An Accurate Bioinformatics Tool For Anti-Cancer Peptide Generation Through Deep Learning Omics

[#]Aman Chandra Kaushik[1], [#]Mengyang Li[2], and Dong-Qing Wei[1*]

[1]State Key Laboratory of Microbial Metabolism and School of life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China

[2]College of Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

[#]Equal Author

[*]Corresponding Author

Aman Chandra Kaushik: amanbioinfo@sjtu.edu.cn

Mengyang Li: Lmyrssjj0311@sjtu.edu.cn

Dong-Qing Wei: dqwei@sjtu.edu.cn

ABSTRACT

The Anti-cancer targets play crucial role in signalling processes of cells. We have developed an Anti-Cancer Scanner (ACS) tool for identification of Anti-cancer targets in form of peptides. ACS tool also allows fast fingerprinting of the Anti-cancer targets of significance in the current bioinformatics research. There are tools currently available which predicts the above-mentioned features in single platform. In the present work, we have compared the features predicted by ACS with other on-line available methods and evaluated the performance of the ACS tool. ACS scanned the Anti-cancer target protein sequences provided by the user against the Anti-cancer target data-sets. It has been developed in PERL language and it is scalable having an extensible application in bioinformatics with robust coding architecture. It achieves a prediction accuracy of 95%, which is much higher than the existing tools.

Keywords: Anti-cancer vaccines, Artificial Intelligence, Machine learning, Artificial Neural Network

INTRODUCTION

Cancer is one of the most devastating diseases responsible for millions of mortalities worldwide. Different types of cancers are abundant in different countries, but lung cancer was reported as the most common cause of death in men while breast cancer was recorded in women. However, stomach, colorectal, liver and prostate cancers are more common in men. Whereas, cervix, lung, colorectal, stomach and breast cancer are uniformly distributed in women [1]. Different risk factors were reported to be associated with cancer but most common cause of this disease is the mutations in functionally significant somatic genes [2]. Due to high rate of morbidity and mortality, diagnosis, treatment and prevention of cancer is the utmost priority in the current area of research [3]. Global efforts revealed different approaches for the treatment of cancer including operational therapy; chemo agents-based treatment, hormonal, radiation and biological therapy. However, these approaches are constrained due to its high financial cost, adverse effects and low therapeutics output [4]. So, this conventional treatment has reduced the success rate of cancer treatment [5]. To tackle this devastating condition, new means of treatments are required.

Recently, anti-cancer therapeutic vaccines have been developed and widely explored[6-10]. Anti-cancer vaccines (ACV), that contained short chain of amino acids usually less than 50 amino acids, were found to be exceptionally better than conventional chemotherapeutic agents [11]. Other advantages over conventional drugs includes specificity towards the target, no or less intracellular toxicity, alteration feasibility and high penetration power have made these vaccines as promising agents over the purposed methods. Due to promising output, such small vaccines have revolutionized the pharmaceutical markets and number of vaccine therapeutics have been

increased in market [12]. Anti-cancer vaccines (ACV) and Antimicrobial vaccines (AMPs) are of same characteristics displaying similar properties. Like cellular surface negativity of bacteria, cancerous cells also possess negative charge and thus, both ACV and AMPs showed broad spectrum of activities. This negatively charge cells are of significant interest, making potential interaction with the cell surface and thus, selective toxicity can be achieved. These specificity properties divide these vaccines into two different categories; one that shows toxicity against all types of cells including bacterial, cancerous and normal cells while other depicts activity against only bacterial and cancerous cells [13-15].

Despite enormous therapeutic significance of Anti-cancer vaccines, till date, no tool for Anti-cancer vaccines and Anti-cancer proteins/peptides has been developed to identify the vaccines from wild and mutated cancerous protein sequences that can be used as anti-cancer vaccines. Large databases are available such as data from Clinical Proteomic Tumor Analysis Consortium (CPTAC)[16] which contains proteomic data from mass spectrometry analysis, and compare expression patterns of proteome and transcript. Also, Cancer Genome Atlas (TCGA)[17] is one of the reservoirs of cancerous datasets that holds RNA-sequence data sets. On the other hand, cancer transcriptomics data is provided by microarray expression analysis. Many different cancer mutations related data such as data from Cancer Genomics Hub[18], Catalogue of Somatic Mutations in Cancer (COSMIC) [19], SNP500Cancer [20], and UCSC Cancer Genomics Browser [21] are available.

Despite the availability of such huge data, no efforts have been made to analyze and retrieve important patterns from this big data that could be clinically significant for the treatment of cancer. Therefore, to support the scientific portfolio developing anti-cancer vaccines; we have developed Anti-Cancer Scanner (ACS) tool which accept the cancer associated proteins data

including information from the above sources to analyze and retrieve important evidence from them and predict Anti-cancer vaccine from their target sequence using machine learning approach as shown in Figure 1. This proposed tool identifies vaccines/peptides from targeted cancerous protein which means it's personalized vaccines for cancer patients. We believe that ACS will be helpful for both bioinformatics and experimental researchers working in the field of Anti-cancer vaccine based therapeutics.

