COMPARISON OF CHEMICAL DESCRIPTORS FOR PROTEIN–CHEMICAL INTERACTION PREDICTION

J. Zhang* and J. Huan**

Abstract

Predicting protein–chemical interaction has been an important and challenging task in the bioinformatics community, and there are many related applications in biomedical research, including QSAR modelling and novel lead discovery. A fundamental hypothesis for predicting protein–chemical interaction is that chemical compounds sharing chemical similarity should also share protein target profiles, and the critical question is hence how to measure the distance (or similarity) between two chemicals. An increasing number of chemical descriptors have been invented in the past decades. As chemical descriptors play a critical role in predicting protein–chemical interaction, it is of great importance to compare chemical descriptors and evaluate their performance in such predictions. In this paper, we reported our case study on comparing the performance of DRAGON descriptors, the frequent subgraph-based descriptors (FFSM), and the signature molecular descriptor on predicting protein–chemical interaction using support vector machines over a large number of data sets. Our experiments demonstrated that FFSM and signature descriptors outperformed most DRAGON descriptor classes, and wisely selecting chemical descriptors will be beneficial for predicting protein–chemical interaction.

Key Words

Protein-chemical interaction, chemical descriptors, molecular graph, cross validation, support vector machines

1. Introduction

One of the critical steps in drug discovery is the identification of chemical compounds with desired and reproducible binding activity against a specific biomolecular target [1]. This has become a significant challenge in the early stage of drug discovery, since any new drug must not only produce the desired medical response to the disease, but should also minimize any side effects [2]. Understanding and predicting the interactions between target proteins and small molecules is hence of great importance in pharmaceutical industry. Our knowledge about the interactions between chemical space and biological space, however, is very limited. For instance, millions of chemicals have been deposited into the NCBI PubChem Databases, but only a very small fraction (less than 2%) have their target protein information linked [3]. In addition, experimental determination of potential protein–chemical interactions remains time-consuming and expensive. Therefore, new *in silico* methods capable of predicting potential protein–chemical interactions efficiently are needed.

There have been many methods that represented each chemical by a set of descriptors based on frequency, molecular properties, topological and geometric substructures [2, 4, 5], e.g., DRAGON descriptors [6], Daylight fingerprints [7], extended connectivity fingerprints (ECFP) [8], Maccs keys [7], cyclic patterns and trees [5], signature molecular descriptors [9] and frequent subgraph-based descrip- tors [10, 11]. All these methods have been well developed and have demonstrated their effectiveness and success in many experiments and applications. In addition, they can also be used to rapidly predict the physical, chemical, and biological properties of small molecules to screen large database and identify suitable drug candidates [12–14].

In general each such method consists of three components: (i) descriptor extraction and selection, (ii) predictive model selection, and (iii) model assessment. First, descriptor extraction methods are used to compute various descriptors for each chemical and convert the chemical to a vector, in which each component is a molecular descriptor, e.g., molecular weight, number of hydrogen bonds, and so on. Moreover, with a descriptor vector and the corresponding class label as a datum point, a set of such points is then divided into three disjoint sets: training, validation, and testing set. Model selection is needed to choose a model of optimal performance, which in practice means selecting best learning parameters from a small set of choices based on training and validation sets. Finally, we applied model assessment techniques to estimate the prediction error (generalization error) of the selected model on the unused testing set.

In the framework of binary classification described above, classifiers played an important role by taking de-

[*] Centre for Bioinformatics, 3000 Becker Drive, University of Kansas, Lawrence, Kansas 66047; e-mail: jtzhang@ku.edu

[**] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, Kansas 66045; e-mail: jhuan@ittc.ku.edu

(paper no. 210-1010)

scriptor vectors and class labels as input and generate prediction models as output. There are many widely used classifiers, e.g. support vector machines (SVM), K-nearest neighbors (KNN), neural network and random forest. As SVM is one of the most widely used kernel-based classifiers in binary classification and regression, we are mainly focusing on protein–chemical interaction prediction using the SVM classifier.

The challenging task then comes to selecting accurate descriptor extraction methods to compute a set of descriptors for each chemical. As different descriptors perform variously, they can significantly affect the resulting prediction accuracy. This is the first motivation for us to do a case study of the performance comparison of different descriptor classes. Our eventual goal is to discover some best performing descriptor sets for predicting protein–chemical interaction based on SVM classification. In this work, we carried out detailed performance comparisons among the 20 classes of DRAGON descriptors [6], the frequent subgraph-based descriptors [10], and the signature molecular descriptors [9] on 14 high-quality chemical data sets. Our results provide important insights on how to select chemical descriptors to achieve optimal prediction performance.

