Heliyon 8 (2022) e12283

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Peptide Utility (PU) search server: A new tool for peptide sequence search from multiple databases

Tanishq Chamoli [a,1], Alisha Khera [b,c,1], Akanksha Sharma [b,d,1], Anshul Gupta [a,1], Sonam Garg [a,1], Kanishk Mamgain [a,1], Aayushi Bansal [a,1], Shriya Verma [a,1], Ankit Gupta [a], Hema K. Alajangi [b,d,**], Gurpal Singh [d,***], Ravi P. Barnwal [a,*]

[a] Department of Computer Science and Engineering, Chandigarh College of Engineering and Technology, Chandigarh, India
[b] Department of Biophysics, Panjab University, Chandigarh 160014, India
[c] National Centre for Cell Science, NCCS Complex, S. P. Pune University Campus, Ganeshkhind, Pune, Maharashtra 411007, India
[d] University Institute of Pharmaceutical Sciences, Panjab University, Chandigarh 160014, India

## HIGHLIGHTS

- A logical approach to finding homologous and identical protein/peptide sequences.
- A Peptide Utility (PU) search webserver is developed for homologous and identical searching from the available sequences.
- Peptide sequences can be searched from the ~0.4 million entries from different databases in both online and offline mode.
- This will open up avenues to explore peptide design for various pharmacological interventions.

## ARTICLE INFO

*Keywords:*
Proteins
Peptide sequence
PDB
Homologous pairs
Identical pairs

## ABSTRACT

Proteins are essential building blocks in humans that have garnered huge attention from researchers worldwide due to their numerous therapeutic applications. To date, different computational tools have been developed to extract pre-existing information on these biological molecules, but most of these tools suffer from limitations such as non-user friendly interface, redundancy of data, etc. To overcome these limitations, a user-friendly interface, the Peptide Utility (PU) webserver (https://chain-searching.herokuapp.com/) has been developed for searching and analyzing homologous and identical protein/peptide sequences that can be searched from approximately 0.4 million sequences (structural and sequence information) in both online and offline modes. The PU web server can also be used to study different types of interactions in PDBSum, identifying the most dominating interface residues, the most prevalent interactions, and the interaction preferences of different residues. The webserver would also pave way for the design of novel therapeutic peptides and folds by identifying conserved residues in the three-dimensional structure space of proteins.

## 1. Introduction

Proteins are the primary workhorses of biological systems, serving a range of functions and possessing antimicrobial, antifungal, antiviral, antiparasitic, and anticancer properties [1, 2, 3, 4]. Peptides have favorable pharmacokinetic properties and a wide range of biological activities which makes them ideal for the treatment of a variety of diseases such as cancer, immune disorders, cardiovascular disease, gastrointestinal dysfunction, hemostasis, and microbial infections [5, 6, 7]. The intermolecular interactions in the proteins can be utilized for pharmacological research [8, 9]. Moreover, detailed knowledge about the dynamics of these biomolecules is needed to understand their functions and intermolecular interactions [10]. It also helps in identifying important functional sites in proteins, sites involved in disease

* Corresponding author.
** Corresponding author.
*** Corresponding author.
*E-mail addresses:* hemasai.biotech@gmail.com (H.K. Alajangi), gurpalsingh.ips@gmail.com (G. Singh), barnwal@pu.ac.in (R.P. Barnwal).
[1] Denotes equal contribution as the first author.

https://doi.org/10.1016/j.heliyon.2022.e12283
Received 23 September 2022; Received in revised form 21 October 2022; Accepted 5 December 2022

manifestation, identifying drug targets, and studying various mutations associated with them [11]. Designed and stabilized α-helical ligands are being explored for their therapeutic applications because they can modulate biomolecular interfaces, have high thermal stability and can act as promising platforms for developing peptide-based pharmaceuticals [12].

### 1.1. Application and use of the database

One of the major challenges faced by researchers and pharmaceutical companies is the interpretation of the data being deposited in various databanks. In the past few years, peptide-based drug development has garnered attention across the globe due to high selectivity, better tolerability, and reduced production costs [13]. The current work involves the development of a web-based interface that searches for homologous and identical peptide sequences from various database repositories. These searches can help in studying the evolution of species and their shared ancestry, evolutionary pathways for proteins, structure prediction, and many more. The application developed consists of approximately 0.4 million entries from various databases which makes the search more exhaustive and meaningful. This application can be used for the analysis of all kinds of peptide sequences deposited in several databases from various source organisms. It provides the user with the option to search for homologous or identical sequences and to select a specific database during identical searches. This application obtains data from various kinds of databases and not from a single database which makes it much more comprehensive as compared to other applications. While designing therapeutic peptides, the researcher faces many challenges like designing sequences for the secondary structure for instance, loops and turns. The modelling and designing of peptides exploits the relationship between structure and peptide sequence [14]. The construction of protein folds depends mainly on the conversion of a linear polypeptide chain into a three-dimensional structure. This can be possible by identifying various homologous and identical sequences for which structural details have already been elucidated. Design of protein folds *de novo* requires a thorough knowledge of protein structure and its stability [15]. This application allows the user to search for 6–7 amino acids long sequences from all the databases or a specific database and with the help of homologous and identical search, the user can identify the lead residues involved in the complex architecture of proteins. This will help the user in the designing of new peptides. The current application can be further modified to include secondary structures of the results obtained in the homologous and identical search from the PDB database. Along with facilitating the study of protein-peptide interactions, the application developed can be used for the designing of new therapeutic molecules as well as studying various conserved domains crucial for deciphering the functions of unknown proteins. It contains information on peptides extracted and studied from a variety of experiments and not a single experiment, thus differentiating it from databases such as Bolt, PepArML, PeptideAtlas, etc. which contain information on peptides obtained through mass spectrometry (MS) experiments [16, 17, 18]. The web-based interface developed in the current work can also be curated to study the peptide structure and function. Furthermore, evolutionary relatedness among a group of organisms that share related peptides from a family can also be analyzed [19].

Searching for homologous sequences using sequence similarity method is one of the first and the most crucial steps in the analysis of new sequences. Since protein databases tend to be large, over 80% of metagenomic sequences have significant similarities to proteins already found in databases [20]. Pattern similarity searching normally performed using BLAST [20] is the most widely employed method for characterizing novel sequences. To identify "homologous" sequences (statistically substantial similarity) by comparing different proteins or gene sequences, similarity searches are performed, which may reflect shared ancestry. Protein structure prediction is primarily focused on modelling proteins to known structures through homology approaches.

### 1.2. Similar databases

Studying the evolution of proteins can only be possible through the identification of identical and homologous sequences. So various research groups across the world have attempted to develop applications or software for sequence database search like PepBank [21], Propedia [22], and PeptideDB [19]. This approach facilitates peptide recognition and identical and homologous searches via the incorporation of a large number of databases [23]. PepBank is a database with 19,792 entries with major databases from public sources like Artificially Selected Proteins/Peptides Database (ASPD) and Universal protein resource (Uniprot). Some of the applications of this database include prediction of binding partners of biological peptides and the development of various peptide-based therapeutics [21]. Propedia is a database that can be used for searching and visualizing protein-peptide complexes, with over 19,000 high-resolution structures from PDB along with their structural and sequence information. Studying protein-peptide interaction can help in the design of new therapeutic molecules [22]. PeptideDB is a database of bioactive peptides with 20,027 entries with different peptide motifs [19]. PeptideAtlas is a compilation of peptides from various species obtained through tandem MS experiments. It incorporates raw data from mass spectrometer output files which can be used for experimental design [16]. Bolt is a cloud-based MS-MS peptide search engine which has the capability to search from 9,00,000 sequences with 41 post-translational modifications (PTMs); as the size of the proteome is increasing at an exponential rate which it can handle in an efficient manner [17]. PepArML meta search peptide identification platform provides a search interface to study tandem MS data via the incorporation of seven search engines which help in the identification of the best peptide for each spectrum [18]. PepServe is a webserver for peptide analysis along with their clustering and visualization, thus making it a useful tool to understand the distribution of features in the peptides. With this application, an analysis of human peptide sequences can be done via UniProt using a set of selected peptide features [24].

