A software tool for data mining of physicochemical properties of peptides

Zhelyazko Terziyski[1], *Margarita* Terziyska[2,*], *Stanka* Hadzhikoleva[3] and *Ivelina* Desseva[4]

[1] Department of Computer Science and Mathematics, Faculty of Economics, Trakia University, 6000 Stara Zagora, Bulgaria
[2] Department of Mathematics, Physics and Information Technologies, Faculty of Economics, University of Food Technologies, 4002 Plovdiv, Bulgaria
[3] Department of Computer Informatics, Faculty of Mathematics and Informatics, Plovdiv University "Paisii Hilendarski", 4000 Plovdiv, Bulgaria
[4] Department of Analytical Chemistry and Physical Chemistry, Technological Faculty, University of Food Technologies, 4002 Plovdiv, Bulgaria

Abstract. Biologically active peptides (BAP) are increasingly in the focus of scientific research due to their widespread use in medicine, food and pharmaceutical industries. Researching and studying the properties of peptides is a laborious and expensive process. In recent years, in silico methods, including data mining or artificial intelligence, have been applied more and more to reveal biological, physicochemical and sensory properties of peptides. This significantly shortens the process of peptide sequences analysis. This article presents a software tool that uses a data mining approach to discover a number of physicochemical properties of a specific peptide. Working with it is extremely simple - it is only necessary to input the amino acid sequence of the peptide of interest. The software tool is designed to generate data in order to increase the classification and prediction accuracy, as well as to leverage the engineering of new amino acid sequences. This way, the proposed software greatly facilitates the work or scientific researchers. The software application is publicly available at www.pep-lab.info/dmpep.

1 Introduction

Peptides, as an intermediate element between amino acids and proteins, have a number of beneficial properties that help in the treatment of severe diseases such as cancer, diabetes, tuberculosis, COVID-19, etc. In addition, they are successfully used as immunostimulators or inhibitors. Because of this, scientists are increasingly interested in studying the properties of peptides – biologically active peptides (BAP) which can be extracted from plants, animals, some of their products, microorganisms, etc., or can be synthesized under laboratory conditions. In principle, peptides are isolated and subsequently studied via in vivo or in vitro procedures. This is time-consuming, labour-intensive, and involves significant resources and expenses. Artificial intelligence techniques have been increasingly used in bioinformatics in recent years to shorten this process and make it much cheaper. Through this kind of technique, the biological activity of the peptide is predicted with a very high probability. The analysis of proteins allows to determine whether peptides with a certain biological activity can be obtained from them. In this way, researchers can target sequencing of precisely defined peptides.

Research has shown that the natural biological properties of peptides are a complex combination of hydrophobicity, charge, molecular mass, reduction of the hydrophobic moment [1], and other physicochemical characteristics which can provide useful information in classifying, predicting, and synthesizing new peptides [2, 3].

Various software applications have been developed to calculate these physicochemical properties. The most widespread ones are web-based platforms such as ExPASy [4], APD [5], Peptide Tools (Peptide 2.0 Inc., Chantilly, VA, USA) and IPC 2.0 [6]. Desktop applications such as EMBOSS [7] are relatively less common. In recent years, open-access packages of R [8], Phyton [9], Perl [10], etc., have also become popular. A large number of these applications, however, cannot claim to be comprehensive. They calculate only a certain part of a peptide's properties, for example only the molecular mass and the isoelectric point [11]. Another group of software applications defines a larger set of characteristics, but each of them is calculated in a different module and it is difficult to combine the data. In addition, some of these applications do not allow downloading the received information in an editable file that is suitable for subsequent use.

The subject of the present study is the development of a software application DMpep (Data Mining from peptides) designed to overcome these shortcomings. DMpep is an open-access web-based tool that aims at a comprehensive analysis and computation of multiple heterogeneous features of peptides. DMpep is implemented as part of the cloud-based Pep-lab project

[*] Corresponding author: mterziyska@uft-plovdiv.bg

which also includes a database of information on food-derived peptides and a module for predicting the type of biological activity of peptides.

DMpep extracts information on multiple physicochemical characteristics – peptide length, molecular weight, hydrophobicity, Grand average of hydropathy index, aliphatic index, Protein-binding Potential index, acidity, polarity, isoelectric point, net charge, etc. The application calculates atomic, amino acid, and cluster amino acid composition, and performs primary analysis of the information. All data are visualized through text and graphics and can be downloaded in a format convenient for further processing and research in the sphere of bioinformatics. The module is intended to work primarily with peptides, but since there is no limit to the length of the sequence analyzed, it can also be successfully used for protein analysis.