The rest of the paper is organized as follows: in Section 2, we present an overview of related work on currently used descriptor extraction methods. Section 3 provides background information about DRAGON descriptors, signature descriptors, graph representation of chemical chemicals, frequent subgraph-based descriptors, and SVM classification. Section 4 describes the data sets and experimental designs in detail. In Section 5, we present our experimental results and give discussions of the performance of various descriptor sets. Finally, we conclude this work with a summary and future plan in Section 6.

2. Related Work

There are two basic hypotheses for protein–chemical interaction prediction: (1) chemicals sharing chemical similarity should also share target proteins; (2) targets sharing similar ligands should share similar biological patterns, or binding sites. The classical paradigm here is as follows: if two chemicals (are considered very similar to each other in some way and we know one of them interacts with a protein, we would expect that it is very likely for the other to interacting with the same target (chemical) too. However, if few or no ligands are known for a target, e.g., orphan G-protein coupled receptors [15], information has to be learned from other related targets with known ligands. In this paradigm, classifiers learn from both proteins and chemicals simultaneously to predict if a pair of protein and chemical is interacting or not. On the other hand, when sufficient ligands of a given target protein are available, it is even more accurate to use only ligand information to make a prediction. In our study, for instance, signature descriptors could give prediction accuracy of up to 90%, but the accuracy decreased to less than 80% when it combined with some protein sequence descriptors [16]. Such strategies are very useful in finding diverse novel lead chemicals in drug discovery and development.

In this paper, we limited our scope on a comprehensive comparison of a few classes of widely used chemical descriptors, and expected to provide insights on their prediction performance. The key question of predicting if a chemical interacts with a given protein is how to measure the distance (or similarity) between two chemicals. As the quality of chemical descriptors that convert each chemical into vectors of numbers play a critical role in such predictions, it is of great importance to investigate which chemical descriptors perform better than the others.

There have been many previous studies on comparing the prediction performance of chemical descriptors. Hert *et al.* [17] compared a range of different 2D fingerprints for similarity-based virtual screening, and found that these fingerprints were notably more effective than fingerprints based on a fragment dictionary. They concluded that the combination of these fingerprints with data fusion based on similarity scores provides an effective virtual screening tool in lead discovery. Gedeck *et al.* [18] analyzed how the quality of QSAR predictions depended on the data sets and descriptor types, and they revealed that none of the descriptors was best for all data sets. Although 2D fragment based descriptors usually performed better than simpler descriptors based on augmented atom types, it was necessary to test them in each individual case.

In addition, Karypis *et al.* [2] introduced some new descriptor sets such as graph fragment based descriptors (GF), and conducted a comprehensive comparison of the performance of the newly developed descriptors with daylight fingerprints [7], extended connectivity fingerprints (ECFP) [8], Maccs keys [7], cyclic patterns and trees [5] in the context of SVM-based chemical compound classification and ranked retrieval. The goal of his work was to analyze what properties of descriptor spaces were important in providing effective representation for molecular graphs, and their experiments demonstrated that descriptor class ECFP and GF consistently and statistically outperformed previously developed all other descriptor sets.

However, some other widely used chemical descriptors were not covered in these previously studies. For instance, DRAGON [6] provided a collection of 20 widely used classes of chemical descriptors. Huan *et al.* [10] developed a fast frequent subgraph mining (FFSM) algorithm to generate frequent subgraph-based descriptors for chemicals. Faulon *et al.* [9] developed an algorithm of signature molecular descriptors for both protein sequences and chemicals. In this paper, we conducted a case study of the performance comparison between these 22 classes of descriptors based on SVM classification, and provided insights on selecting optimal chemical descriptors for predicting protein–chemical interaction.

3. Background

3.1 DRAGON Descriptors

DRAGON is a commercial software package developed by Milano Chemometrics and QSAR Research Group [6] for calculating molecular descriptors that can be used to evaluate molecular structure–activity or property relationships (QSAR/P), as well as for high-throughput virtual screening of chemical databases. Users need molecular structure files (SDF, SMILE, etc.) as input, and are given formatted output files. Although DRAGON can work on 2D structures, only much less descriptors can be calculated in this case. To make full use of DRAGON software, 3D optimized structures with all hydrogen atoms should be used.

DRAGON 5.4, which we used in this work, computes 1664 molecular descriptors that are divided into 20 descriptor sets (or logical blocks): constitutional descriptors (DR01), topological descriptors (DR02), walk and path counts (DR03), connectivity indices (DR04), information indices (DR05), 2D autocorrelations (DR06), edge adjacency indices (DR07), Burden eigenvalues (DR08), topological charge indices (DR09), eigenvalue-based indices (DR10), Randic molecular profiles (DR11), geometrical descriptors (DR12), RDF descriptors (DR13), 3D-MoRSE descriptors (DR14), WHIM descriptors (DR15), GETAWAY descriptors (DR16), functional group counts (DR17), atom-centered fragments (DR18), charge descriptors (DR19), and molecular properties (DR20). For more detailed introduction to DRAGON software and descriptors, refer to the help manual at its homepage (http://www.talete.mi.it/index.htm).

