<th>(L_v)(^c)</th>
<th>Train (r^2)(^d)</th>
<th>Train (q^2)(^e)</th>
<th>Test (r^2)</th>
<th>Test (q^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residue-based</td>
<td>120</td>
<td>16</td>
<td>3</td>
<td>0.481</td>
<td>0.337</td>
<td>0.115</td>
<td>0.002</td>
</tr>
<tr>
<td>Group-based</td>
<td>240</td>
<td>18</td>
<td>3</td>
<td>0.666</td>
<td>0.546</td>
<td>0.286</td>
<td>0.271</td>
</tr>
<tr>
<td>Atom-based</td>
<td>918</td>
<td>32</td>
<td>3</td>
<td>0.699</td>
<td>0.584</td>
<td>0.28</td>
<td>0.26</td>
</tr>
<tr>
<td>Silber et al(^{12})</td>
<td>na(^f)</td>
<td>na(^f)</td>
<td>na(^f)</td>
<td>0.92</td>
<td>na(^f)</td>
<td>0.59</td>
<td>na(^f)</td>
</tr>
</tbody>
</table>

\(^a\) The number of features extracted from original data by different methods  
\(^b\) The number of features selected by GEMPLS  
\(^c\) Latent variable  
\(^d\) The conventional correlation coefficient  
\(^e\) The cross-validated correlation coefficient  
\(^f\) Data not available

Table 4. Performance comparison of GEMPLS and the reference for cyclooxygenase-2

<table>
<thead>
<tr>
<th>Model</th>
<th>Original features(^a)</th>
<th>Selected features(^b)</th>
<th>(L_v)(^c)</th>
<th>Train (r^2)(^d)</th>
<th>Train (q^2)(^e)</th>
<th>Test (r^2)</th>
<th>Test (q^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residue-based</td>
<td>135</td>
<td>21</td>
<td>3</td>
<td>0.504</td>
<td>0.362</td>
<td>0.066</td>
<td>-2.131</td>
</tr>
<tr>
<td>Group-based</td>
<td>270</td>
<td>19</td>
<td>3</td>
<td>0.527</td>
<td>0.385</td>
<td>0.026</td>
<td>-1.736</td>
</tr>
<tr>
<td>Atom-based</td>
<td>1230</td>
<td>62</td>
<td>3</td>
<td>0.688</td>
<td>0.566</td>
<td>0.066</td>
<td>-1.203</td>
</tr>
<tr>
<td>Pei et al(^9)</td>
<td>na(^f)</td>
<td>na(^f)</td>
<td>na(^f)</td>
<td>0.61</td>
<td>na(^f)</td>
<td>0.34</td>
<td>na(^f)</td>
</tr>
</tbody>
</table>

\(^a\) The number of features extracted from original data by different methods  
\(^b\) The number of features selected by GEMPLS  
\(^c\) Latent variable  
\(^d\) The conventional correlation coefficient  
\(^e\) The cross-validated correlation coefficient  
\(^f\) Data not available

We applied the same strategy to the other data sets: GPB\(^{12}\) and COX-2\(^9\), and the results of GPB and COX-2 were shown in Table 3 and Table 4, respectively. The low test \(q^2\) in both data...
sets mean that the descriptors of interaction profiles didn’t have a correlation with activities. There might be some reasons to explain that: 1) Some atom types weren’t defined clearly in GEMDOCK scoring function. For example, the atom type F, was regarded as like C. 2) The hydrogen-bond interaction wasn’t sensitive in GEMDOCK scoring function. In QSAR model, there was little difference in the side chains of ligands, and it was necessary to generate the accurate descriptors of interactions to determine the difference. We would correct the two problems to improve the QSAR model in the recently future.

In summary, we apply GEMQSAR to influenza neuraminidase inhibitor complexes and compare three methods of feature exaction (residue-based, group-based, and atom-based). The results show that the group-based method has better prediction. The important interactions are found in this model, and some suggestions are given to design new inhibitors. In residue-based method, a residue may contain useful information and noise. It will reduce the accuracy of prediction. In atom-based method, there are too many features in training, and it is easily over fitting. Therefore, group-based method is useful to COMBINE model. But we just divide a residue to main chain and side chain as a group-based method. In future, it is necessary to research how to define a “group”. A good definition of group-based method will improve the COMBINE model.

計畫成果自評

We developed a QSAR methodology associating molecular docking and feature selection with PLS. The feature of our model generates from the interaction energies of docked results, named as protein-ligand interaction profile and is extracted as atom-based, group-based and residue-based terms. We applied our QSAR methodology to build inhibitory models of neuraminidase, glycogen phosphorylase b and cyclooxygenase-2. Our results also compare to published references and our performances show more prediction power than other models. In the recently future, we will make more efforts to improving our methodology and combine virtual screen to create high through-put prediction environment.

參考文獻