Databases like UniProt provide extensive information about protein sequences [25], and the Biomolecular Interaction Network Database and related tools elucidate interactions of proteins [26], while databases like ASPD [27] are used specifically for peptides. Despite the availability of numerous such data repositories, a large fraction/chunk of data has not been tapped [28, 29]. PDB is one of the primary data repositories with 3D structural coordinates for macromolecular structures including proteins and nucleic acids determined experimentally using X-ray crystallography, nuclear magnetic resonance (NMR), and electron microscopy. The PDB archive started with fewer structures, and with each passing year, more entries were added [30, 31]. Alpha fold is another artificial intelligence system being used for the prediction of protein structures from amino acid sequences obtained via sequencing techniques. The structures obtained using this software have been observed to be profoundly similar to the experimental structures. Alpha fold has been developed by DeepMind and European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) [32, 33]. To date, training language models for protein sequences has been done at a smaller scale, but the precise details of what biological information they learn on scaling up are not known. For overcoming these limitations, ESMetafold has been recently developed. This enables 3D structure prediction of protein at the atomic level from its sequence. It is comparable to RoseTTAFold and AlphaFold2 in terms of accuracy [34]. In the last few decades, various advanced experimental techniques for determining protein structures have emerged. Several existing databases which have been used in the current study are enlisted in Table 1. The webserver thus developed by our group deploys sequence information from these databases to search for identical and homologous sequences in response to the input sequence of the user.

Rapidly expanding sequence databases include entries for various evolutionary linked proteins, i.e., homologous proteins, grouped into various protein families and aligned via broad multiple sequence

**Table 1.** Peptide and protein databases used in the current study.

| Databanks | Description | URL |
|---|---|---|
| AVPdb | Database of antiviral peptides (experimentally validated) against viruses [38] | http://crdd.osdd.net/servers/avpdb |
| BaAMPs | Database of biofilm active antimicrobial peptides [39] | http://www.baamps.it/ |
| BactPepDB | Database of predicted peptides from prokaryotic genomes [40] | http://bactpepdb.rpbs.univ-paris-diderot.fr/cgi-bin/home.pl |
| CancerPPD | Database of anticancer peptides and proteins [41] | http://crdd.osdd.net/raghava/cancerppd/ |
| Cell-Penetrating peptides (CPP) | CPP site – Database of cell-penetrating peptides which are manually curated [42] | http://crdd.osdd.net/raghava/cppsite/ |
| DADP | Database of anuran defense peptides [43] | http://split4.pmfst.hr/dadp/ |
| DBAASP | Database for antimicrobial peptide and structure of peptides [44] | https://dbaasp.org/ |
| PEPlife | Database of the half-life of peptides [45] | http://crdd.osdd.net/raghava/peplife |
| PeptideDB | Database for bioactive peptides [19] | http://www.peptides.be/index.php?p=home |
| Propepper | Database for food disorders related to wheat [46] | https://propepper.net |
| PlantPepDB | Database of phytopeptides [47] | http://14.139.61.8/PlantPepDB/index.php |
| RCSB PDB | Database for various biomolecules [48] | https://www.rcsb.org |
| SATpdb | Database of therapeutic peptides which are structurally annotated [49] | http://crdd.osdd.net/raghava/satpdb/links.php |
| TopicalPdb | Database for topically delivered peptides [50] | http://crdd.osdd.net/raghava/topicalpdb/ |
| TumorHoPe | Database for tumor homing peptides [51] | http://crdd.osdd.net/raghava/tumorhope/ |
| YADAMP | Yet another database of antimicrobial peptides [52] | http://yadamp.unisa.it/ |

alignments (MSA). Pairwise sequence identities between homologous proteins usually drop to 20–30%, or even lower in several cases [35, 36, 37]. Such low sequence identities are remarkable, given that even a few spontaneous mutations may destabilize or interrupt the function of a protein. The idea was to create the PU search server/platform where all major peptide databases including sequence information from PDB are connected and used to explore specific peptide sequences, whether identical or homologous. This involves as many as 0.4 million entries via online cloud platforms like Selenium framework, MongoDB, and Heroku. We envision that the information fetched through this webserver would open new avenues for designing and exploring novel peptides as therapeutics for pharmacological interventions.

## 2. Methodology

### 2.1. Data collection

Data was collected from various databases, including the RCSB PDB (https://www.rcsb.org/). All the databases (total 16) used for the current study are enlisted in Table 1.

**AVPdb** is a database of antiviral peptides (AVP) that are experimentally verified against approximately 60 viruses using a different database that includes anti-HIV peptides (HIPdb) [38]. **BaAMPs** is a database of antimicrobial peptides (AMPs) with the capability to specifically target the microbial biofilm [39]. **BactPepDB,** a database of predicted peptides from prokaryotic genomes aims at providing reannotation to the genome i.e., chromosomal as well as plasmid DNA obtained from RefSeq which are already defined. It provides key information of peptides, their conservation through evolution, and their biological and structural features [40]. **CancerPPD** is a database of anticancer peptides and proteins which are experimentally determined with data curated from various articles and patents. This database provides information on the tertiary structure of the peptides as well as about various modifications in the amino acids [41]. **CPP site** is a database of cell-penetrating peptides; the recent version of this database contains information on approximately 1850 cell-penetrating peptides. These peptides have the capability to enter eukaryotic cells without causing any major damage to the membranes, so they can be used for therapeutic purposes [42]. **DADP** is a database of anuran defence peptides as anuran tissue-like skin is rich in bioactive peptides e.g., AMPs [43]. **Database of Antimicrobial Activity and Structure of Peptides (DBAASP)** is curated manually for designing compounds with antimicrobial activity. This database has information on the 3D structure of AMPs and their activities [44]. **PEPlife** is a database that contains information about the half-life of peptides as the short half-life of peptides significantly impacts their development as therapeutic peptides. It contains information on about 1193 unique peptides. Bioactive peptides have a crucial role in the regulation of a majority of biological processes in living organisms and several peptides with clinical and industrial importance have been discovered *in vitro* [45]. **PeptideDB** contains information about naturally occurring peptides and includes various AMPs, peptide hormones, etc. [19]. **ProPepper** database consists of prolamin proteins obtained from the family of grasses (Poaceae) along with their peptides. It also contains B- and T-cell specific epitopes responsible for food disorders related to wheat. Datasets are collected from several public databases like NCBI GenBank etc. [46]. **PlantPEPDB** is a database that is curated manually for phytopeptides involved in plant defense mechanisms against pathogens and has potential therapeutic properties with 3848 entries [47]. Research Collaboratory for Structural Bioinformatics (**RCSB**) Protein Data Bank (**PDB**) provides access to information on three-dimensional (3D) structures of various biological molecules like nucleic acids, and proteins. It further supports data deposition and validation of structural data [48]. Database of structurally annotated therapeutic peptides (**SATPdb**) consists of information about antihypertensive, anticancer, and AMPs used for drug delivery with assigned structures. This database has 19,192 entries of experimentally validated therapeutic peptides [49]. **TopicalPdb** is a database for topically delivered peptides. This includes entries for peptides delivered via the eye, nose, and skin. Every entry provides information about the nature of the peptide, its origin, length, various modifications, etc. [50]. Tumor homing peptides are the peptides that play a crucial role in the delivery of anticancer drugs in tumor tissues. These possess high specificity and thus a database containing information about these peptides has been developed with the name **TumorHoPe.** All the details about these peptides and their target cells available on the database are experimentally validated [51]. **YADAMAP** is a database that contains information about 2133 peptides that show activity against bacteria. This database is different from other databases in that it contains information about antimicrobial activity against a variety of bacterial strains [52].

### 2.2. Architecture and interface of the database

Figure 1 showcases an abstract view of the approach used in the creation of this application/web server. As there are exceedingly large numbers of entries, web scraping was performed by using the Selenium framework. Selenium [https://selenium-python.readthedocs.io/] is a web application/website automation research tool that monitors the browser to access the website like a human. Finally, the script was automated to "go to page", extracting the data and then saving it. After extraction and pre-processing of the data, the sequences on the cloud were inserted using MongoDB (https://www.mongodb.com/home).

