Prediction of Protein−Protein Interaction Inhibitors by Chemoinformatics and Machine Learning Methods

Alexander Neugebauer,[†] Rolf W. Hartmann,[†] and Christian D. Klein*,[‡]

Pharmaceutical and Medicinal Chemistry, Saarland University, Saarbrücken, Germany, and Medicinal Chemistry, University of Heidelberg, Heidelberg, Germany

Received May 7, 2007

We describe a collection of structurally diverse inhibitors of protein−protein-interactions (PPIs). This collection is compared against the FDA drug database and a subset of the ZINC database by machine learning methods which rely on classical QSAR descriptors. We obtain a decision tree that contains three descriptors. Of particular importance is a constitutional descriptor related to molecular shape and size. Validation of the decision tree by various procedures indicates that it does not result from chance correlations and has predictive value. We conclude that constitutional descriptors may be valuable tools in the preselection of potential PPI inhibitors from compound databases.

Introduction

Protein−protein interactions (PPI[a]) are responsible for a multitude of biological effects. It is plausible to assume that the number of pharmacologically relevant protein−protein interactions exceeds the number of classical target structures for drug development such as active sites or ligand binding pockets. Therefore, small molecules that specifically interfere with protein−protein recognition processes may become increasingly important in drug development in the future. The challenge of PPIs as target structures is to find ligands that have sufficient affinity toward shallow or superficial binding sites that offer only limited chemical functionality. This stands in marked contrast to active sites that have evolved to bind substrates and transition states with high affinity and, consequently, also have the potential to tightly bind substrate analogs or other small molecules. General reviews on the topic are given by Hamilton and Yin[1] and Fry.[2]

The purpose of the work presented here was to determine if general principles or rules can be defined to quickly determine whether or not a compound is a potential inhibitor of PPIs. Using medicinal-chemical common sense, one can expect that PPI inhibitors are, on average, larger than enzyme inhibitors or receptor ligands; only large molecules can derive sufficient free energy of binding when interacting with shallow or superficial binding sites. However, it is certainly desirable to have a set of differentiating rules that are more specific than “size” (be it molecular weight, volume, or any other bulk property). We, therefore, set out to first compile a collection of known PPI inhibitors and then search for qualitative or quantitative measures that are able to differentiate those compounds from other drugs.

To our knowledge, this work represents the first attempt to determine discrimination rules for PPI inhibitors. The results may be useful for the assessment and prescreening of virtual compound databases, but also for general PPI inhibitor design considerations.

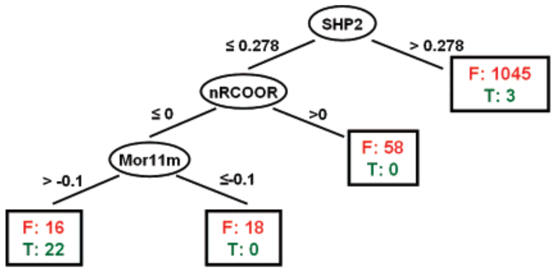

Figure 1. Pruned decision tree of 25 PPI inhibitors (attribute T: true, green) and 1137 non-PPI inhibitors (attribute F: false, red) from the FDA approved drug database. See the text for an explanation of the descriptors. The numbers (e.g., 0.278) are the decision criteria of the corresponding descriptor: at the highest bifurcation of this tree, a large number of non-PPI inhibitors are “successfully” removed from the training set because the value of SHP2 is larger than 0.278

The overall strategy is to extract known PPI data from the literature and calculate descriptors of various types. Machine learning methods are then used to discriminate between PPI inhibitors and non-PPI inhibitors. We decided to construct a decision tree because this type of model yields clear statements about the most relevant descriptors and is, therefore, more transparent and more amenable to interpretation than other techniques such as neural networks or support vector machines. A well-known example of a decision tree are the Lipinski rules for oral bioavailability:[3] this is a collection of criteria such as maximum number of H-bond donors and maximum molecular weight. These criteria are usually not presented in a tree-like format, but work in exactly the same way.

