Sajjadian, 2018. Journal of Genes & Cells, 4(2): p, 40-44
doi:10.15562/gnc.66

REVIEW

Bioinformatics Tools and Methods in Identifying Anticancer Peptides

S.Zahra Sajjadian[1]*

1. Department of Biological Sciences, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan, Iran

ABSTRACT

Usage of therapeutic peptides in cancer treatment has been taking significant attention in the recent years. Identify of anticancer peptides is important for expanding useful anticancer drugs. In this paper, we reviewed some improved predictors to design and identify the anticancer peptides. Consequently, we highlighted that these tools and methods including Pseudo amino acid composition with G-Gap dipeptide mode, Support vector machine (SVM) based models, Jackknife cross-validation, Feature selection, iACP, TRAINER and cancerPPD may provide a convenient platform for the development of anticancer therapeutic peptide.

Keywords: Cancer therapy, stem cell transplantation, cancer statistics, cancer mortality, cancer prevalence

Nowadays, the number of people suffering from a cancer-related disease arises every day so, the development of new anticancer drugs with low cytotoxicity to normal cells and a new mode of mechanism that can prevent multi-drug resistance may supply a novel direction for anticancer therapy. Exploration of cytotoxic agents led the improvement of anticancer compound for several decades. Development in cancer therapy, would be exploited to selectively target tumor cells [1, 2].

In the recent decades many attempts have been dedicated in creating novel treatments that are at the same time more preferable and less deleterious for the cancer patients. Regardless of this, the accessible methods such as surgery and chemotherapy currently, have a relative low success rate as well as they present a risk of reappearance [3]. Chemotherapy is the first line of defense against cancerous cell, where disease return or metastasis appear [4].

In actual, as chemical drugs that are produced to attack the rapidly tumor dividing cells, they are assumed to induce side-effects on normal cells that divide at the equal rate [4]. Furthermore, once many of these drugs pass through the lipid bilayers and enter the cytosol and then, they are transported back to the outside of the cell as a mechanism of resistance from the tumor cells [5]. Therefore, it is urgent to improve novel anticancer agents. Although the advances in cancer therapy, mortality rate because of this lethal disease is still very high [6]. In the last decade, small peptides have emerged as a potential preference approach for cancer treatment

* Correspondence:
szahra.sajjadiyan@gmail.com*

[7]. Peptide-based therapy has many advantages over small molecules that involve high tumor penetration, high specificity, low production cost, simplicity of synthesis and modification etc [8]. The discovery of anticancer peptides (ACPs) has provided an alternative approach to treat cancer. Currently there are about 60 approved peptide drugs in the market for treatment of different diseases such as cancer, diabetes, and cardiovascular [7] However, experimental recognition and improvement of novel ACPs is costly and time-consuming and only few of them have been successfully applied into clinics [7]. Therefore, it seems essential to utilize the computational methods. Hence in order to arrange better therapeutic peptide, a critical and rapid concern on exhibiting the information about the peptides owing any apoptotic domains have been stimulated, which maybe was missing in the previously proposed method AntiCP [9]. In this article, we aim to review some bioinformatics tools, methods and databases which are developed to identify anticancer peptides.

Pseudo amino acid composition with G-Gap dipeptide mode

Given a peptide, we can translate it into a mathematical expression for statistical analysis by using G-Gap dipeptide mode. Obviously, the most straightforward way to formulate a peptide sample P with L residues is to use the sequential model as typically given by

P= R1 R2 R3 …R(L-1) R(L)

which R1 represents the 1st residue in the peptide, R2 the 2nd residue, and so forth. With the sequential model to represent a peptide, all its constituent amino acids and also their sequence order or pattern can be precisely defined [10].

The approximate dipeptide composition has been extensively carried out in mathematical proteomics [11, 12]. Nevertheless, the constitutional characteristic of protein sequences order is mostly revealed by the significantly tier correlation [13] of the component residues because of the long-range interaction. Consequently, instead of the approximate dipeptide composition, the g-gap dipeptide composition would be considered, which has been indicated absolutely hopeful for identifying protein properties [14].

Support vector machine (SVM) based models

The SVM is widely applied to handle large data and it has been displayed to perform well in multiple areas of biological data analysis, including classification, protein functions and type identification. Apart from this prediction study, a tool should provide additional basic information on input sequences [15]. In these regard, people have used to calculate the molecular weight, amino acid residues calculation and percentage of amino acid composition. Therefore, SVM serve for prediction and also to know the basic properties of input sequences [16].

In order to design and predict ACPs, Tyagi et al. [17] developed a webserver, AntiCP based on SVM models using amino acid constitutes and binary profiles as input subsequently, Hajisharifi et al. [18] presented two methods based on Chou's pseudo amino acid and local alignment kernel using SVM. They could predict HIV-1 p24 peptides as new candidates with anticancer effects. Another SVM-based tool, ACPP, particularly predicts anticancer peptides with apoptotic domain [19].

