Effective Predictor of Human Mast Cell Tryptase Inhibitors

SORIN AVRAM[1], STEFANA AVRAM[2]*, CRISTINA DEHELEAN[3]

[1] Romanian Academy, Institute of Chemistry Timisoara, 24 Mihai Viteazul Blvd., 300223, Timisoara, Romania

[2] University of Medicine and Pharmacy Victor Babes Timisoara, Faculty of Pharmacy, Discipline of Pharmacognosy, 2 Eftimie Murgu Sq., 300041, Timisoara, Romania

[3] University of Medicine and Pharmacy Victor Babes Timisoara, Faculty of Pharmacy, Discipline of Toxicology, 2 Eftimie Murgu Sq., 300041, Timisoara, Romania

Mast cells (MCs) play a key role in the immune response to pathogens, in allergic and inflammatory reactions. Recently, increasing evidence has linked degranulated MCs to cancer through the presence of proangiogenic factors. Tryptase, a protease released from activated MCs granules has emerged as a potential target for tumor treatment. In this study, we aimed to develop a random forest model which is able to effectively classify tryptase inhibitors based on two-dimensional pharmacophore fingerprints. The external validation of the predictor demonstrated excellent performance with an AUC value of 0.953. Moreover, we highlight essential variables identified by the algorithm embedded in random forest. The hereby proposed classifier provides new means for the identification and optimization of tryptase inhibitors which are promising anti-angiogenic agents in the treatment of cancers.

Keywords: mast cells, tryptase inhibitor, random forest, prediction model

Mast cells (MCs) are granulocytes found in connective tissues (i.e., blood vessels, lymphatic vessels and nerves) and near skin and mucosa of the gastrointestinal, respiratory, and genitourinary tracts [1]. MCs are commonly known to play an essential role in the pathogenesis of allergies but also in the recognition of pathogens and modulation of appropriate immune responses [2]. Granulated MCs are also able to stimulate an intense angiogenic reaction in the chick embryo chorioallantoic membrane (CAM) assay, linking these cells to cancer [1].

Tryptase, a preformed serine protease stored in MCs secretory granules [1], was characterized as one of the most powerful angiogenic mediators released by human MCs [2]. Out of four types of tryptases, the major protease present in human MCs is β-tryptase [3, 4], which is considered a promising target in adjuvant cancer treatment [5, 6]. Tryptase inhibitors such as gabexate mesylate, nafamostat mesylate and tranilast (fig. 1) are considered promising new antiangiogenic agents in the treatment of various types of cancers, e.g., colon and colorectal, pancreatic, breast, prostate cancer etc [6].

The X-ray structures of β-tryptase, first resolved by Pereira et al. in 1998 [7], facilitated the structure-based discovery [8, 9] and development of several series of new inhibitors [10, 11]. The enrichment in biological activity determinations of various compounds tested against tryptases allows the use of ligand-based computational methods [12] to boost the discovery of novel inhibitors.

To date, ligand-based models have been scarcely applied for the prediction of MCs tryptase inhibitors. In this study, we aimed to generate an efficient ligand-based classification model to identify inhibitors of MCs tryptase. We describe a rigorous performed validation and briefly report important 2D pharmacophore descriptors for the model.