METHODOLOGY

Implementation

Following steps in the methodology were followed in the development of ACS:

Data mining of Anti-cancer targets- Many databases were used for data mining; many cancer databases are a molecular group of precise information scheme that gathers heterogeneous records of Anti-cancer targets.

Data set development of Anti-cancer targets- A dataset of Anti-cancer targets was developed, it includes 100 Anti-cancer targets, and input parameters used for ACS dataset were length of protein, origin character of mutation, mutated characters, and positions.

Development of algorithm for ACS- ACS is based on artificial neural network which is part of nature inspired algorithms that categorised into three important parts (Figure 2)

1. Input layer of node
2. Hidden layer node and
3. Output node.

Each layer in neural network has one parallel node and consequently building the stacking. Output of each neuron is collective information of neurons multiplied by parallel weights with biased value while input value is transformed into output.

Normalization of Anti-cancer targets Dataset- Retrieves data for Anti-cancer targets dataset from various sources. Normalized dataset computed by $V_{new} = (V_{old} – MinV) / (MaxV – MinV) * (D_{max} – D_{min}) + D_{min}$.

Where, MinV is the minimum assessment of variable, MaxV maximum estimation of variable, D_{max} is the maximum estimation succeeding to normalization, D_{min} is the minimum assessment after normalization, V_{new} is new assessment after normalization and V_{old} assessment before normalization.

1. Input the data for training, and executed for training.
2. Set Anti-cancer targets network constraint.
3. Let learning count from 0;
4. Let learning count increase by 1;
5. Training stage iteration begins For ( );
6. Input one sample then assign a new process and assign datasets.
7. Go to step 6 otherwise, (at least one dataset under the process).
8. Verify the sample to the stack for again test or training.
9. Iteration terminate when all samples have been tested
10. Learning error is computed.
11. Go to step 5, if the total learning error is 0.0
12. End

13. Calculate the neurons of output $net_j = \sum_{i=1\sim m} w_{ji}\ x_i + b_j$ where net_j is output neurons, w_{ji} connection weight neurons, x_i is input signal neurons and b_j is bias neurons.
14. Signal of output layers are calculated $net_k = TV_k + \delta_K^L$ where TV_k is target value of output neurons and δ_K^L is the error of neuron.
15. Compute the error of neuron k. Step3 and Step6 are repetitive until $SSE = \sum_{i=1\sim n}(T_i - Y_i)^2$, where T_i is actual assessment and Y_i is estimated assessment.

RESULTS

ACS tool is very useful and reliable tool for Anti-cancer targets analysis. It generates maximum output using minimum input of Anti-cancer targets; using these applications users can determine the Anti-cancer vaccines. ACS tool which accept cancer associated proteins data including data from different cancerous resources; then analyze and retrieve important information from their cancer targets/proteins and predict Anti-cancer vaccine using machine learning approach. This proposed tool identifies vaccines/peptides from targeted cancerous protein which means it's a personalized vaccine for cancer patients.

It is graphical user interface application (Figure 4) where user can easily import cancer target list with mutation information and ACS will predict optimized anti-cancerous peptides from those imported targets using artificial neural network; get retrieves the data for anti-cancer targets dataset and fiddle with defined series of attributes and avoid the infiltration of neurons. Input the cancer target data (cancer proteins) list for training, interrelated values of input cancer data and output are executed for training where anti-cancer targets network constraint like length of protein, origin character of mutation, mutated characters, and positions which is numeral of hidden layers (4 hidden layers produce better union). Assigning a new process and put new

knowledge datasets under this process and analyze the final output of artificial neural network, and verify the sample to stack again for test or training. Calculate the neurons of output, every neuron output signals are calculated shown in Figure 3.

Overview of the procedure

Input/output Instruction

1. Enter the sequence accession number. Many formats can be used: Raw/Plain, EMBL, Pearson/Fasta, etc. for further analysis.

2. Submit the job by clicking on submit button.

3. After you submit your query, a set of scores will be calculated, depending on which option you selected. Based on the scores, the vaccine will be ranked.

Peptide Motif Search

This package allowed users to locate and rank 8-mer, 9-mer, or 10-mer peptides that contain peptide binding motifs.

Input Instructions

1. Choose the gene sequence of interest from list with accession number through menu button. The selection of ID determines which coefficient table in scoring program will be used on the selected sequence.

2. Choose the length of subsequences, program then extracted the information for submitted input sequence and calculated the scores and ranks.

3. Choose whether to display the submitted input sequence on the output page using button.

4. Submit the job by using submit button.

Output Page Returned to the User

Following query submission, a set of scores are calculated for all 8-mer, 9-mer, or 10-mer subsequences contained in the input sequence, depending on selected option. Based on the scores, subsequences are ranked. This task is supposed to complete within a few seconds. After, a display page then returned with following information;

1. A short table of user-input parameters (for verification and later recall), along with some scoring data and other useful info (includes number of scores calculated, number of scores requested, number of scores reported back in the output table, and length of users input sequence)
2. The score output table, where the results of calculations on subsequences are displayed.
3. A listing of submitted input peptide sequence (if you have asked for the sequence to be echoed back)
4. Each row of the score output table consists of four columns. The values for these items represents the ranking of subsequence. The starting position in query protein sequence for the first amino acid residue of subsequence. A residue list 8-mer, 9-mer, or 10-mer peptide subsequence itself.