The following things are important for evaluating the quality of a statistical predictor: (1) what kind of cross-validation method should be adopted to test it; (2) what kind of metrics should be used to measure its accuracy.

Jackknife cross-validation

Among all these methods, the jackknife test is deemed the least arbitrary and most objective because it can always yield a unique outcome for a given benchmark dataset. Accordingly, the jackknife test has been increasingly used and widely recognized by investigators to examine various predictors (see, e.g., [96, 103, 104, 130–134]). In view of this, the jackknife test was also adopted here to examine the proposed model.

Three cross-validation test methods are often adopted in literature to examine a statistical predictor: independent dataset test, sub-sampling (or K-fold cross-validation) test, and jackknife test [20]. Among them, the jackknife test has been progressively applied and widely found out by researchers, [21-23] because, it can always yield a unique outcome for a given benchmark dataset. Accordingly, the jackknife test has been increasingly used and widely recognized by investigators to examine various predictors (see also [21] [22] [23]). Feature selection is a procedure of selecting a subgroup of relevant features for constructing a machine learning or statistical model. Deficient predicted results would be achieved from inclusion of inessential and noisy information. To develop the prediction efficiency, the analysis of variance (ANOVA) procedure was carried out to select the proper features among the g-gap dipeptide compositions. ANOVA has been performed for feature selection in bioinformatical proteomic [24]. The principle of ANOVA is to compute the feature variances by means of measuring the ratio (F-value) of features between categories and within categories [25]. The optimal number of features can be determined by Incremental Feature Selection (IFS). Li and his coworkers could predict protein domain, antimicrobial peptides, and recognize colorectal cancer related genes [26] as well as categorize carcinoma and hepatocellular cirrhosis using [27-29]. During the IFS method features will be ranked in descending order [14].

iACP

The predictor established by going through the above procedures is called iACP, where "i" stands for "identify", and "ACP" for "anticancer peptide". For the convenience of most experimental scientists, a publicly accessible web-server for iACP has been established. A step-by-step guide on how to use the web-server is given bellow.

Step 1. Open the web server at http://lin.uestc.edu. cn/server/iACP and you can see the top page of iACP on your computer, by clicking on the Read Me button then brief introduction about the predictor appear and the caveat when using it.

Step 2. Either type or copy/paste the query peptide sequences into the input. The input sequence that is in FASTA format and it includes a single initial line beginning with a greater-than symbol (">") in the first column, followed by lines of sequence data. The sequence ends if another line starting with a ">" appears; this shows the start of another sequence.

Step 3. Click on the Submit button to see the predicted result.

Step 4. Click on the Data button to download the benchmark dataset or independent dataset used in this study to train and test the iACP predictor.

Step 5. Click on the Citation button to find the relevant papers documenting the detailed development and algorithm of iACP [30].

iACP is a tool to identify anticancer peptides based on its sequence order information lonely. It applied the wrapper-type feature selection method to look for optimized g-gap dipeptide. The predicted outcomes achieved by iACP via the jackknife test, 5-fold cross-validation test, and independent dataset test have revealed that the novel predictor is quite encouraging, or at least, has the capacity to play a complementary role to the existence techniques in this area https://omictools.com/anticancer-peptides-tool.

TRAINER

TRAINER is a new designed tool for predicting ACPs and non-anticancer peptides. It provides a variety of features and supervised machine learning approaches with relevant parameters which can be selected by user [9]. TRAINER is a new online and flexible tool for biosequence analysis. TRAINER interface can provide a user-friendly environment for any basic internet user. This system can respond to thousands of short sequences in seconds and has been shown to produce accurate results for all types of biological sequences [31]. Biological sequences are of varying lengths, cannot be directly fed in to a classifier and need to be represented by a number of numerical features [32].

cancerPPD

It is a database which manually collected experimentally validated ACPs and anticancer proteins. Currently, CancerPPD contains 3491 ACP and 121 anticancer protein entries which provide comprehensive information regarding a peptide such as the source of origin, conformation, N- and C-

terminal modifications, etc. Additionally, it provides users the information of about 249 types of cancer cell lines and 16 various assays used for testing the ACPs.

Conclusion

In spite of different therapy strategies, cancer remains one of the main causes of fatality worldwide. The In silico based anti-cancer peptide identification tools have become essential, since the peptides play a significant role in cancer therapy. According above tools and methods, anti-cancer peptide predictors, with different approaches, which allow the user to predict and design ACP efficiently and presents the inquiry protein to have anticancer function or not.

The creativity and success of the detection attempts discussed above signify well for the future prospects of discovery new therapeutics which could lead into tremendous diminution in therapeutics evolution time.