Results and Discussion

Data Collection. A total of 25 PPI inhibitors were identified from the literature. A compilation of the structures is given in the Supporting Information. Peptides and small proteins like enfuvirtide[4] were not included in this study. It may be noteworthy that none of the PPI inhibitors has a molecular weight below 400 g/mol (cf. Figure 2). A very important point in the preparation of the PPI inhibitor database was to achieve a high structural diversity of the compounds. Therefore, only one representative member from each class of compounds was included in this study. The Food and Drug Administration (FDA) approved drug database was used as a source of reference

* To whom correspondence should be addressed. Dr. C. Klein, Medicinal Chemistry, University of Heidelberg, Im Neuenheimer Feld 364, D-69120 Heidelberg, Germany. Phone: ++49-6221-545824. Fax: ++49-6221-546430. E-mail: c.klein@uni-heidelberg.de.

[†] Saarland University.

[‡] University of Heidelberg.

[a] Abbreviations: PPI, protein−protein interaction; TP, true-positives; FN, false-negatives; FP, false-positives; TN, true-negatives; SHP2, Mor11M, nRCOOR, chemoinformatical descriptors.

10.1021/jm070533j CCC: $37.00
Published on Web 08/17/2007

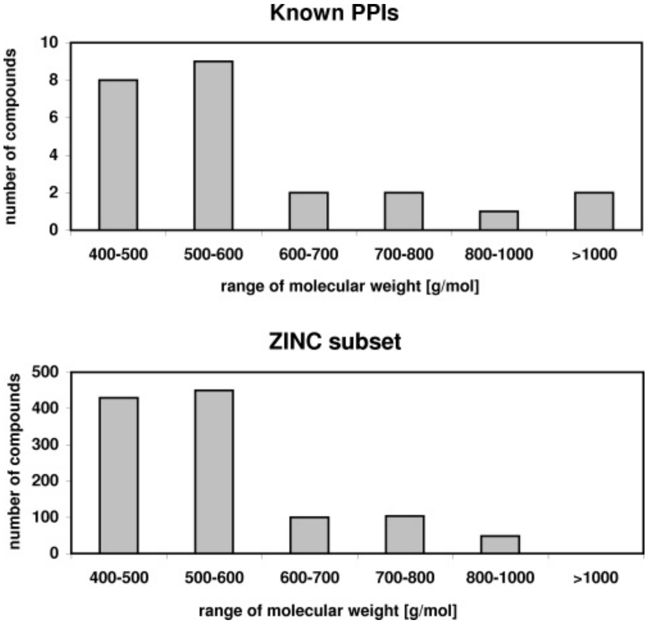

Figure 2. Distribution of molecular weights of the PPI database (upper chart) and the subset of the ZINC database (lower chart). It is interesting to note that none of the known PPI inhibitors has a molecular weight below 400 g/mol.

compounds for model building.[5] Here we considered 1137 compounds. There were 80 compounds from the FDA database that were not considered because the structure files contained more than one compound or it was not possible to calculate all descriptors. The PPI inhibitors from the literature and the FDA database were initially built or received in 2D format (*.sd file) and were converted to 3D format using the Corina[6] program. The ZINC[7] database subset described below was constructed from the complete 3D database that is available from the Shoichet group web site.

Descriptors. Descriptors were calculated by the program DRAGON 5 by Todeschini et al.[8] In summary, 1664 descriptors were calculated from various types like constitutional, molecular profile, functional group count or molecular property descriptors. For a complete list of descriptors, see ref 8. Descriptors were preselected by removing those descriptors from the set that are intercorrelated with a correlation coefficient above 0.9 as well as constant descriptors (descriptors with standard deviation lower than 0.0001) and near-constant descriptors (descriptors with only one value different from the remaining ones). In the end, we obtained 637 descriptors.

Construction of Decision Trees. We employed the divide-and-conquer algorithm C4.5[9,10] revision 8 (J48) implemented in the software package WEKA 3.4.6 by Witten and Frank[10] to construct decision trees. Activity is expressed in a binary manner: T (true) for PPI inhibitor and F (false) for non-PPI inhibitor. We generated a tree using the default parameter settings of the J48 method. The model was evaluated by 10-fold stratified cross-validation. The most important details are given in the Experimental Section.