An estimated numerical score for subsequence (upon which the rank in the first column is based) shown in Figure 4.

Table 1 AutoDock energy for the best ranked complexes where top table represents the cancer targets, ACS suggested vaccines/peptides, global energy in kcal/mol, vdW energy in kcal/mol and H-Bond energy in kcal/mol of cancer targets and ACS suggested vaccines/peptides.

Target	**Vaccine**	**Global Energy**	**vdW energy**	**H-Bond energy**

		(kcal/mol)	(kcal/mol)	(kcal/mol)
PTEN	**EEKKGGGKMCISLRVG**	**-11.06**	**-9.17**	**-6.65**
	EKLLGGVVNAERPNENG	-4.24	-4.11	-2.11
	LLFIFH	-5.12	-5.67	-1.99
	PKIDKGIKD	-4.78	-4.34	-2.11
	LLFIFH	-5.34	-4.45	-1.99
	RHIKLRV	-2.11	-2.07	-1.23
CACNA1S				
	EEKKGGGKMCISLRVG	-8.34	-7.83	0.00
	EKLLGGVVNAERPNENG	-7.36	-6.21	-0.83
	LLFIFH	**-9.34**	**-9.88**	**-4.02**
	PKIDKGIKD	-5.02	-5.22	-0.84
	LLFIFH	-5.34	-5.78	-4.02
	RHIKLRV	-4.40	-5.14	-0.78
GLUK3				
	EEKKGGGKMCISLRVG	-3.11	-3.82	-4.29
	EKLLGGVVNAERPNENG	-5.30	-5.45	0.00
	LLFIFH	**-7.23**	**-7.56**	**-1.24**
	PKIDKGIKD	-5.56	-5.66	-2.71
	LLFIFH	-4.45	-4.4	-1.24
	RHIKLRV	-4.5	-4.65	-1.52
RPL30				
	EEKKGGGKMCISLRVG	-3.10	-2.11	0.00

	EKLLGGVVNAERPNENG	-6.06	-6.95	-3.73
	LLFIFH	**-7.10**	**-7.00**	**0.00**
	PKIDKGIKD	-6.69	-6.60	0.00
	LLFIFH	-5.10	-4.00	0.00
	RHIKLRV	-3.64	-2.05	-0.31
CCLN2				
	EEKKGGGKMCISLRVG	-2.56	-1.90	-0.98
	EKLLGGVVNAERPNENG	-2.74	-2.34	0.00
	LLFIFH	-4.04	-3.80	0.00
	PKIDKGIKD	-6.65	-5.51	-2.84
	LLFIFH	-3.04	-2.80	0.00
	RHIKLRV	**-6.78**	**-6.05**	**-7.65**
RPTN				
	EEKKGGGKMCISLRVG	-3.00	-3.23	-3.59
	EKLLGGVVNAERPNENG	-4.88	-4.62	-5.17
	LLFIFH	-3.16	-2.71	0.00
	PKIDKGIKD	-2.90	-1.00	-0.91
	LLFIFH	-4.16	-3.72	0.00
	RHIKLRV	**-7.55**	**-6.87**	**-4.22**
TM246				
	EEKKGGGKMCISLRVG	-2.63	-1.71	0.00
	EKLLGGVVNAERPNENG	**-9.21**	**-8.66**	**-6.16**
	LLFIFH	-6.77	-6.0	0.00

	PKIDKGIKD	-3.31	-3.44	0.00
	LLFIFH	-4.47	-3.25	0.00
	RHIKLRV	-4.08	-3.86	-1.04
ZN746				
	EEKKGGGKMCISLRVG	-3.54	-3.41	0.00
	EKLLGGVVNAERPNENG	-2.18	-2.3	-1.69
	LLFIFH	**-7.87**	**-6.35**	**0.00**
	PKIDKGIKD	-3.6	-3.6	-1.30
	LLFIFH	-4.5	-4.34	0.00
	RHIKLRV	-4.4	-4.65	-3.57

ACS tool advantages and applications

ACS is based on machine learning approch for identification of Anti-cancer targets based on their binding affinity. There is no such type of tool for identifying or classifying Anti-cancer targets. The advantage of ACS is that it analyses information for Anti-cancer targets which can be further assisted in precision drug discovery, identification of novel drug targets, drug designing for these novel targets as shown in Figure 6.

CONCLUSION

We have proposed machine learning algorithm to identify Anti-cancer targets based on their binding affinity and implemented in artificial neural network. There is no such type of tool for

identifying or classifying Anti-cancer targets using artificial neural network. It is graphical user interface application where user can easily import cancer target list with mutation information and ACS will predict optimized anti-cancerous vaccines/peptides from those imported targets using artificial neural network; get retrieves the data for anti-cancer targets dataset and fiddle with defined series of attributes and avoid the infiltration of neurons. ACS tool is advantageous for finding information about Anti-cancer targets which can be further assisted or employed in precision drug discovery, identification of novel drug targets, drug designing for these novel targets with a prediction accuracy of 95%. Black box approach were used for ACS script testing, where scripts are examining the functionality of ACS tool. ACS also inbuilt white box testing, where users opposed the functionality of ACS.

