Software-aided workflow for predicting protease-specific cleavage sites using physicochemical properties of the natural and unnatural amino acids in peptide-based drug discovery.

Peptide cleavage sites prediction workflow

Tatiana Radchenko [1,2]¶*, Fabien Fontaine [2]&, Luca Morettoni[3]&, Ismael Zamora[1,2]*

[1]Pompeu Fabra University, Barcelona Spain

[2]Lead Molecular Design, S.L, Sant Cugat del Vallés, Spain

[3]Molecular Discovery Ltd, London, UK

*Corresponding author

E-mail: tatiana.radchenko01@estudiant.upf.edu (TR)

E-mail: ismael.zamora@leadmolecular.com (IZ)

¶These authors contributed equally to this work.

&These authors contributed equally to this work.

Abstract

Peptide drugs have been used in the treatment of multiple pathologies. During peptide discovery, it is crucially important to be able to map the potential sites of cleavages of the proteases. This knowledge is used to later chemically modify the peptide drug to adapt it for the therapeutic use, making peptide stable against individual proteases or in complex medias. In some other cases it needed to make it specifically unstable for some proteases, as peptides could be used as a system to target delivery drugs on specific tissues or cells. The information about proteases, their sites of cleavages and substrates are widely spread across publications and collected in databases such as MEROPS. Therefore, it is possible to develop models to improve the understanding of the potential peptide drug proteolysis. We propose a new workflow to derive protease specificity rules and predict the potential scissile bonds in peptides for individual proteases. WebMetabase stores the information from experimental or external sources in a chemically aware database where each peptide and site of cleavage is represented as a sequence of structural blocks connected by amide bonds and characterized by its physicochemical properties described by Volsurf descriptors. Thus, this methodology could be applied in the case of non-standard amino acid. A frequency analysis can be performed in WebMetabase to discover the most frequent cleavage sites. These results were used to train several models using logistic regression, support vector machine and ensemble tree classifiers to map cleavage sites for several human proteases from four different families (serine, cysteine, aspartic and matrix metalloproteases). Finally, we compared the predictive performance of the developed models with other available public tools PROSPERous and SitePrediction.

Introduction

Proteolytic enzymes play critical role in many processes including cell proliferation, immune response, cell death and others. [1] Protease cleavage of peptides is directed by short amino acid motifs, from two to eight amino acids around the scissile bond (site of cleavage, SoC). [2] This specific amino acid sequence is recognized by the active site of a given protease, but the efficiency of a proteolytic activity is also related to the structural properties of the SoC. Recently, it has been shown that the following additional factors to the peptide sequences influence on a cleavage such as unfolding events, allosteric effects, solvent accessibility, secondary structure of the sequence etc. [3]

Nowadays it is known that the new peptide-based drugs must be designed considering the localization of potential protease SoCs to make the compound less susceptible to protease reaction. Firstly, efforts are spent on evaluating a large numbers of lead compounds and their derivatives to reveal motifs recognized by proteases. Secondly, peptide sequences of potent drug candidates are typically chemically modified to eliminate these motifs or to incorporate structural features preventing proteolysis [4]. Various chemical strategies have been developed to overcome the low availability issue of peptides, including cyclization, changing the stereochemistry of an amino acid, substitution of natural amino acids by unnatural ones and others. [5] These changes are applied during the design-make-test drug discovery cycle, with hopes of improving the physicochemical and pharmacokinetics properties of the compound of interest. Therefore, it is crucial to evaluate these properties rapidly in early development and suitable in vitro techniques providing reliable predictions of in vivo performance are required.

The information about proteases and their sites of cleavage is widely spread across publications and databases such as MEROPS [6], CutDB [7] and Proteasix [2]. The MEROPS database integrates available information about proteolytic sites and, consequently, proteases from different organisms, their experimentally identified or predicted SoCs with their sequences, peptidase substrates and inhibitors. This information can be used to develop models to identify possible labile residues in the candidate peptide for individual proteases. Although useful, these databases still have several limitations. For example, none of the available resources allow to add new information in an automatic way to enrich the database information.

This article presents a new approach which aim is to derive cleavage site appearance rules for the specific peptide family or for specific experimental condition (i.e. individual protease). The system considers the following steps: firstly, the information coming from liquid chromatography mass spectrometry (LC-MS) based experimental data or from external sources such as MEROPS database is stored in a chemically aware database (i.e. WebMetabase [8]). We represented each amino acid as a vector of physicochemical properties. We used validated Volsurf molecular descriptors [10] to describe these properties. Finally, the cleavage site-specific descriptors are calculated as a combination of the individual amino acid descriptors for the residues surrounding the cleavage site. Each stored peptide substrate and cleavage site are described as a combination of pharmacophoric and physicochemical properties of each amino acid contained in the sequence. This database can be used to perform frequency analysis (FA) to discover the most frequent SoC [9] within the experimentally derived and/or public database. FA results can be used to create a set of empirically derived rules based on molecular properties of the cleavage sites. At the next step, results of the frequency analysis for several individual proteases were used to train a Logistic Regression (LR), Support Vector Machine (SVM) and Ensemble Trees (ET) classifier prediction

models. We trained models for eighteen proteases from four protease families: serine, cysteine, aspartic and matrix metalloproteases.

Nevertheless, the proposed methodology could be applied in the case of non-natural amino acid and/or cyclic peptides. Moreover, since the system used to derive the cleavage site appearance rules (frequency analysis, FA) could be linked to the software assisted metabolite structure elucidation based on MS data, the database is automatically enriched with the new experiments. Rules can then be refined to tune the system for the experimental conditions and/or peptide families of interest. This knowledge can be applied during the design-make-test drug discovery cycle.

Materials and Methods

In-silico MEROPS dataset

In this study we exported peptide cleavage data from the MEROPS database (version 11 1/09/2017) [6] for all the available proteases in MySQL format using our own script that converts each organism/protease/peptide information in a single xml file inside a folder data structure. This data consists of substrate sequences, the cleaving protease, and the cleavage sites. Each peptide sequence was converted to a chemical structure and saved in MDL mol format. [33] For our analysis, we focused on peptides having less than 200 amino acids. Several peptides were excluded from analysis because it was not possible to restore their chemical structure. The reasons could be, for example, the lack of an appropriate defined structure for some encoded abbreviations of amino acids in a sequence.

Two different strategies were followed depending if the source organism for the protease was indicated in the MEROPS database or it was not. The summary information extracted from the MEROPS database includes: substrate peptide sequence, parent sequence peptide range, substrate Uniprot number, cleavage site position in substrate, metabolite peptide sequence, protease name, protease source organism and references list related to the protease. All the information was collected and written in a xml file. The xml has four blocks of information:

1. Properties: The Matrix (name of the protease), the MEROPS Peptidase Code, the MEROPS Peptide Uniprot_acc, the MEROPS Peptide Range, the MEROPS Organism and the MEROPS Peptidase Reference.
2. Parent: the parent structure in sdf format, the InchiKey, the molecular formula and the m/z.
3. Metabolized Parent: as the previous one without the InchiKey.
4. Metabolites: There is a new block for each metabolite. It contains the metabolite name and the structure. Each structure has the structure ID inside the xml file, the sdf (V3000 or V2000 format), the formula, the m/z, the formula difference compared to the parent structure, the metabolic mechanism, the m/z difference compared to the parent, the collection of atoms in the parent that suffer the metabolic transformation and the list that contains the isomorphism between the number of the atom in the parent and the number of atom in the metabolite.

After the xml files were imported into WebMetabase the same steps to annotate the information for each file that were described elsewhere for the case of experimentally derived approach were followed [10]. Enabling in this way the combination of the external and internal sources of information.

WebMetabase

All files extracted from MEROPS database were uploaded into the web application "WebMetabase 3.2.12" (Molecular Discovery Ltd, Middlesex, UK). WebMetabase imports the data from external sources in a predefined xml format where the structure of the peptides introduced in MDL mol format (V3000 or V2000). This conversion to a structure format is necessary to feed the chemically aware system. A complete list of uploaded substrates for each protease grouped by organism can be found at http://webmetabase.com:8182/WebMetabaseBioinformatics/.

Peptide database and search algorithm based on similarity.

Once the experimental results and/or data from external database were interpreted and approved in WebMetabase, the parent and metabolite structures were stored in the database. Each peptide structure was annotated by the Structural Blocks (SBs) defined as the structural fragment between amide bonds and their connectivity. SBs in the substrate sequence were numbered as …-P4-P3-P2-P1-/-P1'-P2'-P3'-P4'-…, with the cleaved bond located between the P1 and P1' sites. [12, 13]

In this article we described a new approach implemented in WebMetabase to store each peptide described as a combination of physicochemical properties of each SB contained in the sequence. These physicochemical properties of the SBs were represented through molecular descriptors calculated by the cheminformatics library Volsurf. [9] Therefore, each SoC P1-P1' was also characterized as a combination of physicochemical properties of the two SBs. Moreover, similarity matrix for Volsurf descriptors was calculated for all SB stored in the database.

The annotation of the peptide information in this manner enables doing an exact substructure and similarity-based substructure search inside the database without being limited to any type of peptides and/or amino acid (cyclic/linear, natural/synthetic). This methodology can be used to perform a structural search on the exact sequences identifying the experiments where a certain bond of interest participated in a metabolic reaction. Also, the described algorithm implemented in WebMetabase includes a system to perform searches based on similarity of the molecular descriptors. As an output, searches give a list of experiments that fulfill the search criteria. [10]

Volsurf descriptors

We used the cheminformatics toolkit Volsurf [9, 14] to calculate the physicochemical properties for the individual SBs. Peptide descriptors are then built from the descriptors of the SB in the peptide. Volsurf+ creates 128 molecular descriptors from 3D Molecular Interaction Fields (MIFs) produced by our software GRID [15, 16]. The Volsurf descriptors summarize the MIF information in a few variables easy to understand and to interpret. Basically, it quantitatively characterizes size, shape, polarity, hydrophobicity and the balance between them.

Frequency analysis

The algorithm can perform frequency analysis (FA) based on exact match of the sequence of interest (simple frequency analysis, sFA) and/or on the similarity based on the molecular descriptors (similarity frequency analysis, SFA). A simple peptide frequency analysis of the cleaved amide bonds can be performed for the entire database or for the selected set of the approved experiments in WebMetabase. The algorithm collects information about all cleaved

peptide bonds that were involved in the metabolic reactions of interest. Herein site of cleavage (SoC) considers the two SBs involved in the amide cleavage containing the C-terminal and N-terminal of the SBs. We define a potential SoC (pSoC) as the pair of structural blocks (referred as P1 and P1') that may and may not be involved in the catalysis. The frequency analysis refers to the number of times that a pSoC is observed in the parent structure and how many times this is an actual SoC (aSoC).

The output of the frequency analysis is done by protease and provides two tables. The first table contains the counts of a pSoC found and the counts that it was an aSoC. The second table contains the counts a pSoC found and if it was cleaved once, twice or more than twice. A frequency analysis of the aSoCs depending on the protease can be done to create a set of empirically derived rules that later can be used to predict the metabolic liability of different amide bonds in a new non-tested peptide. WebMetabase automatically generates an excel file in temporary folder to store FA results. Filename format starts with string "soc" and after some random number, e.g. soc7511073558888689580.xls. This file can be used for the following analysis if needed. Moreover, frequency analysis results can be persisted in the database for the further usage.

