Collaboration across Disciplines for the Health of People, Animals and Ecosystems
L. Stoicu-Tivadar et al. (Eds.)

doi:10.3233/SHTI241068

Leveraging Cancer Therapy Peptide Data: A Case Study on Machine Learning Application in Accelerating Cancer Research

Georgios FERETZAKIS[a], Athanasios ANASTASIOU[b,1], Stavros PITOGLOU[e], Aikaterini SAKAGIANNI[c], Zoi RAKOPOULOU[c], Konstantinos KALODANIS[d], Vasileios KALDIS[c], Evgenia PAXINOU[a], Dimitris KALLES[a] and Vassilios S. VERYKIOS[a]

[a] *School of Science and Technology, Hellenic Open University, Patras, Greece*
[b] *Biomedical Engineering Laboratory, National Technical University of Athens, Athens, Greece*
[c] *Sismanogleio General Hospital, Marousi, Greece*
[d] *Harokopio University of Athens, Kallithea, Greece*
[e] *Computer Solutions SA, Athens, Greece*

ORCiD ID: Georgios FERETZAKIS https://orcid.org/0000-0002-3597-1187,
Athanasios ANASTASIOU https://orcid.org/0000-0001-6679-3590,
Stavros PITOGLOU https://orcid.org/0000-0002-5309-4683,
Aikaterini SAKAGIANNI https://orcid.org/0000-0002-3199-9158,
Konstantinos KALODANIS https://orcid.org/0000-0003-2456-9261,
Vasileios KALDIS https://orcid.org/0000-0001-6416-2293,
Evgenia PAXINOU https://orcid.org/0000-0002-9910-8569,
Dimitris KALLES https://orcid.org/0000-0003-0364-5966,
Vassilios S. VERYKIOS https://orcid.org/0000-0002-9758-0819

Abstract. This study leverages the DCTPep database, a comprehensive repository of cancer therapy peptides, to explore the application of machine learning in accelerating cancer research. We applied Principal Component Analysis (PCA) and K-means clustering to categorize cancer therapy peptides based on their physicochemical properties. Our analysis identified three distinct clusters, each characterized by unique features such as sequence length, isoelectric point (pI), net charge, and mass. These findings provide valuable insights into the key properties that influence peptide efficacy, offering a foundation for the design of new therapeutic peptides. Future work will focus on experimental validation and the integration of additional data sources to refine the clustering and enhance the predictive power of the model, ultimately contributing to the development of more effective peptide-based cancer treatments.

Keywords. Machine learning, PCA, Cancer therapy peptides, Peptide design, K-means clustering

[1] Corresponding Author: Athanasios ANASTASIOU, PhD, Biomedical Engineering Laboratory, National Technical University of Athens, Greece; E-mail address: aanastasiou@biomed.ntua.gr.

1. Introduction and Background

Treating cancer remains one of the most significant challenges in medical and biomedical research fields, globally, necessitating continuous advancements in therapeutic approaches. Traditional treatments such as chemotherapy, radiation, and surgery often come with severe side effects and limitations in targeting specificity [1]. In recent years, peptides have emerged as promising therapeutic agents due to their high specificity, efficacy, and relatively low toxicity [2]. Peptides can interact with proteins and other macromolecules, playing crucial roles in various cellular functions such as cell signaling and immune modulation. Studies have reported that 15-40% of all protein-protein interactions in human cells are mediated by peptides [3]. As a result, these short chains of about 2-50 amino acids long, linked by peptide bonds, can disrupt specific molecular pathways involved in cancer progression, offering improved tumor penetration and lower immunogenicity, compared to conventional drugs [3]. Peptides, are used to deliver carriers and therapeutic cargoes to tumors and form peptide-drug conjugates (PDCs), enhancing both delivery and efficacy [4]. They selectively bind to cell surface receptors and intracellular proteins, blocking activity or disrupting interactions. This multifunctionality makes peptides highly promising for cancer therapy [3].

The development of peptide-based therapies relies heavily on comprehensive datasets, detailing peptide sequences, structures, and biological activities. Datasets provided by Mendeley Data, PeptideAtlas, MassIVE, etc. are available for download as well as for online browsing of submitted identifications. In this study, we used the DCTPep (Data of cancer therapy peptides), an open data repository of cancer therapy peptides, composed of two sub-libraries: Peptide Library and Drug Library. It has been developed to provide scientists with the information for designing new anticancer peptides and targeted peptide-conjugated anticancer agents with a high selectivity. The DCTPep database is an invaluable resource, providing extensive information on cancer therapy peptides, including their physicochemical properties and structural annotations [5].