For instance, in a molecule with known molecular composition and atom connectivities, functional group counts (DR17) are simply defined as the number of specific functional groups, and atom-centered fragments (DR18) are defined as the number of specific atom types. For each atom-centered fragment, its frequency occurring in the chemical structure is counted, and so far 120 atom-centered fragments defined by Ghose and Crippen [19] are included. Finally, molecular properties (DR20) include a set of heterogeneous molecular descriptors describing physico-chemical and biological properties as well as some molecular characteristics, such as hydrophilic factor, octane–water partition coefficient, molar refractivity, etc. As the charge descriptors (DR19) are not available to many chemicals in our data sets, they will be skipped in our study and a combination of all other 19 descriptor sets (denoted as "DRAL") will be used instead.

3.2 Signature Molecular Descriptors

Faulon *et al.* [9] developed an algorithm of signature molecular descriptors by enumerating all molecular signatures with a given height from chemical structures. Specifically, the signature of a molecule is a vector whose components are counts of the number of occurrences of a particular atomic signature in the molecule. An atomic signature is a canonical representation of the subgraph surrounding a particular atom. This subgraph includes all atoms and bonds up to a predefined distance, called signature height, from a given atom. The optimal signature height for chemical is usually in range of 1–5.

To generate the signature lists of a chemical, a signature translation program named "translator" was downloaded from the homepage of Faulon *et al.* [9]. Given a chemical structure with a predefined signature height, a list of available signatures and their occurring frequencies can be generated very efficiently. The running time for generating the signatures of most chemical structures is only up to a minute. Given a data set with m chemicals, a list of available signatures for each chemical are generated, and then all distinct signatures present in all chemicals (the union of all signature lists, e.g., totally n signatures) can be obtained. Finally, each chemical will be associated with an n-dimensional vector, in which each component is the number of occurrence of each signature in the chemical.

3.3 Graph Representation of Chemical Structures

Chemical compounds have well-defined geometric structures that can be easily converted into a connected, labelled and undirected graph representation. Each chemical has a number of atoms represented as vertices and a number of bonds between atoms represented as edges in the molecular graph. Usually vertices are labelled with the atom element type(atomic symbol or number, e.g., carbon atoms are labelled with C or 6), and edges are labelled with the bond type (bond order or separate integers, use 1, 2, 3, 4 for single, double, triple, and aromatic bonds, respectively). Edges in a graph are undirected because chemical bonds have no associated directionality. Figure 1 shows an example of a chemical structure and the corresponding graph representation.

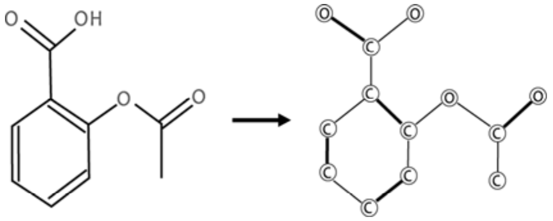


Figure 1. A chemical structure and the corresponding graph representation.

3.4 Frequent Subgraph-Based Descriptors

Frequent subgraph mining is widely studied since frequent subgraphs are believed to be related to some structural or functional motifs in chemical and biological structures. Huan *et al.* [10] developed a depth-first search algorithm for fast frequent subgraph mining (FFSM), which identified all connected subgraphs that occurs more frequently than a predefined frequency threshold σ called *support threshold* in a graph database. Each chemical compound is represented by a binary vector with length equal to the number of all mined subgraphs, and then each subgraph is mapped into a specific vector index. If a chemical compound contains a subgraph then the corresponding bit is set to one, otherwise it is set to zero. [10, 20]

By mining all frequent subgraphs from a chemical database, FFSM creates an n-dimensional descriptor vector for each chemical, and hence provides frequent subgraph-based descriptors. A potential disadvantage of this method is that it is unclear how to select a suitable value of the *support* σ for a given problem. A very high value will fail to discover important sub-structures whereas a very low value will result in combinatorial explosion of frequent subgraphs. In Fig. 2, a graph database with 3 graphs is in the top row, and the returned frequent subgraphs are listed in the bottom row with support = 2/3.

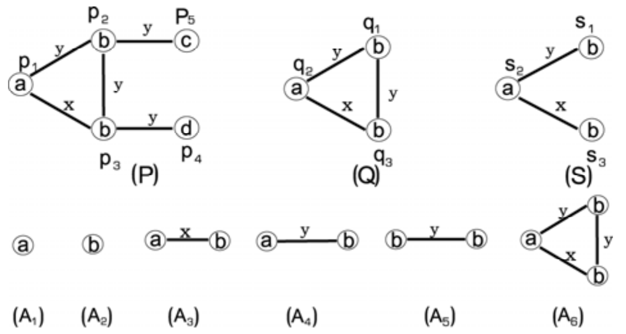


Figure 2. A graph database with 3 graphs in upper row, and frequent subgraphs returned by FFSM algorithm with support = 2/3 in lower row.

