International Journal of Control and Automation
Vol. 9, No. 10 (2016), pp.123-134
http//dx.doi.org/10.14257/ijca.2016.9.10.11

GSVM-Based Proteochemometrics Modeling for Prediction of Kinase-inhibitor Interaction

Yiqi Wang[1], Qingfeng Chen[2*], Chaohong Wang[3*] and Yan Liang[4]

[1]*School of Computer, Electronic and Information, Guangxi University, Nanning, 530004, China*
[2]*School of Computer, Electronic and Information, Guangxi University, Nanning, 530004, China and State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, Guangxi University, China*
[3]*School of Automation and Information Engineering, Qingdao University of Science and Technology, China, 266061*
[4]*School of Computer, Electronic and Information, Guangxi University, Nanning, 530004, China*
[1]*yiqi_wong@163.com*, [2]*qingfeng@gxu.edu.cn*, [3]*wangchaohongqd@163.com*, [4]*419471767@qq.com*

Abstract

Life activity is closely related to the dynamic change of protein. Protein Phosphorylation is one of the most important proc GSVM-Based Proteochemometrics Modeling (PCM) for Prediction of Kinase-inhibitor Interaction within the protein modification after translation. It is found that more than 30% proteins can be phosphorylated. Abnormal protein kinases can lead to diverse diseases, such as cancers. Kinase inhibition is an effective method for disease treatment. However, some inhibitors are able to interact with several kinases that hidden but interesting kinase/inhibitor relationships may be included. Use of multi-targeted mining that select inhibitors act on a group of kinases increases the chance to achieve clinical antitumor activity. Proteochemometrics is a novel technology to predict inhibitor-kinase interactions from the chemical properties of kinase inhibitors which can help design more selective treatment and show better curative effect and low toxicity. This article uses a novel machine learning method called granular support vector machines (GSVM) to correlate the descriptors of kinase inhibitors and kinases to the interaction activities. GSVM develops on the basis of statistical learning theory and granular computing theory and thus provides an interesting new mechanism to address complex classification problems. Compared with other algorithms, GSVM gets better predictive abilities whose q^2=0.89.

Keywords: *Kinase inhibitors; Kinase/inhibitor inference; Proteochemometrics; GSVM*

1. Introduction

Protein kinome plays a crucial role in cellular responses or other cellular processes, including metabolic pathways, cell growth, differentiation, survival, and apoptosis. Currently, 518 genes that can produce protein kinases have been identified in the human genome. Due to kinases' overexpression or mutations, these abnormal kinases may lead to development of many serious diseases, such as cancer, diabetes, inflammatory, *etc* [1]. In order to treat these diseases, kinase inhibitors can be employed to disclose the biological consequences of the inactivation of their targets. Generally, kinase inhibitors are ATP-mimetic compounds. According to the ATP binding sites, inhibitors can be divided into

* Corresponding Author

ISSN: 2005-4297 IJCA

five types (type I, II, III, IV, and V). If inhibitors occupy directly the ATP binding sites, they belong to type I class, then if inhibitors target the ATP binding sites but extend also to an allosteric pocket adjacent to that it will belong to type II class, the other ones with non-ATP-mimetic inhibitor classes belong to type III, IV, and V [2]. However, there are still some disadvantages for the ATP-binding site classification methods, the main problem faced is to design a drug selective for only one kinase at a time. Thus, the problem on how to achieve specificity in kinase inhibition becomes the major challenge in the development of kinase inhibitors.

There have been a number of methods developed to detect the inhibitors of dysfunctional kinases by biological experiments, such as quantitative structure-activity relationship (QSAR). However, in QSAR, it considers only properties of compunds that make the interaction of ligand groups become a single target which makes this algorithm become low-throughput [3]. LBVS (ligand-based virtual screening), which is a relatively new algorithm, is regarded as the improvement of QSAR. It can be implemented in large-scale data sets and the main procedure for LBVS is shown in Figure1. The current method is aimed to derive a simpler and sturdier approach and applied to large-scale data sets. As to multi-target mining procedures, the most general approach is proteochemometric modeling (PCM). It uses inhibition data from groups of proteins tested against groups of inhibitors based on numerical descriptors, and seeks possible non-linear relationships. In terms of large-scale data sets, thanks to the emerging of chemical genomics makes exploration of novel bioactive molecules become possible, it helps many researchers to collect large-scale data sets of inhibitor-kinase pairs.