### 2.3. Database design: data scraping and database insertion

Selenium employs a web-driver package that can take control of the browser and imitate user-oriented behavior. An HTTP request was made

to fetch data from the RCSB website. Thus, a list of PDBs on that webpage was fetched and stored in a file and this process was repeated until the data from the last page was fetched. Then, at that point, the server traversed through the downloaded PDB list and made an HTTP request to extract the required FASTA sequence that was appended to the list of PDBs. After this extraction process, a JSON file was created which had a key (the PDB code), the URL from where the PDB entry was downloaded, and the FASTA sequence of the PDB entry. The URL contains basic information about the PDB entries, such as the biological assembly, digital object identifier (DOI) number, classification, organism(s), expression system, and mutations. The data was then embedded into MongoDB, which is an open-source cross-platform, document-oriented software application. It is a NoSQL database software that works with JSON-like documents and optional schemas using threading (after checking whether all threads are completed or not). The whole process was optimized using threading. But if the threads were not completed, then the whole process after traversal through the PDB list is repeated (Figure 2).

### 2.4. Web interface and search modules

To incorporate the search queries for both identical and homologous sequences, an application was developed that was deployed using Heroku used for deploying websites (https://www.heroku.com/). This application program has two separate functions, one for identical searching and another for homologous searching. Thus a website with two search bars was created. At the point when a user enters a specific sequence in the search bar, the query parameters are forwarded to the application program, which further calls the respective function. For example, if a user enters the sequence for homologous searching, the application program will call the function corresponding to homologous search and the same applies to identical searching. In the case of the identical chain function, the input provided by the user was used, however, in the case of the homologous chain function a list of similar chains was prepared by referring to the inset table in Figure 3, and afterward making a list of unique chains. A schematic visual Figure 3 was added for better clarity of the concept of

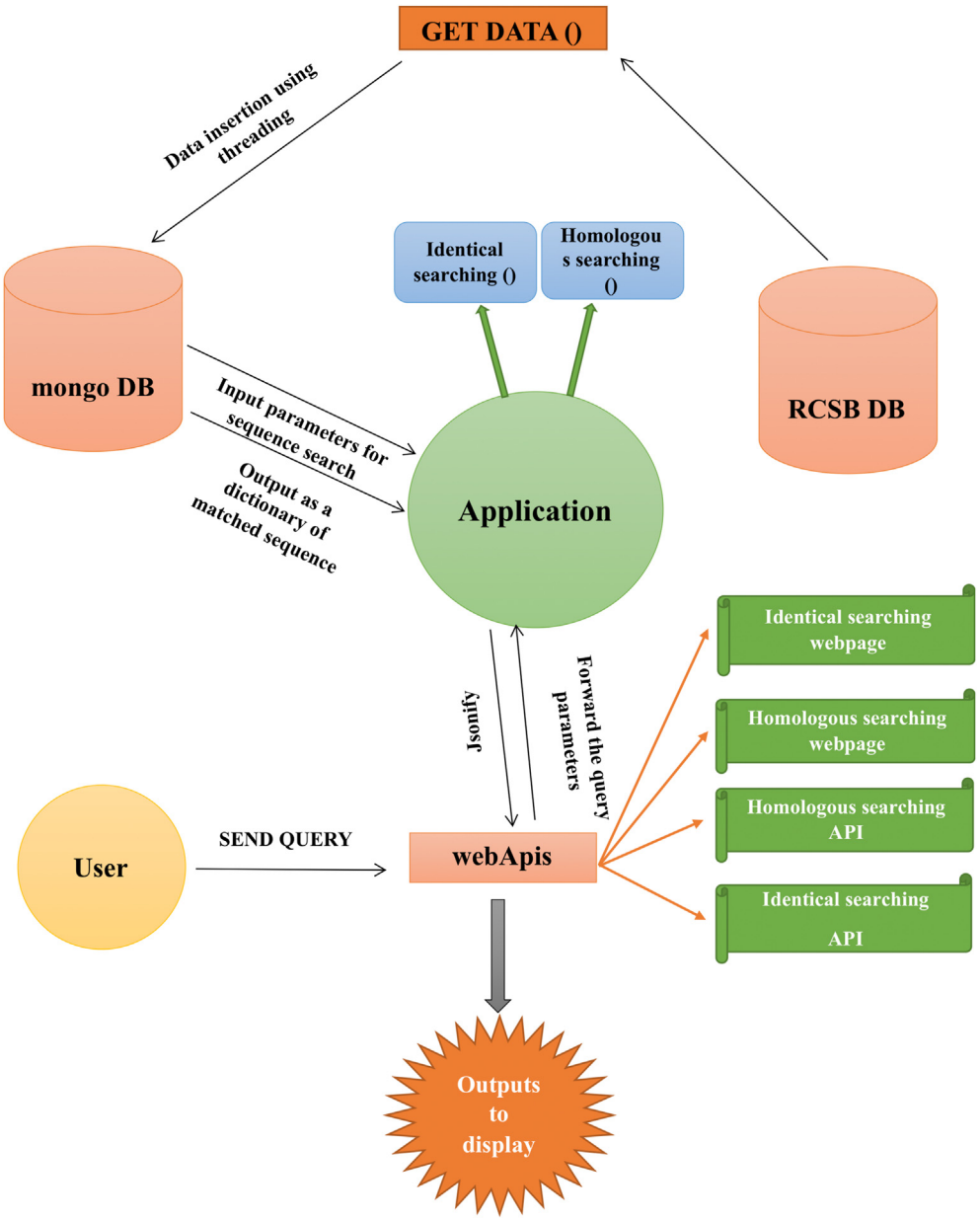



**Figure 1. Flowchart for application development** – Flowchart representing the abstract view for the creation of the application/web server.