3.5 Support Vector Machines

As a supervised learning method widely used for classification and regression, support vector machines (SVM) [21] view input data as two sets of vectors in an n-dimensional space, each with different class labels. SVM constructs a separating hyperplane in that space by maximizing the margin between the two data sets. To calculate the margin, two additional parallel hyperplanes are constructed, one on each side of the separating hyperplane, and are "pushed up" against the two sets of data points respectively during the optimization process. Intuitively, a good separation is achieved by the two parallel hyperplanes with the largest distance to each other.

We downloaded the state-of-the-art implementation named LIBSVM [22] of the SVM classifier. Signature, DRAGON, and FFSM descriptors take chemical compounds in SDF format as input, convert them into n-dimensional descriptor vectors. Then LIBSVM treats each n-dimensional vector as a point in n-dimensional space, and build a decision boundary between actives and inactive samples in that space.

4. Experimental Study

4.1 Data Sets

We selected 14 protein–chemical binding data sets from different sources, and the characteristics of these data sets are listed in Table 1. The first data set consists of 279 Factor Xa inhibitors and 156 inactives [23]. The next four data sets includes a number of inhibitors and approximately equal number of inactives to each of four target proteins: (1) ACE; (2) COX2; (3) DHFR; and (4) THR [24]. Chemicals with $\mathrm{IC}_{50} < 10\,\mathrm{nM}$ ($\mathrm{pIC}_{50} > 8$) are defined as actives and $>1\,\mu\mathrm{M}$ (or $\mathrm{pIC}_{50} < 6$) as inactives.

The other nine data sets were extracted manually from the BindingDB database [25, 26]. This database contained more than 450 target proteins and their binding chemicals. Two types of binding activity parameters K_i and IC_{50} were provided, and both of them measured the inhibition power of a chemical compound to a specific target protein [27, 28]. We manually selected nine target proteins and enough binding chemicals with known K_i values for each: (1) Androgen Receptor (AR); (2) Collagenase (ChC); (3) DPP-IV; (4) Factor VIIa (FVII); (5) Factor Xa (FXa); (6) Fatty Acid Amide Hydrolase (FAAH); (7) HCV NS3-NS4A Serine Protease (HCV); (8) HIV-1 Protease (HIVP); and (9) HIV-1 Protease B Subtype (HIVB). As the BindingDB database provided only real-valued binding activity parameters, we need to find a way to define class labels for each chemical. We used a strategy presented in Smalter *et al.* [20], in which all K_i value are first sorted in non-descending order, and then the top 37.5% data are defined as actives and the bottom 37.5% are as inactives, hence the middle 25% are thrown out to impose some separation between the two classes.

4.2 Experimental Methods

The first experiment was to use all 1,664 descriptors from DRAGON 5.4 (some descriptors may not be available to all

Table 1
Characteristics of the 14 data sets

Data sets	FFX	ACE	COX2	DHFR	THR	AR	ChC
Number of Actives	279	65	219	203	68	60	69
Number of Inactives	156	49	103	194	20	60	69
Avg. Number of Atoms	59.8	42.4	41.9	40.6	68.1	38.3	41.7
Data Sets	DDP4	FVII	FXa	FAAH	HCV	HIVP	HIVB
Number of Actives	82	53	144	41	79	140	136
Number of Inactives	85	53	144	41	79	140	138
Avg. Number of Atoms	43.8	57.2	60.0	46.7	99.6	86.6	83.8

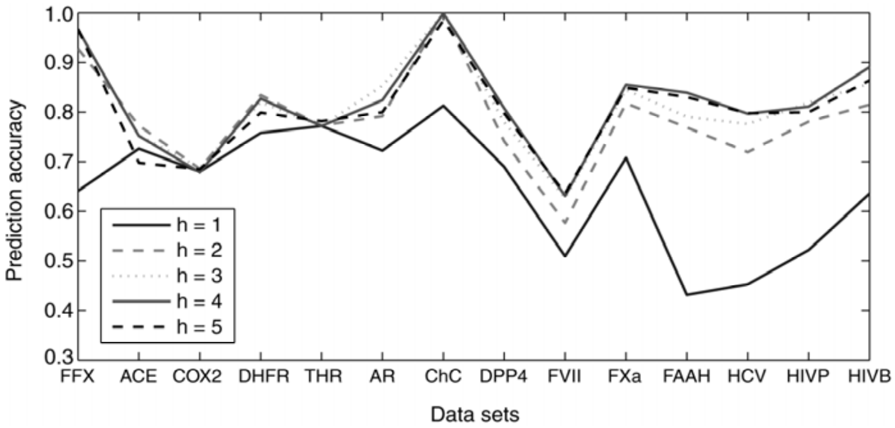


Figure 3. Prediction accuracy of signature descriptors over signature heights.