We obtained a decision tree that employs eight descriptors and has a very high true-positive rate, that is, the tree performs a nearly perfect classification of the training set compounds (cf. Table 1, "initial decision tree"). However, the true-positive rate is much lower in the cross-validation runs, indicating that the initial decision tree has been overfitted and does not have sufficient predictive power. We therefore decided to limit the size of the tree to the most relevant descriptors (pruning). When only the three most relevant descriptors of the initial tree are used in the model building, the resulting decision tree performs somewhat worse in the classification of the training set, but has a much higher predictive power, as indicated by the cross-validation results.

Table 1. Confusion Matrices and True-Positive Rates of the Initial and the Pruned Decision Trees

<table>
<tr><th></th><th>Initial decision tree</th><th>Pruned decision tree</th></tr>
<tr><td>Number of descriptors for training</td><td>637</td><td>3
(SHP2, Mor11m, nRCOOR)</td></tr>
<tr><td>Number of descriptors used in the decision tree</td><td>8</td><td>3</td></tr>
<tr><td>Confusion matrix of the actual tree</td><td><table><tr><td></td><td colspan="2">Classified as</td></tr><tr><td></td><td>PPI</td><td>Non-PPI</td></tr><tr><td>PPI</td><td><b>22</b></td><td>3</td></tr><tr><td>Non-PPI</td><td>0</td><td><b>1137</b></td></tr></table></td><td><table><tr><td></td><td colspan="2">Classified as</td></tr><tr><td></td><td>PPI</td><td>Non-PPI</td></tr><tr><td>PPI</td><td><b>22</b></td><td>3</td></tr><tr><td>Non-PPI</td><td>16</td><td><b>1121</b></td></tr></table></td></tr>
<tr><td>Confusion matrix from cross-validation</td><td><table><tr><td></td><td colspan="2">Classified as</td></tr><tr><td></td><td>PPI</td><td>Non-PPI</td></tr><tr><td>PPI</td><td><b>9</b></td><td>16</td></tr><tr><td>Non-PPI</td><td>23</td><td><b>1114</b></td></tr></table></td><td><table><tr><td></td><td colspan="2">Classified as</td></tr><tr><td></td><td>PPI</td><td>Non-PPI</td></tr><tr><td>PPI</td><td><b>21</b></td><td>4</td></tr><tr><td>Non-PPI</td><td>17</td><td><b>1120</b></td></tr></table></td></tr>
<tr><td>True-positive rate from actual tree</td><td>1.00</td><td>0.88</td></tr>
<tr><td>True-positive rate from cross-validation</td><td>0.36</td><td>0.84</td></tr>
</table>

Table 2. Correlation between Descriptors Used for Discrimination of PPI and Non-PPI Inhibitors and Average Molecular Weight[a]

	SHP2	nRCOOR	Mor11m	MW
SHP2	1			
nRCOOR	−0.598	1		
Mor11m	0.180	−0.229	1	
MW	−0.697	−0.613	−0.592	1

[a] All values are correlation coefficients, *R*; $n = 1137$.

The pruned decision tree is shown in Figure 1. SHP2, the most relevant descriptor at the top of the decision tree, denotes a molecular shape descriptor introduced by Randić.[11] The descriptor SHP2 is an average of various lower-level descriptors that are derived from the interatomic distances of the atoms at the periphery of a molecule. Therefore, it is related to molecular properties like shape, size, and extension. nRCOOR is a functional group count descriptor: nRCOOR denotes the number of ester functions in the molecule. The nRCOOR descriptor branching point serves to exclude ester functions. The ester functionality is usually not considered as drug-like and is, therefore, not present in most of the more recently developed PPI inhibitors. Mor11m is a 3D-MoRSE ("Molecule Representation of Structures based on Electron diffraction") descriptor developed by Gasteiger et al.[12] Mor11m is a representation of the three-dimensional structure of a molecule.