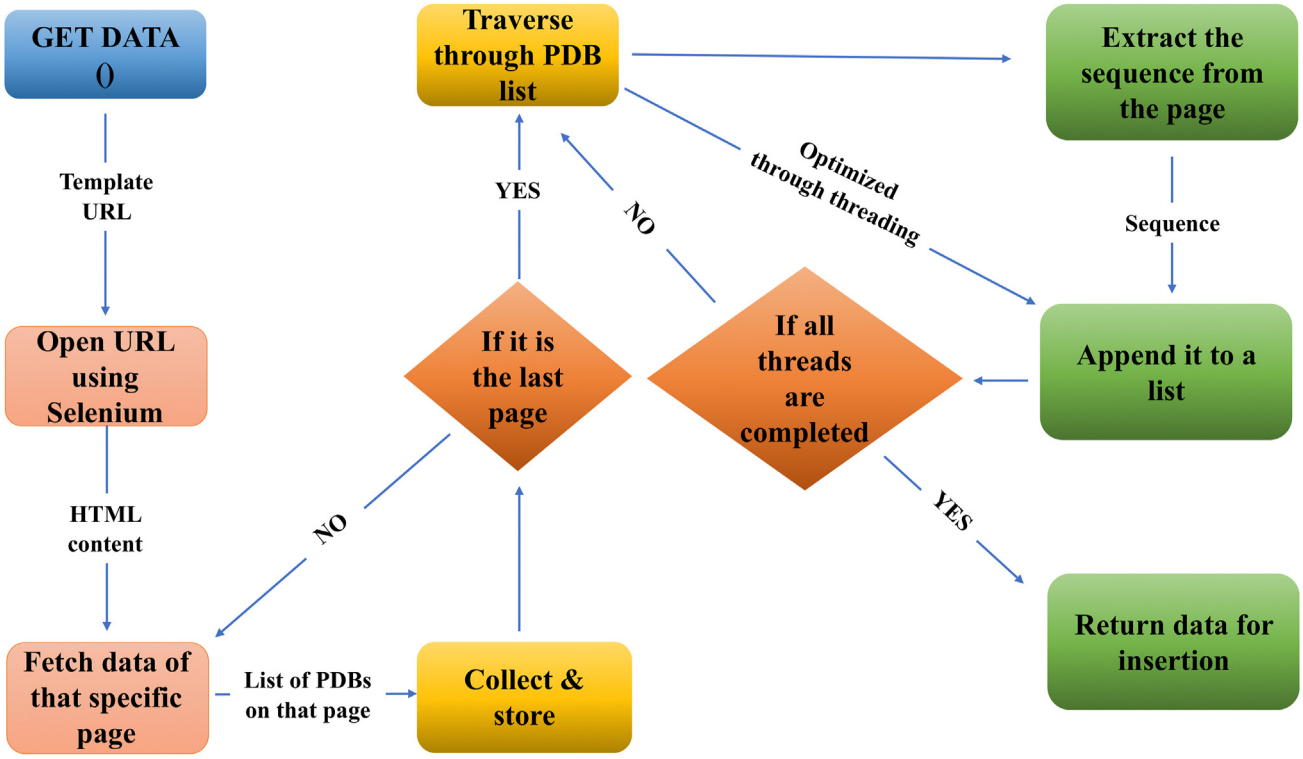
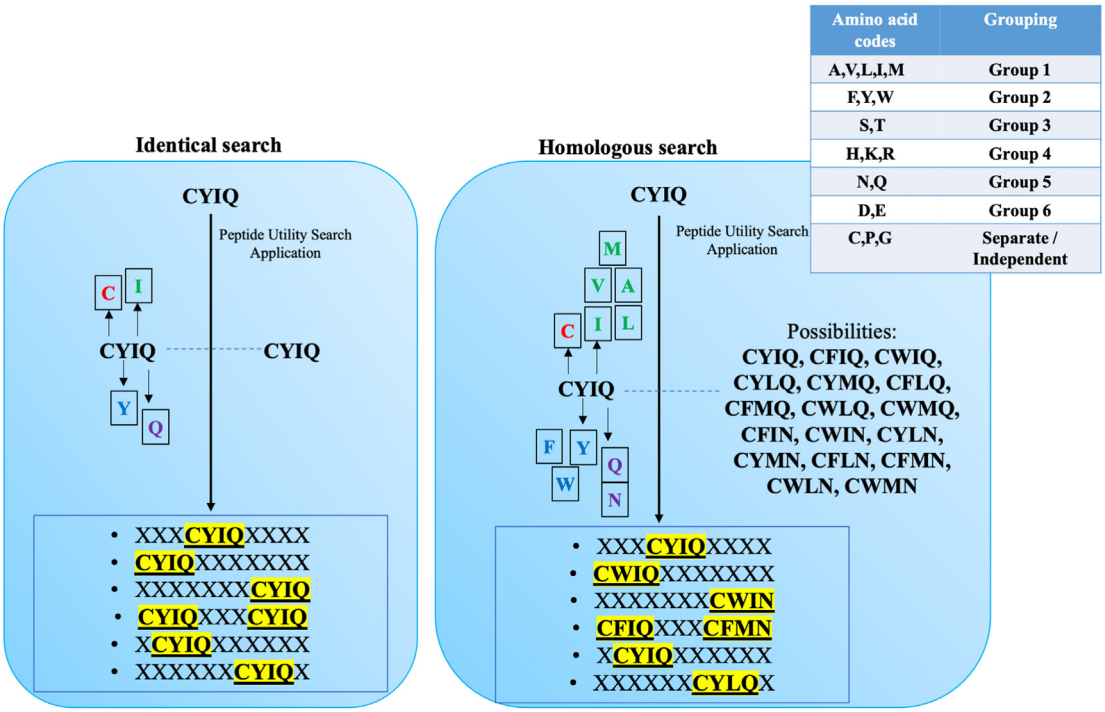



**Figure 2. Pre-processing and data collection** – Flow of pre-processing and data collection for the PU search server.

identical and homologous search outcomes. The application program then forwards these queries for sequence search to MongoDB. MongoDB at that point processes these requests and checks for similar sequences in the database. All such sequences which are similar to the one entered by the user are shown as output in the form of a dictionary that consists of a collection of key-value pairs. Each key-value pair maps the key to its associated value. An API has also been developed for terminal users (Figure 4). Flask was used for the development of the backend of the code (https://flask.palletsprojects.com/en/2.0.x/) and the limit was kept to 100 sequences. A GUI program was also developed for exactly the same thing, and since it is local, all the data can be obtained without rate limitation and it is stored locally in a JSON formatted file. As a result, an offline




| Amino acid codes | Grouping |
|---|---|
| A,V,L,I,M | Group 1 |
| F,Y,W | Group 2 |
| S,T | Group 3 |
| H,K,R | Group 4 |
| N,Q | Group 5 |
| D,E | Group 6 |
| C,P,G | Separate / Independent |

**Figure 3. Schematic representation of Identical and Homologous Search in PU search application** – Difference between the outcome of identical and homologous search in the search application is provided with an inset "table" representing the grouping logic for different amino acids in homologous pairs search.

tool, as well as an online tool that uses data from the cloud and permits internet connection, was developed.

### 2.5. Availability and weblinks

The PU application is freely available as an online and offline tool with different search options displayed on the homepage (https://chain-searching.herokuapp.com/). The homepage provides the user with a link on top for identical (https://chain-searching.herokuapp.com/identical) and homologous search (https://chain-searching.herokuapp.com/homologous) separately. This application also provides the user with an option for identical search from a specific database under the DB search tab (https://chain-searching.herokuapp.com/dbSearch) on the homepage and also with a Web API tab (https://chain-searching.herokuapp.com/api_web) which allows the user to download the results in API format to use this application in their project which makes the application much more user friendly.

## 3. Results

### 3.1. Rationale behind development

The PU search application is a web-based interface to search for identical and homologous peptide and protein sequences from various databases (https://chain-searching.herokuapp.com/). This sequence search tool has been developed to search for identical or homologous sequence pools. This is based on a web format, which is easily accessible to users around the globe. It can be modified for use in predicting the binding partners of biologically significant peptides, creating peptide-based therapeutic or diagnostic agents, and predicting molecular targets or binding specificities of peptides selected by phage screen. Data from different databases (as listed in Table 1) was collected and stored in one place so the user does not need to navigate different websites to obtain data. In this interface, a user can input a particular sequence and search for identical and homologous sequences in a single application. Figure 5 shows the home page of the interface which has many tabs that the user can utilize as per their requirement.

### 3.2. Identical and homologous search functions in the application

The home page provides a tutorial for users describing the working of the application. Identical and homologous search for a sequence can be done directly through the tabs on the top, while if the user wants to access a specific database for identical search; it can be done by using the DB-Search tab which provides a list of various databases before searching the sequence. Two different formats were created, one is a web API and the other is a JSON API (Figure 6). This allows the user to choose either one as per the need under the DB-Search tab. In identical searching, the FASTA sequences generated as output will include the exact input sequence. The output of the identical search contains 4 attributes, the key value for which is the name or PDB code of that protein (if any), the FASTA sequence, the URL of the sequence diverted to the page which contains information of a particular output, and the database name from where it is fetched. In homologous searching, the outputs have a relation with the input sequence, which implies they have a common progenitor as per grouping in the inset table in Figure 3. Unlike identical searching, the output FASTA sequences in homologous searching may or may not possibly contain the exact input sequences. The output of the homologous search contains 5 attributes, the key value for which is the name or PDB code of that protein, the FASTA sequence, the URL of the page which contains information about that particular output, and the database name from where it is fetched and the homologous sequence which it was flagged for.

### 3.3. Case studies for identical and homologous search

#### 3.3.1. Identical search

For example, Oxytocin is a peptide of nine amino acids. The FASTA sequence of Oxytocin is CYIQNCPLG. At the point when this sequence is taken as input for identical searching, it displays 6 unique identical FASTA sequences along with their key values and database names which imply that this same sequence will be present in all those 6 FASTA sequences (Figure 7A). Highlighted portion in each result depicts the sequence which was used for the search and where it is present in the peptide sequence of the results.