Model implementation

The information registered in the database and FA results based on MEROPS dataset for eighteen proteases were selected to make some examples of the methodology. These proteases from four different families (serine, cysteine, aspartic, matrix metalloproteases) from Homo sapiens were chosen to develop classifier models. They included granzyme A, granzyme B, granzyme B (rodent-type), granzyme M, thrombin, trypsin1, caspase-1, caspase-2, caspase-3, caspase-6, caspase-7, cathepsin E, cathepsin D, cathepsin L, MMP2, MMP3, MMP8, MMP9.

Trypsin1, MMP2, caspase-6 and cathepsin L were selected because the number of the identified aSoCs was the highest for proteases imported in WebMetabase. Others were selected as a representative of the respective protease family. For each selected protease we split all extracted substrates in train and external dataset (ED). Train dataset contained experiments with 80% of extracted cleavages events and was used to develop models and to perform the cross-validation. ED was used only as a final assessment of model predictive performance quality for each protease and to perform comparison with public available PROSPERous [17] and SitePrediction [18] approaches. In accordance with the literature it was revealed that increasing of the investigated window's size around cleavage site could improve the performance of the model [17]. Due to this fact, all classifier models were built using two sequence windows, consisting of the P1-P1' (SoC2) and P4-P4'(SoC8). SoC2 considers the two SBs P1 and P1' involved in the amide cleavage containing the C-terminal and N-terminal of the SBs. SoC8 considers the eight SBs on positions P4-P4'. S1 Table summarizes the information regarding selected proteases, protease family, number of selected substrate peptides, number of cleavage events exported from MEROPS, number of unique aSoCs from two amino acids (P1-P1'), number of unique aSoCs from eight amino acids (P4-P4') and number of substrates excluded from the exported dataset to perform external validation (EV) of the models.

To derive a training and external dataset for each of the selected proteases, we cut out all possible 2 SBs and 8 SBs sequences from substrate peptides. Thus, a substrate with 100 amino acids would lead to 93 and 99 sequences of length 8 and 2, respectively. Each sequence than was characterized as a combination of physicochemical properties described by Volsurf descriptors of the SBs in the sequence. These sequences were then classified as either cleaved (aSoC2 and aSoC8) or non-cleaved sequences (pSoC2 and pSoC8). Cleaved sequence for aSoC2 was the one

involved in metabolic reaction. In cleaved sequence for aSoC8 the proteolytic event happened between fourth and fifth residue. Substrate cleavage site prediction was formulated as a binary classification problem for two classes cleaved SoCs so-called positive examples or non-cleaved SoCs so-called negative examples. We calculated number of times each pattern was cleaved and number of times it was met. For the SoC2 it was done during sFA in WebMetabase, for the SoC8 additional script was applied. Since the size of positive dataset was significantly smaller than the number of non-cleaved motifs, we split all negative dataset in groups with number of objects equal to the number of cleaved patterns, than combined positive dataset with each of the received negative subsets, therefore obtained balanced datasets to derive models. In details, we decomposed the non-cleaved objects dataset into X partitions, where X depended on the number of cleaved sequences and combined all the metabolized SoCs with each partition of pSoCs to be an individual subset. Thus, we generated multiple training subsets with better class distribution and each non-cleaved object occurred in at least one training subset (Fig 1). Moreover, we performed directed oversampling for aSoCs and pSoCs to give a weight to the motifs that appeared in the dataset and were metabolized more frequently [28, 29]. We gave them an over representation in the dataset for certain number of times. This number was calculated in accordance with the following formula (2). S2 Table summarizes the information regarding number of subsets for each protease.

$$N = (M * 11) \div (100 + M) \quad (2)$$

N is the number of times sequence pattern will be repeat in a final dataset. M is the number of times the sequence pattern was met in the initial exported dataset.

Figure 1. Architecture of the classifier ensembles;

The next step, for each protease we applied multiple learning algorithms to each training subset of sequences described by Volsurf descriptors to train prediction models independently. In this study, we used Logistic regression (LR) [16, 22], Support Vector Machine Classifier (SVC) [17, 22] and Random Forest Classifier (RFC) [18, 22] and Gradient Boost Classifier (GBC) [19, 22] to build the models to estimate the cleavage site probability of substrates for selected proteases. For SVC we selected radial basis kernel function (RBF) [20, 22] and applied Grid Search CV [21, 22] to search best parameters of the classifier. For all other classifiers default parameters were used.

While models were generated, we performed 5-fold cross-validation (5CV). The predictive performance of each model was evaluated using the accuracy, sensitivity, specificity, area under the curve (AUC) for the receiver–operator characteristic curve (ROC) using these measures based on 5-fold cross-validation. Finally, we combined all models measures for each learning algorithm by calculating the mean for each measure.

To evaluate the model's performance in a real scenario and to inspect the potential over-fitting effect on these algorithms, we also performed prediction on EDs, peptides that had not been included in the model building. Prediction was performed for each peptide from the external dataset independently by each model. To complete this goal each peptide from external dataset also was split in motifs and each motif was characterized as a combination of physiochemical properties of its SBs. To combine the results from all models for each learning method we calculated a mean of predicted probabilities for each bond in the peptide, using the predicting bond breaking probability we evaluated ranking performance for each classifier. The procedures are described in detail below.

Finally, we performed independent tests and compared the predictive performance by ranking on ED between selected LR and RFC model classifiers and the other public available tools such as SitePrediction and PROSPERous.

Evaluation of model performance

Performance evaluation comparison between different classifiers

We performed comprehensively evaluation and comparison of the predictive performance between different classifiers (LR, SVC, RFC and GBC) on the external dataset (ED) in each protease family combining the prediction for all proteases inside the family. Based on the intermediate results presented below we decided to proceed with classifiers using local window P4-P4'. To complete this task, we cut out each external peptide in subsequences of length 8, calculated Volsurf descriptors for each of the SB in this peptide and characterized each sequences of 8 SBs as a combination of physicochemical properties of SBs contained in the subsequence. We used these combinations of Volsurf descriptors to perform prediction on each of them. Since the final goal of the model was to select which of the bonds was more likely to be broken in a given sequence, an analysis of the rank (sorted by the predicted SoC) was performed. Therefore, when the probability for each bond was computed by the model, we sorted each sequence by their value and registered a rank of each predicted cleavage site. Since peptides differ in a size ranking was normalized by the amount of possible cleavage sites. A recovery analysis of the rank value for the known SoC was done and finally we accumulated all ranking positions for each external peptide and each protease inside the family. Moreover, we calculated the best and the random ranking. The random ranking was done by assigning a random number for the probability to be broken for each SoC and continuing in the same way as in the model case. In the case of the best ranking positions, the known

SoCs in a peptide were assigned to 1,2,3, etc. rank for each of the known SoCs. As a summary of the results, we plotted the cumulative ranking score for the recovered known SoCs for each family including the best and the random cumulative ranking.

Performance evaluation comparison with other prediction tools (SitePrediction and PROSPERous)

We performed comprehensively evaluation and comparison of the predictive performance between selected LR and RFC model classifiers and other public available tools such SitePrediction and PROSPERous. Comparison with the external methodologies was performed only for the proteases in each family where models were available. Models for trypsin1, granzyme M and caspase-2 were not available in both methods, also for granzyme A, granzyme B, granzyme B (rodent-type) and thrombin were not available in SitePrediction approach. To complete this task, we used as an input for SitePrediction and PROSPERous FASTA format sequences of the peptides in the external dataset and exported prediction results. In SitePrediction we selected models trained for each protease for Homo sapiens. In PROSPERous we selected logistic regression models and local window P4-P4'. After each potential cleavage site was sorted in accordance with the predicted scores from each method. For the SitePrediction case where a score was not predicted it was assigned to 0. After we followed the same rank strategy explained for the models developed in this article, registering the rank of each predicted cleavage site and normalized it by the amount of possible cleavage sites for each peptide. Finally, we plotted the cumulative ranking score for each family for our models and for each method including the best and the random cumulative ranking.

Results and Discussion

In this section, we present the results of the application of our approach in the analysis of the MEROPS. All files extracted from MEROPS database were uploaded into the web application "WebMetabase 3.2.12" (Molecular Discovery Ltd, Middlesex, UK). WebMetabase imports the data from external sources in a predefined xml format. Metabolites structures extracted from MEROPS were directly uploaded into WebMetabase and automatically approved. After approval parent peptide and metabolites were persisted into the database and annotated as described before. Finally, the methodology was applied to perform a sFA on the stored information for selected proteases: granzyme A, granzyme B, granzyme B (rodent-type), granzyme M, thrombin, trypsin1, caspase-1, caspase-2, caspase-3, caspase-6, caspase-7, cathepsin E, cathepsin D, cathepsin L, MMP2, MMP3, MMP8, MMP9. The sequence pattern around the potential cleavage site was represented as a combination of Volsurf descriptors. Moreover, an additional script was applied on the database to perform sFA for the sites of cleavage with the local window P4-P4'. The results from both FA were used to build predictive models for the selected proteases enabling the cleavage site prioritization based on the physicochemical properties for the cleavage sites. The MEROPS/WebMetabase workflow used is shown in Fig 2.

Figure 2. MEROPS/WebMetabase workflow from experimental data to searchable information manageable by in silico analysis tools.;

In total we extracted 18760 xml files and they contained information for 18760 substrate peptides and 21804 metabolites. For each selected protease we split all extracted experiments in train and external dataset. Train dataset contained experiments with 80% randomly selected of the extracted identified cleavages events and it was used to develop models and perform cross-validation. External dataset was used to evaluate the model performance and compare it with other

methods. In both datasets each SoC was represented as a combination of the Volsurf descriptors of the SBs in the sequence. We performed a simple frequency analysis of the metabolized chemical moieties for each protease in the train and external datasets. The summary on the amount of identified individual metabolized bonds by frequency analysis is presented in S3 Table. Full information of identified cleavage sites frequency can be found at http://webmetabase.com:8182/WebMetabaseBio informatics.

Using the prepared substrate training datasets, we analyzed the statistical distributions of cleavage sites for the eighteen proteases analyzed (S1 Table). The amino acid occurrences in the P4-P4' positions for the SoCs of all selected proteases were calculated to generate heat map. The heat maps for the selected caspases and trypsin1 are presented as an example in Fig 3, for all other proteases they are provided in S2-4 Figs. These heat maps help to identify conserved and frequently occurring amino acids at scissile bonds (Fig 3, S2-4 Figs). For example, as shown in Fig 3, one of the main characteristics of caspase-3 specificity is that this protease preferentially cleaves after the aspartic acid (Asp, D) at P1 and P4 position. We identified Asp in position P1 in 100% of analyzed SoCs and in position P4 in 57.1%. Also, we noted that glutamic acid (Glu, E) presented in 28.6% at P3 and valine (Val, V) presented at position P2 in 30.95%. Also, glycine (Gly, G) at P1' in 30.95% and alanine (Ala, A) in 23.8% or leucine (Leu, L) in 21.4% at position P2'. Therefore, a good substrate for caspase-3 should contain sequence like D-E-V-D+G-A or D-E-V-D+G-L. These results agree with literature where it was described caspase-3 preferentially cleaved at D-E-V-D+X-X-X [28, 29], Comparison of different caspases reveals differences between preferential cleavage motifs, for example, phenylalanine at P2' for caspase-1, serine at P1' for caspase-2 and caspase-6, threonine at P2 and leucine at P4' for caspase-7. These results also agreed with specificity matrix presented in MEROPS DE-E-V-D+GS---. This indicates that,

although we reduced our dataset by selecting only peptides of small size limited to 200 of amino acids, we still got representative dataset of substrates for each protease.