chemicals in a data set) a single descriptor class, and the prediction accuracy was denoted in 2 as "DRAL". Then each class of DRAGON descriptors was used individually and the results were listed as DR01, DR02, ..., DR18, and DR20. We omitted DR19 since this descriptor class was not available to many chemicals in the data sets. Second, the signature molecular descriptors were used to extract descriptors from the chemicals. As signature height can significantly influent the prediction accuracy, we first optimized the signature height from 1 to 5, and the results were shown in Fig. 3. Finally, we used FFSM descriptors to mine frequent subgraphs from each of the 14 data sets and the results were listed in Table 2 and denoted as "FFSM". In this paper, the support σ was set at 30% for all experiments.

A server with 96 Intel Celeron 1.6 Ghz processors and 30 GB memory was used for FFSM descriptors, and most calculations can be finished within a few minutes. All calculations for DRAGON and signature descriptors were performed on a notebook computer with a 1.6 GHz Intel CoreDuo processor and 2 GB memory within tens of minutes per data set.

4.3 Model Selection and Assessment

To select a model with the best balance of inductive bias and optimal complexity, we first randomly selected 30% of each data set as testing set, and the rest 70% data will be used as training and validation sets. We then used LIBSVM and RBF kernels to train our classification model with training sets and validated it with validation sets. A computational grid of parameters were searched to identify the best parameter set using a standard 10-fold cross-validation process.

Our experiments showed that the classification model was not very sensitive to a small change of C and γ, therefore this optimal parameter set was applied to all data sets for simplicity. The reason might be that our data sets are all balanced. For all experiments, the optimal model was obtained when $C = 1.0$ (train error/margin tradeoff) and $\gamma = 0.5$ (inverse kernel width).

After finding the optimal model, we then evaluated our model on the testing set that was never used in training. Prediction accuracy was defined here as $(TP + TN)/(TP + TN + FP + FN)$ (TP: true positive, TN: true negative, FP: false positive, FN: false negative). To obtain stable results, we permutated each data set randomly and selected the testing set, and then repeated the same experiments for 10, 20, 30 times. Our results revealed that the number of repeating times did not affect the results significantly. The final accuracy and standard deviations were computed by averaging over 20 repeated experiments.

5. Results and Discussions

We conducted experiments for signature molecular descriptors in the signature height range 1–5 and the results are shown in Fig. 3. It was obviously that when the prediction accuracy converges when the height is in 3–5, therefore we selected signature height = 3 for all following experiments (denoted as "Sign3").

In Table 2, we presented classification results for the 22 descriptor sets, and also plotted the prediction accuracy versus the 14 data sets for each specific descriptor class (figure not shown). As a first impression, the performance of each individual descriptor class differed significantly, and some descriptor sets consistently outperformed other descriptor sets over almost all the data sets. In Fig. 4, we plotted some best performing descriptor sets for further comparisons. In all figures, each curve represented the performance of a specific descriptor class, and was discriminated from each other with different colours and line styles.

A clear trend in Fig. 4 was that the prediction accuracy of most descriptor sets was better on data sets that were careful curated (set 1–5) than on the rest (set 6–14). One of the possible reasons was that our labelling strategy can conveniently produce binary data sets with equal proportion of positive/negative classes, but may not accurately reflect true biological activities of chemicals.

Combining Table 2 and Fig. 4, we found that FFSM descriptors and signature descriptors (height = 3) performed

Table 2
Prediction Accuracy on Testing Set with the Optimal Model for all 14 Data Sets. Star(*) Denotes the Descriptor Class that Yields the Best Accuracy for a Given Data Set