2 Materials and methods

Peptides are chains of amino acids joined together by a peptide bond whose molecular mass is less than 10kDa. Larger structures are defined as proteins. Following this requirement, peptides can contain from 2 to about 50 amino acids. These structures have specific physicochemical characteristics that determine their properties and biological activity. Through DMpep, using only the amino acid sequence, information on a number of features can be retrieved.

2.1 Peptide length

One of the main functionalities of the application is the determination of the peptide length. It is determined by the number of amino acids contained in the peptide.

2.2 Atomic composition (AC)

Amino acids may contain carbon (C), hydrogen (H), nitrogen (N), oxygen (O), and sulphur (S) atoms. The sum of all atoms of the corresponding species represents the atomic composition. Each of these sums is an element of the vector AC_{1x5} represented by eq. (1), where L is the number of amino acids in the concrete peptide.

$$AC = [\sum_{i=1}^{L} C_i \;\; \sum_{i=1}^{L} H_i \;\; \sum_{i=1}^{L} N_i \;\; \sum_{i=1}^{L} O_i \;\; \sum_{i=1}^{L} S_i] \quad (1)$$

2.3 Molecular weight (MW)

Molecular weight is the sum of the masses of all the atoms contained in the specific amino acid sequence. It is directly related to the length and is measured in units called Daltons (Da). As already mentioned, peptides weigh less than 10 kDa (10000 Da). The variant for calculating the molecular weight used here sums up the molecular weights of the amino acids that make up the peptide, taking into account the weights of the molecules released when the peptide bonds are formed. MW is calculated according to eq. (2), where L is the number of amino acids, $AAMW_i$ is the molecular weight of the amino acid, and MW_{H2O} is the molecular weight of water.

$$MW = (\sum_{i=1}^{L} AAMW_i) - (L-1).\mathrm{MW}_{\mathrm{H2O}} \quad (2)$$

2.4 Grand average of hydropathy index (GRAVY)

The hydrophobicity of peptides is one of their most important characteristics. This is the force that determines the folding of the structure and changes depending on the solvent in which the peptide is present. The hydrophobicity index (GRAVY) is calculated according to eq. (3), where L is the number of amino acids, and H_i is the hydrophobicity of the amino acid.

$$GRAVY = \frac{\sum_{i=1}^{L} H_i}{L} \quad (3)$$

There are over 40 hydrophobicity scales for peptides [12]. Of those, 7 scales are used most often – Kyte-Doolittle, Engelman, Eisenberg, Hopp-Woods, Cornette, Rose, Janin. To calculate the hydrophobicity index in this part, the most popular scale is used - Kyte-Doolittle [13], where the value for the individual amino acid ranges from 4.5 to -4.5.

2.5 Aliphatic index (AI)

Aliphatic amino acids Alanine (A), Isoleucine (I), Leucine (L), and Valine (V) are responsible for the thermal stability of proteins. The aliphatic index is calculated according to eq. (4), proposed by Ikai [14], in which X_A, X_V, X_I, and X_L are the mole percentages (100 x mole fraction) of Alanine, Isoleucine, Leucine, and Valine. The coefficients a and b are the relative volume of Valine side chain and the Isoleucine/Leucine side chain respectively.

$$AI = X_A + a.X_V + b.(X_I + X_L) \quad (4)$$

2.6 Acidic/ Basic index (ABI)

Peptides have two amino acids with acidic properties – Aspartic acid (D) and Glutamic acid (E). Three others have the characteristics of a base – Arginine (R), Lysine (K), and Histidine (H). The peptide acidity is determined by calculating the values of the different groups, after which the property of the peptide is found by eq. (5), where ABI_A and ABI_B are the mole percentages of the amino acids with acidic and base properties in the peptide.

$$ABI = \begin{cases} Acidic, & ABI_A > ABI_B \\ Basic, & ABI_A < ABI_B \\ Neutral, & ABI_A = ABI_B \end{cases} \quad (5)$$

2.7 Protein-binding potential index (Boman index)

The Boman index [15] shows the protein-binding potential of the peptide. If it has a high value, it means that the peptide is likely to have several biological activities due to its ability to interact with a wide variety of proteins.