Conflict of Interest:

Author declare no conflict of interest.

References:

1. Raz A, Bucana C, McLellan W, Fidler I. Distribution of membrane anionic sites on B16 melanoma variants with differing lung colonising potential. 1980.
2. Cragg G, Newman D. Nature: a vital source of leads for anticancer drug development. Phytochemistry reviews. 2009;8(2):313-31.
3. Harris F, Dennison SR, Singh J, Phoenix DA. On the selectivity and efficacy of defense peptides with respect to cancer cells. Medicinal research reviews. 2013;33(1):190-234.
4. Riedl S, Zweytick D, Lohner K. Membrane-active host defense peptides–challenges and perspectives for the development of novel anticancer drugs. Chemistry and physics of lipids. 2011;164(8):766-81.
5. Perez-Tomas R. Multidrug resistance: retrospect and prospects in anti-cancer drug treatment. Current medicinal chemistry. 2006;13(16):1859-76.
6. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA: a cancer journal for clinicians. 2013;63(1):11-30.
7. Thundimadathil J. Cancer treatment using peptides: current therapies and future prospects. Journal of amino acids. 2012;2012.
8. Vlieghe P, Lisowski V, Martinez J, Khrestchatisky M. Synthetic therapeutic peptides: science and market. Drug discovery today. 2010;15(1):40-56.
9. Poorinmohammad N, Mohabatkar H. A Comparison of Different Machine Learning Algorithms for the Prediction of Anti-HIV-1 Peptides Based on Their Sequence-Related Properties. International Journal of Peptide Research and Therapeutics. 2015;21(1):57-62.
10. Chou K-C, Shen H-B. Recent progress in protein subcellular location prediction. Analytical biochemistry. 2007;370(1):1-16.
11. Chen W, Lin H. Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine. Computers in biology and medicine. 2012;42(4):504-7.
12. Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. Journal of microbiological methods. 2011;84(1):67-70.
13. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001;43(3):246-55.
14. Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7(13):16895-909.
15. Muthukrishnan S, Garg A, Raghava G. Oxypred: prediction and classification of oxygen-binding proteins. Genomics, proteomics & bioinformatics. 2007;5(3):250-2.
16. Muthukrishnan S, Puri M, Lefevre C. Support vector machine (SVM) based multiclass prediction with basic statistical analysis of plasminogen activators. BMC research notes. 2014;7(1):63.
17. Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava G. In silico models for

designing and discovering novel anticancer peptides. Scientific reports. 2013;3.
18. Hajisharifi Z, Piryaiee M, Beigi MM, Behbahani M, Mohabatkar H. Predicting anticancer peptides with Chou′ s pseudo amino acid composition and investigating their mutagenicity via Ames test. Journal of Theoretical Biology. 2014;341:34-40.
19. Vijayakumar S, Lakshmi P. ACPP: a web server for prediction and design of anti-cancer peptides. International Journal of Peptide Research and Therapeutics. 2015;21(1):99-106.
20. Chou K-C, Zhang C-T. Prediction of protein structural classes. Critical reviews in biochemistry and molecular biology. 1995;30(4):275-349.
21. Mohabatkar H, Beigi MM, Esmaeili A. Prediction of GABA A receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. Journal of Theoretical Biology. 2011;281(1):18-23.
22. Khan ZU, Hayat M, Khan MA. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. Journal of Theoretical Biology. 2015;365:197-203.
23. Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′ s general PseAAC. Journal of Theoretical Biology. 2015;364:284-94.
24. Lin H, Chen W, Ding H. AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PloS one. 2013;8(10):e75726.
25. Lin H, Ding H. Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. Journal of Theoretical Biology. 2011;269(1):64-9.
26. Li B-Q, Huang T, Liu L, Cai Y-D, Chou K-C. Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PloS one. 2012;7(4):e33393.
27. Li B-Q, Hu L-L, Chen L, Feng K-Y, Cai Y-D, Chou K-C. Prediction of protein domain with mRMR feature selection and analysis. PloS one. 2012;7(6):e39308.
28. Wang P, Hu L, Liu G, Jiang N, Chen X, Xu J, Zheng W, Li L, Tan M, Chen Z. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PloS one. 2011;6(4):e18476.
29. Huang T, Wang J, Cai Y-D, Yu H, Chou K-C. Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. PloS one. 2012;7(4):e34460.
30. Chen W, Ding H, Feng P, Lin H, Chou K-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016;7(13):16895.
31. Ogul H, T Kalkan A, U Umu S, S Akkaya M. TRAINER: A General-Purpose Trainable Short Biosequence Classifer. Protein and peptide letters. 2013;20(10):1108-14.
32. Hajisharifi Z, Mohabatkar H. In silico prediction of anticancer peptides by TRAINER tool. Molecular Biology Research Communications. 2013;2(1):39-45.