Descriptors	FFX	ACE	COX2	DHFR	THR	AR	ChC
FFSM	0.893	0.712	0.701	0.800	0.808	0.764	0.970
Sign3	0.963*	0.753	0.688	0.820	0.772	0.853*	0.999*
DRAL	0.645	0.583	0.680	0.483	0.773	0.450	0.426
DR01	0.705	0.604	0.664	0.708	0.748	0.632	0.605
DR02	0.645	0.584	0.680	0.484	0.773	0.682	0.428
DR03	0.702	0.607	0.672	0.542	0.726	0.750	0.627
DR04	0.753	0.616	0.655	0.695	0.726	0.710	0.537
DR05	0.660	0.584	0.671	0.496	0.773	0.698	0.549
DR06	0.945	0.740	0.710*	0.758	0.812	0.810	0.934
DR07	0.882	0.693	0.665	0.770	0.750	0.817	0.918
DR08	0.884	0.776	0.700	0.797	0.822*	0.816	0.804
DR09	0.882	0.841*	0.674	0.692	0.773	0.785	0.752
DR10	0.657	0.584	0.686	0.497	0.773	0.684	0.464
DR11	0.872	0.742	0.686	0.687	0.766	0.695	0.724
DR12	0.645	0.583	0.680	0.484	0.773	0.450	0.427
DR13	0.645	0.583	0.680	0.485	0.773	0.451	0.429
DR14	0.649	0.582	0.673	0.509	0.773	0.491	0.429
DR15	0.645	0.583	0.680	0.498	0.773	0.465	0.433
DR16	0.653	0.593	0.670	0.562	0.773	0.553	0.450
DR17	0.822	0.797	0.690	0.821*	0.755	0.773	0.909
DR18	0.720	0.615	0.651	0.756	0.773	0.726	0.749
DR20	0.700	0.589	0.652	0.686	0.759	0.688	0.626
Descriptors	DPP4	FVII	FXa	FAAH	HCV	HIVP	HIVB
FFSM	0.773	0.681*	0.868*	0.795	0.756	0.741	0.790
Sign3	0.780	0.627	0.845	0.791	0.776	0.819*	0.857*
DRAL	0.461	0.429	0.457	0.401	0.440	0.458	0.461
DR01	0.542	0.443	0.533	0.762	0.525	0.568	0.602
DR02	0.468	0.453	0.465	0.401	0.440	0.465	0.460
DR03	0.631	0.660	0.549	0.628	0.588	0.517	0.599
DR04	0.605	0.511	0.665	0.796	0.571	0.583	0.637
DR05	0.507	0.510	0.481	0.483	0.443	0.486	0.483
DR06	0.814*	0.607	0.834	0.796	0.742	0.803	0.802

(Continued.)

Table 2
Continued.

DR07	0.754	0.589	0.797	0.760	0.815	0.748	0.775
DR08	0.765	0.578	0.801	0.708	0.818*	0.726	0.775
DR09	0.764	0.582	0.734	0.553	0.784	0.747	0.780
DR10	0.496	0.484	0.484	0.484	0.460	0.493	0.478
DR11	0.653	0.516	0.645	0.542	0.523	0.561	0.570
DR12	0.461	0.429	0.457	0.401	0.440	0.458	0.461
DR13	0.461	0.437	0.460	0.402	0.439	0.458	0.461
DR14	0.469	0.431	0.460	0.451	0.441	0.458	0.465
DR15	0.464	0.431	0.458	0.403	0.440	0.458	0.461
DR16	0.516	0.417	0.481	0.438	0.458	0.461	0.500
DR17	0.733	0.633	0.822	0.841*	0.746	0.702	0.746
DR18	0.671	0.578	0.641	0.777	0.612	0.610	0.646
DR20	0.529	0.460	0.522	0.706	0.612	0.553	0.547

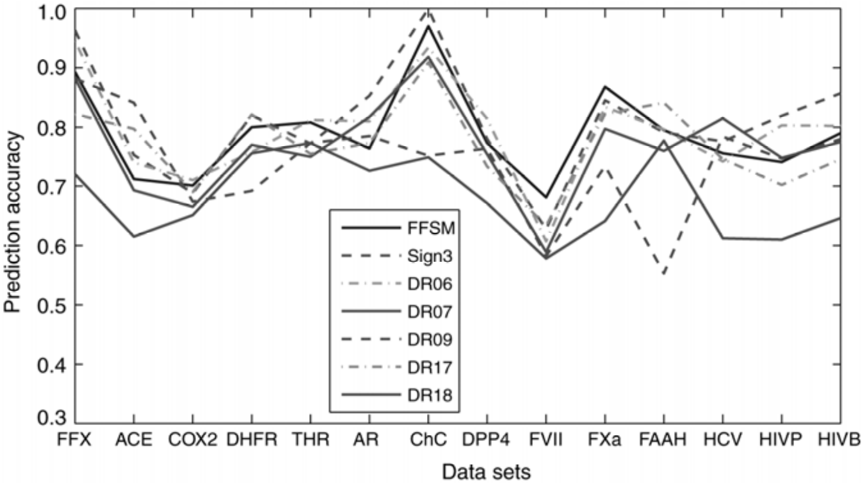


Figure 4. SVM prediction accuracy for some "best" descriptor sets: FFSM, Signature ($h = 3$), DRAGON descriptor class 06, 07 and 18.

better than 17 DRAGON descriptors and equivalently to the other three: DR06(2D autocorrelations), DR07(edge adjacency indices), and DR17(functional group counts). In addition, some single DRAGON descriptor sets consistently performed better than other descriptor sets over almost all the data sets. For instance, descriptor class DR06, DR07, and DR17 generally outperformed all other DRAGON descriptor sets. DR09 (topological charge indices) and DR18 (atom-centered fragments) also showed decent accuracy compared to other DRAGON descriptors. Finally, using any single DRAGON descriptor class almost always yielded better performance than using all DRAGON descriptors together, and the reason might be due to over-fitting.