LIST OF ABBREVIATIONS

Anti-cancer vaccines (ACV), Antimicrobial vaccines (AMPs), Clinical Proteomic Tumor Analysis Consortium (CPTAC), Cancer Genome Atlas (TCGA), Catalogue of Somatic Mutations in Cancer (COSMIC), Artificial neural network (ANN), European Molecular Biology Lab (EMBL).

FUNDING

This work is supported by the grants from the Key Research Area Grant 2016YFA0501703 of the Ministry of Science and Technology of China, the National Natural Science Foundation of China (Contract no. 61832019, 61503244), the State Key Lab of Microbial Metabolism and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2017ZD14).

ACKNOWLEDGMENTS

The simulations in this work were supported by the Center for High Performance Computing, Shanghai Jiao Tong University.

AUTHORS CONTRIBUTIONS

Conceptualization: DQW, ACK

Data curation: ACK

Formal analysis: ACK

Funding acquisition: DQW

Investigation: DQW, ACK

Methodology: ACK

Project Administration: DQW

Resources: ACK, DQW

Supervision: DQW

Validation: DQW, ACK, ML

Writing-original draft: ACK

Writing-review draft: ACK, ML

AVAILABILITY OF DATA AND MATERIAL

The data during and/or analyzed during the current study available from the corresponding author on request.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

REFERENCES

1. Fitzmaurice, C.; Allen, C.; Barber, R.M.; Barregard, L.; Bhutta, Z.A.; Brenner, H.; Dicker, D.J.; Chimed-Orchir, O.; Dandona, R.; Dandona L.; *et al*. **Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study**. *JAMA oncology* 2017, **3**(4):524-548.
2. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A.; Kinzler, K.W. **Cancer genome landscapes**. *Science (New York, NY)* 2013, **339**(6127):1546-1558.
3. Biemar, F.; Foti, M. **Global progress against cancer-challenges and opportunities**. *Cancer biology & medicine* 2013, **10**(4):183-186.
4. Mahassni, S. H.; Al-Reemi, R.M. **Apoptosis and necrosis of human breast cancer cells by an aqueous extract of garden cress (Lepidium sativum) seeds**. *Saudi journal of biological sciences* 2013, **20**(2):131-139.
5. Holohan, C.; Van Schaeybroeck, S.; Longley, D.B.; Johnston, P.G. **Cancer drug resistance: an evolving paradigm**. *Nature reviews Cancer* 2013, **13**(10):714-726.
6. Barras, D.; Widmann, C. **Promises of apoptosis-inducing peptides in cancer therapeutics**. *Current pharmaceutical biotechnology* 2011, **12**(8):1153-1165.

7. Li, Z. J.; Cho, C.H. **Peptides as targeting probes against tumor vasculature for diagnosis and drug delivery**. *Journal of translational medicine* 2012, **10 Suppl 1**:S1.

8. Boohaker, R.J.; Lee, M.W.; Vishnubhotla, P.; Perez, J.M.; Khaled, A.R. **The use of therapeutic peptides to target and to kill cancer cells**. *Current medicinal chemistry* 2012, **19**(22):3794-3804.

9. Shapira, S.; Fokra, A.; Arber, N.; Kraus, S. **Peptides for diagnosis and treatment of colorectal cancer**. *Current medicinal chemistry* 2014, **21**(21):2410-2416.

10. Gautam, A.; Kapoor, P.; Chaudhary, K.; Kumar, R.; Raghava, G.P. **Tumor homing peptides as molecular probes for cancer therapeutics, diagnostics and theranostics**. *Current medicinal chemistry* 2014, **21**(21):2367-2391.

11. Thundimadathil, J. **Cancer treatment using peptides: current therapies and future prospects**. *Journal of amino acids* 2012, **2012**:967347.

12. Vlieghe, P.; Lisowski, V.; Martinez, J.; Khrestchatisky, M. **Synthetic therapeutic peptides: science and market**. *Drug discovery today* 2010, **15**(1-2):40-56.

13. Hoskin, D.W.; Ramamoorthy, A. **Studies on anticancer activities of antimicrobial peptides**. *Biochimica et biophysica acta* 2008, **1778**(2):357-375.

14. Mader, J.S.; Hoskin, D.W. **Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment**. *Expert opinion on investigational drugs* 2006, **15**(8):933-946.

15. Schweizer, F. **Cationic amphiphilic peptides with cancer-selective toxicity**. *European journal of pharmacology* 2009, **625**(1-3):190-194.

16. Edwards, N.J.; Oberti, M.; Thangudu, R.R.; Cai, S.; McGarvey, P.B.; Jacob, S.; Madhavan, S.; Ketchum, K.A. **The CPTAC Data Portal: A Resource for Cancer Proteomics Research**. *Journal of proteome research* 2015, **14**(6):2707-2713.

17. Cancer Genome Atlas Network. **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**(7407):330-337.

18. Wilks, C.; Cline, M.S.; Weiler, E.; Diehkans, M.; Craft, B.; Martin, C.; Murphy, D.; Pierce, H.; Black, J.; Nelson, D. *et al*. **The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data**. *Database : the journal of biological databases and curation* 2014, **2014**.