The molecular shape descriptor SHP2 rejects most of the non-PPI inhibitors (1045 compounds). This represents a nearly 10-fold enrichment from 2.15% positives in the original dataset to 19.3%. Only three PPI inhibitors were falsely eliminated by applying the decision rule based on this molecular shape descriptor. It is not surprising to observe the importance of molecular shape in the context of protein−protein interactions, but we did not expect to observe such a pronounced enrichment by applying a single, simple filter rule. Additionally, 58 non-PPI inhibitors were rejected by the nRCOOR descriptor. The MoRSE descriptor eliminates 18 non-PPI inhibitors. In the last branch of the decision tree, the enrichment of PPI inhibitors has reached 57.9%.

The correlation of SHP2 with descriptors describing molecular size like molecular weight, mean atomic van der Waals volume, or number of carbon atoms is low to medium (Table 2). PPIs often have comparably high molecular weights, on average

Table 3. Validation of the Proposed Model by Permutation Testing[a]

			confusion matrix from cross-validation[b]				true-positive rate[c]	
No.	No. of leafs	main descriptor	TP	FN	FP	TN	F	T
0 (initial tree)	9	SHP2	9	16	23	1114	0.997	0.360
0 (pruned tree)	4	SHP2	22	3	16	1121	0.986	0.880
1	1		0	25	3	1134	0.997	0.000
2	1		0	25	0	1137	1.000	0.000
3	1		0	25	3	1134	0.997	0.000
4	1		0	25	0	1137	1.000	0.000
5	1		0	25	3	1134	0.997	0.000
6	3	nRCO	0	25	1	1136	0.999	0.000
7	1		0	25	0	1137	1.000	0.000
8	1		0	25	2	1135	0.998	0.000
9	1		0	25	0	1137	1.000	0.000
10	1		0	25	0	1137	1.000	0.000
11	1		0	25	0	1137	1.000	0.000
12	1		0	25	0	1137	1.000	0.000
13	1		0	25	4	1133	0.996	0.000
14	1		0	25	1	1136	0.999	0.000
15	1		0	25	1	1136	0.999	0.000
16	1		0	25	0	1137	1.000	0.000
17	1		0	25	0	1137	1.000	0.000
18	5	X1A	0	25	2	1135	0.998	0.000
19	1		0	25	0	1137	1.000	0.000
20	12	Mor03m	0	25	8	1129	0.993	0.000
21	1		0	25	0	1137	1.000	0.000
22	1		0	25	0	1137	1.000	0.000
23	1		0	25	0	1137	1.000	0.000
24	1		0	25	0	1137	1.000	0.000
25	1		0	25	0	1137	1.000	0.000
26	4	nR09	0	25	3	1134	0.997	0.000
27	1		0	25	0	1137	1.000	0.000
28	1		0	25	2	1135	0.998	0.000
29	3		0	25	2	1135	0.998	0.000
30	1		0	25	1	1136	0.999	0.000
31	1		0	25	4	1133	0.996	0.000
32	1		0	25	1	1136	0.999	0.000
33	1		0	25	2	1135	0.998	0.000
34	1		0	25	1	1136	0.999	0.000
35	1		0	25	0	1137	1.000	0.000
36	1		0	25	0	1137	1.000	0.000
37	3	S107	0	25	7	1130	0.994	0.000
38	1		0	25	3	1134	0.997	0.000
39	1		0	25	0	1137	1.000	0.000
40	1		0	25	1	1136	0.999	0.000
41	1		0	25	0	1137	1.000	0.000
42	1		0	25	3	1134	0.997	0.000
43	1		0	25	1	1136	0.999	0.000
44	3	GGI1	0	25	0	1137	1.000	0.000
45	1		0	25	0	1137	1.000	0.000
46	1		0	25	1	1136	0.999	0.000
47	1		0	25	0	1137	1.000	0.000
48	1		0	25	0	1137	1.000	0.000
49	1		0	25	0	1137	1.000	0.000
50	1		0	25	1	1136	0.999	0.000

[a] For an explanation of descriptors other than SHP2, see ref 8. [b] TP, true-positives; FN, false-negatives; FP, false-positives; TN, true-negatives. [c] See Experimental Section.

ranging from 400 to 1000 g/mol. However, molecular weight alone is not a useful descriptor for the discrimination of PPI inhibitors and non-PPI inhibitors. We were unable to build meaningful decision trees or other models based on molecular weight or similar bulk descriptors.