In biomedical research, machine learning techniques are increasingly used to analyze large datasets and uncover patterns that traditional methods might miss [6]. Machine learning techniques enable highly accurate data modeling without relying on strong assumptions about the system being modeled. It often outperforms biomedical models in data description, offering both practical engineering solutions and a crucial benchmark. These techniques are particularly valuable in peptide research for predicting peptide activity, classifying peptides, and identifying key therapeutic features [7].

Principal Component Analysis (PCA) is a statistical method that reduces data dimensionality while preserving essential variance, aiding in visualization and pattern identification [8]. K-means clustering is a widely used unsupervised machine learning algorithm designed to partition a dataset into a specified number of clusters [9]. Its objective is to group similar data points together, thereby uncovering underlying patterns or structures within the data. When combined with K-means clustering, PCA can categorize peptides based on their physicochemical properties, highlighting characteristics that influence their biological activity. Applying these advanced techniques to the DCTPep dataset can significantly accelerate the discovery and optimization of cancer therapy peptides. By identifying natural groupings and key differentiating features, researchers can streamline the peptide design process and prioritize those with the highest therapeutic potential [10]. This paper aims to leverage machine learning, specifically PCA and K-means clustering, to analyze the DCTPep

dataset, identify meaningful patterns, and enhance our understanding of the correlation between peptide properties and their therapeutic dynamic.

2. Materials and Methods

This study leverages the DCTPep database to explore the effectiveness of machine-learning techniques in selecting the most effective peptides for cancer treatment. Considering that peptide functionality is determined by three parameters—the identity of amino acids, the sequence of amino acids, and the shape of the peptide—DCTPep is an ideal resource for specific analysis [5]. Initially, the dataset contained several entries labeled as 'Not available' or 'Not Applicable,' which were replaced with NaN (Not a Number) to facilitate numerical analysis. Relevant numeric features were identified, including Sequence Length, isoelectric point (pI), Net Charge, Mass, Boman Index, and Aliphatic Index, crucial physicochemical features for supporting drug development and ongoing quality control. To handle missing values in the dataset, mean imputation was employed, wherein missing values were replaced with the mean of the respective feature. This approach ensures that the imputed values are within the range of observed data, maintaining the dataset's integrity [11]. Subsequently, the data was standardized using StandardScaler to ensure that all features contribute equally to the analysis [12]. Python was utilized as the programming language for data analysis, employing libraries such as pandas for data manipulation, scikit-learn for machine learning algorithms, and Matplotlib and Seaborn for data visualization [13].

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining the most significant variance. This step aids in simplifying the complex dataset, making it easier to visualize and identify patterns. PCA reduced the data to two principal components, which were then used for clustering analysis [8].

K-means clustering was employed to categorize the peptides based on the PCA-transformed data. The algorithm partitions the data into k-clusters, where each data point belongs to the cluster with the nearest mean value [9]. In this study, three clusters were identified, providing a clear segmentation of the peptides based on their physicochemical properties [14].

3. Results

The PCA reduced the DCTPep dataset to two principal components, capturing approximately 72.5% of the variance. This high percentage indicates that the majority of the information contained in the original high-dimensional dataset is preserved in these two components [8]. The scatter plot (Figure 1) illustrates the distribution of cancer therapy peptides in the reduced two-dimensional PCA space. Each point represents a peptide, colored according to its assigned cluster by the K-means algorithm (Cluster 0 in purple, Cluster 1 in teal, and Cluster 2 in yellow). Table 1 shows cluster 0 includes peptides with moderate sequence length, higher pI, and net charge. These properties suggest that the peptides have stronger interactions with negatively charged cell membranes, potentially influencing their efficacy and specificity [3]. Cluster 1 contains peptides with the longest sequence, highest mass, and a low Boman Index. These peptides which might have reduced off-target effects and improved stability, could be suitable for specific targeting [2]. Cluster 2 consists of the shortest peptides,

characterized by lower pI and net charge, and the highest affinity. These peptides may be highly effective in binding to their targets, contributing to their rapid penetration and action within target cells [4].

Our clustering results align well with the annotations in DCTPep, validating the effectiveness of our machine-learning approach. Peptides in Cluster 2, for instance, show high affinity and specific physicochemical properties consistent with known high-efficacy peptides in the database [5]. This analysis not only confirms existing knowledge but also provides a framework for discovering new insights and guiding future peptide design and development efforts [10].