In addition, for trypsin1 we identified arginine (Arg, R) in position P1 in 36.8% and lysine (Lys, K) in 63.2%. These results agreed with literature where it was described trypsin preferentially cleaved at Arg and Lys in position P1 [30] and with MEROPS specificity matrix.

Figure. 3. Substrate cleavage sites for P4–P4' position for caspases: a) caspase-1; b) caspase-2; c) caspase-3; d) caspase-6; e) caspase-7; f) trypsin1. The cumulative amino acid occurrences in P4–P4' were calculated and displayed in the form of a two-dimensional heat map. The scissile peptide bond was between sites P1 and P1'.

To evaluate the performance of the selected classifier models for cleavage site prediction of multiple proteases, we carried out a 5-fold cross validation (5CV) test on each of selected proteases under investigation in this study. Moreover, the predictive performance on several external datasets was explored. Firstly, we trained and compared the predictive performance of models with two learning methods LR and SVC using two sizes of local windows P1-P1' and P4-P4' For evaluation of predictive performance during 5CV test the following criteria were calculated: accuracy, ROC AUC, sensitivity and specificity. For evaluation of the predictive performance on external dataset we calculated the accuracy, sensitivity and specificity. Fig 4 summarizes results for the predictive performance evaluation for LR and SVC classifiers trained using local windows P1-P1' and P4-P4' for 5CV and external dataset validation.

Figure 4. Results for the predictive performance evaluation for LR and SVC classifiers trained using local windows P1-P1' and P4-P4' for 5CV and external dataset validation: a) predictive performance for LR classifier evaluated by 5CV; b) predictive performance for

SVC evaluated by 5CV; c) predictive performance for LR classifier evaluated on external dataset; d) predictive performance for SVC evaluated on external dataset.

For caspases-1, 3 and 7 all performance measures decreased for both classifiers when window was widened, on the other hand for caspases-2 and 6 they increased. For example, for caspase-1 and 2 comparing LR classifier performance based on AUC for the 5CV results for window P1-P1' was 0.89 and 0.88 and for P4-P4' was 0.75 and 0.95, respectively. Comparing results for LR and SVC specificity was higher for LR classifier in both P1-P1' and P4-P4'. Meanwhile, when we compared predictive performance on external dataset of the models trained using two local windows accuracy and sensitivity for P4-P4' were higher for caspases-2, 3, 6, 7 and cathepsin L in both classifiers. For caspase-1 accuracy was lower that also could be related with small amount of training data. It can be explained by the fact that models trained for window P4-P4' are more complex and involve higher number of variables to be defined during the training process. We suppose that because the amount of data is not sufficient to properly train the model for window P4-P4' for caspases-1, 3, and-7. This theory is also confirmed by the fact that accuracy, specificity and sensitivity drastically increased when P4-P4' dataset was used to train both learning methods for cathepsin L. Based on this analysis results we decided to continue using local window sequence P4-P4' and excluded SVC learning method from analysis.

In addition, we compared the difference of amount of recovered known cleavage sites for all selected proteases for the P1-P1' and P4-P4' LR models for the 10 top positions. To complete this task, we calculated a difference and normalized it by total amount of known cleavage sites. Results are shown in Fig 5. P4-P4' models recovered higher amount of known SoCs for each of the all top 10 rank positions.

Figure 5. Comparison results of accuracy for all investigated classifiers for caspases using local window P4-P4'

Different learning approaches methods were used to train the models in other tools [17, 31-33]. For example, in the PROSPERous approach the best performing models are based on logistic regression. [17] At the next step, we trained and compared predictive performance of models with the following learning methods: LR, RFC and GBC. Fig 6 summarizes information regarding accuracy evaluated by 5CV test for all investigated classifiers comparing predictive performance of these models for caspases using window around cleavage site P4-P4'. Both ensemble tree classifiers outperformed logistic regression for all caspases just by comparing accuracy. All other metrics (AUC, sensitivity and specificity) for all investigated classifiers comparing predictive performance of these models for caspases using window P4-P4' are shown in S4 Table. In addition, predictive performance of RFC and GBC models was equal for caspases-2 and 6, meanwhile RFC demonstrated higher accuracy for caspases-3 and 7 and accuracy for GBC model was higher for caspase-1 that also can be related with the size of the training dataset.

Figure 6. Comparison results of accuracy for all investigated classifiers for caspases using local window P4-P4'

Lastly, we evaluated predictive performance of LR and RFC models on external dataset to understand if trained models were able to predict known cleavage sites with the top-ranking positions for the protease of interest. To complete this task, we explored the ranking positions of the probable cleavage sites sorted by predicted probability for each learning approach. Since peptides differ in a size, ranking was normalized by the amount of possible cleavage sites. Therefore, we completed a recovery analysis of the rank value for the known SoCs and calculated

the sum of all ranking positions for each external peptide for each protease inside the family. Moreover, we calculated the best and random ranking. The cumulative percentage ranking score for the recovered sites of cleavage for caspases is demonstrated in Fig 7.

Figure 7. The cumulative ranking score in percentage for the recovered known sites of cleavage for cysteine proteases in range of ranking positions from 0 to 100%

We noted that the best ranking recovered 100% of known SoC reaching 20% of the normalized ranking positions. Fig 8 summarizes percentage recovered by each learning algorithm at 20% of the normalized ranking positions.

Figure 8. The cumulative ranking score in percentage for the recovered known sites of cleavage for cysteine proteases in range of ranking positions from 0 to 20%

All classifiers were able to recover all known cleavage sites better than random but at lower percentage of ranking positions than the best. LR models were the best considering they recovered all known SoCs at 21% of the ranking positions, while the best recovered known sites of cleavage at 20%. RFC recovered all SoCs from external dataset for caspases at 60% of the ranking positions, accordingly. Moreover, we compared percentage of recovered known SoCs at the percentage ranking positions when the best ranking collected 100% of known sites of cleavage for each protease family selected for this study, results are presented in S5 Table. For all protease families except of cysteine the highest percentage of correctly predicted cleavage sites was reached by GBC models and the lowest by SVC. S6 Table summarizes the normalized ranking positions reached when 100% of known cleavage sites were recovered for all classifiers for all selected protease families.

Finally, we compared the predictive performance by ranking on external dataset between LR and RFC classifiers and other tools such as SitePrediction and PROSPERous. For following proteases predictive performance was compared with PROSPERous and SitePrediction: caspase-1, caspase-3, caspase-6, caspase-7, cathepsin E, cathepsin D, cathepsin L, MMP2, MMP3, MMP8, MMP9. Also, PROSPERous contained models trained for serine proteases: granzyme A, granzyme B, granzyme B (rodent-type), thrombin. Our LR and RFC models as well as PROSPERous correctly predicted 376 known cleavage sites, while 306 known cleavage sites were correctly predicted by both PROSPERous and SitePrediction. Fig 9 summarizes the cumulative percentage score for the recovered known sites of cleavage between our LR and RFC models compared to PROSPERous and SitePrediction performance on external dataset. Detailed summary of comparison performed for each protease family is shown in S1 Fig.

Figure 9. The cumulative ranking score in percentage for the recovered known sites of cleavage for selected protease families compared with PROSPERous and SitePrediction for two ranges of ranking positions percentage. a) comparison results with PROSPERous and SitePrediction for 0-100% range of ranking positions percentage; b) comparison results with PROSPERous and SitePrediction for 0-20% range of ranking positions percentage; c) comparison results with PROSPERous only for 0-100% range of ranking positions percentage; d) comparison results with PROSPERous only for 0-30% range of ranking positions percentage;

Our models were able to collect the same amount of known cleavage sites as PROSPERous and SitePrediction. SitePrediction is based on statistical scoring methods, while PROSPERous is based on SVM learning and LR approach. In the dataset of proteases compared to SitePrediction

and PROSPERous LR and RFC models recovered higher percentage of known SoCs at the ranking percentage position equal to 30% but reaching 60% the percentage of recovered known sites of cleavage was the same for our models, PROSPERous and SitePrediction. In the dataset of proteases compared only to PROSPERous RFC model performed worse than PROSPERous and recovered less percentage of known cleavage sites. On the other hand, LR model recovered higher percentage of known SoCs at the ranking percentage position equals to 30% but reaching about 60% of ranking position percentage of recovered known sites of cleavage was the same for LR and PROSPERous. While we performed detailed analysis of ranking position percentage between 0 and 30% LR outperformed RFC and PROSPERous, while last ones performed at the same level. We can suppose that LR model works better for the peptides of the smaller size.

Conclusions

Predicting possible sites of cleavage for individual proteases is an important task to be completed during drug-design process of peptide therapeutics to improve their stability and availably as a promising drug. In this study we presented a new approach in WebMetabase that helps to predict cleavage site for the specific peptide family or for specific experimental condition (i.e. individual protease). One of the main advantages of this approach is that it generates a searchable database for the information coming from LC-MS based experimental data or from external sources such as MEROPS database. In this database each amino acid is described as a vector of physicochemical properties, Volsurf molecular descriptors. Thus, the sequence pattern around the potential cleavage was represented as combination of Volsurf descriptors. The proposed methodology can be applied in the case of non-natural amino acid. Comparing to MEROPS this database type can be enriched with new experimental or external data. This way to store the data

can be utilized to perform frequency analysis to discover the most frequent scissile bonds within the generated database. The FA results can be used to derive a cleavage site appearance rules based on molecular properties of the cleavage sites. To demonstrate this, we trained several models using Logistic Regression (LR), Support Vector Machine (SVM) and Ensemble Trees (ET) classifier learning approaches for eighteen proteases from four protease families: serine, cysteine, aspartic and matrix metalloproteases. In the training dataset each sequence pattern around the potential cleavage site and actual site of cleavage was represented as a combination of Volsurf descriptors that characterized physicochemical properties of the SBs in the sequence. We compared predictive performance of the models trained with different learning approaches applying 5-fold cross validation test and external dataset validation test. Moreover, we examined the influence of the local window sequence size around the site of cleavage by comparing the models trained for P1-P1' and P4-P4' range. We revealed that LR and RFC models trained using window P4-P4' outperformed other learning methods and the models trained using P1-P1' window. We noted that training dataset size influenced on the predictive performance of the models analyzing data for caspases. Finally, we compared the predictive performance of LR and RFC models with other approaches such as PROSPERous and SitePrediction tools. LR model recovered higher percentage of the known cleavage site in the first 30% of the ranking positions comparing to the other tools. It can be explained by the fact that it performs better prediction on smaller peptides. Moreover, since the system can be linked to the software assisted metabolite structure elucidation based on MS data, the database is automatically enriched with the new experiments. Nevertheless, models can be re-trained with updated dataset and derived rules can be refined to tune the system for the experimental conditions and/or peptide families of interest. This knowledge can be applied during the design-make-test drug discovery cycle.