The Boman index is determined by summing the water solubility values of each amino acid and dividing the resulting number by the number of amino acids in the peptide chain. This is represented by f eq. (6), in which L is the number of amino acids, and S_i is the solubility according to the scale of Radzicka and Wolfenden [16].

$$BI = \frac{\sum_{i=1}^{L} S_i}{L} \quad (6)$$

2.8 Net charge

The side chains of some amino acids have a positive electrical charge, while others have a negative one. This charge has different values that depend on the pH values. The sum of the charges on all amino acids and the end terminals is called net charge. Positively charged are Arginine (R), Lysine (K), and Histidine (H), and the negatively charged are Aspartic acid (D), Glutamic acid (E), Cysteine (C), and Threonine (T).

The net charge can be calculated by using eq. (7) which is a variant of the Henderson-Hasselbalch equation presented by Moore [17].

$$Q = \sum Q^{-} + \sum Q^{+} \quad (7)$$

Using the eq. (8), we can calculate Q^{-} for each negatively charged amino acid and the C-terminal charge, and Q^{+} is found for each positively charged amino acid and the N-terminal charge by eq. (9).

$$Q^{-} = \frac{(-1)}{1+10^{(pKa-pH)}} \quad (8)$$

$$Q^{+} = \frac{(+1)}{1+10^{(pH-pKa)}} \quad (9)$$

In these formulas, pK_a signify the acid dissociation constant K_a as a negative decimal logarithm. There are different scales for pK_a values, and the ones used in this paper have been taken from the National Center for Biotechnology Information with the National Institutes of Health in the USA [18].

2.9 Isoelectric point (pI)

The isoelectric point is the pH value at which the net charge equals zero. In this case, the peptide is electrically neutral – the negative and positive charges are equal. At pH less than the isoelectric point, the peptide is positively charged, and at greater pH, it is negatively charged. The isoelectric point can be determined by calculating the net charge *Q* at different pH values. DMpep calculates the net charge at values of pH from 2 to 14.

2.10 Amino acid composition (AAC)

The frequency distribution shows the frequency of repetition of each of the 20 basic amino acids in the composition of the given peptide. Vector AAC_{1x20} is obtained, described by the eq. (10), each AA_i value of the vector representing the percentage ratio of a particular amino acid contained in the peptide, calculated by eq. (11).

$$AAC = [\mathrm{AA}_1 \ \mathrm{AA}_2 \ \dots \ \mathrm{AA}_{20}] \quad (10)$$

$$AA_i = \frac{Frequency\ of\ AA_i}{L} \quad (11)$$

2.11 Grouped amino acid composition (GAAC)

Grouped amino acid composition shows the distribution of amino acids grouped by different physicochemical characteristics. The percentage ratio of different groups of amino acids in the peptide is calculated according to their properties listed in Table 1.

Table 1. Physicochemical properties and the amino acids having the relevant property

Physicochemical property	Amino acids
Polar	S, T, C, N, Q, Y
Hydrophobic	F, I, L, M, V, W, A, P
Charged	K, H, R, D, E
Positively charged	K, H, R
Negatively charged	D, E
Aliphatic	I, L, V, A
Aromatic	F, H, W, Y
Tiny	G, A, S, P, V, T
Small	C, I, L, N, D, Q, K, E, M, H
Large	F, R, Y, W

As a result of GAAS, a 10-dimensional vector eq. (12) is obtained, the values of which are calculated by the eq. (13), where L is the number of amino acids in the peptide, and PCP is the percentage ratio of the amino acids belonging to the relevant group that are contained in the peptide.

$$GAAC = [PCP_1 \ \ PCP_2 \ \dots \ PCP_{10}] \quad (12)$$

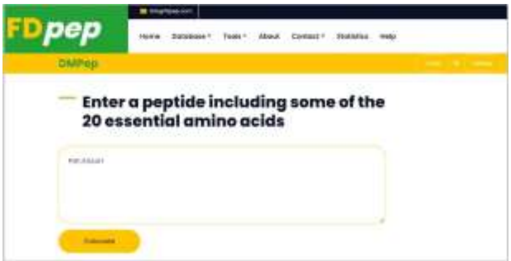
$$PCP_i = \frac{Frequency\ of\ PCP_i}{L} \quad (13)$$

3 Results and discussion

Several programming languages were used to implement DMpep - PHP 7.4, HTML 5, CSS 3, JavaScript, MySQL 8.0 database, Bootstrap 5.1 framework. The design is tested on the most popular browsers. It is implemented in a fully responsive and multi-column layout variant, so that it can be visualized in an appropriate way both on a computer and on small-screen devices such as a tablet and a phone.

Thanks to the Highcharts graphic libraries, all charts are dynamic and visualize additional data. It is possible to save the graphics in one of the popular raster and vector graphics formats, as well as save the data in tabular form.