The correlation between the descriptors forming the decision tree is generally weak. Table 2 shows the correlation coefficients R ($n = 1137$) of descriptors used for discrimination of PPI and non-PPI inhibitors and molecular weight.

Validation: Y-Scrambling. Because the validation of the proposed model is of utmost importance, we applied another independent validation strategy: The PPI classification values (true, false) of the 25 PPI inhibitors (classification true) were reordered in a random manner (y scrambling) over the data set. Afterward, attempts were made to build decision trees with the scrambled activity data. If the classification rules were based on chance correlations, then the resulting random datasets should yield models that are of similar significance as the one that is based on the real data. A total of 50 randomization runs were performed. Table 3 shows the results of the permutation test. In all cases, the obtained models do not have any useful predictive power. In particular, the models are unable to reliably identify the true positives from the dataset, which would be the prime interest in using a PPI prediction model.

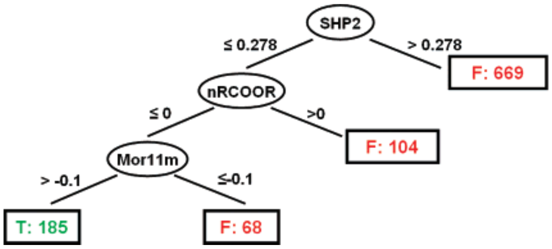

Figure 3. Application of the model to the ZINC subset. A total of 185 compounds from the ZINC subset are predicted to be PPI inhibitors.

ZINC Subset as Test Case. We next studied the performance of the model against a "background" database that is more typical of high-throughput screening compound collections than the FDA approved drug database.

There were 1130 molecules that were extracted from the ZINC[7] release 2007 database to obtain a collection of compounds with a molecular weight distribution similar to the 25 PPI inhibitors. Figure 2 shows the molecular weight distributions of the PPI inhibitors and the ZINC subset. The application of the model to the ZINC subset leads to the classification shown in Figure 3. There are 185 compounds that are predicted to be potential PPI inhibitors, a number that is considerably larger than in the training dataset, where 38 compounds were predicted to be active (16 false-positives and 22 true-positives). The high number of "false-positive" substances (if one assumes that none of these compounds is active) in this test case may indicate structural differences between the ZINC database and the FDA-approved drug database: the ZINC database contains more compounds that, judged by their chemical properties and molecular shape, resemble known PPI inhibitors.

When the ZINC subset is limited to compounds with molecular weights from 400 to 600 g/mol, the number of "false-positives" decreases to about 130 (data not shown). This could indicate that the proposed model works somewhat better for smaller compounds. In the context of lead discovery and druglikeness, this feature may be of particular value when compound collections are prescreened using the chemoinformatical descriptors described here.

Conclusion

In our opinion, the significance of the results lies only partially in the nature of the actual descriptors used. We do not intend to propose the particular decision tree shown here as the "one and only" method for identifying potential PPI inhibitors by database screening. However, the fact that a 10-fold enrichment can be achieved by a single descriptor clearly indicates that it may be possible to perform an effective virtual screening or pre-selection of compounds based on classical, low-dimensional QSAR descriptors in the search for PPI inhibitors. Given the results of the validation experiments, we do not believe that the obtained decision tree is the result of chance correlations.