Acknowledgments

This work has been partially supported by Doctorats Industrials, AGAUR, Generalitat de Catalunya. We would like to thank Xavier Pascual from Lead Molecular Design for the programming support and Dr. Manuel Pastor from Barcelona Biomedical Research Park, Universidad Pompeu Fabra for his contributions and help in statistical analysis discussions.

References

1. Lopez-Otin C, Bond JS. Proteases: multifunctional enzymes in life and disease. J Biol Chem. 2008;283(45):30433–7. Epub 2008/07/25. pmid:18650443; PubMed Central PMCID: PMC2576539
2. Klein J, Eales J, Zurbig P., Vlahou A, Mischak H, Stevens R. Proteasix: a tool for automated and large-scale prediction of proteases involved in naturally occurring peptide generation. Proteomics. 2013 Apr;13(7):1077-82. doi: 10.1002/pmic.201200493. Epub 2013 Feb 26
3. Belushkin AA, Vinogradov DV, Gelfand MS, Osterman AL, Cieplak P, Kazanov MD. Sequence-derived structural features driving proteolytic processing. Proteomics. 2014;14(1):42–50. Epub 2013/11/15. pmid:24227478.
4. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug Discov Today. 2010;15(1/2):40-56. doi: 10.1016/j.drudis.2009.10.009.

5. Werle M, Bernkop-Schnurch A. Strategies to improve plasma half life time of peptide and protein drugs. Amino Acids. 2006; 30: 351–367. https://doi.org/10.1007/s00726-005-0289-3 PMID: 16622600
6. Rawlings ND, Waller M, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. 2014; 42:503-509. doi: 10.1093/nar/gkt953
7. Igarashi Y, Eroshkin A, Gramatikova S, Gramatikoff K, Zhang Y, Smith JW, et al. CutDB: a proteolytic event database. Nucleic Acids Res. 2007; 35:546–549. doi: 10.1093/nar/gkl813
8. Brink A, Fontaine F, Marschmann M, Steinhuber B, Cece EN, Zamora I, et al. Post-acquisition analysis of untargeted accurate mass quadrupole time-of-flight MS(E) data for multiple collision-induced neutral losses and fragment ions of glutathione conjugates. Rapid Commun Mass Spectrom. 2014;28(24):2695-703. doi: 10.1002/rcm.7062
9. Cruciani G., Pastor M., Guba W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. (2000) European Journal of Pharmaceutical Sciences 11 Suppl. 2 (2000) S29–S39
10. Radchenko T, Brink A, Siegrist Y, Kochansky C, Bateman A, Fontaine F, et al. (2017) Software-aided approach to investigate peptide structure and metabolic susceptibility of amide bonds in peptide drugs based on high resolution mass spectrometry. PLoS ONE 12(11): e0186461. https://doi.org/10.1371/journal.pone.0186461
11. Dalby A, Nourse J. G, Hounshell W. D, Gushurst A. K. I, Grier D. L, Leland B. A, et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. Journal of Chemical Information and Modeling. 1992;32 (3): 244.doi:10.1021/ci00007a012.

12. Schechter I, Berger A. On the size of the active site in proteases. I. Papain. Biochem Biophys Res Commun. 1967 Apr 20;27(2):157-62. PMID: 6035483
13. Schechter I, Berger A. On the size of the active site in proteases. 3. Mapping the active site of papin; specific peptide inhibitors of papain. Biochem Biophys Res Commun. 1968 Sep 6;32(5):898-902. PMID: 5682314
14. Cruciani G, Crivori P, Carrupt P.-A, Testa B. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. Journal of Molecular Structure (Theochem). 2000; 503: 17–30
15. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J. Med. Chem. 1985, 28, 849-857.
16. Boobbyer, D. N, Goodford, P. J, McWhinnie, P. M, Wade, R. C. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. J. Med. Chem. 1989, 32, 1083-1094
17. Song J, Li F, Leier A, Marquez-Lago T. T, Akutsu T, Haffari G, et al. PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy, Bioinformatics. 2018 Feb 15; 34 (4): 684–687. https://doi.org /10.1093/bioinformatics/btx670
18. Verspurten J, Gevaert K, Declercq W, Vandenabeele P. SitePredicting the cleavage of proteinase substrates. Trends Biochem Sci. 2009;34(7):319-23. doi: 10.1016/j.tibs.2009.04.001.
19. Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering 30.1 (2006): 25-36.

20. Rong Yan, Yan Liu, Rong Jin, Hauptmann A. On predicting rare classes with SVM ensembles in scene classification. Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on, 2003, pp. III-21-4 vol.3. doi: 10.1109/ICASSP.2003.1199097.
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python, JMLR 2011; Oct 12: 2825-2830
22. Scikit-learn. Pedregosa et al. Logistic Regression Classifier. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
23. Scikit-learn. Pedregosa et al. Support Vector Machine Classifier. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
24. Scikit-learn. Pedregosa et al. Random Forest Classifier. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
25. Scikit-learn. Pedregosa et al. Gradient Boosting Classifier. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
26. Scikit-learn. Pedregosa et al. Support Vector Machine Classifier kernels. Available from: http://scikit-learn.org/stable/modules/svm.html#svm-kernels
27. Scikit-learn. Pedregosa et al. Grid Search Cross-Validation. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
28. Pop C, Salvesen G.S. Human caspases: activation, specificity, and regulation. J Biol Chem. 2009 Aug 14;284(33):21777-81. doi: 10.1074/jbc.R800084200. PMID: 19473994 PMCID: PMC2755903. Epub 2009 May 26.

29. Talanian R. V, Quinlan C, Trautz S, Hackett M. C, Mankovich J. A, Banach D et al. Substrate specificities of caspase family proteases. J Biol Chem. 1997 Apr 11;272(15):9677-82. PMID: 9092497.
30. Keil B. Specificity of proteolysis. 1st ed. New York: Springer-Verlag Berlin-Heidelberg-NewYork; 1992: 335.
31. Wang M, Zhao X.-M, Tan H, Akutsu T, Whisstock C.J, Song j. Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets, Bioinformatics.2014 Jan 1; 30 (1): 71–80. https://doi.org/10.1093/bioinformatics/btt603
32. Yang Z.R. Prediction of caspase cleavage sites using Bayesian bio-basis function neural networks, Bioinformatics. 2005 May 1;21(9):1831-7. Epub 2005 Jan 25. PMID: 15671118 DOI: 10.1093/bioinformatics/bti281.
33. Song J, Tan H, Perry AJ, Akutsu T, Webb GI, Whisstock JC et al. PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. PLoS ONE. 2012;7(11): e50300. doi: 10.1371/journal.pone.0050300.

Supporting information

S1 Figure. The cumulative ranking score in percentage for the recovered known sites of cleavage for all selected protease families compared with PROSPERous and SitePrediction: a) Serine proteases; b) Cysteine proteases; c) Aspartic proteases; d) Matrix metalloproteases.

S1 Table. Summary on the extracted information from MEROPS database.

S2 Figure. Substrate cleavage sites for P4–P4' position for cathepsins. a) cathepsin D; b) cathepsin E; c) cathepsin L. The cumulative amino acid occurrences in P4–P4' were

calculated and displayed in the form of a two-dimensional heat map. The scissile peptide bond was between sites P1 and P1'.

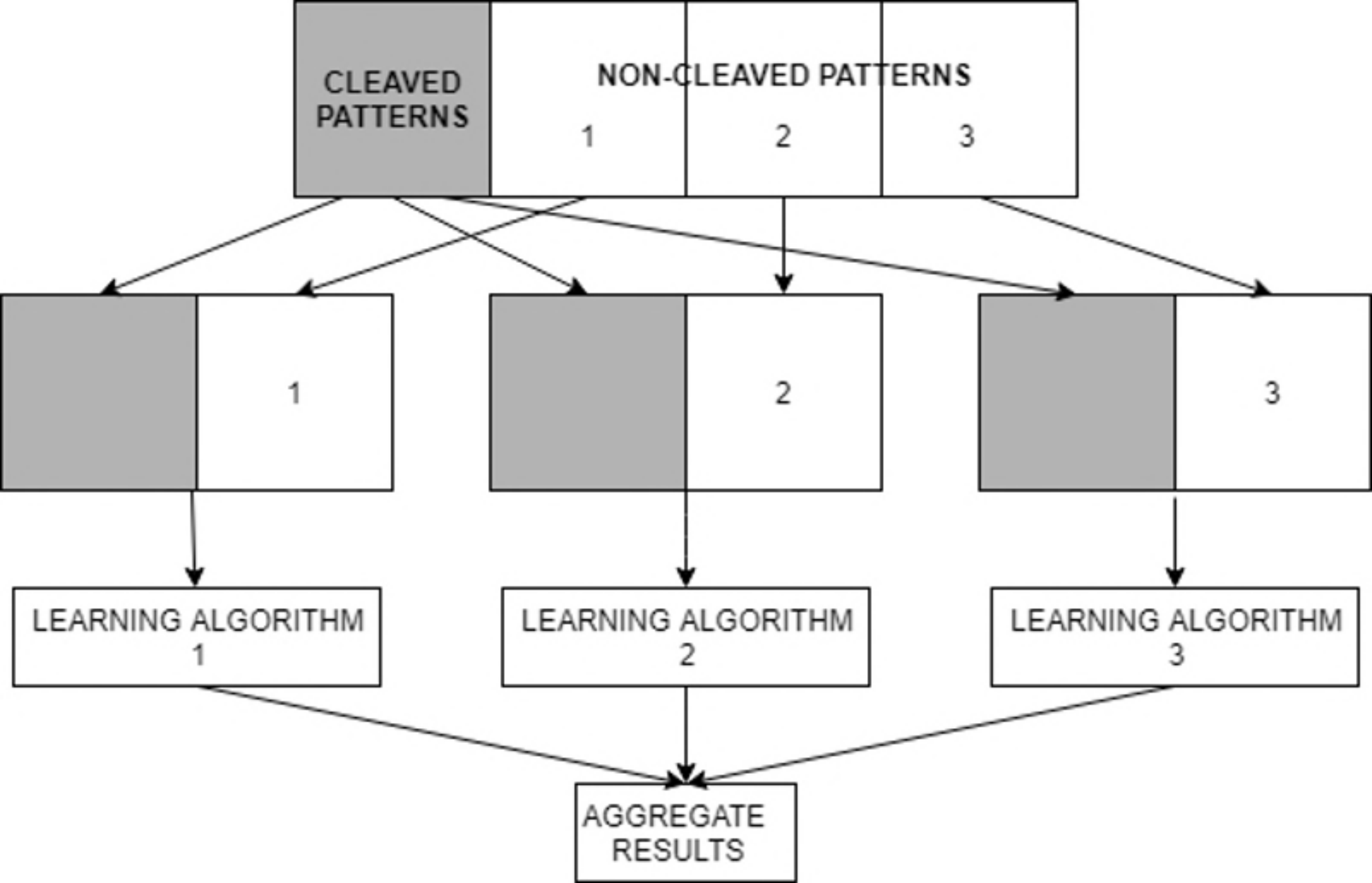
S2 Table. Summary on the model's number for each protease for cleavage window P1-P1' and P4-P4'.