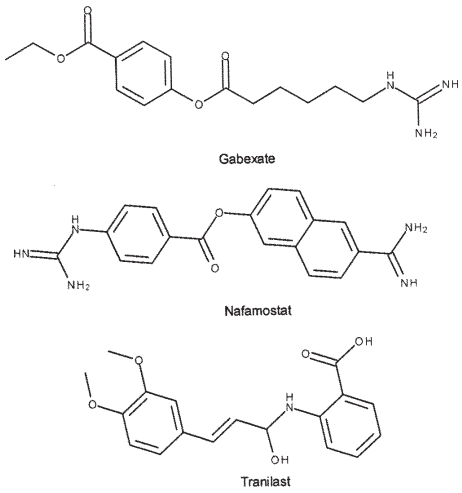


Fig. 1. Chemical representation of three tryptase inhibitors

Experimental part

Data set preparation

A set of 205 compounds with inhibitory data of type K_i was downloaded for tryptase alpha/beta-1 (tryptase-1, target CHEMBL2617, Uniprot ID Q15661; EC:3.4.21.59) from ChEMBLdb. Multiple activity values per compounds were resumed to the average value, thereby leaving each compound with a single K_i pair. Compounds with K_i < 100 nM were considered active (denominated further as actives) and those with K_i > 1000 nM were considered inactive (further denominated as inactives). Ideal classification modelling requires balanced label sets. Thus, due to a much small number of inactives (37 compounds) compared to the actives (103) we included another 61 inactives (23 with K_i > 1000 nM, and 38 with IC_{50} > 1000 nM) tested against target ID CHEMBL2095193 which includes besides tryptase-1, also serine protease 31 (or tryptase gamma; UniProt ID Q9NRR2) and tryptase-3 (UniProt ID Q9BZJ3). The remaining sets of 103 actives

* *email: stefana.avram@umft.ro*

Set	Number of actives	Number of inactives
Training set	77	70
Test set	26	24
Entire set	103	94

Table 1
DESCRIPTION OF ACTIVE AND INACTIVE SETS

and 94 inactive were standardize and set to their major tautomeric form at *p*H 7.4 using ChemAxon JChem API package (version 6.1.0, 2013, http://www.chemaxon.com).

Molecular descriptors

The chemical structures were described by two-dimensional fuzzy pharmacophore fingerprints (FPFs) generated by ChemAxon JChem API package (version 6.1.0, 2013, http://www.chemaxon.com). FPFs encoded here are pharmacophore smoothened counts generated for cations, anions, acceptors, donors, hydrophobic and aromatic rings with path lengths of 1 to 10 (between pairs of pharmacophoric points). Before suppling to model training and testing, FPFs have been standardized to size 210 using a self-written Java script.

Training and test set

The available set of compounds was split into a training set (75% of the data) and a test set (25% of the data) using function *createDataPartition* in package "caret" [13] available in R statistical software [14]. The function performs a random split of a proportional number of representatives of each label as shown in table 1.

Random forest

Function *cForest*,[15-18] in package "party"[15] available in R statistical software [14] was employed to generate classification models to discriminate between active and inactive compounds tested against tryptase-1. The number of trees grown was set to 500 and the number of randomly preselected variables has been varied for optimization. The response variable is predicted as an average vote of the predictions of all trees grown in the forest [17]. Thus, the predicted class probabilities are returned from the conditional distribution of the response [19].

The importance of the variables were assessed by the AUC variant of the variables important measure (VIM) implemented in the same package, i.e., "party". The AUC-based VIM has been demonstrated to be less-biased compared to the standard permutation VIM [20]. Here, we used function "varimpAUC" with the following parameters: replacement = false and permutations = 50.

Clustering

ChemAxon's Library MCS (version 6.1.0, 2013, http://www.chemaxon.com) was applied to perform a structure based clustering of the 103 tryptase-1 inhibitors. The program implements the maximum common substructure concept in a fast, hierarchical clustering method which identifies the largest substructure shared by several molecular structures.

Evaluation parameters

Well-established parameters were employed to evaluate the discriminative power of the models: sensitivity (*Se*, i.e., the ratio of correctly predicted actives or positives, to the total number of actives available), specificity (Sp, i.e., the ratio of correctly predicted inactives or negatives, to the total number of inactives available), accuracy (Acc, the ratio of correctly classified compounds in the dataset). The receiver operating curve (ROC) plot was generated based on the predicted probabilities computed by the random forest models. The area under ROC (AUC) was also computed to reflect the capacity of the models to separate actives from inactives without imposing a class-membership threshold value [21, 22].