19. Forbes, S.A.; Beare, D.; Gunasekaran, P.; Leung, K.; Bindal, N.; Boutselakis, H.; Ding, M.; Bamford, S.; Cole, C.; Ward, S. *et al*. **COSMIC: exploring the world's knowledge of somatic mutations in human cancer**. *Nucleic acids research* 2015, **43**(Database issue):D805-811.

20. Packer, B.R.; Yeager, M.; Burdett, L.; Welch, R.; Beerman, M.; Qi, L.; Sicotte, H.; Staats, B.; Acharya, M.; Crenshaw A. *et al*. **SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes**. *Nucleic acids research* 2006, **34**(Database issue):D617-621.

21. Goldman, M.; Craft, B.; Swatloski, T.; Cline, M.; Morozova, O.; Diekhans, M.; Haussler, D.; Zhu, J. **The UCSC Cancer Genomics Browser: update 2015**. *Nucleic acids research* 2015, **43**(Database issue):D812-817.

Figure Legends

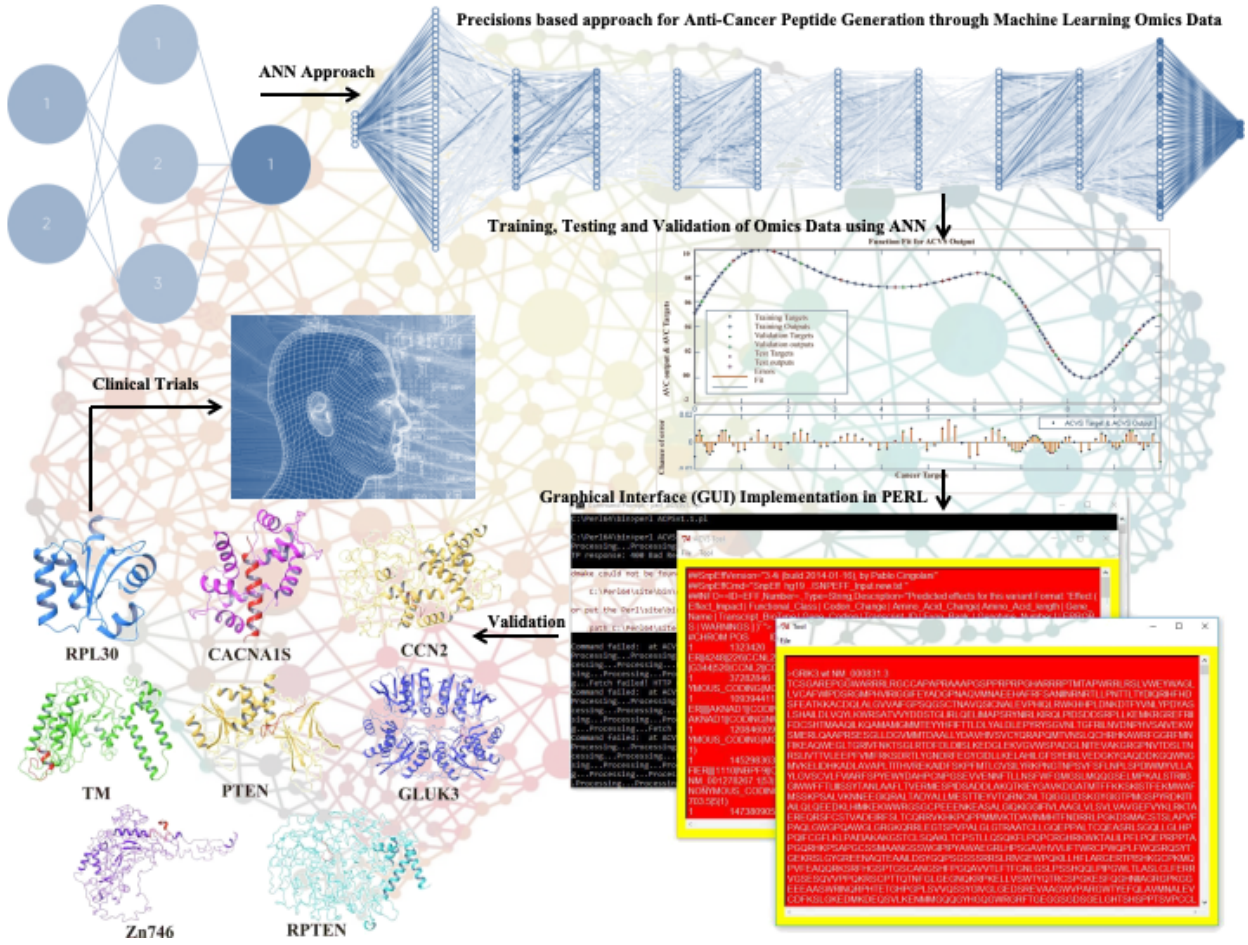

Figure 1. Schematic plan shows six steps employed of precisions based approach for anti-cancer peptide generation through machine learning omics data in form of agonist as promising strategy in the treatment of cancer