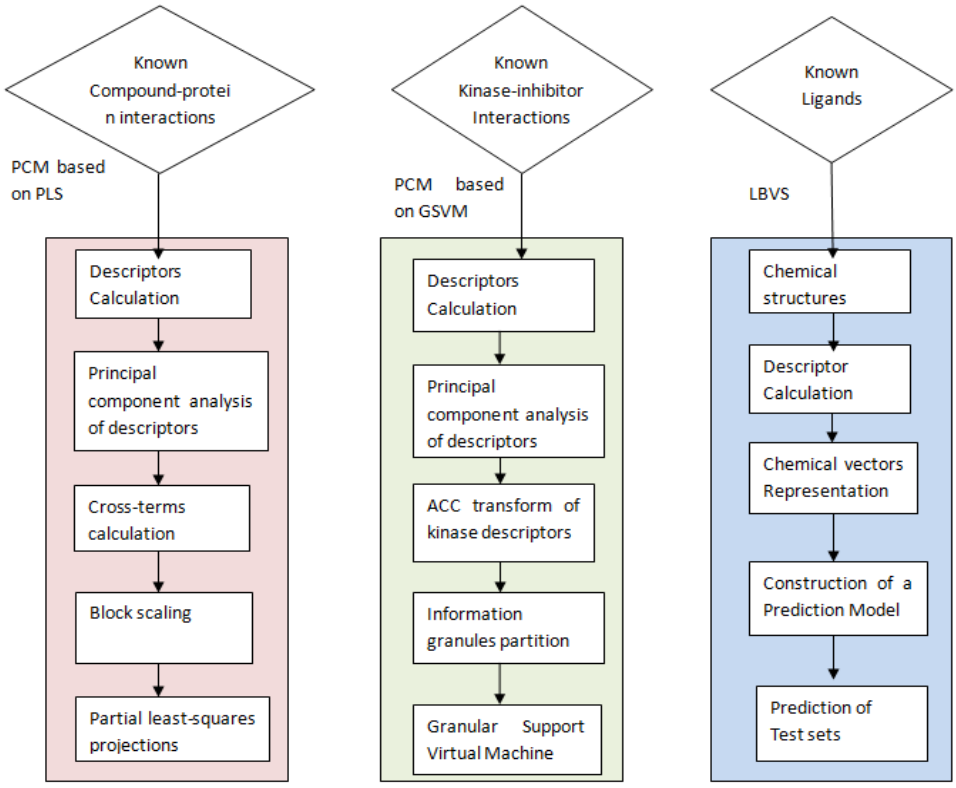

Figure 1. Comparison of the Strategies Used for Three Algorithms (Traditional PCM, GSVM Based PCM and Ligand-Based Virtual Screening (LBVS))

Proteochemometric modeling (PCM), whose key step is to search the similarity of a group of inhibitors and kinases, can be considered as ligand-target interaction space to some extent. Compared with QSAR, PCM can not only describe various inference

between a series of compounds and a series of targets but also work on single target. Further, PCM can also connect neighboring QSAR data sets makes it quite similar to inductive learning. Contrary to other strategies with numbers of models to predict numbers of outputs, there is only a single model extracting a single output variable of the interaction in PCM. For example the affinity is very diverse between different inhibitors or kinases but we can still extract a statistically solid model. Occasionally, it is unavailable to find useful 3D information from inhibitors or 3D strategy is unreliable, PCM still can be applied in these situations [4].

The reasons for many promising kinase inhibitors are abandoned because of the lack of clinical efficacy due to toxicity [5]. In modern research, people try to extract multi-targeted compounds that can selectively inhibit a specific group of kinases might increase the chance to achieve clinical antitumor activity [6]. Unfortunately, multiple structurally different compounds have been shown to bind the same protein or express similar biological activities [7]. In general, the PCM strategies need descriptors extracted from the two sets of interacting entities then create models to predict activities of all untested inhibitor-kinase combinations. Thus extracting the optimal descriptors for distinguishment is one of the important tasks in our modeling algorithm. In terms of descriptors extraction, both inhibitors and kinases need to be described. In the description of the inhibitor compounds, we applied chemical descriptors that contain some information unique in other descriptors. It helps us to improve the specificity of inhibitor descriptors. As to the description of the kinase, alignment-independent method with transformation is adopted instead of alignment-dependent method to avoid the alignment errors.

Apart from statistical methods, both linear machine learning techniques and non-linear machine learning techniques can be and have been used in PCM. Such as, Partial Least Squares (PLS) has been created on a large number of PCM models and specifies the relationship between an output variable Y and a set of predictor variables Xi [8]; Naive Bayesian (NB) shows good interpretability when modeling the datasets [9]; Support Vector Machines (SVM) has been proven to be very robust of modeling data sets [10]; Neural Net (NN) is known for their ability to handle complex inputoutput relationships [11]; and Decision Tree (DT) makes this algorithm highly interpretable [12]. However, there are different drawbacks on these algorithms, e.g. it is needful for the leaner methods PLS and NB to produce cross-terms which is time-consuming, then the SVM is low interpretable, while the descriptors in NN are high dimensional. Hence all currently machine learning methods that have been used in PCM have both advantages and disadvantages, we still need to find a better approach used in PCM.