#### 3.3.2. Homologous search

For example, if the input sequence for homologous searching is CYIQ, the result would include related PDBs and the connection between the outputs and input sequences can be found as mentioned in the inset table in Figure 3. As shown in the table, only cysteine (C), glycine (G), and proline (P) are treated as separate characters, so they will not be mapped with any other characters. Moving to the next character 'Y', it can be seen that it belongs to group 2 which has total 3 characters, 'F', 'Y', and 'W'; this means that 'Y' can be mapped to 'F' and 'W', therefore, CFIQ and CWIQ are

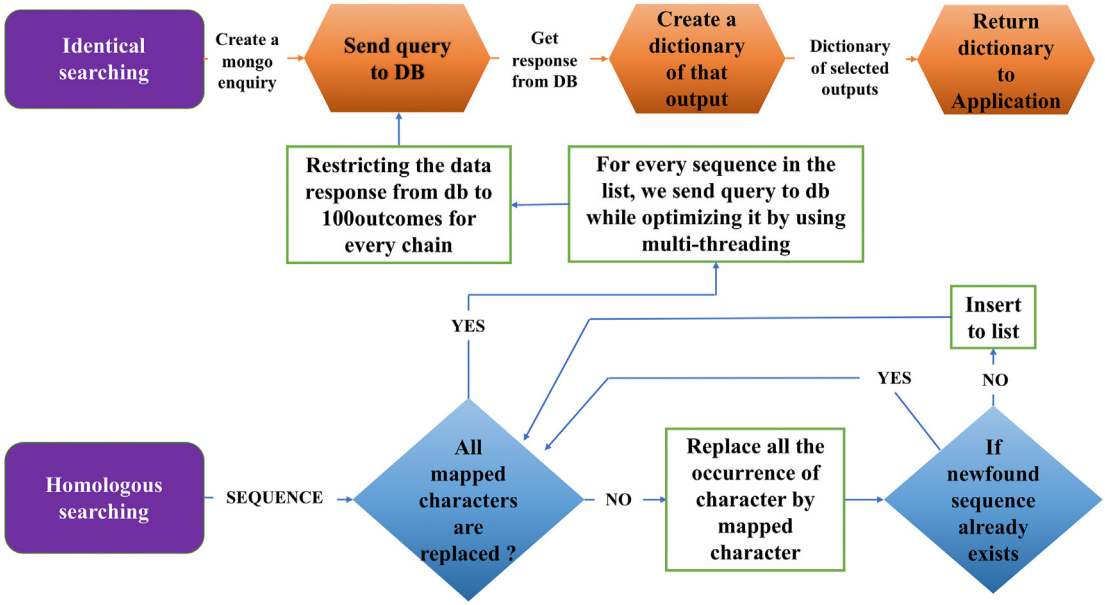
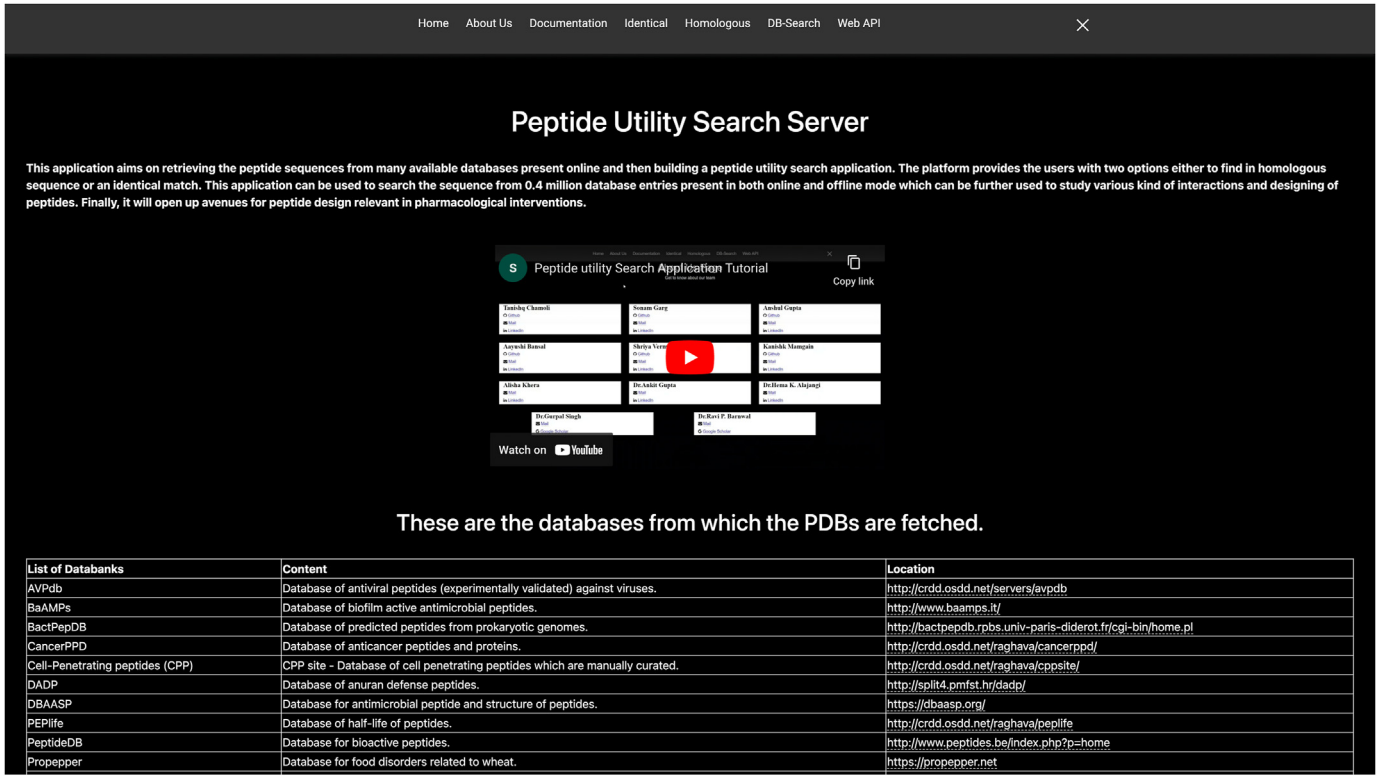



**Figure 4. Mechanism of search** – Flow of Identical and Homologous search mechanism.




| List of Databanks | Content | Location |
|---|---|---|
| AVPdb | Database of antiviral peptides (experimentally validated) against viruses. | http://crdd.osdd.net/servers/avpdb |
| BaAMPs | Database of biofilm active antimicrobial peptides. | http://www.baamps.it/ |
| BactPepDB | Database of predicted peptides from prokaryotic genomes. | http://bactpepdb.rpbs.univ-paris-diderot.fr/cgi-bin/home.pl |
| CancerPPD | Database of anticancer peptides and proteins. | http://crdd.osdd.net/raghava/cancerppd/ |
| Cell-Penetrating peptides (CPP) | CPP site - Database of cell penetrating peptides which are manually curated. | http://crdd.osdd.net/raghava/cppsite/ |
| DADP | Database of anuran defense peptides. | http://split4.pmfst.hr/dadp/ |
| DBAASP | Database for antimicrobial peptide and structure of peptides. | https://dbaasp.org/ |
| PEPlife | Database of half-life of peptides. | http://crdd.osdd.net/raghava/peplife |
| PeptideDB | Database for bioactive peptides. | http://www.peptides.be/index.php?p=home |
| Propepper | Database for food disorders related to wheat. | https://propepper.net |

**Figure 5. Home-page of the webserver** – Home-page for the interface showing different tabs which can be employed by the user.

homologous to the input sequence CYIQ. Thus, all those FASTA sequences containing CFIQ and CWIQ will be shown as output for the input sequence CYIQ. Similarly, checking for the next character, 'I', which belongs to group 1 containing five amino acids, signifies 'I' can be mapped to other five aliphatic amino acids. Now combining this mapping along with the previous mapping of Y, 8 such sequences were obtained after the search, which include CFIQ, CWIQ, CYLQ, CYMQ, CFLQ, CFMQ, CWLQ, and CWMQ. Similarly, including the options for 'Q', a total of 16 sequences were obtained: CFIQ, CWIQ, CYLQ, CYMQ, CFLQ, CFMQ, CWLQ, CWMQ, CFIN, CWIN, CYLN, CYMN, CFLN, CFMN, CWLN, and CWMN. This implies that the output FASTA sequences will contain any of the aforementioned 16 sequences and CYIQ in the exact format (Figure 7B). For the input sequence CYIQ, the output of identical and homologous searching generates 83 identical PDBs and 914 FASTA sequences, which contain these homologous sequences.