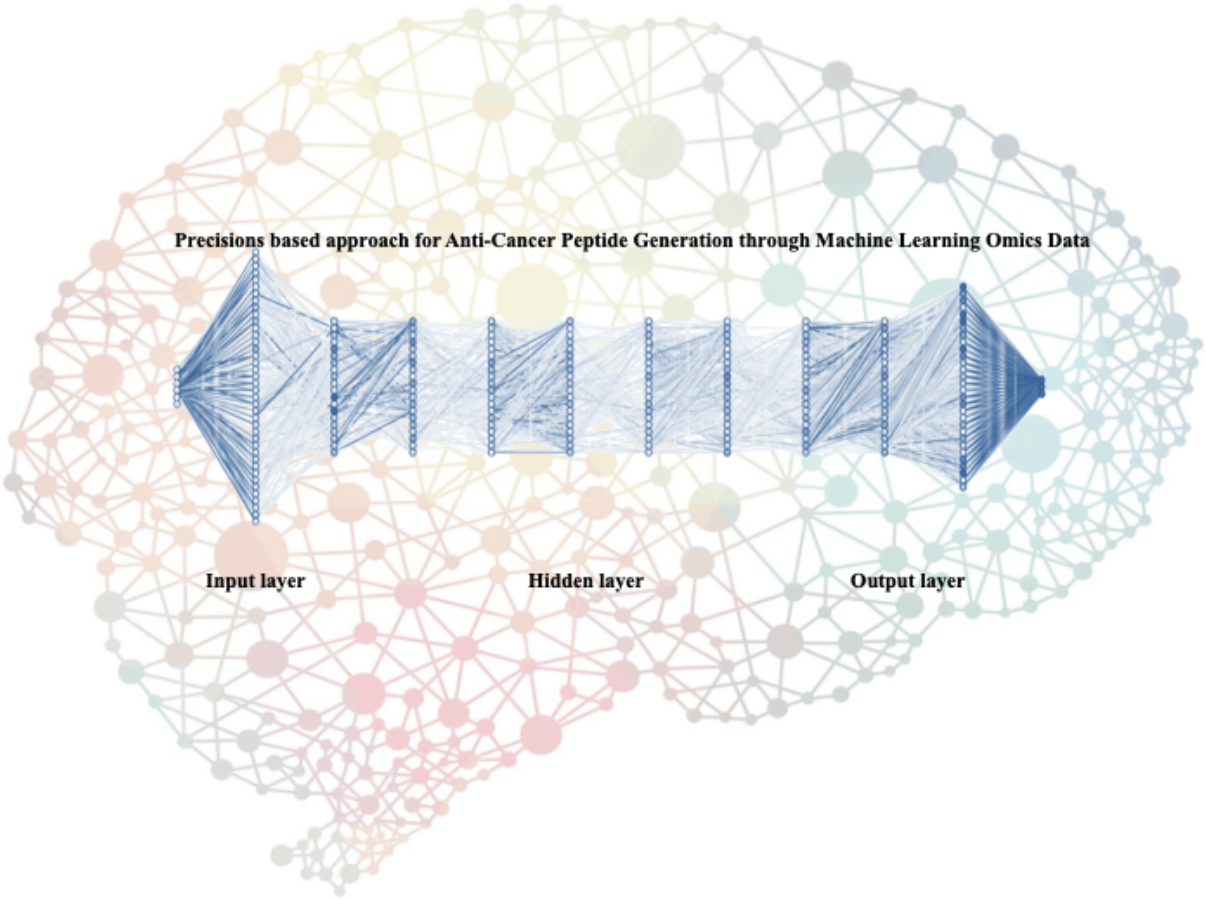

Figure 2: Neural network-based model consisting of connected submodels in where circles represent the input layer, hidden layer and output layer