Results and discussions

Model optimization

Standard tuning parameters in random forest are the number of trees grown (i.e., *ntree*) and the number of randomly preselected variables (i.e., *mtry*). The first parameter was set to 500, which we considered large enough for the number of samples available. In the case of *mtry,* the square root value of the number of predictor variables is commonly used. In this case, for the 210 pharamcophore features *mtry* would give a value of 15. Here, we aimed to optimize *mtry* and generated, using the training set, ten RF models by varying the value of *mtry* between 3 and 30 by a step of 3. The resulted models were evaluated internally by means of the 'out-of-bag' (OOB) accuracy. In RF models every classification tree is grown on random samples of compounds representing 63% of the available entries. The rest of the samples serve as an external data set which is predicted by the model and subsequently evaluated. The ensemble of these evaluations provides the OOB error, which is believed to

Parameter	Formula[a]
Sensitivity	$Se = TP / A$
Specificity	$Sp = TN / I$
Accuracy	$Acc = (TP + TN) / (A + I)$
Area under the receiver operating curve	$AUC = 1 - \frac{1}{A}\sum_{i=1}^{A} FPR_i$

Table 2
EVALUATION PARAMETERS USED TO MEASURE THE DISCRIMINATIVE CAPACITY OF THE MODELS

[a] TP = true positive, TN = true negative, FPR_i = false positive rate at the i^{th} active found in the ranked list, A = total number of actives and I = total number of inactives

provide a realistic image of the performance of the model. Here, the model constructed with *mtry* = 12 was superior to the others with evaluation parameters shown in table 2.

Model Vvalidation

The wining model obtained the in optimization step, i.e. *mtry* = 12, was further challenged to classify the 26 actives and 24 inactives in the external set. As shown in table 3,

Table 3
INTERNAL AND EXTERNAL EVALUATION RESULTS

Set	Se	Sp	Acc
Internal	0.885	0.889	0.887
External	0.960	0.870	0.917

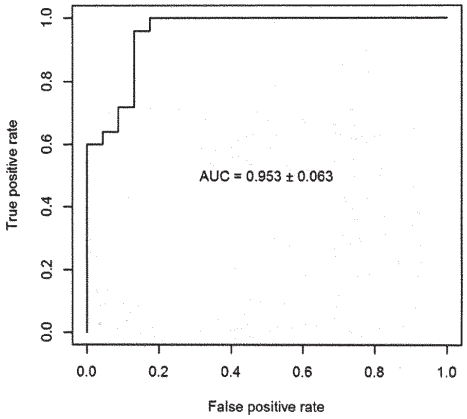


Fig. 2. External test set evaluation. Receiver operating curve (ROC) and the area under the ROC