Although there is low-interpretability when using SVM, it still achieves the best prediction performance than others' as the data shown in literature. Thus SVM has been employed in many PCM algorithms. Simultaneously, Granular computing is a new theory currently in the field of intelligent information processing to simulate of human thinking and to solve the problem of large-scale complex. Combine both effectively, may make SVM become a good machine learning method to be applied to solve practical problems better. A new learning model called granular support vector machines (GSVM) has been proposed by Yuchun Tang in 2005. GSVM systematically and formally combines the principles from statistical learning theory and granular computing theory and thus provides an interesting new mechanism to address complex classification problems. It works by building a sequence of information granules and then building support vector machines (SVM) in some of these information granules on demand. A good granulation method to find suitable granules is crucial for modeling a GSVM with good performance. In this paper, we also propose an association rules-based granulation method. For the granules induced by association rules with high confidence and significant support, we leave them because of significant effect on simplifying the classification task. For every other granule, a SVM is modeled to discriminate the corresponding data. In this way, a complex classification problem is divided into multiple smaller problems that the learning

task is simplified. Due to most general machine learning algorithm in PCM [13] is PLS, we compared GSVM based PCM modeling framework with that of PLS whose main steps comparison between our methods and others are shown in Figure1.

In GSVM based PCM modeling algorithm, the first task is kinases and inhibitors description. We have experimented two types of inhibitor descriptors, chemical descriptors and fragment descriptors, where the chemical descriptors whose predictive abilities (q2=0.81) showed better prediction performance. Further, three types of kinase descriptors, alignment-based, alignment-independent and ACC transform descriptors are compared that the ACC-descriptors performed the best in large-scale data sets. GSVM based PCM model is compared with the other three algorithms. The results demonstrate the performance of our methods (q2=0.89) is better than the other three machine learning algorithm in PCM.

2. Methods

2.1. Data Sets

We used the dataset published by GVK Biosciences kinase inhibitor database. All the data sets was provided by literature [14]. There are about 15,616 inhibitor-kinase pairs (including 143 kinases with corresponding 8830 inhibitors). However, the data downloading in GVK Biosciences kinase inhibitor database is rechargeable, thus we collect the data set from Yabuuchi *et al* [14]. There are two parts of data: Protein and Compound. In terms of protein (kinase), we have used two columns. The first variable is SwissProt_ID which can be used for kinase feature acquisition, the other one is kinase subfamily which used for kinase group partition. However, as to the inhibitor compounds, the column of SMILES format data is the most important part that contributes to descriptors extraction.

2.2. Inhibitor and Kinase Descriptors Extraction

The structures of kinase inhibitors were drawn by ChemBioDraw of ChemBioOffice software and converted to 3D structure by the ChemBio3D. Then run MOPAC on ChemBio3D to calculate the dipole moment and polarizability. In this way, the inhibitors' descriptors can be calculated. Two applications (DragonX and AFGEN) with different operating principles are used for inhibitors descriptor detection that the chemical descriptors and fragment descriptors can be extracted simultaneously.

The inhibitors' chemical descriptors calculation applied the E-DRAGON1.0 (http://www.vcclab.org/lab/edragon/). E-dragon is the electronic remote version of the well-known software DRAGON5.4, which is the general application for calculation of compound descriptors. The DRAGON software can provide more than 1,600 molecular descriptors which are divided into 20 logical blocks. Then chose the following descriptor blocks including constitutional descriptors, groups and atom-centered fragments, geometrical descriptors, counts of functional, charge and aromaticity indices, empirical descriptors, edge adjacency indices and molecular properties. When pairwise r2 between the two descriptors were exceeded 0.9, the one got the lowest correlation with other descriptors will be retained [15]. In this way, 169 molecular descriptors were obtained for each inhibitor for the modelling.

The other fragment descriptors were extracted according to AFGEN. It is a program that takes a set of chemical compounds as input and generates their vector-space representation based on the set of fragment-based descriptors [16]. The descriptor space that consists of graph fragments can have three different types of topologies: paths (PF), acyclic subgraphs (AF), and arbitrary topology subgraphs (GF). In this way, 2715 fragment descriptors for every inhibitor were gained.

It reveals that all the descriptors can be denoted in matrix, thus they must be mean centred and scaled to unit variance prior to use in modelling, using Eqn1.

$$\frac{1}{N_b^{1/2}} \tag{1}$$

Nb represents the quantity of descriptors in block b. It is a formula to scale one block of descriptors. Further, after modeling the PCM, a biased model may be created. In order to prevent this phenomenon, block scaling should be employed. Scaling the descriptors is able to make descriptors compatible.

Kinase sequences were downloaded from the Swiss-Prot database (http://www.ebi.ac.uk/swissprot). There are two types of kinase descriptors: alignment based and alignment independent. There are both advantages and disadvantages between this two methods. In terms of alignment based descriptors, the descriptors are obtained according to the conserved amino acid positions [17]. However, if the kinases sequences aligned wrongly, the descriptor extraction results will be inaccurate and there are also alignment gap(sequence stretches that located far from the ATP binding site) existed. However, when using alignment independent descriptors, a large number of descriptors may be produced that the complexity for algorithm may increase seriously. For example, PROFEAT Webserver [18], a common application used for alignment independent descriptors extraction contains 1497 features. Thus, principal component analyze(PCA) is needed to compress a large input matrix which will be introduced in next section.According to PCA process, the Z-scales introduced by Sandberg [21] are received to describe the properties of the amino acids.