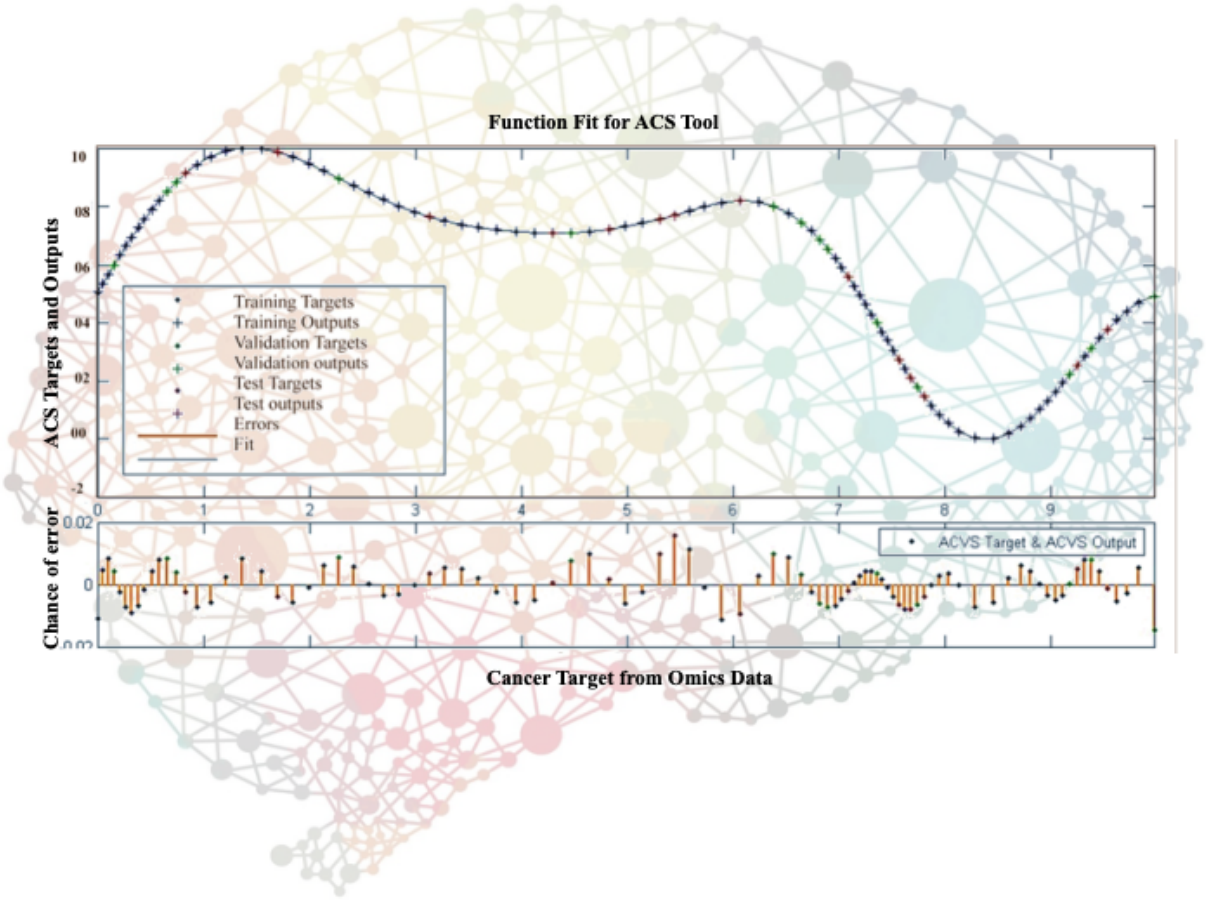

Figure 3: ACS performance. Graph represents the training, output and validation of targets; test output of targets; chance of error and function fit using artificial neural network approach, where upper X axis represents function fit and Y axis represents the ACS output and target dataset performance while lower X axis represents input data and Y axis represents the chance of performance

