Journal of Computer-Aided Molecular Design, **16:** 803–807, 2002.
KLUWER/ESCOM

Classification scheme for the design of serine protease targeted compound libraries

Stanley A. Lang*, Andrey V. Kozyukov, Konstantin V. Balakin, Andrey V. Skorenko, Andrey A. Ivashchenko & Nikolay P. Savchuk
Chemical Diversity Labs, Inc., 11558 Sorrento Valley Road, San Diego, CA 92121, USA

Received 24 September 2002; Accepted 8 January 2003

Key words: library design, neural network approach, sensitivity analysis, serine protease inhibitors, virtual screening

Summary

The development of a scoring scheme for the classification of molecules into serine protease (SP) actives and inactives is described. The method employed a set of pre-selected descriptors for encoding the molecular structures, and a trained neural network for classifying the molecules. The molecular requirements were profiled and validated by using available databases of SP- and non-SP-active agents [1,439 diverse SP-active molecules, and 5,131 diverse non-SP-active molecules from the Ensemble Database (Prous Science, 2002)] and Sensitivity Analysis. The method enables an efficient qualification or disqualification of a molecule as a potential serine protease ligand. It represents a useful tool for constraining the size of virtual libraries that will help accelerate the development of new serine protease active drugs.

Introduction

Serine proteases are a family of proteolytic enzymes that have been implicated in many important biological processes, including blood coagulation, digestion, inflammation, metastasis, and cellular invasion [1]. Regulation of their biological action by small molecule inhibitors is a potentially attractive way of treating many disease states. Despite extensive efforts spanning several decades, there are few small molecule inhibitors of serine proteases widely available as drugs. The reason for the scarcity of SP-targeted drugs is a specific set of stringent criteria required for a therapeutically useful small molecule inhibitor of these enzymes [1–5]. Many potent serine protease inhibitors are large flexible molecules with many highly polar (often cationic) groups, and, therefore, they are often compromised by their instability, low bioavailability, and poor pharmacological profiles. To be effective drugs, protease inhibitors need to have high stability to nonselective proteolytic degradation, long lifetimes in the bloodstream and in cells, low susceptibility to elimination, high selectivity for a target enzyme, and good bioavailability (preferably by oral delivery). Thus efficient approaches are needed for the high-throughput search for novel candidates to be assayed as SP targeted drugs.

A promising solution to a more efficient search strategy of SP inhibitors is the filtering of a database through the use of a computational approach based on a quantitative discrimination function. Such a function would permit the selection of a series of compounds to be assayed for biological activity. Recently, several examples of a successful application of neural network classification methodology for enhancement of target-specific content of compound selections were reported [6, 7]. In this work, we have applied this approach for the development of an effective scoring scheme for the classification of molecules into SP-actives and SP non-actives.

*To whom correspondence should be addressed. E-mail: slang@chemdiv.com

Table 1. Target-specific classes of SP inhibitors studied in this work.

No.	Enzyme	Compounds
1	Chymase	91
2	Chymotrypsin	9
3	Complement convertase	1
4	Dipeptidyl-Peptidase IV	58
5	Elastase	1
6	Factor VIIa	23
7	Factor Xa	235
8	HCV NS3 Protease	73
9	Plasmin	3
10	Prolyl Endopeptidase	110
11	Serine Protease	312
12	Thrombin	480
13	Trypsin	41
14	Tryptase	97
15	Urokinase	71
Total*		1,439

*The total number of compounds is not equal to the sum of the shown values, as some compounds are not selective and manifest activity against more than one target.

Table 2. Compounds at various development stages for SP(+) and SP(−) databases.

Development stage	Compounds	
	SP(+)	SP(−)
Biological testing	1,214	0
Preclinical	187	3,566
Clinical	34	1,518
Launched	4	47
Total	1,439	5,131

Databases

1,439 known SP ligands belonging to 15 different SP classes (Table 1) were used as a positive training set, SP(+). For comparison, a subset of 5,131 compounds, representing over 150 various non-protease activities were used as a negative training set, SP(−). All compounds were selected from the Ensemble database (Prous Science, 2002) of known pharmaceutical agents. The distribution of compounds through the development stages for these data sets is shown in Table 2. The series of SP-inhibitors included anionic (9.7%), cationic (36.3%), neutral zwitter-ionic (10.1%) and non-ionized (43.9%) molecules (status at physiological pH). The distribution was 16.7% of anionic, 9.7% of cationic, 6.3% of neutral zwitter-ionic and 67.3% of non-ionized compounds in the series of non-protease ligands. Structures were extracted according to the assigned target-specific activity class. An assumption was made that a molecule is SP-active, if it contains the indication on a SP protein in the 'mechanism of action' field. All other compounds were considered as SP-inactive. Active agents to proteases, other than serine (such as aspartic, cysteine and metallo), were excluded from both sets. All molecules were filtered based on molecular weight range (150–850) and atom type content (only C, N, O, H, S, P, F, Cl, Br, and I were permitted).