Although z-scales are useful and convenient for kinase descriptors extraction, the results will be thwarted if sequences are aligned wrongly. Therefore, auto- and cross-covariance (ACC) transform method can be used to avoid the alignment step and transform sequence descriptions directly into uniform matrices [20]. This method can describe changes in some properties or some property combinations over sequence stretches of different lengths. The equations are as follows:

$$AC_{z,lag} = \sum_{i=1,2\cdots,N-lag} \frac{V_{z,i} * V_{z,i+lag}}{N-lag} \tag{2}$$

$$CC_{z_a \neq z_b,lag} = \sum_{i=1,2\cdots N-lag} \frac{V_{z_a,i} * V_{z_b,i+lag}}{N-lag} \tag{3}$$

where AC represents auto-covariances of the same z-scale and CC represents the cross-covariances of different z-scales, and where z = 1, 2, 3 (Z is the number of z-scales), i=1, 2, ..., N-lag (i is the amino acid position in the sequence and N the total number of amino acids), lag=1, 2, ..., L (L is the maximum lag, *i.e.* the longest sequence stretch used, which can be up to the length of the shortest sequence in the dataset), and V is the z-scale value. The total value of ACC terms depend on the chosen of maximum lags. In our experiment, the maximum lags are set to 10, 20, 35, and 50.

2.3. Descriptors Principle Component Analysis (PCA)

As can we seen from the descriptor results, the descriptor groups can be seen as a matrix. The block numbers can be referred to as the rows and the values for corresponding descriptors can be referred to as the columns of the matrix. It is assumed that the quantity of descriptors is n with p variables in each sample that an $n \times p$ matrix can be constructed [20].

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{4}$$

When there is a high p value, it is very difficult to process data in high dimensional space. Hence dimensionality reduction can be used to make less variables instead of original variables indicators. Further, the new variables should express the original information as much as possible and isolated each other. We use $x_1, x_2, \cdots, x_p$ to denote

the original features, while $z_1, z_2, \cdots, z_m (m \leq p)$ can be seen as the principal component for xi ($1 \leq i \leq p$). The formula between these two feature groups is:

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases} \quad (5)$$

where $l_{ij}\ (i = 1,2,\cdots,m;\ j = 1,2,\cdots,p)$ is the load of principal component $z_i (i = 1,2,\cdots,m)$ under original variables $x_j (j = 1,2,\cdots,p)$. Load lij is feature vector among the greater characteristic values under matrix m.

(1) Solve characteristic equations $|\omega I - R| = 0$, then the eigenvalues can be extracted which are ordered by size. $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_p \geq 0$

(2) Calculate the feature vector $l_i (i = 1,2,\cdots,p)$, corresponding to eigenvalues ω_i. It is notable that the $\|l_i\| = 1$, in other words, $\sum_{j=1}^{p} l_{ij}^2 = 1$ denote the jth component li in vector lij.

(3) Compute the contribution and accumulative contribution rate according to Eqn6 and Eqn7.

$$\frac{\omega_i}{\sum_{k=1}^{p}\omega_k} (i = 1,2,\cdots,p) \quad (6)$$

$$\frac{\sum_{k=1}^{i}\omega_k}{\sum_{k=1}^{p}\omega_k} (i = 1,2,\cdots,p) \quad (7)$$

In general, we choose the eigenvalues $\omega_1 \geq \omega_2 \geq \cdots \geq \omega_p \geq 0$ whose accumulative contribution rate occupied more than 85% as corresponding $z_1, z_2, \cdots, z_m\ (m \leq p)$ principal components. We use the z-scales, z1-z3, derived by Sandberg *et al*.[21] to describe the aligned positions where then described by amino acid physico-chemical properties encapsulated in the three. Z-scales are quantitative descriptors obtained from principal component analysis for 26 physico-chemical properties of corresponding amino acids. There are a total of five z-scales (z1-z5), the z-scales are computed according to principle component analysis (PCA) which will be introduced in next section. In the PCM modeling, most descriptors need mean-centered and uncorrelated each other, while adopting z-scales description can obtain such descriptors directly. This three z-scales can be interpreted as eflecting hydrophobicity (z1), steric properties (z2) and polarity (z3)[22]. In this way, kinase sequences properties were encoded by $141 \times 3 = 423$ descriptors.

2.4. GSVM and PCM Model Validation

We adopted the Association Rules based Granular SVM (AR-CSVM). It is a machine learning to compute appropriate size and purity in granularity level by defining the support degree and relational two metrics. And on the basis the samples are mapped from the original space to high-dimensional space which make both granule division and SVM training are implemented in high dimensional space. Thus the consistency of the data distribution can be ensured that the generalization ability of the algorithm is improved. Further, the basic approach for GSVM is using association rule mining method to divide sample space into multiple granules(subspace) including a serial of pure granules(all samples belong to the positive or negative class) and a serial of impure granules(samples with both positive and negative class). Remove the pure granules with highest correlation at every turn, then train the mixed granules using SVM, loop execution until the classification accuracy is no longer changed.