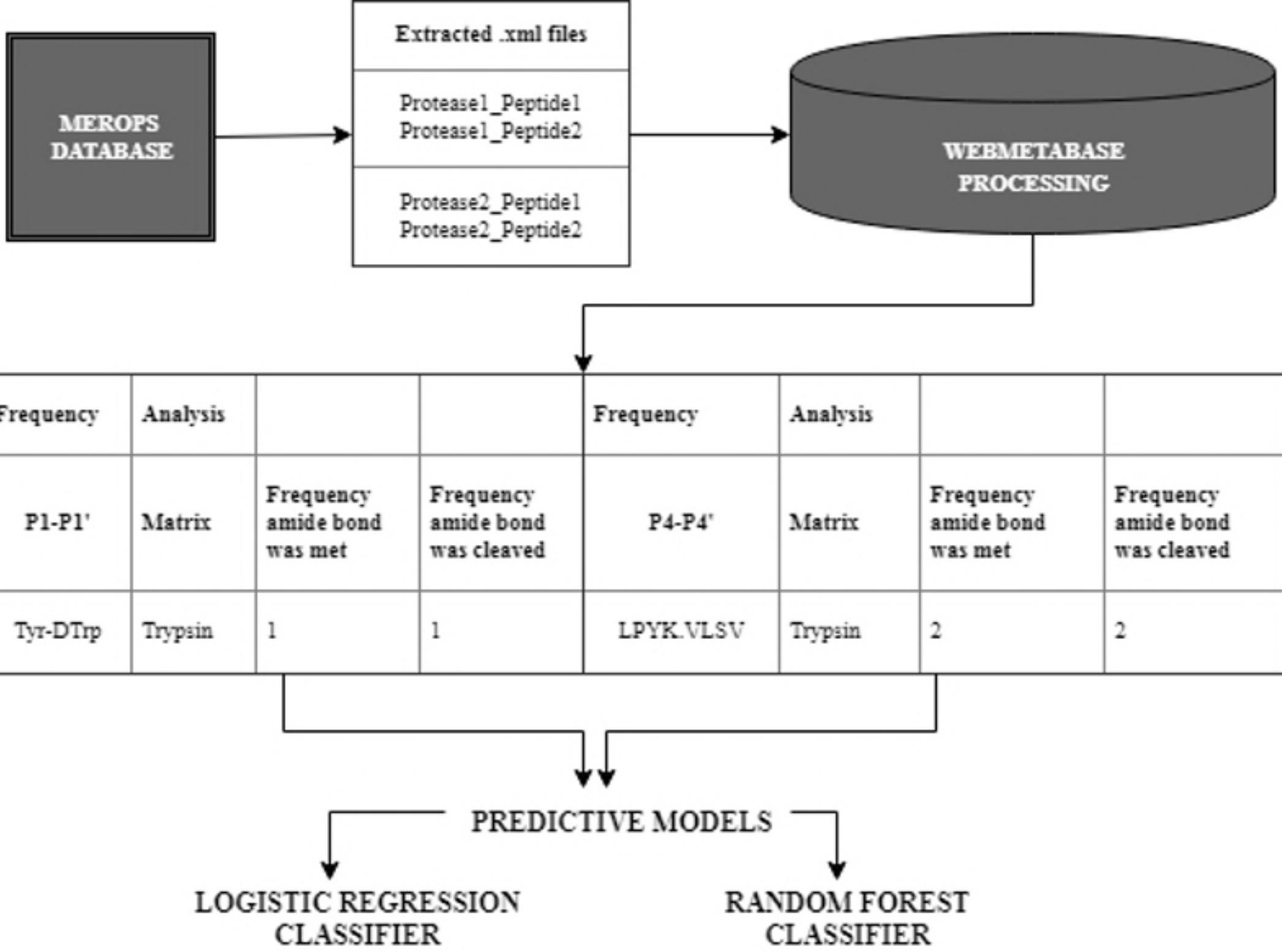
S3 Figure. Substrate cleavage sites for P4–P4' position for granzymes. a) granzyme A; b) granzyme B; c) granzyme B (rodent-type); d) granzyme M. The cumulative amino acid occurrences in P4–P4' were calculated and displayed in the form of a two-dimensional heat map. The scissile peptide bond was between sites P1 and P1'.

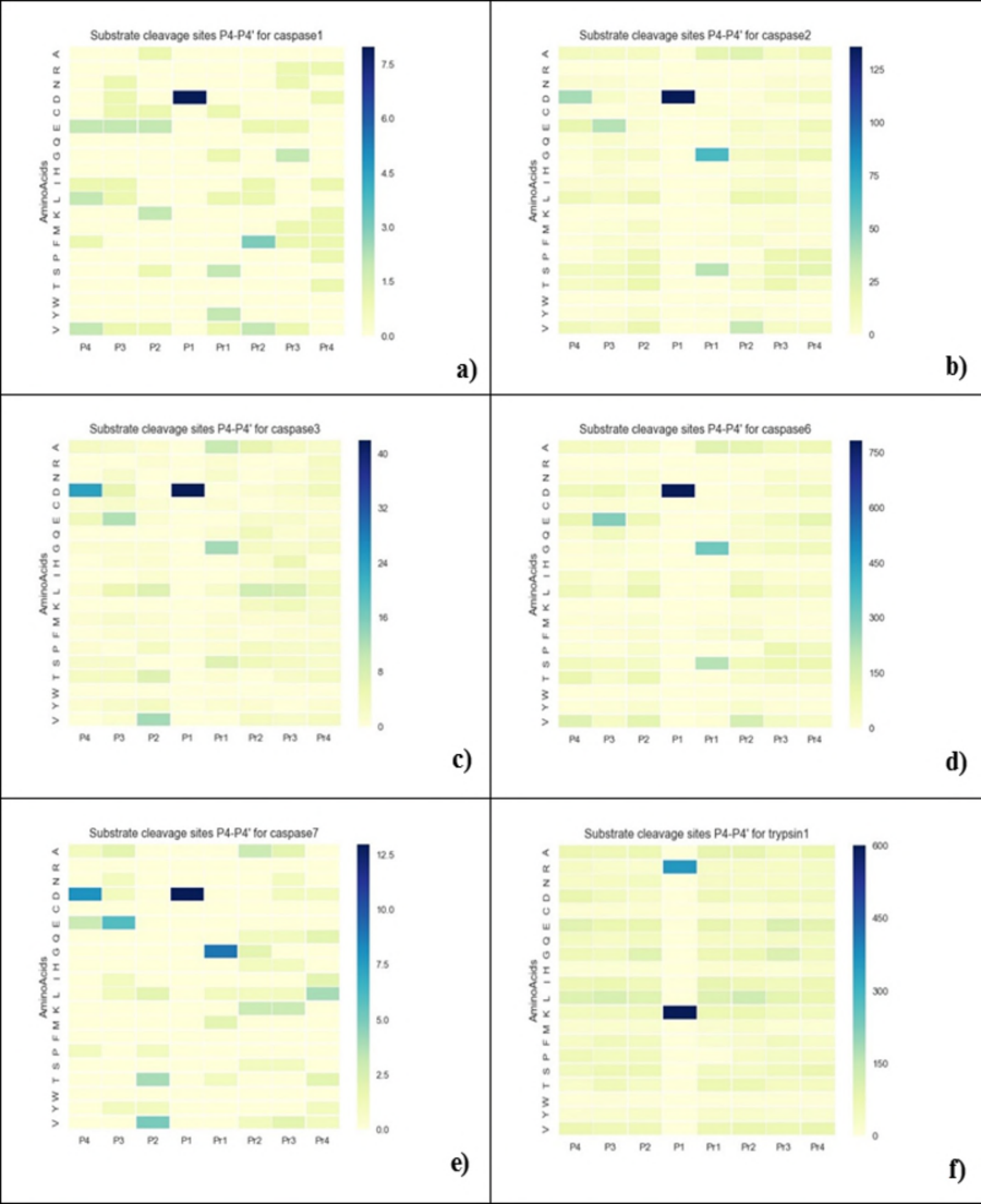
S3 Table. Summary on count of individual SoC2 and SoC8 calculated by frequency analysis.

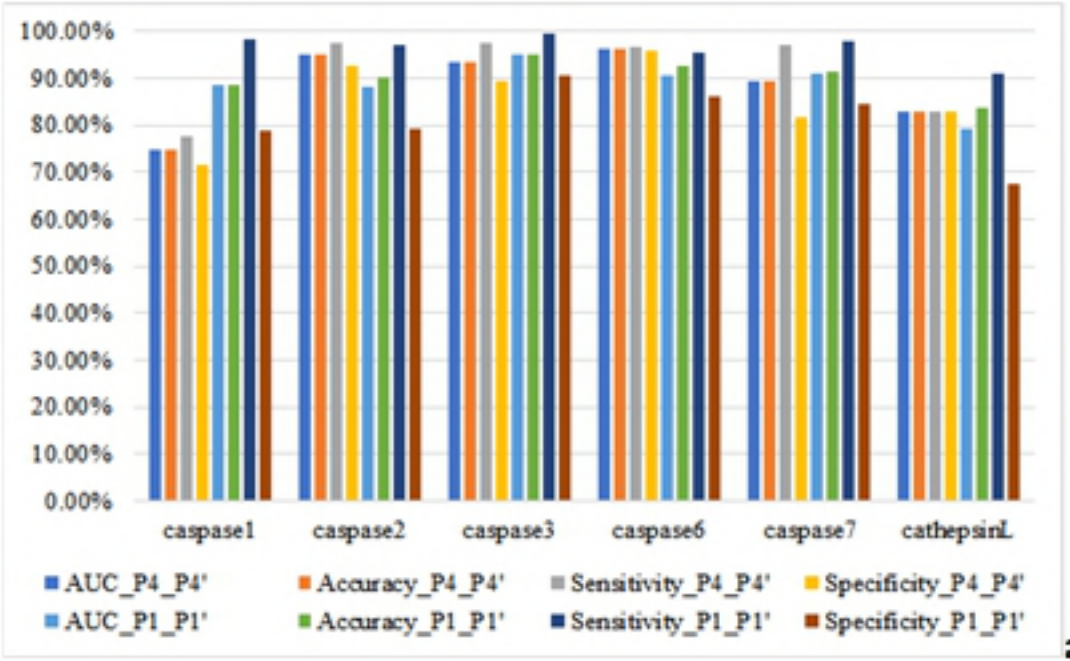
S4 Figure. Substrate cleavage sites for P4–P4' position for matrix metalloproteases (MMP). a) MMP2; b) MMP3; c) MMP8; d) MMP9. The cumulative amino acid occurrences in P4–P4' were calculated and displayed in the form of a two-dimensional heat map. The scissile peptide bond was between sites P1 and P1'.

S4 Table. Accuracy, AUC, sensitivity and specificity for all investigated classifiers comparing predictive performance for caspases using local window P4-P4'.