## 4. Discussion

The PU search tool (https://chain-searching.herokuapp.com/) offers wide applicability for biological and pharmaceutical research domains. The data was gathered from various sources and the whole process was optimized using threading. Besides this, a separate application-programming interface (API) was also incorporated for the terminal users. For the development of the backend of the code, Flask has been used. Even when the cost of design and development of novel peptides would reduce, executing the code of this program would still significantly help in managing the budget for novel peptide development and it will be profitable to search in this web crawler database. As the size of different peptides gets reduced, the cost will also be reduced making it economically feasible for users from different fields.

### 4.1. Comparison with other databases PU database interests

The developed application is easy to execute, is freely available to all the users and is robust due to a large number of entries. As also discussed in the introduction section, this application includes information on peptides obtained experimentally, and can be used to study evolutionary relatedness. Like various applications, in the PU webserver, the searched results from the databases can be downloaded in API format and can be utilized by the user.

### 4.2. Limitations

The application was developed using only the freely available resources with low computational power, so it has a limit of 6 amino acids search at a time. The application has the limit of fetching 100 sequences in case of homologous searching due to the large size of the database and also because of the limitation of 30 seconds timeout offered by the deployment. For a single search, the server scans through the entire database, making it an exhaustive search, which is not only time-consuming but also puts constraints on the server. Hence, the first 100 results which appear are not similarity-based. In the event of homologous scanning, because the application runs into an exceptionally enormous database and most deployment sites have a timeout of 30 seconds, henceforth, the data constraint was necessary. Heroku's system resources are not sufficient enough to perform high-level calculations or search efficiently via a large database. For homologous searching, the proposed application is not very effective at swiftly searching the database and is limited to a set of parameters or characters. In Heroku, the web process has a 30-second window to respond to HTTP queries with response data (either the completed response or some amount of response data to indicate that the process is active). If a process does not send response data within the first 30-second timeframe, an H12 error will appear in the logs. This is dependent on multiprocessing and calculation, which further rely on the system configuration; which then determines how many threads are allowed and how quickly homologous chain calculations can be performed. Due to limited resources, the application cannot continue data crawling regularly to keep the database updated and in sync with other databases.

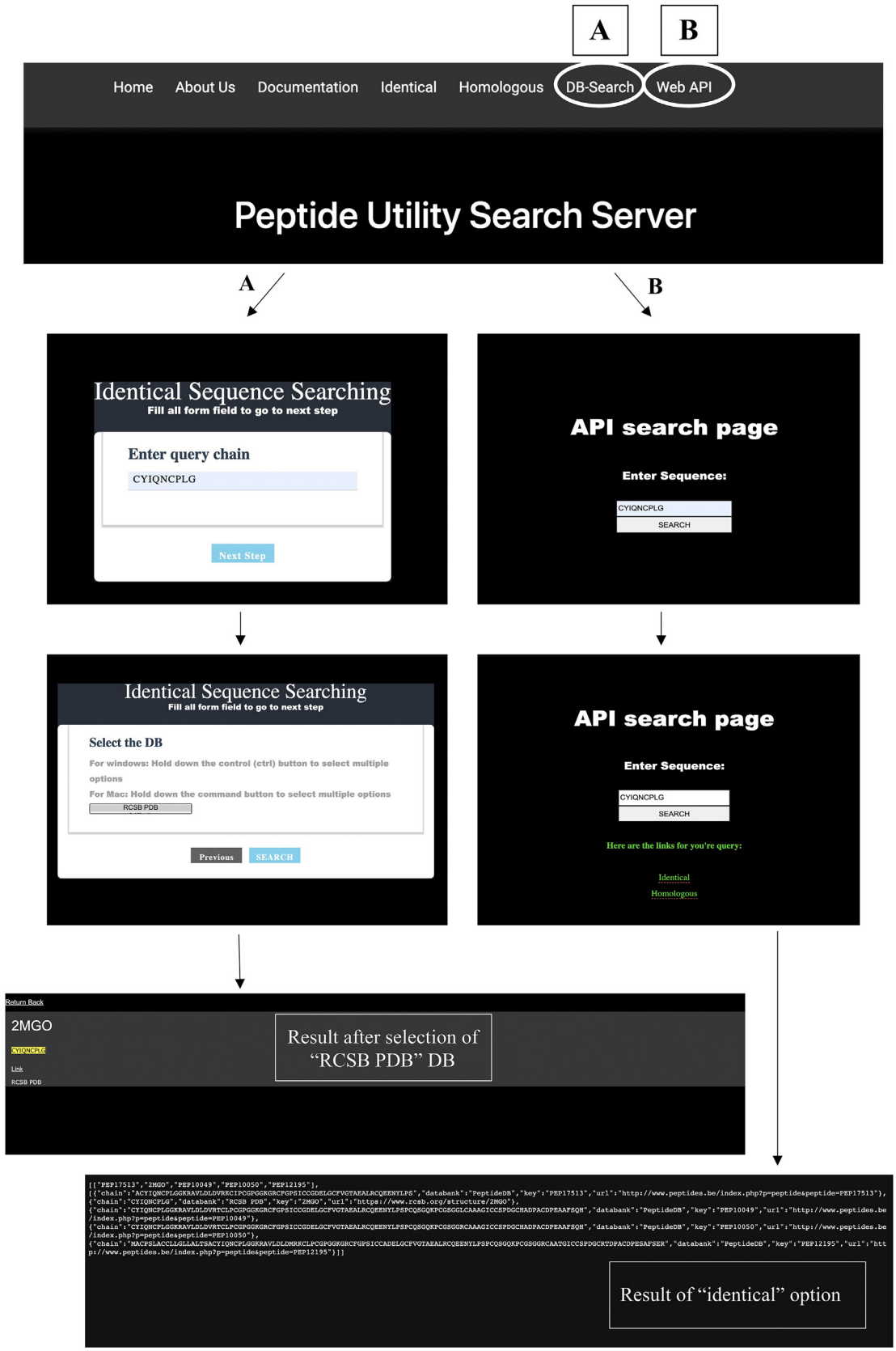



**Figure 6. DB-Search and Web API** – (A) JSON output of identical search under DBsearch for sequence "CYIQNCPLG" from the RCSB PDB database that was specifically chosen from the drop-down menu, and (B) Web API output of identical search under Web API for sequence "CYIQNCPLG".