These pairs can be seen as a binary classification problem with continuous features, which the experiment kinase-inhibitor pairs are used as pure granules while the other combination pairs can be set as mixed granules. Then the SVM calculations used LIBSVM toolbox (http://www.csie.ntu.edu.tw/~cjlin/). However, in terms of granule mining, the distance measurements need to be redefined. Assume the samples that input

to space is $x_i \in \{R^N\}, i = 1,2,\ldots n$, then the x_i map to a feature space H by nonlinear mapping function to obtain $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n)$, thus the dot product that input the space using Kernel function is expressed as Eqn8:

$$K(x_i, x_j) = (\Phi(x_i) \cdot \Phi(x_j)) \quad (8)$$

Euclidean distance in the kernel characteristic space can be expressed as Eqn9.

$$d_{H(x,y)} = \sqrt{\|\Phi(x) - \Phi(y)\|^2} = \sqrt{\Phi(x) \cdot \Phi(x) - 2\Phi(x)\Phi(y) + \Phi(y) \cdot \Phi(y)} \quad (9)$$

Due to the PCM model employs descriptors, the the number of variables may rise with the descriptor space expands. It may lead to both chance correlations and model overfitting. However, the key goal for us is to get a reliable estimate of the model quality and applicability. Thus, cross validation should be accommodated. There are three forms of cross validation, namely Leave-One-Out (LOO)) cross validation, n-fold cross validation and double loop cross validation [23]. Nowadays, the most popular validation approach belongs to n-fold cross validation. In our experiment, we chose the n to 5. In 5-fold validation the training set is divided into 5 equal amounts. Subsequently a model is trained on four of the subsets and used the remaining one to predict the activity of the data points. This process is repeated until all subsets have been left out of the training and the plots of these iterations are used to calculate q^2 [23]. The formula is:

$$q^2 = 1 - \frac{\sum_{n=1}^{N}(\gamma_{predicted,n} - \gamma_{measured,n})^2}{\sum_{n=1}^{N}(\gamma_{measured,n} - \bar{\gamma})^2} \quad (10)$$

Where γ represents the values under corresponding condition, and ($\bar{\gamma}$) is the average of the measured outcome values for the N objects in the dataset.A $q^2 > 0.4$ is generally considered acceptable for modelling biological data [23]. q^2 can be directly computed by software called Multivariate infometric analysis (S.r.l., www.miasrl.com) software.

3. Results and Discussion

3.1. Theoretical Framework for Algorithm

Then main step for our algorithm are shown in Figure1. There are five steps in our algorithm: data set collection, feature extraction, ACC transformation, data scaling, machine learning algorithm construction and testing data prediction. Thereinto, step1, We used the dataset published by GVK Biosciences kinase inhibitor database. There are about 15,616 inhibitor-kinase pairs (including 143 kinases with corresponding 8830 inhibitors). In step2, Inhibitor structures and kinase sequences were converted into numerical descriptors for PCM. In terms of kinase inhibitors description, were drawn by ChemBioDraw of ChemBioOffice software and converted to 3D by the ChemBio3D unit. Inhibitor structures were then characterized by various molecular descriptors using Dragon5 software. All descriptors were mean centred and scaled to unit variance prior to use in modeling. Then we go to the next step for kinase descriptors extraction. In step3, we use auto- and cross-covariance (ACC) transform method to avoid the alignment step and transform sequence descriptions directly into uniform matrices. In step4, we used block scaling to construct kinase-inhibitor vectors. In step4, the modeling learning techniques should be used in PCM, in our experiment, concatenated vectors for positive samples and negative samples were input into a Granular SVM (GSVM). We can split the whole feature space into a set of subspaces (information granules) and then build a SVM for some mixed ones of the subspaces which is implemented in next step. In final step, a serial of pure granule (all samples belong to the positive or negative class) and a serial of impure granules (samples with both positive and negative class) are obtained. Remove the pure granule with highest correlation at every turn, and then train the impure granule using SVM.

3.2. Theoretical Framework for Algorithm

The features of inhibitors can be described in different ways. We adopted two types of descriptors for comparison with the help of conventional q2 parameter to assess the predictive abilities of the model. The descriptor results are given in different forms between these two applications. In terms of the descriptors given by DRAGON6, the attributes are denoted directly by its descriptor classes, such as geometrical descriptors, charge and aromaticity indices, empirical descriptors, and molecular properties *etc*. While the descriptors in AFGEN2.0 are labeled in fragment-ID(sorted in increasing fragment-ID order). The kinases group whose subfamily is STE is chosen as the training set. These two kinds of descriptors are evaluated on training set. And the prediction performance comparison is displayed in Table 1. As can we see from this table, the chemical descriptors show better performance in kinase-inhibitor interaction inference whose q2 is up to 0.81. Thus the chemical descriptors are chosen for inhibitor description in the following steps.