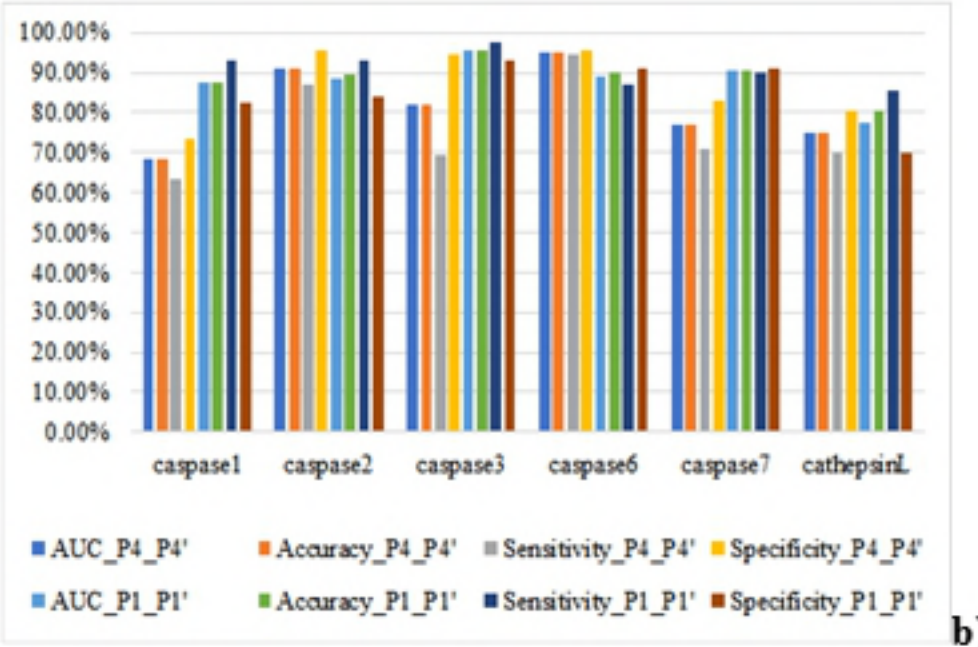

CLEAVED PATTERNS
NON-CLEAVED PATTERNS
1
2
3
1
2
3
LEARNING ALGORITHM 1
LEARNING ALGORITHM 2
LEARNING ALGORITHM 3
AGGREGATE RESULTS


MEROPS DATABASE
Extracted .xml files
Protease1_Peptide1
Protease1_Peptide2
Protease2_Peptide1
Protease2_Peptide2
WEBMETABASE PROCESSING
Frequency
Analysis
Frequency
Analysis
P1-P1'
Matrix
Frequency amide bond was met
Frequency amide bond was cleaved
P4-P4'
Matrix
Frequency amide bond was met
Frequency amide bond was cleaved
Tyr-DTrp
Trypsin
1
1
LPYK.VLSV
Trypsin
2
2
PREDICTIVE MODELS
LOGISTIC REGRESSION CLASSIFIER
RANDOM FOREST CLASSIFIER

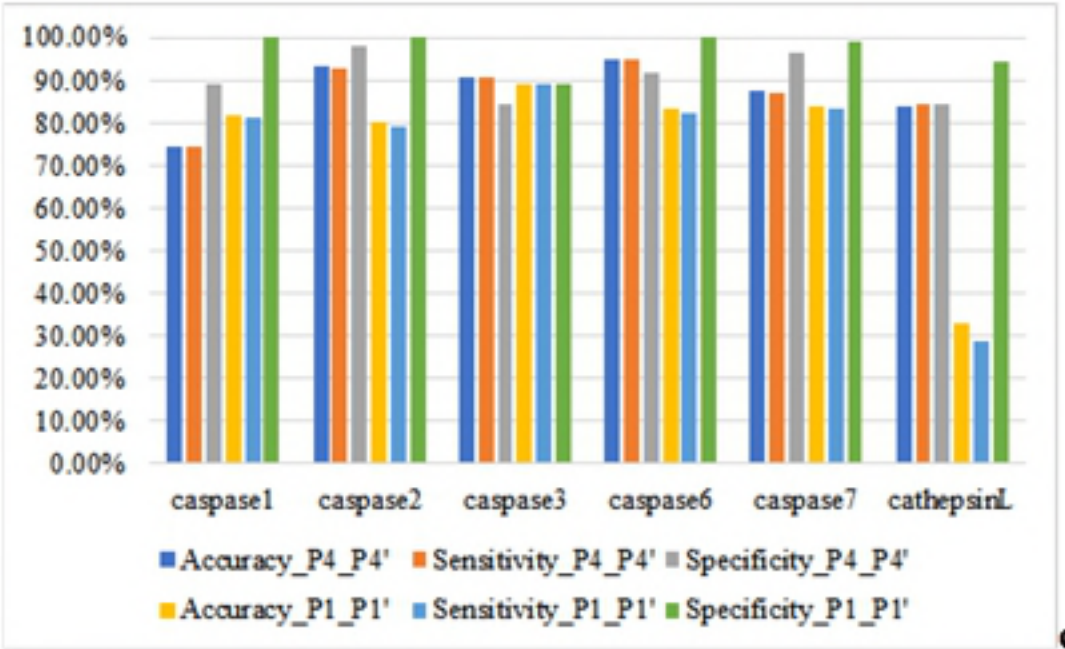

Substrate cleavage sites P4-P4' for caspase1
AminoAcids
V Y W T S P F M K L I H G Q E C D N R A
P4 P3 P2 P1 Pr1 Pr2 Pr3 Pr4
7.5 6.0 4.5 3.0 1.5 0.0
a)
Substrate cleavage sites P4-P4' for caspase2
AminoAcids
V Y W T S P F M K L I H G Q E C D N R A
P4 P3 P2 P1 Pr1 Pr2 Pr3 Pr4
125 100 75 50 25 0
b)
Substrate cleavage sites P4-P4' for caspase3
AminoAcids
V Y W T S P F M K L I H G Q E C D N R A
P4 P3 P2 P1 Pr1 Pr2 Pr3 Pr4
40 32 24 16 8 0
c)
Substrate cleavage sites P4-P4' for caspase6
AminoAcids
V Y W T S P F M K L I H G Q E C D N R A
P4 P3 P2 P1 Pr1 Pr2 Pr3 Pr4
750 600 450 300 150 0
d)
Substrate cleavage sites P4-P4' for caspase7
AminoAcids
V Y W T S P F M K L I H G Q E C D N R A
P4 P3 P2 P1 Pr1 Pr2 Pr3 Pr4
12.5 10.0 7.5 5.0 2.5 0.0
e)
Substrate cleavage sites P4-P4' for trypsin1
AminoAcids
V Y W T S P F M K L I H G Q E C D N R A
P4 P3 P2 P1 Pr1 Pr2 Pr3 Pr4
600 450 300 150 0
f)


100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
caspase1
caspase2
caspase3
caspase6
caspase7
cathepsinL
AUC_P4_P4'
Accuracy_P4_P4'
Sensitivity_P4_P4'
Specificity_P4_P4'
AUC_P1_P1'
Accuracy_P1_P1'
Sensitivity_P1_P1'
Specificity_P1_P1'

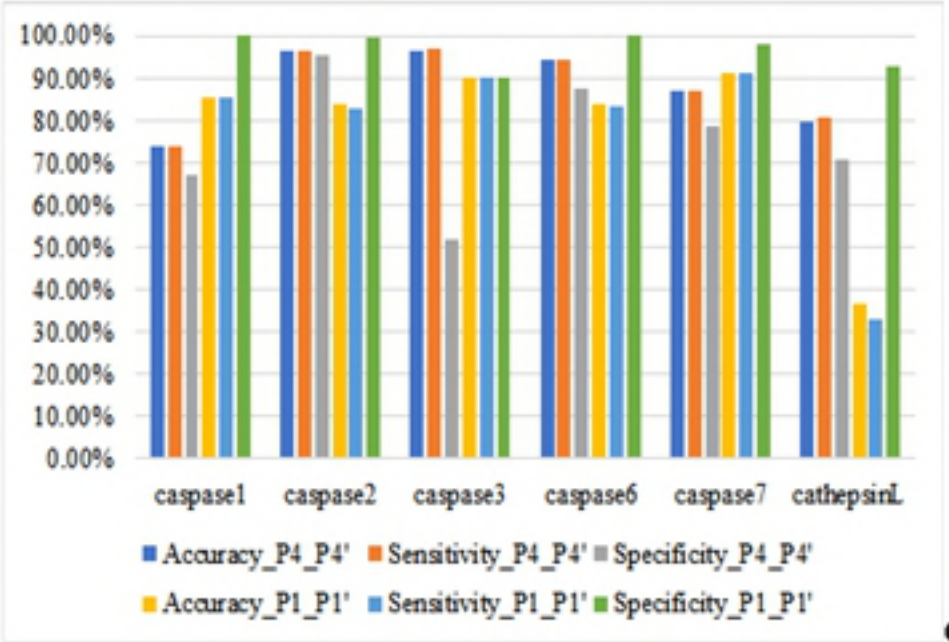
a)


100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
caspase1
caspase2
caspase3
caspase6
caspase7
cathepsinL
AUC_P4_P4'
Accuracy_P4_P4'
Sensitivity_P4_P4'
Specificity_P4_P4'
AUC_P1_P1'
Accuracy_P1_P1'
Sensitivity_P1_P1'
Specificity_P1_P1'

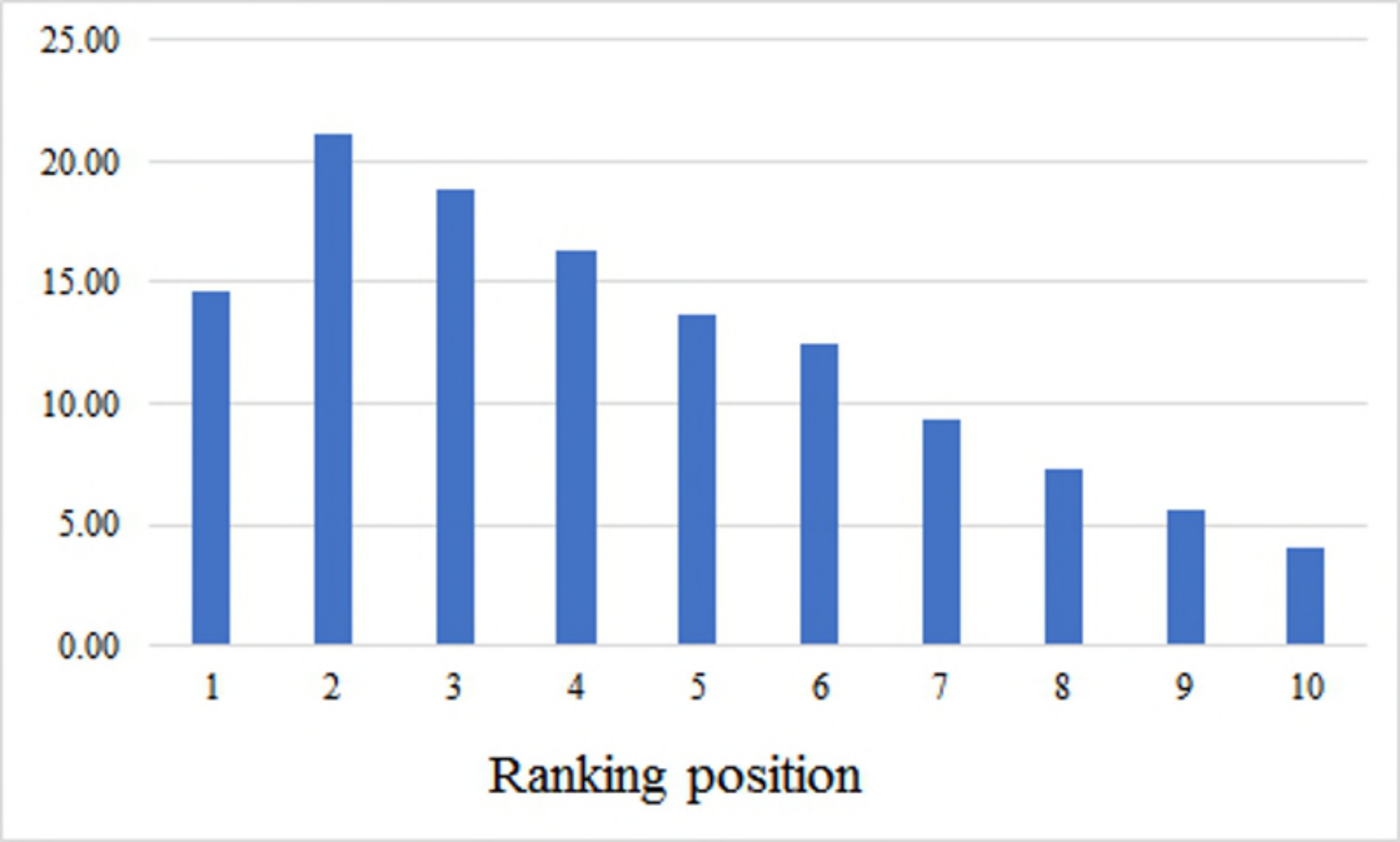
b)