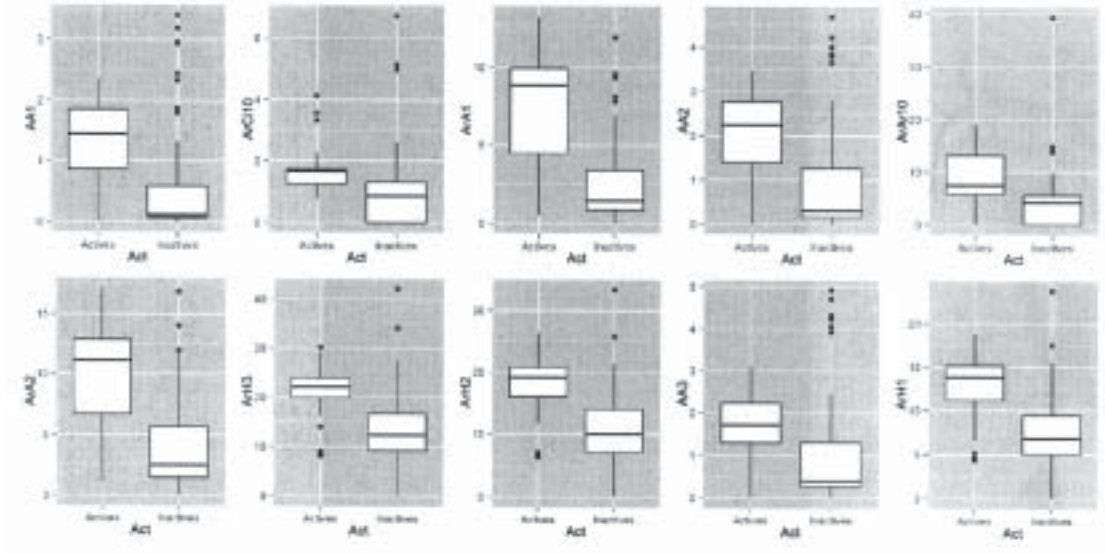
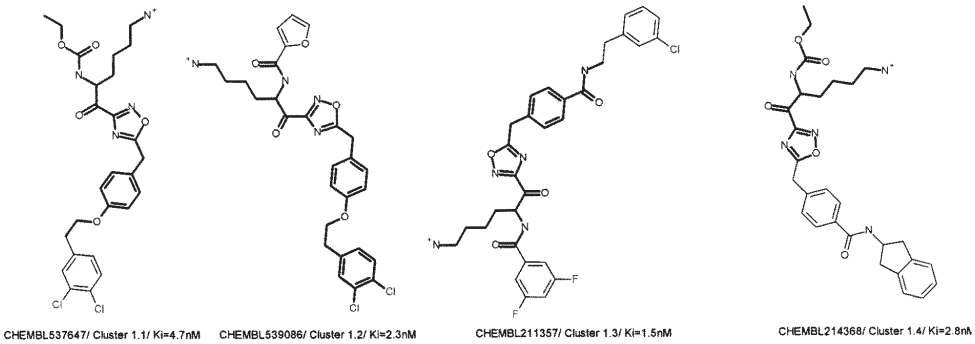


Fig. 3. Examples of descriptors which are able to discriminate effectively actives from inactives



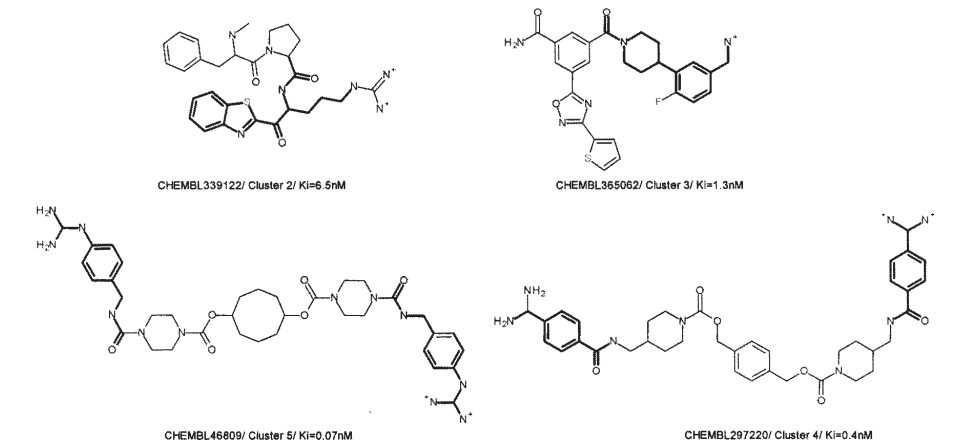


Fig. 4. Maximum common substructure clustering of the inhibitors of human mast cell tryptase used in the current study

the RF predictor successfully classified both actives (Se = 0.960) as well as inactives (Sp = 0.870) in a very similar percent as encountered in the internal evaluation. The ROC curve (fig. 2) also confirms a very good separation between the two classes based on the predicted class-probabilities, with an AUC value of 0.953 ± 0.063.