One noteworthy phenomena was that DRAGON descriptors performed significantly worse when they were normalized than when they were not (data not shown), with an average of 10–15% accuracy difference for most data sets, and this difference was stable when we changed the experimental repeating times from 10 to 30. One of the reasons might be that each DRAGON descriptor has its chemical, physical, or biological meaning, and normalization will cause the descriptor to lose its real meaning, and hence cannot characterize chemical compounds very well. Please note that our results for DRAGON descriptors in Table 2 were obtained without normalization.

This work demonstrated that it was really important to carefully select DRAGON descriptor sets. Simply using all of them or randomly selecting some of them will diminish the prediction performance. A combination of DR06, DR07, DR17 and a few other DRAGON descriptor sets would be a robust and optimal selection. In addition, the signature and FFSM descriptors were satisfactory candidates for virtual screening and protein–chemical interaction prediction.

A major reason that the FFSM descriptors outperformed most DRAGON descriptors was that FFSM has implicit functions of descriptor selection, since it only counted frequent subgraphs and infrequent descriptors were removed. In many cases, frequent subgraphs have been proven to correspond to some biological or chemical structural motifs. FFSM descriptors can be a better method if approximate matching of subgraphs was allowed. In addition, signature descriptor also performed very excellent after careful tuning of the parameters, partially since they completely characterized a chemical by including geometrical and topological descriptors, atomic properties, chemical bonding, and hydrodization information. Finally, through our experimental results we should keep in mind that no descriptors can be claimed the best for all data sets and all situations, and our comparisons and conclusions were made on a statistically average basis.

Our results and conclusions may be biased since only the SVM classifier was used, therefore our study only focused on limited scope of the subject of SVM-based protein–chemical interaction prediction. Although using other popular classifiers such as K-nearest neighbours, neural network, and random forest is beyond the scope of this paper, our study would be more complete and valuable if comparisons based on those classifiers are included.

6. Conclusions

Prediction protein–chemical interaction is a challenging *in silico* problem in bioinformatics and cheminformatics research. Chemical descriptors are of key importance for converting chemicals into descriptor vectors to be understood by computers. In this paper, we conducted a comprehensive comparison of the performance of various chemical descriptors based on SVM classification on 14 high quality data sets. Our results shed light on selecting descriptor extraction methods wisely to obtain best prediction performance. Our experiments also demonstrated that FFSM descriptors consistently outperform 17 of the 20 DRAGON descriptor sets, and performed equivalently to signature descriptors and the other three DRAGON descriptor sets. With wise selection of some DRAGON descriptor sets, we can earn robust and optimal performance for protein–chemical interaction prediction. In the future, we would perform similar studies for protein descriptors and also on other classifiers such as random forest, K-nearest neighbours, and neural network.

Acknowledgements

This work has been partially supported by an NIH grant #R01 GM868665 and NSF grant IIS 0845951.

References

[1] A.R. Leach, Molecular modeling: Principles and applications (Englewood Cliffs, NJ: Prentice Hall, 2001).

[2] N. Wale, I. Watson, & G. Karypis, Comparison of descriptor spaces for chemical compound retrieval and classification, *Knowledge and Information Systems*, *14*(3), 2007, 1–29.

[3] G. Paolini, R. Shapland, W. van Hoorn, J. Masonn, & A. Hopkins, Global mapping of pharmacological space, *Nature Biotechnology 24*, 2006, 805–815.

[4] M. Deshpande, M. Kuramochi, N. Wale, & G. Karypis, Frequent substructure-based approaches for classifying chemical compounds, *IEEE TKDE*, *17*(8), 2005, 1036–1050.

[5] T. Horvath, T. Grtner, & S. Wrobel, Cyclic pattern kernels for predictive graph mining, *SIGKDD* 2004, 158–167.

[6] DRAGON User Manual, Milano, http://www.talete.mi.it/products/products.htm.

[7] Daylight User Manual, Daylight Inc., http://www.daylight.com.

[8] D. Rogers, R. Brown, & M. Hahn, Using extended-connectivity fingerprints with laplacian-modified bayesian analysis in high-throughput screening, *Journal of Biomolecular Screening*, *10*(7), 2005, 682–686.

[9] J. Faulon, M. Collins, & R. Carr, The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *Journal of Chemical Information and Computer Science 44*(2), 2004, 427–436.