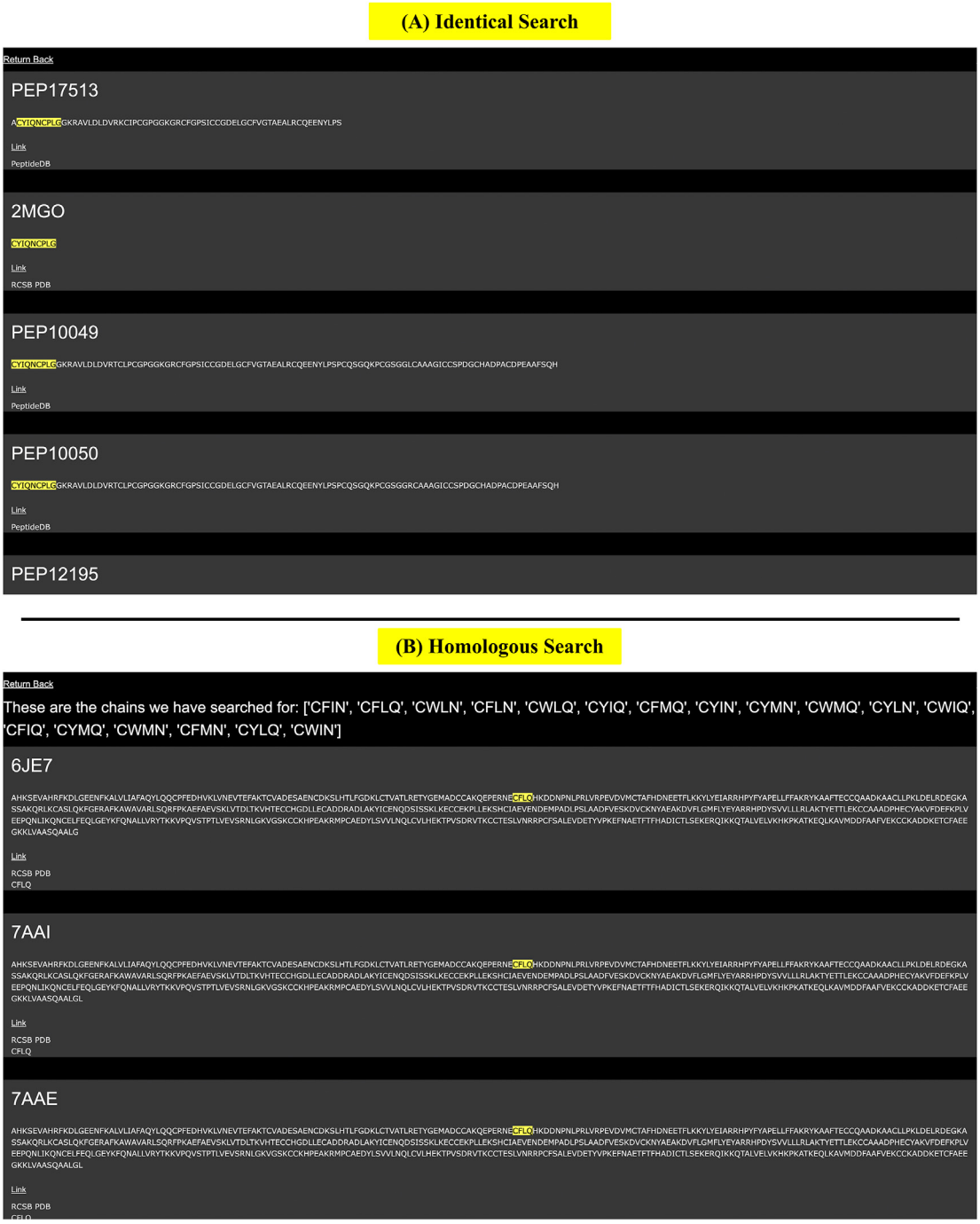



**Figure 7. Screenshots of different searches (A) Identical search** – Identical search for sequence "CYIQNCPLG" and **(B) Homologous search** – Homologous search for "CYIQ" with homologous sequences as output is shown in the highlighted part (only first few are shown).

### *4.3. Future work*

A Graphical User Interface (GUI) application capable of performing the same tasks for local computers, especially the high-end systems can be created which would be devoid of rate limitation on data and the data will be stored in JSON formatted files locally. With the use of high-end computers, this application can also be further extended to include more amino acids in a single search. With the inclusion of more resources, the 30-second window time frame can also be removed. Here, both offline and online tools were created for the users with an option to extend (on faster networks/computers), which uses the data from cloud and would require an internet connection. Henceforth, a generically accessible tool has been developed for global researchers, who wish to search for identical and homologous chain pools.

## 5. Conclusions

Peptide sequence information is crucial not just for predicting structure and function, but also for several other therapeutic applications. Using available platforms like Selenium framework, MongoDB, and Heroku, two sections were created by means of two search bars i.e., one for identical (https://chain-searching.herokuapp.com/identical) and the other for homologous searching (https://chain-searching.herokuapp.com/homologous). The dynamic website thus created made the tedious

task of searching simpler by providing most of the necessary information in one place. With the inclusion of different databases, the user is also provided with an option to search a sequence from a specific database under the DBsearch tab. Following this, the search queries are implemented for both identical and homologous sequences. After query building, two separate functions are executed. The identical search function is based on input provided by the user, while the homologous search function is based on a list of similar chains created by referring to the inset table in Figure 3, which was then queried to the databases. The final application has been developed using Heroku. Overall, this PU webserver/tool (https://chain-searching.herokuapp.com/) provides reliability and stability even for long sequence searches and can reduce the pre-processing work for peptide development.

## Declarations

### *Author contribution statement*

Tanishq Chamoli; Alisha Khera; Akanksha Sharma; Anshul Gupta; Sonam Garg; Kanishk Mamgain; Aayushi Bansal; Shriya Verma; Ankit Gupta, PhD; Hema Kumari Alajangi, PhD; Gurpal Singh, PhD; Ravi Pratap Barnwal, PhD: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

### *Funding statement*

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### *Data availability statement*

Data associated with this study has been deposited at Webservers developed

1.Identical multi db

https://chain-searching.herokuapp.com/dbSearch

2.API data

https://chain-searching.herokuapp.com/identical-api/ravi and https://chain-searching.herokuapp.com/homologous-api/ravi

3.Homepage

https://chain-searching.herokuapp.com/
https://github.com/TanishqChamoli/Peptide_Utility_Search_Application

### *Declaration of interest's statement*

The authors declare no competing interests.

### *Additional information*

No additional information is available for this paper.

## Acknowledgements

Our lab is supported by SERB, DBT and ICMR, Govt of India grants which are duly acknowledged for the work. The authors also thank RPB and GS lab members for their input throughout the project development. We would like to dedicate this paper to Prof. KVR Chary, Professor and founding Director, IISER Berhampur, whose sudden demise is a loss for the scientific community across the world.