Table 1. Comparison between Two Kinds of Inhibitor Descriptors

Application	Descriptors	Descriptors Quantity	q2
AFGEN2.0	Fragment descriptors	2715	0.69
DRAGON	Chemical descriptors	169	0.81

3.3. Optimal Lags Extraction for ACC Transform Method

In the last section, we used Eqn3 to identify the optimal complexity of the ACC descriptions. However, it is shown that the variable lag L determine the value of ACC transformation. In other words, we need to find the maximum lag L, up to which descriptors contribute to improved separation of kinase groups. As described in Methods, covariances over long distances are less helpful in finding physico-chemical similarities in related protein sequences due to the differences in the length of segments that connect their functional units. Use of very many ACC terms with large lags may then give rise to chance correlations, deteriorating the resolution of any mathematical models created from them.In the ACC transform method introduced in previous section, we had set the maximum lags as 10, 20, 35 and 50. Subsequently, the method for selection of optimal lags will be tested in our experiment. The kinases in AGC group were chosen for training set. And the comparison of q2 under these four lags are displayed in Figure2. As can we see from this figure, the q2 bottomed at lag=10 whose value is 0.67; after the lag increased to 20, q2 got a little improvement to 0.75; the q2 peaked at 0.87 when the lag become 35. Finally, increasing the lag to 50, the q2 started falling to 0.81. Hence we chose maximum lag 35 as the optimal lag.

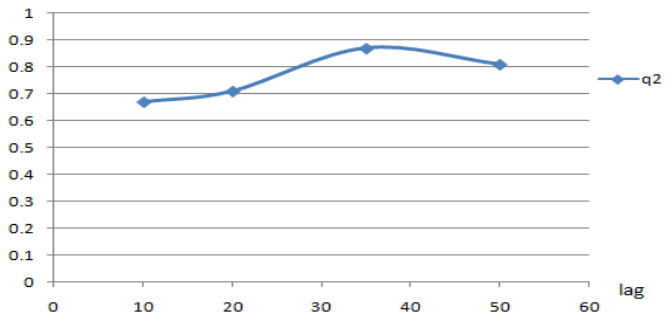

Figure 2. Prediction Performances under Different Values of Lags

In the experiment, we had adopted two kinds of method:alignment-based and alignment-independent kinase descriptors[19]. In terms of alignment-based, the auto- and cross-covariance (ACC) transform is also employed. We have used these three methods, and the comparison of these are described in Table 2. Among these three approaches, alignment-independent kinase description is the easiest one to operate, while excessive amounts of descriptors increase the risk of chance correlation models. While the alignment-based kinase description method is the most popular method in PCM modeling besides the risk if sequences are aligned wrongly. Thus we adopted the ACCs transform method that can avoid alignment step and obtain uniform matrices. The comparison for performance of these three kinds of descriptors in PCM modeling are shown in Figure 3. As can we see from this figure, performance for these three methods all exceeded 0.75 which are 0.76, 0.86, 0.87 respectively. The performance for using ACCs is not very extrude because the raw data we used is high quality data set.

Table 2. Comparison of Descriptors Extraction Methods We Used

Method	Main Step	Advantage	Disadvantage
Alignment-independent	Using PROFEAT web server	Easy to extract	Mass of description variables
Alignment-based	Aligned positions description	Compress input matrix by PCA	Alignment errors can influence results
ACCs	auto- and cross-covariance (ACC) transform	avoid the alignment step	lags may too long

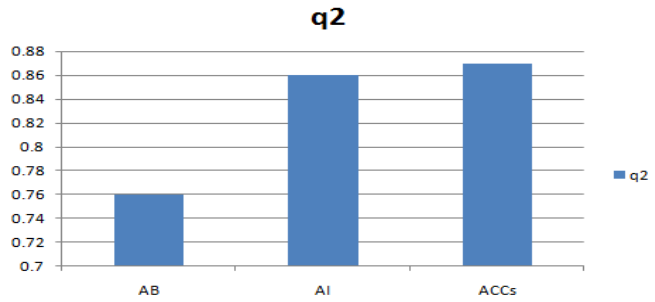

Figure 3. Comparison of Kinse Descriptors

3.4. GSVM Predict Novel Kinase-Inhibitor Associations

When using GSVM, the sample space are divided into multiple granules(subspaces), after pure granules(samples that belong to the effective classes) and mixed granules(samples that belong to both effective and negative classes) extraction. The mixed granules are training in the SVM according to delete the pure granule with highest correlation degree successively, looping through this process until the classification accuracy is no longer improved. The formula to calculate classification accuracy is Eqn11. Where where TP, FP, TN, and FN represent true positive, false positive, true negative and false negative, respectively. Further, we have listed kinase-inhibitor interactions in top ten classification accuracies in Table 3.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (11)$$