Neural networks

The NeuroSolution^{TM} 4.0 program (NeuroDimension, Inc., 2001) was used for all neural network operations. Feed-forward nets were constructed that consist of input neurons, one hidden layer, and two output neurons. The back-propagated networks were trained following the "momentum learning rule" as implemented in the NeuroSolution^{TM} software. The training was performed over 1000 iterations. For the generation of neural network model, the total of 6,570 molecules of the SP(+) set and SP(−) set were subdivided into three categories: (1) a training set (50% of the total number of compounds), (2) a cross-validation set (25%), and (3) a test set (25%). The cross-validation set was used to avoid over-training during the development of neural network models.

Descriptors

Sixty molecular descriptors encoding important molecular properties, such as lipophilicity, charge distribution, topological features, steric and surface parameters, were explored. These descriptors were calculated from 2D molecular representations with the Cerius^2 (Accelrys, Inc., 2000) and ChemoSoft^{TM} (Chemical Diversity Labs, Inc., 2002) software tools. After removing the low-variable and highly correlated ($R > 0.9$) descriptors, their total number was reduced to 37. To further reduce the number of descriptors that could contain redundant information, Sensitivity Analysis [8] of the generated neural network models was performed. The descriptor dataset is extended with an additional random "phantom" variable to scale the sensitivities. The underlying assumption is that

descriptors with sensitivities less than this random variable are not important for the model. The random variable was obtained from a normal distribution. The values of all descriptors were normalized.

Sensitivity analysis

Feature selection for this work is based on Sensitivity Analysis [8]. The main objective of the Sensitivity Analysis is to determine the saliency of each of the features in a model and to reduce the number of features. The testing process provides a measure of the relative importance among the inputs of the neural model. The first input is varied between its mean ± standard deviation while all other inputs are fixed at their respective means. The network output is computed for a defined number of steps above and below the mean. This process is repeated for each input and the variation of each output is measured as a sensitivity coefficient (SC) that is the standard deviation of each output divided by the standard deviation of the input which was varied.

After a neural network has been trained, a sensitivity measure per feature is obtained, and then this procedure is repeated for three times. These sensitivities are then combined as the average of three runs in order to obtain the final sensitivity value for each feature. These sensitivities are then sorted in ascending order and all features with sensitivities smaller than the random 'phantom' variable are dropped. This elimination process is done in successive iterations for feature reduction stages, constructing a new model based on the new reduced feature set. This iterative feature elimination process is halted when no more features can be dropped (there are no more features with a sensitivity below the sensitivity of the random scale variable). Three final descriptor sets corresponding to three independent randomizations were generated using the described selection procedure. The final set of descriptors remaining in at least two final sets is shown in Table 3.

Detailed analysis of the selected descriptors is not reported here because of the volume of the data. Some notes, however, should be made. Recently, it has been demonstrated that the protease family proteins have some specific conformational, mechanistic and inhibitor composition features (reviewed in [1]). Thus a key *conformational* requirement of all the proteases is their universal recognition of peptides, substrate analogues, and inhibitors in an extended backbone conformation [2, 3]. This structural motif is unlike all other

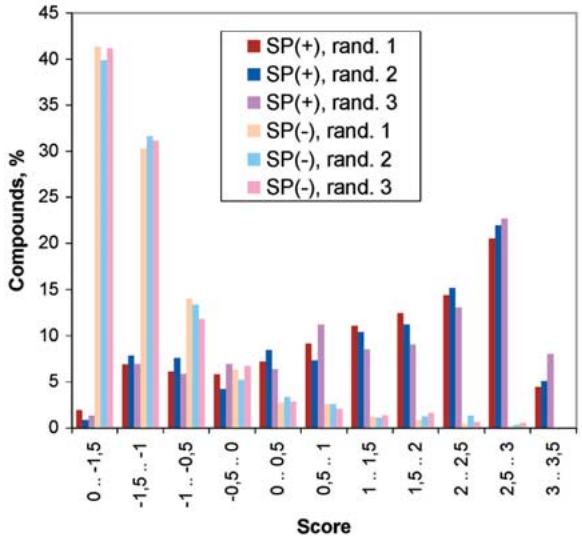

Figure 1. Compound distributions on the scale of prediction scores for three test sets.

elements of secondary structure in providing maximum separation and exposure of all component amino acids for interaction with solvent or protease. A key *mechanistic* feature of many protease inhibitors is the presence of an electrophilic transition-state isostere that simulates to some extent the transition state of amide bond hydrolysis [4, 5]. *Inhibitor composition* is another specific feature [1]. Thus for optimal oral and systemic delivery, the most successful drugs based upon protease inhibition have tended to be small organic molecules (MW < 300–500 Da) with few or no peptide bonds, sufficient lipophilicity, and a high selectivity for the specific protease. Analysis of the selected descriptors shows that they adequately encode these key molecular features. For example, one of the most discriminative descriptors, HOMO (energy of highest occupied molecular orbital), measures the nucleophilicity of a molecule. Not surprisingly, a strong negative correlation between the values of this descriptor and the neural network scores is observed, since the SP inhibitors usually have an electrophilic nature. The importance of such descriptors as the number of H-bond donors/acceptors and several members of Jurs-family describing surface charge distribution, is in full agreement with the necessity of stabilization of the substrate's position by forming hydrogen and ionic bonds with the catalytic center of the enzyme active site. LogP, logSw, TPSA can be thought of as molecular parameters reflecting the specific pharmacokinetic behavior of SP ligands. The importance of

Table 3. The final set of descriptors.