To demonstrate the work with the DMpep module, let's look at a practical example with the amino acid sequence PSELSGAAH, extracted from the legume plant *Lupinusmutabilis* [19]. This is a peptide with ACE inhibitory and DPP-IV inhibitory biological activity. The results obtained are presented in the sections below.



Enter a peptide including some of the 20 essential amino acids

Fig. 1. DMpep home screen

The initial screen of the application requires the user to enter a peptide sequence (Fig. 1), after which the information about it is displayed.

3.1 Sequence information

Initially, the application displays general information – the amino acid sequence itself and its length, as well as a graphical representation of AAC in mole percentages (Fig. 2.).

3.2 Atomic information

The information about the atomic and molecular structure of the peptide generated by the application includes a molecular mass, a chemical formula that describes it, and an AC graph (Fig. 3.), as follows:

3.3 Net charge

The next section provides information about the charge. The charge values at pH=7 and the pH of the isoelectric point are calculated. Graphical visualization is provided for the change in charge at pH changing its value from 2 to 14 (Fig. 4.). The change step is one-tenth and each of the 120 charge values is displayed by hovering over the respective column of the graph. The specific results of the investigated peptide sequence PSELSGAAH are as follows:

3.4 Acidic/ Basic

Acidic/Basic calculates and visualizes the mole percentages of acidic, basic, and neutral amino acids.

Sequence: PSELSGAAH
Length: 9

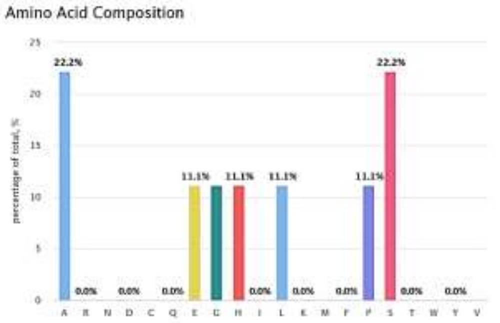


Fig. 2. Frequency distribution of amino acids

Molecular Weight (Da): 867.92
Chemical Formula: $C_{36}H_{57}N_{11}O_{14}$

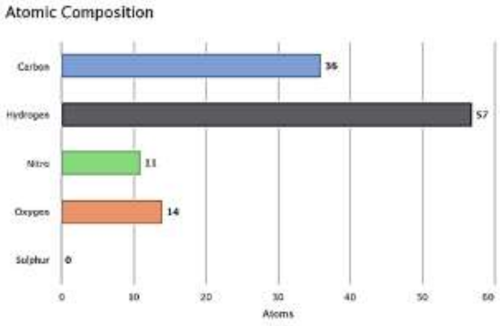


Fig. 3. Frequency distribution of the number of atoms in the peptide

Isoelectric Point (pI): 5.11

Net Charge (pH=7): -0.913

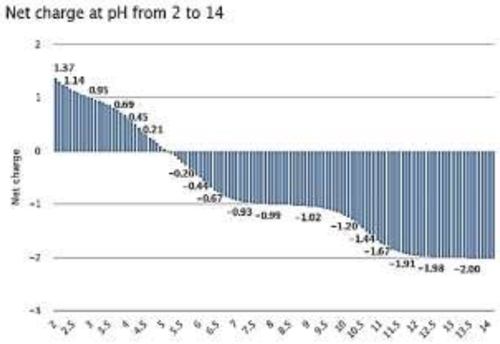


Fig. 4. Charge values at pH 2 to 14

Thanks to them, the status of the peptide is predicted. In the example elaborated above, the following data is obtained:

Status: NEUTRAL
Acidic: 11.11%
Basic: 11.11%
Neutral: 33.34%
Hydrophobic: 44.44%

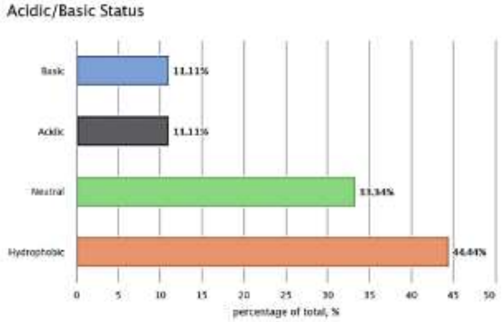


Fig. 5. Distribution of amino acids of acidic, basic, neutral and hydrophobic nature

3.5 Indexes

This functionality calculates the values of three important indices determining the physicochemical properties of the peptide – Grand average of hydropathy (GRAVY), aliphatic index, and Protein-binding Potential (Boman index). In this case, they are:

Hydrophobicity index (GRAVY): -0.3222
Aliphatic index: 65.56
Boman index: 0.9767

3.6 Grouped amino acid composition

Another important piece of information that the module can visualize is a graphical representation of Grouped amino acid composition (Fig. 6.).