Table 3. Top 10 Classified Kinase Inhibitor Interactions

Protein	Inhibitor(GVK_ID)	Classification Accuracy
YES_HUMAN	14973	0.986
STK4_HUMAN	192492	0.982
MK03_HUMAN	233053	0.981
PAK4_HUMAN	212360	0.972
HIPK3_HUMAN	56743	0.966
MP2K2_HUMAN	56743	0.965
PAK2_HUMAN	212360	0.960
BMX_HUMAN	218610	0.954
CDK6_HUMAN	233432	0.949
PASK_HUMAN	64540	0.943

3.5. Comparison between Different Algorithms

There are a number of machine learning method have been used in PCM model as exhibited in last section. In this section, we compare the performance between these software. Besides the data we computed in our experiment, the values of prediction performances are quoted from [22], including k-NN, SVM and PLS. While the performance for random forest (RF) are downloaded from literature[20]. These four machine learning methods are the most general approach in PCM, and it has been proved that all these three algorithms achieved high-performance. We have compared these algorithms in two sides. The positive and negative consequences of these algorithms are shown in Table 4 while the prediction attributes are exhibited in Figure 4. As the short summary shown in Table 4, there are both merits and drawbacks existed in these algorithms. We need to choose the appropriate method for corresponding dataset. The kinase-inhibitor pairs whose kinases' subfamily is TKL are used as training set. As the comparative results shown in Figure 4, besides property of RF(0.90) is slightly higher, the prediction ability for GSVM is preferable with 0.89 compared with SVM, k-NN and PLS with 0.83, 0.82 and 0.79 respectively.

Table 4. Advantages and Disadvantages of Different Algorithms

Method	Linear?	Advantage	Disadvantage
SVM	Non-linear	Robust on complex data	Low-interpretability
k-NN	Non-linear	Appropriate on complex data	High Dimensionality
PLS	Linear	High-interpretability	Needs Cross-Terms
GSVM	Non-linear	Performs well on large-scale data	Need information granules partition

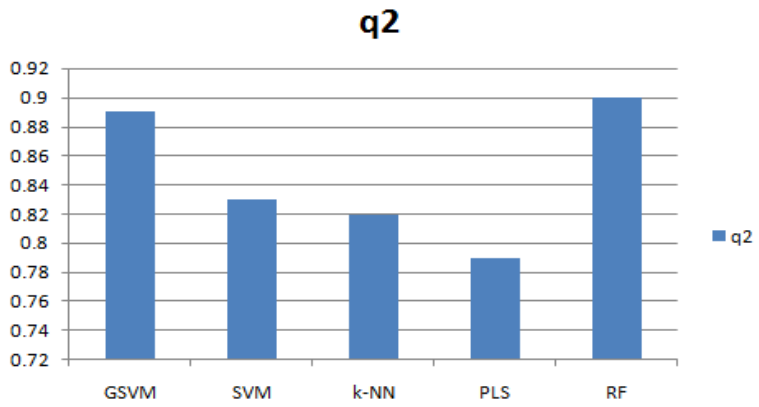

Figure 4. Performance Using Different Methods

4. Conclusion

The analysis of kinase-inhibitor interaction has become one of the most difficult and important challenges in bioactive molecules and medicine. PCM is a relatively new technique that, by including target descriptors in addition to ligand descriptors. It is an improvement for modeling of data sets that could previously only be modeled separately using conventional QSAR based techniques. At the same time, GSVM provides a new mechanism to address complex classification problems, which are common in medical or biological information processing applications. The modeling method for a GSVM proposed here is just one step into this interesting research topic. Thus the GSVM approach presented here can satisfy the need for original and effective computational methods to unravel the rich and complex kinase/inhibitor relationships systematically measured in inhibition profiling panels, which can have significant implications in understanding the reasons of the inhibition, helping in the rational design of bioactive molecules, and can be used for the in silico prediction of inhibition for those neglected kinases for which no systematic analysis has been carried yet, and for the selection of inhibitors with desired promiscuity. Additionally, a better understanding of the kinase determinants of inhibition can help in apprehending the different response of individual patients to treatment, such as inhibitor resistance due to specific mutations, moving toward a more personalized treatment.