Variable importance

For each split in a decision, variables are selected only from a small random subset of the predictor variables which permits the computation of variable importance measures over the entire ensemble of trees. Stobl et al [17] and Janitza et al [16] demonstrated several advantages of AUC-based VIM implemented in package "party" over the original random forest VIM.

We generated a new random forest model using the entire set of compounds (training and test set) to study the importance of the variables. We computed the AUC-based VIM using 50 and report the first ten most important descriptors in figure 3. The distribution of the actives compared to the inactives reveal significant class-separation achieved by all descriptors. This finding underlines the efficiently of the Chemaxon fuzzy pharmacophore fingerprints in pursuing new molecules for drug discovery.

The diversity of the chemical structures of the tryptase-1 inhibitors is remarkably high but essential functionalities find correspondence in the important pharmacophore variables in figure 3, e.g., ArCi10 underlies the presence of a polar (basic) group, which interacts with Asp189 in the catalytic site (a specific interaction for almost all trypsin-like serine proteases [10]), the hydrophobic molecular fragment is also required as highlighted by the aromatic-hydrophobic (ArH) variables, a complex network of H-bond interactions is in the binding site is suggested by the presence of frequent acceptors (AA). In figure 4 we show the result of a clustering attempt to group tryptase-1 inhibitors based on the concept of maximum common substructure. We obtained five clusters (ignoring singletons) of size 65, 10, 12, 8 and 5. The above mentioned pharmacophore feature-pairs are present in the representatives shown fig. 4: α-keto-[1,2,4]-oxadiazoles [11], benzothiazole ketones [23], pseudo-bivalent dibasic ligands [24, 25] etc.

Conclusions

The hereby generated classifiers demonstrate the successful use of 2D fuzzy pharmacohore fingerprints in random forest modelling with applications in drug discovery. A variety of cheminformatics tools are applied to aid decision making in medicinal chemistry. The classification model developed herein is able to efficiently recognize trypase-1 inhibitors and support the discovery and optimization of novel molecules with therapeutic potential for treating allergic and inflammatory disorders as well as various types of cancers.

Acknowledgments This paper was published under the frame of European Social Found, Human Resources Development Operational Program 2007-2013, project no. POSDRU/159/1.5/S/136893 through the contribution of Stefana Avram. Sorin Avram is indebted to ChemAxon Ldt for providing access to their software.