Figure 1. Principal Component Analysis (PCA) Results with K-means Clustering

Table 1. Cluster Summary Statistics

Cluster	Sequence_Length	pI	Net_charge	Mass	Boman_Index	Aliphatic_Index	Affinity	PCA 1	PCA 2
0	16.95	11.77	5.65	230342.31	-3820.23	84.87	62.32	0.59	-0.90
1	31.93	8.31	3.02	421543.48	-8710.33	58.45	61.31	1.93	1.50
2	10.84	7.01	0.31	129484.38	-220.13	98.78	63.23	-1.60	0.33

4. Discussion and Conclusions

The application of PCA and K-means clustering has enabled us to categorize cancer therapy peptides into three distinct clusters based on their physicochemical properties, providing valuable insights into the key features that influence peptide efficacy and guiding the design of new therapeutic peptides. By reducing the dimensionality of the dataset, PCA helps in visualizing complex relationships between different peptide properties, while K-means clustering identifies natural groupings within the data [8].

Understanding the physicochemical properties that differentiate effective peptides from less effective ones can significantly inform the development of new peptides with enhanced therapeutic potential. Our findings suggest that specific characteristics, such as sequence length and net charge, are critical for peptide activity. Peptides in Cluster 1, characterized by the longest sequences, might be designed for stability and targeted delivery, which are important for maintaining therapeutic efficacy [2]. Peptides with higher net charges (Cluster 0) might interact more effectively with negatively charged

cell membranes, enhancing their ability to penetrate cells and exert their therapeutic effects [3]. The high affinity and specific mass properties in Cluster 2 suggest that these peptides are optimized for rapid and specific interactions with their targets, which is crucial for effective cancer therapy [4]. These insights can help in tailoring peptide sequences to enhance their binding affinity, stability, and overall therapeutic efficacy, ultimately leading to more effective cancer treatments. However, there are some limitations in this study. The dataset used has inherent limitations due to missing values and potential biases in the data sources. The use of mean imputation for missing values may not fully capture the true variability in the data. Additionally, the findings from this study may not generalize well to other datasets or peptide types without further validation. Future research should focus on validating these findings through experimental studies and integrating additional data sources to refine clustering and enhance model accuracy [10]. Conducting in vitro and in vivo experiments to test the efficacy and specificity of peptides from different clusters can validate the predictive power of the PCA and clustering models. Including additional datasets, such as peptide-drug conjugates and predicted 3D structures, can provide a more comprehensive view of peptide properties and improve model accuracy. Using advanced machine learning techniques and integrating multi-omics data can further enhance the understanding of the relationships between peptide properties and their therapeutic potential. By continuing to refine these models and integrating more data, researchers can develop more accurate and effective peptide-based therapies for cancer treatment.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin. 2020;70(1):7-30. doi:10.3322/caac.21590.
[2] Hamley IW. Introduction to Peptide Science. Wiley; September 2020. ISBN 978-1-119-69817-3.
[3] Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, et al. Therapeutic peptides: current applications and future directions. Signal Transduct Target Ther. 2022;7(1):48. doi:10.1038/s41392-022-00904-4.
[4] Bader JE, Enright HA, Seelig LL. Peptide-drug conjugates for targeting cancer: Current status and future prospects. Curr Pharm Des. 2011;17(28):3105-3121. doi:10.2174/138161211797931282.
[5] Sun X, Liu Y, Ma T, Zhu N, Lao X, Zheng H. DCTPep, the data of cancer therapy peptides. Sci Data. 2024;11(1):541. doi:10.1038/s41597-024-03388-9.
[6] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z.
[7] Kording KP, Benjamin AS, Farhoodi R, Glaser JI. The Roles of Machine Learning in Biomedical Science. In: National Academy of Engineering. Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2017 Symposium. Washington (DC): National Academies Press (US); 2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK481619/.
[8] Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202. doi:10.1098/rsta.2015.0202.
[9] Hartigan JA, Wong MA. Algorithm AS 136: A K-means clustering algorithm. J R Stat Soc Ser C Appl Stat. 1979;28(1):100-108. doi:10.2307/2346830.
[10] Schneider P, Walters WP, Plowright AT. Rethinking the medicinal chemistry toolbox: Novel ways to address biological complexity. J Med Chem. 2019;62(21):10796-10810. doi:10.1021/acs.jmedchem.9b00397.
[11] Little RJ, Rubin DB. Statistical Analysis with Missing Data. 3rd ed. Wiley; 2019. doi:10.1002/9781119482260.
[12] Shanker M, Hu MY, Hung MS. Effect of data standardization on neural network training. Omega. 1996;24(4):385-397. doi:10.1016/0305-0483(96)00010-2.
[13] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-2830. doi:10.48550/arXiv.1201.0490.
[14] Lloyd SP. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982;28(2):129-137. doi:10.1109/TIT.1982.1056489.