Quite obviously, there are certain steric properties that are found more frequently in PPI inhibitors than in "traditional" ligands or inhibitors, and these steric properties can be described using abstract descriptors of molecular shape such as SHP2. In contrast to other more detailed and less abstract ways of searching large databases (e.g., docking or calculation of receptor-independent 3D pharmacophoric models), the use of low-dimensional QSAR descriptors has the great advantage of being fast and very robust. It may be used for the effective generation of subsets that may be further pruned by more “sophisticated” modeling techniques. In the future, especially when more and more PPI inhibitors are described, widening the statistical basis for model development, it may be possible to develop rules for the identification of PPI inhibitors that resemble the Lipinski criteria[3] for bioavailability. Given the results described here, we are optimistic that it will be possible to define such rules, and low-dimensional QSAR descriptors will probably by useful in the process.

Experimental Section

Decision Trees. Decision trees were generated by the data mining software package WEKA by Witten and Frank using the C4.5 algorithm J48 implemented in WEKA.[10] The most important parameters are confidenceFactor, minNumObj, numFolds, and the pruning methodology. ConfidenceFactor is a parameter used for pruning the decision tree. Smaller values lead to more pruning. We used the default value of 0.25. MinNumObj is the minimum number of instances per leaf. It is set to two. NumFolds determines the amount of data used for pruning. One fold is used for pruning and the rest is used for growing the tree. In our study, we used numFolds = 3. WEKA allows that use of two pruning methodologies, the reduced-error pruning and the C4.5 pruning. The latter method was used in this study.

There is no different cost associated with the different types of misclassifications. This means that we assigned the same cost to each type of error. The a priori probability for the two classes is 2.15% for class T and 97.85% for class F.

True-Positive Rate. The accuracy of the derived model is calculated by true-positive rates of the different classes. The true-positive rate tp of the T and F classes is calculated by

$$\mathrm{tp}(T) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{tp}(F) = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

where TP denotes the true-positives, FN denotes the false-negatives, and TN denotes the true-negatives classification.

Cross-Validation. Cross-validation procedures eliminate one or several data sets (instances) from the training set, derive a quantitative model from the remaining instances, and predict the PPI classes for one or several instances that were not included in the derivation of the model. All decision trees were computed from scratch in the cross-validation. This means that the whole tree is built from all the available data, and no incremental decision tree induction is used. A 10-fold stratified crossvalidation was performed. Random seeds were set arbitrarily.

Overfitting can be prevented by pruning the tree. Therefore, the concept space is examined starting from the easiest complex description toward more complex concept descriptions (simplest-first ordering). The simplest-first search and breakup at a sufficient complex concept description is a reliable way to avoid overfitting.[9] The C4.5 algorithm J48 implemented in WEKA is robust concerning overfitting.[10]

Acknowledgment. We would like to thank Dr. Eibe Frank for helpful discussions concerning the WEKA machine learning program. This work was supported by the Deutsche Forschungsgemeinschaft as part of the Clinical Research Group KFO 129. C.K. gratefully acknowledges financial support by the Emmy-Noether-Programm (DFG KL 1356) and the Fonds der Chemischen Industrie.

Supporting Information Available: Structures of known protein−protein interaction inhibitors. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Yin, H.; Hamilton, A. D. Strategies for targeting protein−protein interactions with synthetic agents. *Angew. Chem., Int. Ed.* **2005**, *44*, 4130−4163.
(2) Fry, D. C. Protein−protein interactions as targets for small molecule drug discovery. *Biopolymers* **2006**, *84*, 535−552.
(3) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(4) Cullell-Young, M.; Leeson, P. A.; del Fresno, M.; Bayés, M. Enfuvirtide. *Drugs Future* **2002**, *27*, 450−457.
(5) http://www.epa.gov/ncct/dsstox/sdf_fdamdd.html.
(6) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comp. Method* **1990**, *3*, 537−547.
(7) Irwin, J. J.; Shoichet, B. K. ZINC−A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(8) Todeschini, R. Talete srl, *DRAGON* (for Windows), software for molecular descriptor calculations, version 5.4, 2006 (http://www.talete.mi.it/).
(9) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Francisco, CA, 1993.
(10) I. H. Witten, E. F. *Data Mining. Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann Publishers: San Francisco, CA, 2005.
(11) Randic, M. Molecular Shape Profiles. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 337−382.
(12) Schuur, J. H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1995**, *36*, 334−344.

JM070533J