References

1. RIBATTI, D.CRIVELLATO, E., Biochim. Biophys. Acta, **1822**, no. 1, 2012, p. 2.
2. ABRAHAM, S.N.ST. JOHN, A.L., Nat. Rev. Immunol., **10**, no. 6, 2010, p. 440.
3. HALLGREN, J.PEJLER, G., FEBS J., **273**, no. 9, 2006, p. 1871.
4. CAIRNS, J.A., Pulm. Pharmacol. Ther., **18**, no. 1, 2005, p. 55-66.
5. RIBATTI, D.RANIERI, G., Exp. Cell. Res., **332**, no. 2, 2015, p. 157-62.
6. AMMENDOLA, M., LEPORINI, C., MARECH, I., GADALETA, C.D., SCOGNAMILLO, G., SACCO, R., SAMMARCO, G., DE SARRO, G., RUSSO, E.RANIERI, G., Biomed. Res. Int., **2014**, 2014, p. 154702.
7. PEREIRA, P.J., BERGNER, A., MACEDO-RIBEIRO, S., HUBER, R., MATSCHINER, G., FRITZ, H., SOMMERHOFF, C.P.BODE, W., Nature, **392**, no. 6673, 1998, p. 306.
8. AVRAM, S., PACUREANU, L.M., SECLAMAN, E., BORA, A.KURUNCZI, L., J. Chem. Inf. Model., **51**, no. 12, 2012, p. 3169.
9. CRISAN, L., BORA, A., PACUREANU, L., AVRAM, S.L., K., REV. CHIM. (Bucharest), **63**, no. 5, 2015, p. 481.
10. LIANG, G., ALDOUS, S., MERRIMAN, G., LEVELL, J., PRIBISH, J., CAIRNS, J., CHEN, X., MAIGNAN, S., MATHIEU, M., TSAY, J., SIDES, K., REBELLO, S., WHITELY, B., MORIZE, I.PAULS, H.W., Bioorg. Med. Chem. Lett., **22**, no. 2, 2012, p. 1049.
11. LEE, C.S., LIU, W., SPRENGELER, P.A., SOMOZA, J.R., JANC, J.W., SPERANDIO, D., SPENCER, J.R., GREEN, M.J.MCGRATH, M.E., Bioorg. Med. Chem. Lett., **16**, no. 15, 2006, p. 4036.
12. AVRAM, S.I., CRISAN, L., BORA, A., PACUREANU, L.M., AVRAM, S.KURUNCZI, L., Bioorg. Med. Chem., **21**, no. 5, 2013, p. 1268.
13. "KUHN, M.C.F.W., J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A.; COOPER, T.; MAYER, Z.; KENKEL, B.; THE R CORE TEAM; BENESTY, M.; LESCARBEAU, R.; ZIEM, A.; SCRUCCA, L. Caret: Classification And Regression Training 2015, R Package Version 6.0-41, http://cran.r-project.org/package=caret"
14. R Development Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2012.
15. STROBL, C., HOTHORN, T.ZEILEIS, A., The R Journal, **1**, no. 2, 2009, p. 14.
16. STROBL, C., MALLEY, J.TUTZ, G., Psychol. Methods, **14**, no. 4, 2009, p. 323.
17. STROBL, C., BOULESTEIX, A.L., ZEILEIS, A.HOTHORN, T., BMC Bioinformatics, **8**, 2007, p. 25.
18. HOTHORN, T., BUHLMANN, P., DUDOIT, S., MOLINARO, A.VAN DER LAAN, M.J., Biostatistics, **7**, no. 3, 2006, p. 355.
19. BREIMAN, L., Machine Learning, **45**, no. 1, 2001, p. 5.
20. JANITZA, S., STROBL, C.BOULESTEIX, A.L., BMC Bioinformatics, **14**, no. 119, 2013, p. 1.
21. JIN, H.LING, C.X., IEEE Transactions on Knowledge and Data Engineering, **17**, no. 3, 2005, p. 299.
22. FAWCETT, T., Pattern Recognition Letters, **27**, no. 8, 2006, p. 861.
23. COSTANZO, M.J., YABUT, S.C., ALMOND, H.R., JR., ANDRADE-GORDON, P., CORCORAN, T.W., DE GARAVILLA, L., KAUFFMAN, J.A., ABRAHAM, W.M., RECACHA, R., CHATTOPADHYAY, D.MARYANOFF, B.E., J. Med. Chem., **46**, no. 18, 2003, p. 3865.
24. VAZ, R.J., GAO, Z., PRIBISH, J., CHEN, X., LEVELL, J., DAVIS, L., ALBERT, E., BROLLO, M., UGOLINI, A., CRAMER, D.M., CAIRNS, J., SIDES, K., LIU, F., KWONG, J., KANG, J., REBELLO, S., ELLIOT, M., LIM, H., CHELLARAJ, V., SINGLETON, R.W.LI, Y., Bioorg. Med. Chem. Lett., **14**, no. 24, 2004, p. 6053.
25. DENER, J.M., RICE, K.D., NEWCOMB, W.S., WANG, V.R., YOUNG, W.B., GANGLOFF, A.R., KUO, E.Y.L., CREGAR, L., PUTNAM, D.WONG, M., Bioorg. Med. Chem. Lett., **11**, no. 13, 2001, p. 1629.

Manuscript received: 15.06.2015