Descriptor	Definition	Average SC
Included into three final models		
HBD	no. of H-bond donors	0.318
HOMO	energy of highest occupied molecular orbital	0.318
HBA	no. of H-bond acceptors	0.181
Apol	sum of atomic polarizabilities	0.168
Density	molecular density	0.148
LogP	log of 1-octanol–water partition coefficient	0.146
TPSA	total polar surface area	0.128
Jurs-FPSA-3	atomic charge weighted negative surface area divided by TPSA	0.101
$\log S_w$	log of water solubility	0.092
RGyr	radius of gyration	0.073
RotB	number of rotatable bonds	0.065
Energy	energy of the currently selected conformation	0.035
Included into two final models		
Jurs-PNSA-1	partial negative surface area	0.114
Jurs-FNSA-3	atomic charge weighted negative surface area divided by TPSA	0.100
Jurs-DPSA-1	partial positive surface area minus partial negative surface area	0.075
PMI	principal moment of inertia	0.057
Jurs-FPSA-1	partial positive surface area divided by TPSA	0.047
Jurs-PNSA-2	total charge weighted negative surface area	0.044
MW	molecular weight	0.043

Table 4. Results of the final test sets classification with the neural net model.

	randomizations, SP(+)				randomizations, SP(−)			
	1	2	3	Mean	1	2	3	Mean
Correct predictions, %	85.0	83.7	85.9	84.9	85.6	84.8	84.1	84.8

conformational attributes, RotB and RGyr can be related to the fact that most of the known serine protease inhibitors are relatively flexible molecules.

Predictive modeling

Three additional independent randomizations were generated and final training–testing experiments were carried out with the final 19-descriptor set. The prediction quality was approximately the same in each of these 3 independent cycles: up to 85% of SP ligands and 85% of non-SP ligands (Table 4) were correctly predicted in the corresponding test sets. The separation threshold is set to the score value equal to 0. Figure 1 shows the distributions of the prediction scores for the test sets. The distinctively different distribution of the compounds with known SP-activity and compounds with other activities demonstrates the high discriminative power of the trained network.

Conclusions

Our observations indicate the presence of a combination of some specific physicochemical features that distinctly differentiate the serine protease inhibitors from the compounds belonging to other target specific classes. Using these findings, we created a neural network classification model with an excellent discrimination power. It is anticipated that this general model will be an extremely useful method in constraining the size of combinatorial libraries and in the collection, manipulation, and use of the data that are

generated. In general, this neural network based filter is most useful in conjunction with the drug/non-drug filter reported earlier [9, 10]. In practice, ligand structure and property based approaches are to be used in combination, taking into account as much information as possible. For example, a useful extension of the proposed methodology would include, as an initial stage, selecting of substructure similar and bioisosteric analogs of known serine protease inhibitors. Such a combined strategy could significantly increase the hit rate in the bioscreening programs by up to two orders of magnitude.

References

1. Leung, D., Abbenante, G. and Fairlie, D.P., J. Med. Chem., 43 (2000) 305.
2. Fairlie, D.P., Tyndall, J.D.A., Reid, R.C., Wong, A.K., Abbenante, G., Scanlon, M.J., March, D.R., Bergman, D.A., Chai, C.L.L. and Burkett, B.A., J. Med. Chem., 43 (2000) 1271.
3. Tyndall, J.D.A. and Fairlie, D.P., J. Mol. Recognit., 12 (1999) 1.
4. Andreasen, P.A., Kjoller, L., Christensen, L. and Duffy, M.J., Int. J. Cancer, 72 (1997) 1.
5. Henkin, J. Annu. Rep. Med. Chem., 28 (1993) 151.
6. Ajay, Bemis, G.W. and Murcko, M.A., J. Med. Chem., 42 (1999) 4942.
7. Balakin, K.V., Tkachenko, S.E., Lang, S.A., Okun, I., Ivashchenko, A.A. and Savchuk, N.P., J. Chem. Inf. Comput. Sci., 42 (2002), in press.
8. LeCun, Y., Denker, J.S. and Solla, S.A., In Touretzky, D. (Ed.), Advances in Neural Information Processing Systems, Morgan Kaufmann, San Mateo, CA, 1990, pp. 598–605.
9. Sadowski, J. and Kubinyi, H., J. Med. Chem., 41 (1998) 3325.
10. Ajay, A., Walters, W.P. and Murcko, M.A., J. Med. Chem., 41 (1998) 3314.