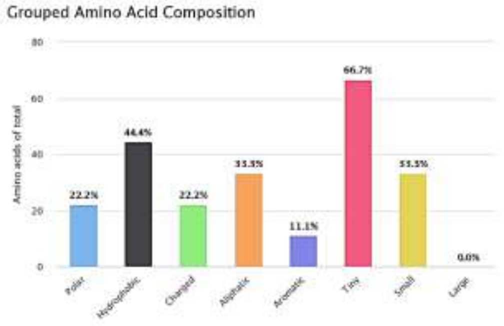


Fig. 6. Distribution of amino acids according to physicochemical properties

3.7. Download

The application provides the option to download all 46 calculated and analyzed features - *Sequence, Length, AAC* (20 features), *Molecular Weight, Carbon, Hydrogen, Nitro, Oxygen, Sulphur, Isoelectric Point, Acidic, Basic, Neutral, Hydrophobic,* GRAVY, *Aliphatic index, Boman index and GAAC (10* features*)*. This information can be used by researchers for further processing, for example, in predicting the biological activity of peptides.

Very often, researchers need features for more than one peptide, which is why the software module provides an option to analyse a data set of amino acid sequences. The generated file is a dimensional matrix: 46 x (length of data set).

4 Conclusion

This paper presents the software module DMpep which uses the data mining method to extract information from a given amino acid sequence and generate a large set of physicochemical characteristics, indices, and compositions.

The results of the application's work are presented in text and graphic form, but there is an option to save the data in a structured format in the form of a vector of 46 features. An option is provided for obtaining information about the physicochemical properties of both an individual peptide and a data set.

The information obtained from the DMpep module can be used by researchers as a basis for predicting the biological activity of peptides with artificial intelligence models.

The authors would like to acknowledge the support of the Research and International Sector at the University of Food Technologies (UFT), Plovdiv, Bulgaria under research project № 08/21-H funded through the UFT's Science Fund.

References

1. C. D. Fjell, J. A. Hiss, R. E. Hancock, G. Schneider, Nature Rev. Drug Discov. **11**, 37 (2012)
2. B. Manavalan, S. Basith, T. H. Shin, S. Choi, M.O. Kim, G. Lee, Oncotarget. **8**, 77121 (2017)
3. S. Akbar, M. Hayat, M. Tahir, K.T. Chong, IEEE Access. **8**, 131939 (2020)
4. P. Artimo, M. Jonnalagedda, K. Arnold, D. Baratin, G. Csardi, E. De Castro, H. Stockinger, Nucl. Acids Res. **40**, W597 (2012)
5. G. Wang, X. Li, Z. Wang, Nucl. Acids Res. **44**, D1087 (2016)
6. L.P. Kozlowski, Nucl. Acids Res. **49**, W285 (2021)
7. A. L. Lamprecht, S. Naujokat, T. Margaria, B. Steffen, J. Biomed. Seman. **2**, 1 (2011)
8. D. Osorio, P. Rondón-Villarreal, R. Torres, Small. **12**, 44 (2015)
9. L. P. Kozlowski, Biology Direct. **11**, 1 (2016)
10. J. E. Stajich, An Introduction to BioPerl. In: *Plant Bioinformatics* (Humana Press, Totowa 2007)
11. I. J. Brum, D. Martins-de-Souza, M. B. Smolka, J. C. Novello, E. J. Galembeck, Comput. Sci. Syst. Biol. **2**, 093 (2009)
12. S. Simm, J. Einloft, O. Mirus, Schleiff, Biol. Res. **49**, 1 (2016)
13. J. Kyte, R. F. J. Doolittle, Mol. Biol. **157,** 105 (1982)
14. A. Ikai. J. Biochem. **88,** 1895 (1980)
15. H. G. Boman, J. Internal Med. **254**, 197 (2003)

16. A. Radzicka, R. Wolfenden, Biochem. **27**, 1664 (1988)
17. D. S. Moore. Biochem. Edu. **13**, 10 (1985)
18. S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, E. E. Bolton, Nucl. Acids Res. **47**, D1102 (2019)
19. E. B., Munoz, D. A. Luna-Vital, M. Fornasini, M. E. Baldeón, E. G. J. de Mejia, Func. Foods. **45**, 339 (2018)