## References

[1] D.J. Dietzen, Amino Acids, Peptides, and Proteins, 2018. Elsevier Inc.
[2] V. Kašička, Amino acids, peptides and proteins, Electrophoresis 34 (2013) 2603.
[3] C. Bonduelle, Secondary structures of synthetic polypeptide polymers, Polym. Chem. 9 (2018) 1517–1529.
[4] A. Thakur, A. Sharma, H.K. Alajangi, P.K. Jaiswal, Y. Lim, G. Singh, R.P. Barnwal, In pursuit of next-generation therapeutics: antimicrobial peptides against superbugs, their sources, mechanism of action, nanotechnology-based delivery, and clinical applications, Int. J. Biol. Macromol. 218 (2022) 135–156.
[5] F. Plisson, O. Ramírez-Sánchez, C. Martínez-Hernández, Machine learning-guided discovery and design of non-hemolytic peptides, Sci. Rep. 10 (2020) 1–19.
[6] I.W. Hamley, Small bioactive peptides for biomaterials design and therapeutics, Chem. Rev. 117 (2017) 14015–14041.
[7] M.H. Baig, K. Ahmad, M. Saeed, A.M. Alharbi, G.E. Barreto, G.M. Ashraf, I. Choi, Peptide based therapeutics and their use for the treatment of neurodegenerative and other diseases, Biomed. Pharmacother. 103 (2018) 574–581.
[8] B. Kuhlman, P. Bradley, Advances in protein structure prediction and design, Nat. Rev. Mol. Cell Biol. 20 (2019) 681–697.
[9] H. Geng, F. Chen, J. Ye, F. Jiang, Applications of molecular dynamics simulation in structure prediction of peptides and proteins, Comput. Struct. Biotechnol. J. 17 (2019) 1162–1170.
[10] S.A. Hollingsworth, R.O. Dror, Molecular dynamics simulation for all, Neuron 99 (2018) 1129–1143.
[11] C. Pál, B. Papp, M.J. Lercher, An integrated view of protein evolution, Nat. Rev. Genet. 7 (2006) 337–348.
[12] D. Kim, S. Han, H. Park, S. Choi, M. Kaur, E. Hwang, S. Han, J. Ryu, H. Cheong, R.P. Barnwal, Y. Lim, Pseudo-isolated α-helix platform for the recognition of deep and 2 narrow targets, J. Am. Chem. Soc. (2022).
[13] K. Yan, H. Lv, Y. Guo, Y. Chen, H. Wu, B. Liu, TPpred-ATMV: therapeutic peptide prediction by adaptive multi-view tensor learning model, Bioinformatics 38 (2022) 2712–2718.
[14] P. Zhou, C. Wang, Y. Ren, C. Yang, F. Tian, Computational peptidology: a new and promising approach to therapeutic peptide design, Curr. Med. Chem. 20 (2013) 1985–1996.
[15] J. Venkatraman, S.C. Shankaramma, P. Balaram, Design of folded peptides, Chem. Rev. 101 (2001) 3131–3152.
[16] E.W. Deutsch, The PeptideAtlas project, Methods Mol. Biol. 604 (2010) 285–296.
[17] A. Prakash, S. Ahmad, S. Majumder, C. Jenkins, B. Orsburn, Bolt: a new age peptide search engine for comprehensive MS/MS sequencing through vast protein databases in minutes, J. Am. Soc. Mass Spectrom. 30 (2019) 2408–2418.
[18] N.J. Edwards, PepArML: A Meta-Search Peptide Identification Platform, 2013.
[19] F. Liu, G. Baggerman, L. Schoofs, G. Wets, The construction of a bioactive peptide database in metazoa, J. Proteome Res. 7 (2008) 4119–4131.
[20] G.D. Stormo, An Introduction to Sequence Similarity ("homology") Searching, 2009.
[21] T. Shtatland, D. Guettler, M. Kossodo, M. Pivovarov, R. Weissleder, PepBank - a database of peptides based on sequence text mining and public peptide data sources, BMC Bioinf. 8 (2007) 1–10.
[22] P.M. Martins, L.H. Santos, D. Mariano, F.C. Queiroz, L.L. Bastos, I. de S. Gomes, P.H.C. Fischer, R.E.O. Rocha, S.A. Silveira, L.H.F. de Lima, M.T.Q. de Magalhães, M.G.A. Oliveira, R.C. de Melo-Minardi, Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm, BMC Bioinf. 22 (2021) 1–20.
[23] J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G.A. Lajoie, B. Ma, PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification, Mol. Cell. Proteom. 11 (2012) 1–8.
[24] A. Alexandridou, N. Dovrolis, G.T. Tsangaris, K. Nikita, G. Spyrou, PepServe: a web server for peptide analysis, clustering and visualization, Nucleic Acids Res. 39 (2011) 381–384.
[25] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S.L. Yeh, The Universal Protein Resource (UniProt), Nucleic Acids Res. 33 (2005) 154–159.
[26] C. Alfarano, C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M.J. Dumontier, M.R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J.P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J.J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B.F.F. Ouellette, C.W.V. Hogue, The biomolecular interaction network database and related tools 2005 update, Nucleic Acids Res. 33 (2005) 418–424.
[27] V.P. Valuev, D.A. Afonnikov, M.P. Ponomarenko, L. Milanesi, N.A. Kolchanov, ASPD (Artificially Selected Proteins/Peptides Database): a database of proteins and peptides evolved in vitro, Nucleic Acids Res 30 (1) (2002) 200–202.
[28] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, W. Helmberg, Y. Kapustin, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 34 (2006) 3–6.

[29] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, H. Hermjakob, IntAct - open source resource for molecular interaction data, Nucleic Acids Res. 35 (2007) 561–565.
[30] W.G. Touw, C. Baakman, J. Black, T.A.H. Te Beek, E. Krieger, R.P. Joosten, G. Vriend, A series of PDB-related databanks for everyday needs, Nucleic Acids Res. 43 (2015) D364–D368.
[31] S. Khan, DeepAcid: Classification of Macromolecule Type Based on Sequences of Amino Aacids, 2019 arXiv:1907.03532.
[32] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021).
[33] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W.R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning, Nature 577 (2020) 706–710.
[34] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. Santos, M. Ai, F. Team, Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction, 2022, pp. 1–31.
[35] M. Figliuzzi, P. Barrat-Charlaix, M. Weigt, How pairwise coevolutionary models capture the collective residue variability in proteins? Mol. Biol. Evol. 35 (2018) 1018–1027.
[36] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, S.C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G.A. Salazar, J. Tate, A. Bateman, The Pfam protein families database: towards a more sustainable future, Nucleic Acids Res. 44 (2016) D279–D285.
[37] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling, Bioinformatics 22 (2006) 195–201.
[38] A. Qureshi, N. Thakur, H. Tandon, M. Kumar, AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses, Nucleic Acids Res. 42 (2014) 1147–1153.
[39] M. Di Luca, G. Maccari, G. Maisetta, G. Batoni, BaAMPs: the database of biofilm-active antimicrobial peptides, Biofouling 31 (2015) 193–199.
[40] J. Rey, P. Deschavanne, P. Tuffery, Bact Pep DB: a database of predicted peptides from a exhaustive survey of complete prokaryote genomes, Database 2014 (2014) 1–9.
[41] A. Tyagi, A. Tuknait, P. Anand, S. Gupta, M. Sharma, D. Mathur, A. Joshi, S. Singh, A. Gautam, G.P.S. Raghava, CancerPPD: a database of anticancer peptides and proteins, Nucleic Acids Res. 43 (2015) D837–D843.
[42] P. Agrawal, S. Bhalla, S.S. Usmani, S. Singh, K. Chaudhary, G.P.S. Raghava, A. Gautam, CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides, Nucleic Acids Res. 44 (2016) D1098–D1103.
[43] M. Novković, J. Simunić, V. Bojović, A. Tossi, D. Juretić, DADP: the database of anuran defense peptides, Bioinformatics 28 (2012) 1406–1407.
[44] M. Pirtskhalava, A.A. Amstrong, M. Grigolava, M. Chubinidze, E. Alimbarashvili, B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D.E. Hurt, M. Tartakovsky, DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics, Nucleic Acids Res. 49 (2021) D288–D297.
[45] D. Mathur, S. Prakash, P. Anand, H. Kaur, P. Agrawal, A. Mehta, R. Kumar, S. Singh, G.P.S. Raghava, PEPlife: a repository of the half-life of peptides, Sci. Rep. 6 (2016) 1–7.
[46] A. Juhász, R. Haraszi, C. Maulis, ProPepper: a curated database for identification and analysis of peptide and immune-responsive epitope composition of cereal grain protein families, Database 2015 (2015) 1–16.
[47] D. Das, M. Jaiswal, F.N. Khan, S. Ahamad, S. Kumar, PlantPepDB: a manually curated plant peptide database, Sci. Rep. 10 (2020) 1–8.
[48] C. Zardecki, S. Dutta, D.S. Goodsell, M. Voigt, S.K. Burley, RCSB Protein Data Bank: a resource for chemical, biochemical, and structural explorations of large and small biomolecules, J. Chem. Educ. 93 (2016) 569–575.
[49] S. Singh, K. Chaudhary, S.K. Dhanda, S. Bhalla, S.S. Usmani, A. Gautam, A. Tuknait, P. Agrawal, D. Mathur, G.P.S. Raghava, SATPdb: a database of structurally annotated therapeutic peptides, Nucleic Acids Res. 44 (2015) D1119–D1126.
[50] D. Mathur, A. Mehta, P. Firmal, G. Bedi, C. Sood, A. Gautam, G.P.S. Raghava, TopicalPdb: a database of topically delivered peptides, PLoS One 13 (2018) 1–9.
[51] P. Kapoor, H. Singh, A. Gautam, K. Chaudhary, R. Kumar, G.P.S. Raghava, Tumorhope: a database of tumor homing peptides, PLoS One 7 (2012).
[52] S.P. Piotto, L. Sessa, S. Concilio, P. Iannelli, YADAMP: yet another database of antimicrobial peptides, Int. J. Antimicrob. Agents 39 (2012) 346–351.