References

[1] C. Madhusudan S and T.S. Ganesan, "Tyrosine kinase inhibitors in cancer therapy", Clin Biochem, vol. 37, **(2004)**, pp. 618-635.
[2] F. Ferr*, A. Palmeri and M. Helmer-Citterich, "Computational methods for analysis and inference of kinase/inhibitor relationships", Genetics, vol. 00196, **(2014)**.
[3] D. M. Goldstein, N. S. Gray and P. P. Zarrinkar, "High-throughput kinase profiling as a platform for drug discovery", Nat. Rev. Drug Discov., vol. 7, 391397, doi: 10.1038/nrd2541, **(2008)**.
[4] H. Yabuuchi, S. Niijima, H. Takematsu and T. Ida "Analysis of multiple compound Protein interactions reveals novel bioactive molecules", Molecular Systems Biology, vol. 7, **(2011)**, pp. 472-484.
[5] N. Wale, I. Watson and G. Karypis, **"**Comparison of Descriptor Spaces for Chemical Compound Retrieval and Classification", Knowledge and Information Systems, vol. 408, **(2007)**, pp. 297315.
[6] J. Bain, L. Plater, M. Elliott, N. Shpiro, C. J. Hastie and H. McLauchlan, "The selectivity of protein kinase inhibitors: a further update.Biochem", J., doi: 10.1042/BJ20070797, vol. 408, **(2007)**, pp. 297315.
[7] S. Niijima, A. Shiraishi and Y. Okuno, "Dissecting kinase profiling data to predict activity and understand cross-reactivity of kinase inhibitors", J. Chem. Inf. Model, doi: 10.1021/ci200607, vol. 52, **(2012)**, pp. 901312.
[8] D. S. Cao, G. H. Zhou, S. Liu, L. X. Zhang, Q. S. Xu and M. He, "Large-scale prediction of human kinase/inhibitor interactions using protein sequences and molecular topological structures", Anal. Chim. Acta, doi: 10.1016/j.aca.2013.07.003, vol. 792, **(2013)**, pp. 1038.

[9] S. C.Scherer and S. M. Muskal, "Kinome-wide activity modeling from diverse public high-quality data sets", J. Chem. Inf. Model, doi:10.1021/ci300403k, vol. 53, **(2013)**, pp. 278.
[10] G. Scapin, "Protein kinase inhibition: different approaches to selective inhibitor design", Curr Drug Targets, vol. 7, **(2006)**, pp. 1443-1454.
[11] A. Kamb, S.Wee and C. Lengauer, "Why is cancer drug discovery so difficult?", Nat Rev Drug Discov, vol. 6, **(2007)**, pp. 115-120.
[12] D.W. Young, A. Bender, J.Hoyt, E. McWhinnie, G.W. Chirn, C.Y. Tao, J.A. Tallarico, M. Labow, J.L. Jenkins, T.J. Mitchison and Y. Feng, "Integrating high-content screening and ligand-target prediction to identify mechanism of action", Nat Chem Biol, vol. 4, pp. 59, **(2008)**.
[13] M. Lapins and J. E. Wikberg, "Research article Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques", BMC Bioinformatics, vol. 11, **(2010)**, pp. 339-354.
[14] M. Lapinsh, P. Prusis, S. Uhlen and J. E. S. Wikberg, "Improved approach for proteochemometrics modeling: application to organic compoundmine G protein-coupled receptor interactions", Bioinformatics, vol. 21, **(2005)**,pp. 4289-4296.
[15] N. Weill and D. Rognan, "J. Chem. Development and Validation of a Novel Protein Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands", Inf. Model, vol. 49, **(2009)**, pp. 1049：1062.
[16] H. Geppert, J. Humrich, D. Stumpfe, T. Gaertner and J. Bajorath, "J. Chem Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors", Inf. Model, vol. 49, . **(2009)**, pp. 767:779
[17] M. Fernandez, L. Fernandez, J. Caballero, J. I. Abreu and G. Reyes, "Chem. Proteochemometric modeling of the inhibition complexes of matrix metalloproteinases with N-hydroxy-2-(phenylsulfonyl)amino acetamide derivatives using topological autocorrelation interaction matrix and model ensemble averaging", Biol. Drug Des, vol. 72, no. 65, **(2008)**, pp. 7 8.
[18] J. P. Gerard van Westen, J. K. Wegner, A.P. IJzerman, H. W. T. van Vlijmen and A. Bender, "Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets", The Royal Society of Chemistry, vol. 2, **(2010)**, pp. 16.
[19] Z.R. Li, H.H. Lin L.Y. Han, L. Jiang, X. Chen and Y.Z. Chen, "PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence", Nucleic Acids Res34:W32W 37, **(2006)**.
[20] J.M. Baldwin, "An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors", J. Mol. Biol., vol. 272, no. 144, **(1997)**, pp. 164.
[21] Y. Tang, B. Jin and Y.-Q. Zhang, "Granular support vector machines with association rules mining for protein homology prediction. Artifical Intelligence in Medicine", vol. 35, no. 1, **(2005)**, pp. 121-134.
[22] E. Freyhult, P. Prusis, M. Lapinsh, J. ES Wikberg, V. Moulton and M. G. Gustafsson, "Unbiased descriptor and parameter selection confirms the potential of proteochemometric modeling", BMC Bioinformatics, vol. 2005, **(2005)**, pp. 6.
[23] A. Golbraikh and A. Tropsha, "Beware of q2! J Mol Graph Model", vol. 2002, no. 20, **(2002)**, pp. 269-276.

Author

Wang Yiqi, she received the B.Sc. degree in computer science from Guangxi University, China, in 2013. She is now postgraduate studying in Guangxi University. Her research interests include data mining and bioinformatics, especially prediction algorithm and semi-supervised learning algorithm.