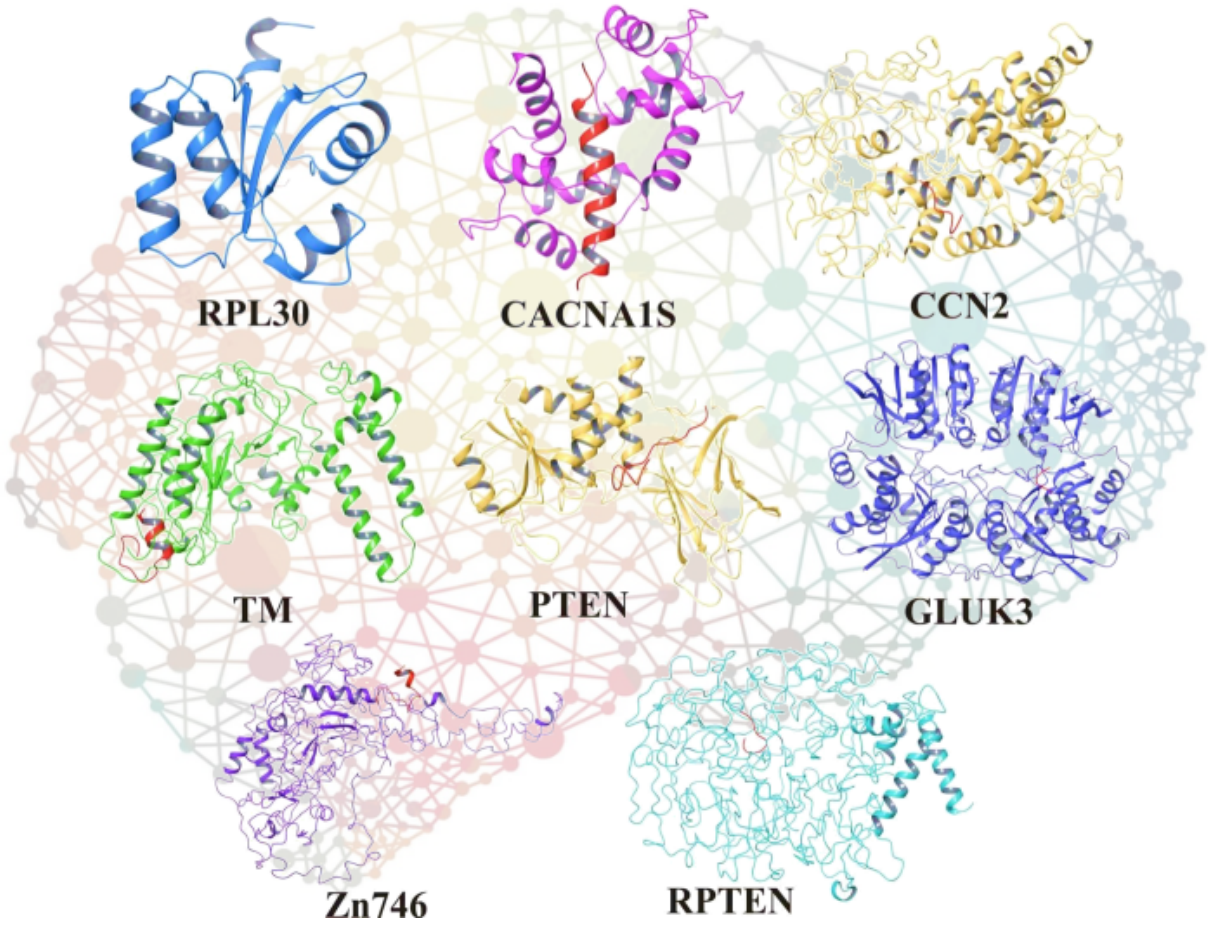

Figure 4: Front view of Anti-Cancer Scanner Tool through Machine Learning Omics Data

Validation. For the validation of ACS tool, we took ten random cancer targets from dataset as a target and ten ACS generated optimized cancer vaccines/peptides (outputs) against same cancer targets, and applied molecular docking for each cancer targets and ACS suggested peptides **(Figure 5 and Table 1)**.

Figure 5: The figure is showing the interaction predicted peptides from their respective protein sequences. AutoDock were used to dock and define the interactions

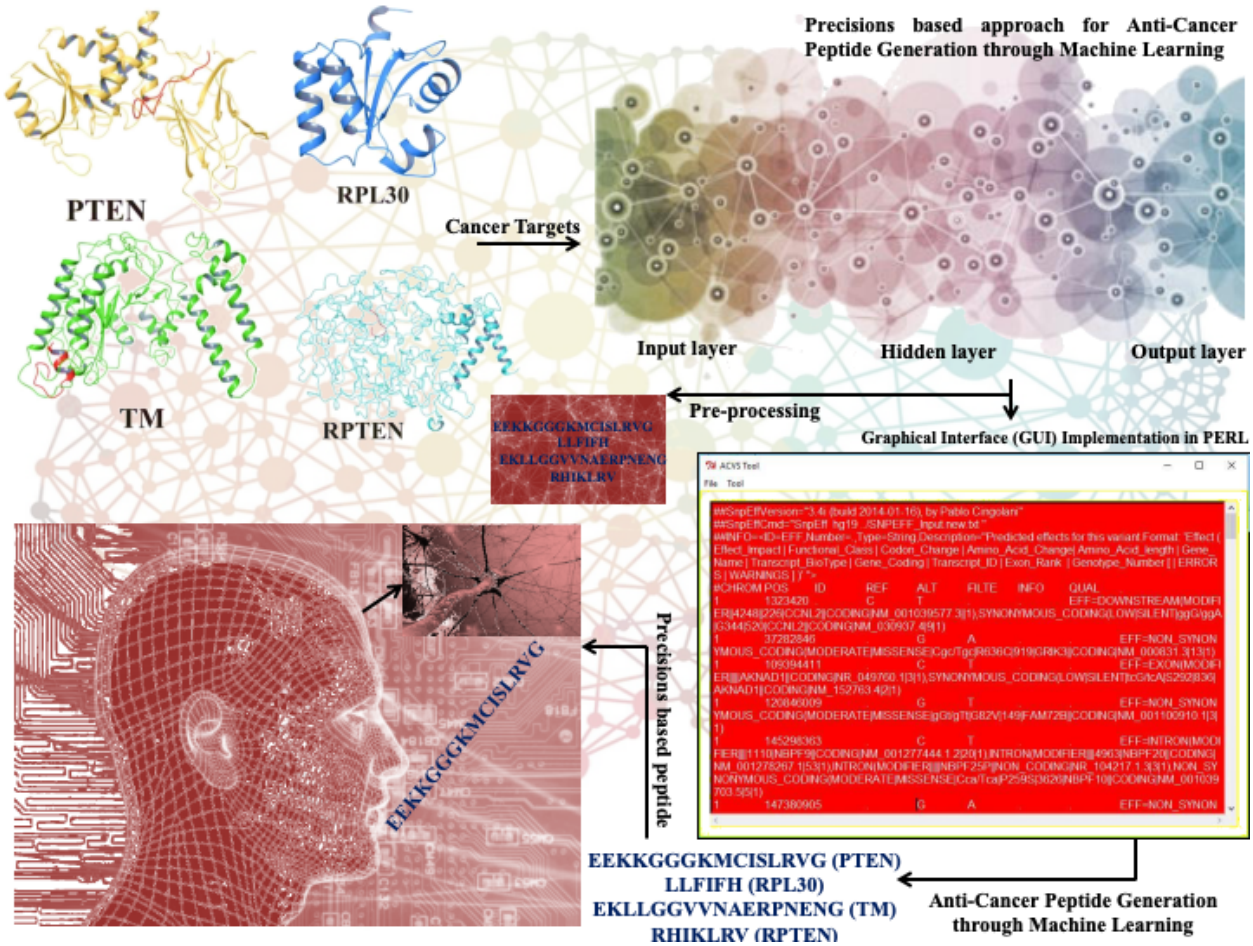

Figure 6. Schematic plan depicts the employed of precisions based peptides for anti-cancer peptide generation through machine learning