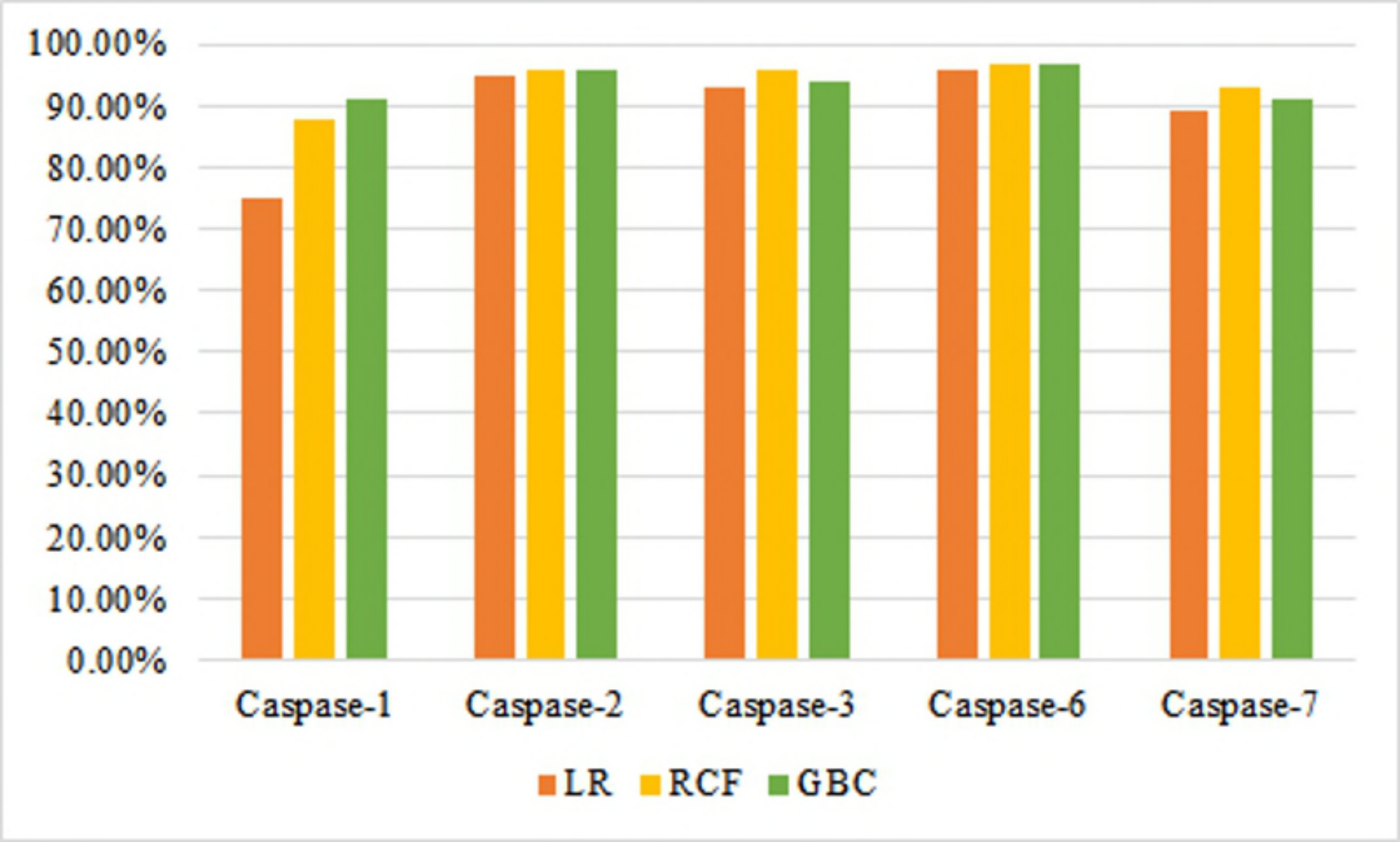

100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
caspase1
caspase2
caspase3
caspase6
caspase7
cathepsinL
Accuracy_P4_P4'
Sensitivity_P4_P4'
Specificity_P4_P4'
Accuracy_P1_P1'
Sensitivity_P1_P1'
Specificity_P1_P1'

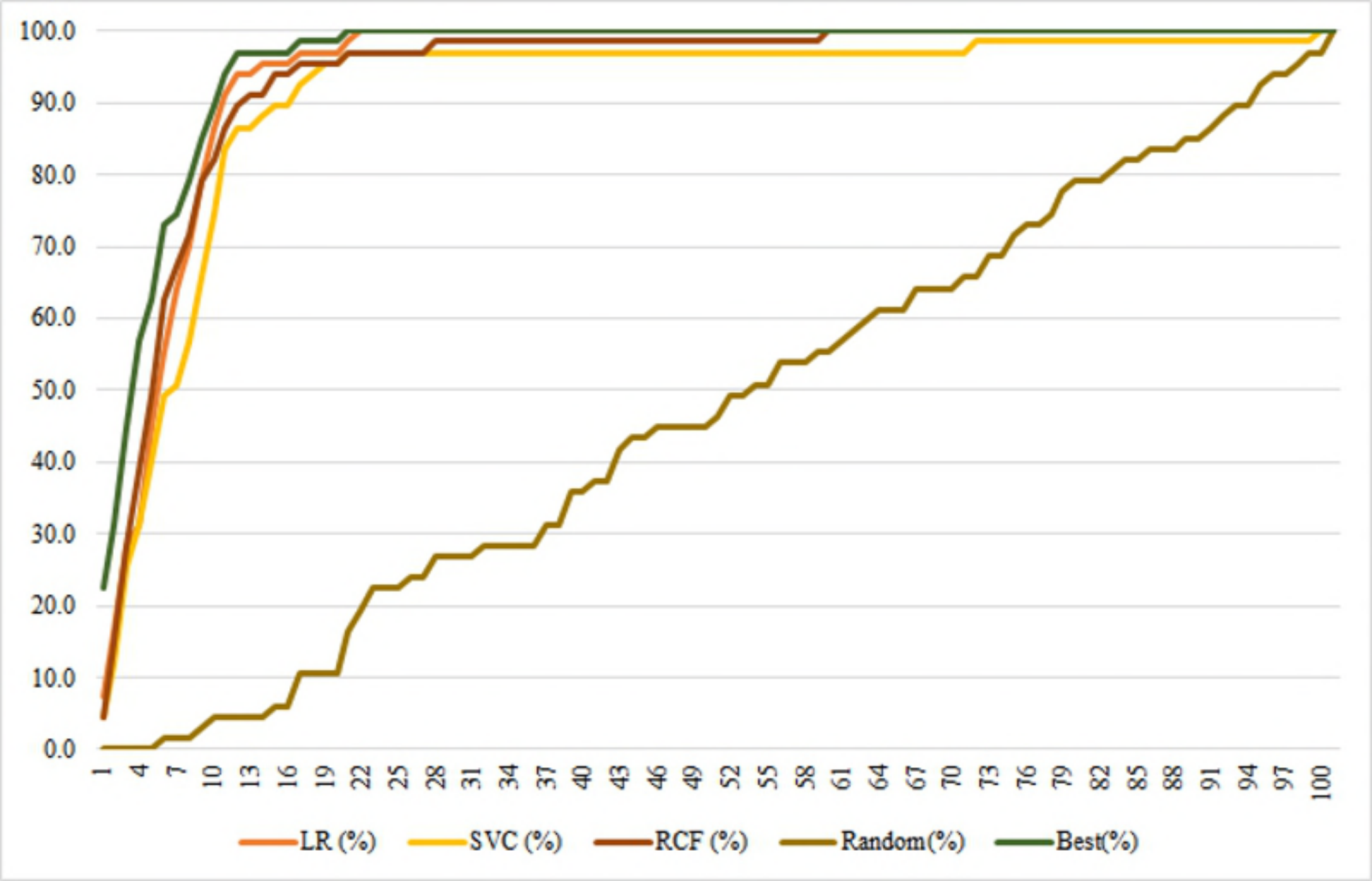
c)