[10] J. Huan, W. Wang, & J. Prins, Efficient mining of frequent subgraph in the presence of isomorphism, *Proc. 3rd IEEE Int. Conf. on Data Mining (ICDM)* 2003, 549–552.

[11] M. Kuramochi & G. Karypis, An efficient algorithm for discovering frequent subgraphs, *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 2004, 1038–1051.

[12] D. Agrafiotis, V. Lobanov, & F.R. Salemme, Combinatorial informatics in the post-genomics era, *Natural Reviews Drug Discovery*, *1*, 2002, 337–346.

[13] C. Lipinski & A. Hopkins, Navigating chemical space for biology and medicine, *Nature*, *432*, 2004, 855–861.

[14] C.M. Dobson, Chemical space and biology, *Nature*, *432*, 2004, 824–828.

[15] J. Ballesteros & K. Palczewski, G protein-coupled receptor drug discovery: Implications from the crystal structure of rhodopsin, *Current Opinion in Drug Discovery Development*, *4*, 2001, 561–574.

[16] J. Faulon, M. Misra, S. Martin, K. Sale, & R. Sapra, Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor, *Bioinformatics*, *24*(2), 2008, 225–233.

[17] J. Hert, P. Willett, D. Wilton, P. Acklin, K. Azzaoui, & E. Jacoby, A. Schuffenhauer, Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures, *Organic and Biomolecular Chemistry 2*(22), 2004 3256–3266.

[18] P. Gedeck, B. Rohde, & C. Bartels, QSAR – how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets, *Journal of Chemical Information and Modeling*, *46*(5), 2006, 1924–1936.

[19] V. Viswanadhan, A. Ghose, G. Revankar, & R. Robins, Atomic physicochemical parameters for three-dimensional structure directed quantitative structure-activity relationships, *Journal of Chemical Information and Computer Science 29*, 1989, 163–172.

[20] A. Smalter, J. Huan, & G. Lushington, Pattern diffusion graph kernel for chemical compound classification from Lasso regression to feature vector, *IWDMB08*, 2008.

[21] C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, *2*, 1998, 121–167.

[22] C. Chang & C. Lin, LIBSVM: A library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[23] F. Fontaine, M. Pastor, & I. Zamora, *et al.*, Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent descriptors. *Journal of Medicinal Chemistry 48*(7), 2005, 2687–2694.

[24] J. Sutherland, L. O'Brien, & D. Weaver, A comparison of methods for modeling quantitative structure–activity Relationships, *Journal of Medicinal Chemistry 47*(22), 2004, 5541–5554.

[25] T. Liu, Y. Lin, X. Wen, R. Jorrisen, & M. Gilson, BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities, *Nucleic Acids Research*, *35*, 2007, D198–D201.

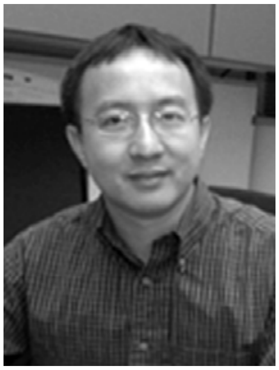
[26] X. Chen, Y. Lin, & M. Gilson, The binding database: Overview and user's guide. *Biopolymers Nucleic Acid Science 61*, 2002, 127–141.

[27] X. Chen, Y. Lin, M. Liu, & M. Gilson, The binding database: Data management and interface design, *Bioinformatics*, *18*, 2002, 130–139.

[28] X. Chen, M. Liu, & M. Gilson, Binding DB: A web-accessible molecular recognition database, *Journal Combinational Chemistry High-Throughput Screen*, *4*, 2001, 719–725.

Biographies

Jintao Zhang has been a Ph.D. student in the Centre for Bioinformatics at University of Kansas since Fall 2006, and joined Dr. Huan's research group in summer 2007 as a teching/research assistant, with research interest on data mining in bioinformatics and chemical biology. He received his Bachelor's degree in Chemical Physics from the University of Science & Technology of China in 2001, and a Master's degree in Chemistry from University of California, Riverside in 2005.



Jun (Luke) Huan has been an assistant professor in the Electrical Engineering and Computer Science department at the University of Kansas since 2006. He is an affiliated member of the Information and Telecommunication Technology Center (ITTC), Bioinformatics Center, Bioengineering Program, and the Center for Biostatistics and Advanced Informatics and all KU research organizations. He received his Ph.D. in Computer Science from the University of North Carolina at Chapel Hill in 2006. Before joining KU, he worked at the Argonne National Laboratory (with Ross Overbeek) and GlaxoSmithKline (with Nicolas Guex). He was a recipient of the NSF Faculty Early Career Development (CAREER) Award in 2009. He serves on the program committees of leading international conferences including ACM SIGKDD, IEEE ICDE, ACM CIKM, IEEE ICDM, and IEEE BIBM.