DOI 10.18699/SBB-2022-202

Development of a searching algorithm for new cell-penetrating peptides based on machine learning methods

Serebrennikova M.*, Grafskaia E., Latsis I., Lazarev V.

Federal Research and Clinical Center of Physical-Chemical Medicine of Federal Medical Biological Agency, Moscow, Russia

** maria.serebrennikova.msu@gmail.com*

Key words: cell-penetrating peptides, machine learning algorithm, prediction

With the development of medicine, systems for the intracellular delivery of biologically active molecules become an increasingly urgent task. From this point of view, cell-penetrating peptides (CPPs) are attractive candidates for new approaches in various therapeutic areas, including cancer treatment. This is a class of peptides, from 5 to 60 amino acids in length, which are the most promising vectors for transporting various therapeutic agents. The presented study aimed to develop the predictive algorithm based on the ML techniques for the identification of new CPPs.

The first stage of the study was to create a sample of all currently known CPPs and non-CPPs sequences. Those that had a high level of homology or contained non-natural amino acid residues and synthetic modifications were excluded. Based on this, the training set consists of 2398 molecules ranging in length from 5 to 61 amino acid residues.

Further, molecular descriptors, that characterize the unique properties of each peptide, were determined for each sequence. Using dimensionality reduction methods, 20 optimal parameters were identified to optimize the predictive ability of the models.

At the following stage, using k-fold cross-validation and grid search, we determined the optimal values of incoming hyperparameters for each of the following algorithms: k-nearest neighbors, logistic regression, gradient boosting, random forest, and support vector model.

According to the results of evaluating the quality of the obtained predictive models searching for penetrating peptides, we choose the three best models based on k-nearest neighbors, gradient boosting, and random forest algorithms. Notice, that the resulting predictive method, in comparison with existing predictive models, demonstrates comparable results.

The predictive method was applied to search for new CPP using an independent test set of data obtained from open sources. A list of candidate CPPs was formed and their structural and functional characteristics were assesed.

Acknowledgements: The research was supported by the Russian Science Foundation (project No. 20–15–00270).