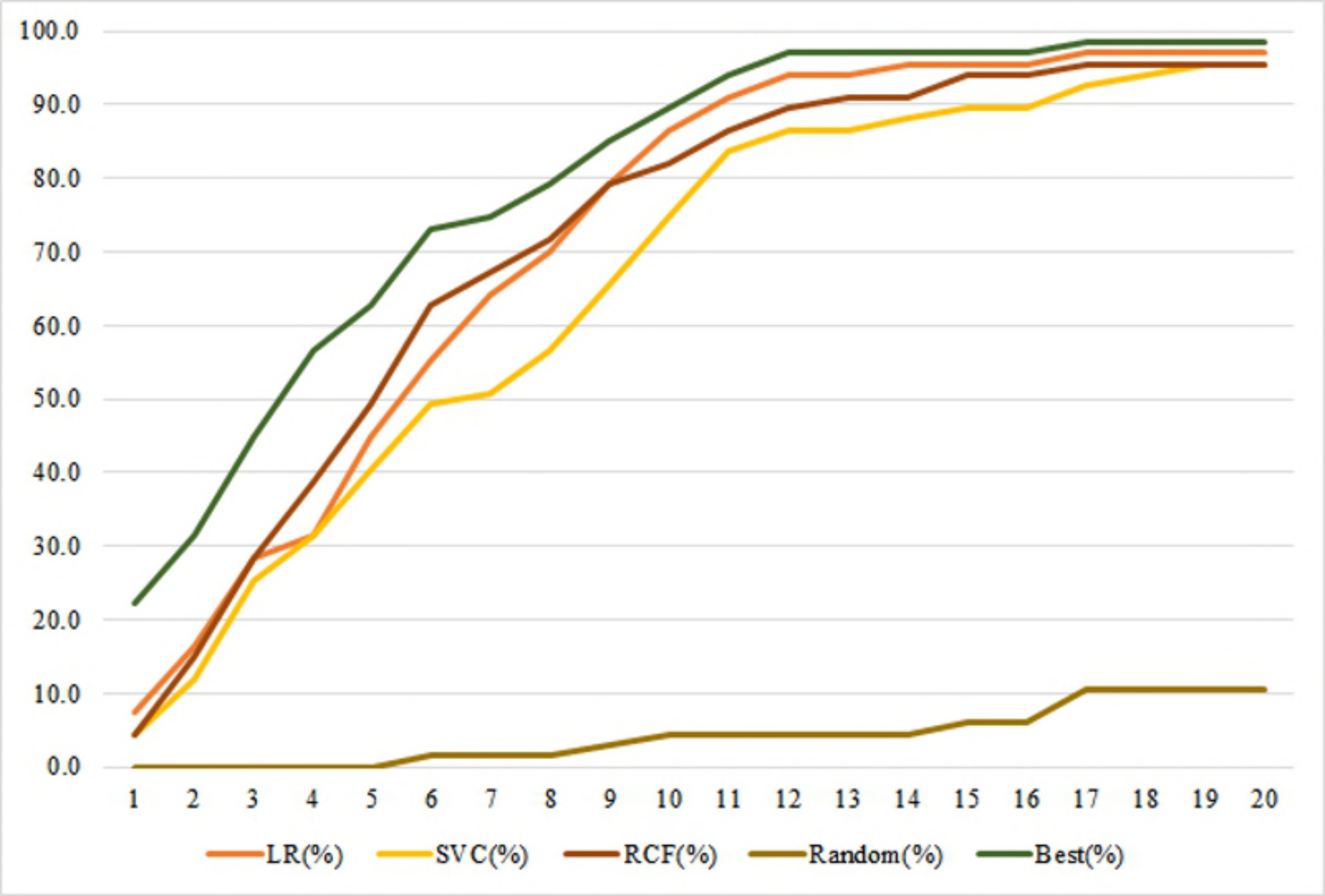

100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
caspase1
caspase2
caspase3
caspase6
caspase7
cathepsinL
Accuracy_P4_P4'
Sensitivity_P4_P4'
Specificity_P4_P4'
Accuracy_P1_P1'
Sensitivity_P1_P1'
Specificity_P1_P1'

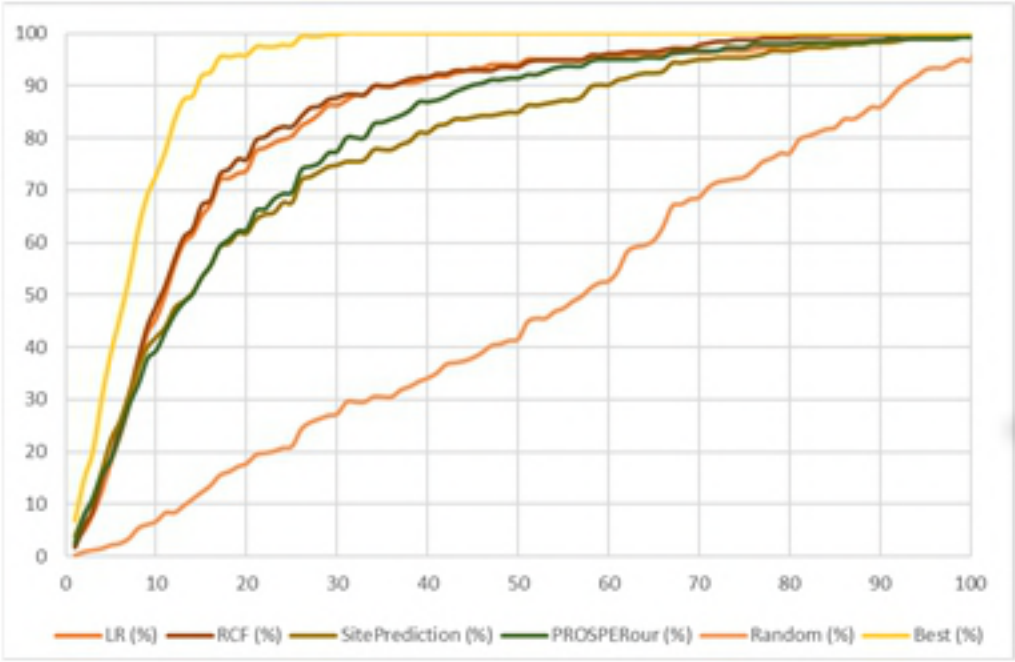
d)


25.00
20.00
15.00
10.00
5.00
0.00
1
2
3
4
5
6
7
8
9
10
Ranking position

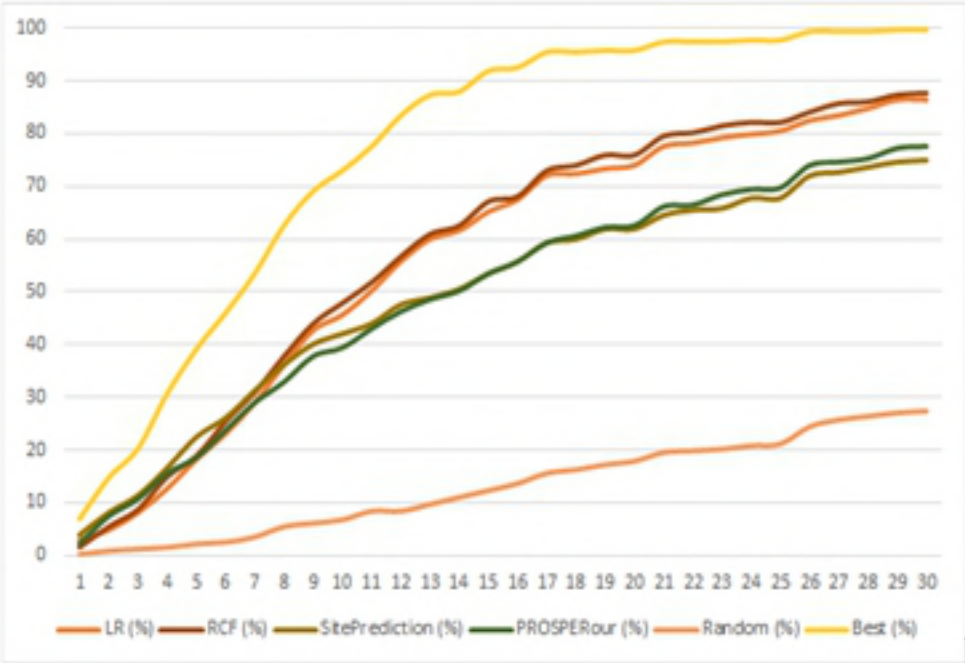

100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Caspase-1
Caspase-2
Caspase-3
Caspase-6
Caspase-7
LR
RCF
GBC


100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97 100
LR (%)
SVC (%)
RCF (%)
Random(%)
Best(%)

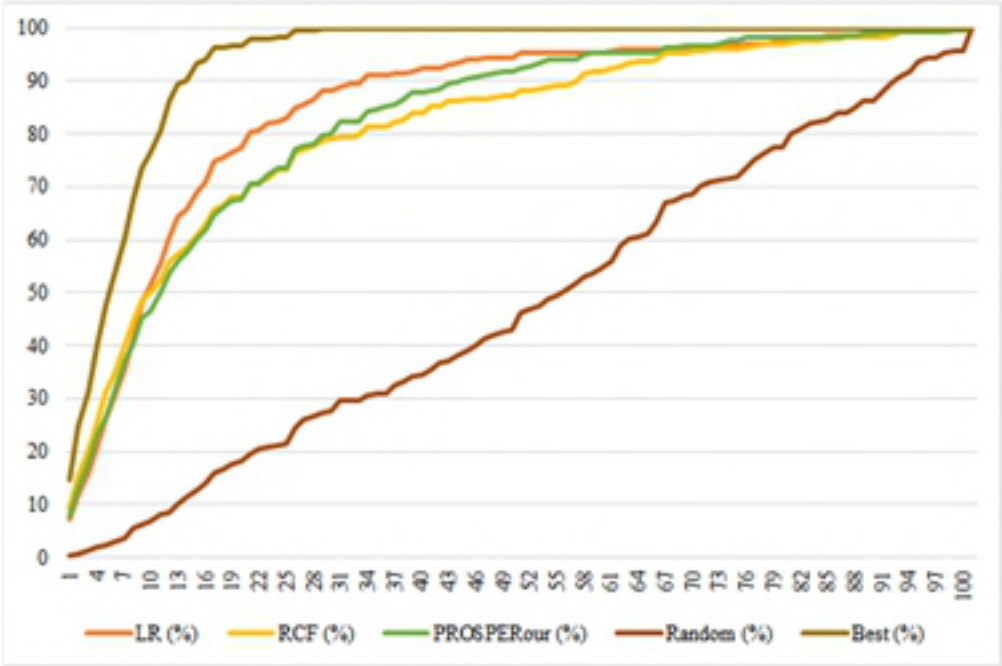

100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
LR(%)
SVC(%)
RCF(%)
Random(%)
Best(%)


LR (%)
RCF (%)
SitePrediction (%)
PROSPERour (%)
Random (%)
Best (%)

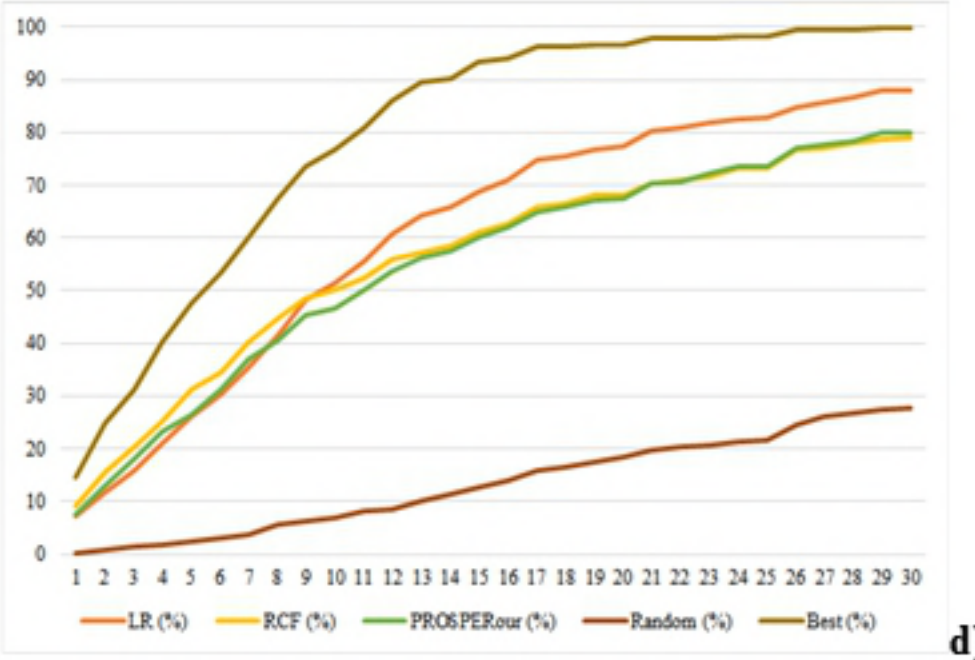
a)


LR (%)
RCF (%)
SitePrediction (%)
PROSPERour (%)
Random (%)
Best (%)

b)


LR (%)
RCF (%)
PROSPERour (%)
Random (%)
Best (%)

c)


LR (%)
RCF (%)
PROSPERour (%)
Random (%)
Best (%)

d)