Chapter 4

Computer-Aided Virtual Screening and Designing of Cell-Penetrating Peptides

Ankur Gautam, Kumardeep Chaudhary, Rahul Kumar, and Gajendra Pal Singh Raghava

Abstract

Cell-penetrating peptides (CPPs) have proven their potential as versatile drug delivery vehicles. Last decade has witnessed an unprecedented growth in CPP-based research, demonstrating the potential of CPPs as therapeutic candidates. In the past, many in silico algorithms have been developed for the prediction and screening of CPPs, which expedites the CPP-based research. In silico screening/prediction of CPPs followed by experimental validation seems to be a reliable, less time-consuming, and cost-effective approach. This chapter describes the prediction, screening, and designing of novel efficient CPPs using "CellPPD," an in silico tool.

Key words Cell-penetrating peptides, Drug delivery system, Machine learning approach, Virtual screening, Support vector machine, Prediction

1 Introduction

Therapeutic peptides have emerged as a very promising area of modern research. Previously, peptides were not considered as ideal therapeutic molecules, though the therapeutic peptide market emerged almost 40 years back in 1974, when Novartis launched the first therapeutic peptide-based drug, Lypressin, a vasopressin analog. In the last decade, a number of peptide databases [1–7] have been developed suggesting a restoration of interest in therapeutic peptides as potential drug candidates [8–10]. In this context, cell-penetrating peptides (CPPs) have attracted a significant scientific attention [11]. An increasing number of patents and research articles pertaining to CPPs have demonstrated how popular CPPs are as therapeutic peptides.

CPPs are small peptides (<50 amino acids) and have an inherent ability to penetrate a variety of cells [12]. In addition, CPPs are capable of transporting different types of cargoes like small molecules, nanoparticles, peptides, proteins, and nucleic acids and

Ülo Langel (ed.), *Cell-Penetrating Peptides: Methods and Protocols*, Methods in Molecular Biology, vol. 1324,
DOI 10.1007/978-1-4939-2806-4_4,

thus are being used widely in various drug delivery applications [13]. Since their inception in 1988, several hundreds of novel CPPs and their derivatives have been identified so far [14]. However, rational designing and identification of novel highly efficient CPPs is still a very challenging task.

In general, in the wet lab, identification of novel CPPs by high-throughput screening of modified analogs of existing CPPs is a time-consuming and labor-intensive approach. On the other hand, in silico screening/prediction of modified CPPs followed by experimental validation (integrative approach) is a more reliable approach, which has higher success rate, as compared to the traditional wet lab approach. In the past, substantial efforts have been made to develop in silico methods for the prediction of CPPs [15–19]. In 2005, Hallbrink et al. developed the first in silico prediction method. Since then, seven prediction methods have been developed (Fig. 1), but most of these methods are not freely available, and no web interface has been provided.

Two methods, CellPPD [20] and CPPpred [21], have been published recently and both are freely available to public. CPPpred is a neural network-based in silico method, which can predict cell-penetrating potential of peptides having length between 5 and 30 amino acids. But it cannot help in designing novel CPP analogs, while CellPPD can predict, as well as is helpful in designing novel efficient CPP analogs of parent CPP. This chapter describes the various modules of CellPPD for the prediction and designing of CPPs.

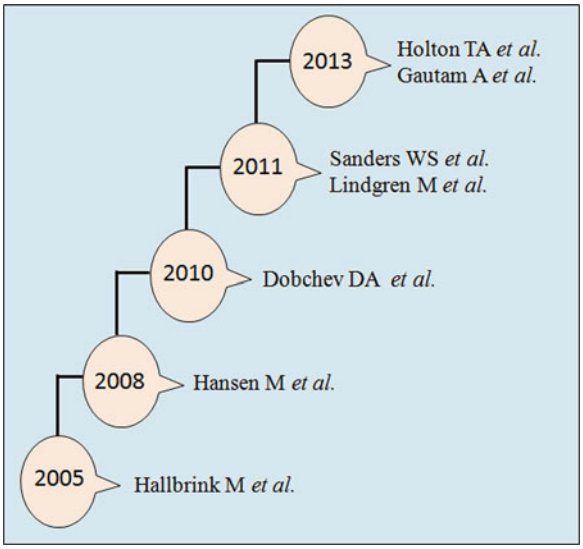

Fig. 1 Progress in the development of in silico tools for CPP prediction

2 Materials

2.1 CPPsite

CPPsite is a unique and the only repository of experimentally validated CPPs [14]. This database was developed with an aim to provide comprehensive information on experimentally tested CPPs. Apart from CPP sequences, it provides information about uptake mechanism, intracellular localization of CPPs, tested cell lines, uptake efficiency, etc. One of the essential features of this database is that it also provides predicted structures of all CPPs. Various tools for data retrieval and analysis were integrated, which makes it a very user-friendly resource. One of the major aims of this database is to provide latest datasets of CPPs for the analysis and development of CPP prediction methods. Two methods, CellPPD and CPPpred, which have been developed recently, have utilized the datasets generated from CPPsite.

3 Methods

3.1 CellPPD: Tool for the Prediction and Designing of Cell-Penetrating Peptides

CellPPD is a support vector machine-based in silico tool, which can predict whether a peptide will be CPP or non-CPP [20]. CellPPD is freely available at http://crdd.osdd.net/raghava/cellppd/. Prediction using CellPPD was based on peptide sequence information like amino acid composition, dipeptide composition, binary patterns, and motif information. Apart from prediction, CellPPD also provides the facility to design novel efficient CPPs. There are four modules: Design Peptide, Multiple Peptides, Protein Scanning, and Motif Scanning. The description of these modules is as follows.

3.2 Prediction and Designing of CPPs

One of the modules in CellPPD is "Design Peptide," which allows prediction of a query peptide as CPP or non-CPP according to the threshold cutoff chosen by the user (Fig. 2). In this module, user can submit single peptide at a time as a query in single-letter code. For the in silico prediction of the input sequence, there are two different prediction models provided (*see* **Note 1**), namely "SVM" and "SVM + motif" (hybrid model) based. Since the hybrid model is based on both motif information and SVM-based model, user can also select hybrid model for more reliable prediction. The selection of model leads to further more filters like setting of SVM threshold and *E*-value (*see* **Notes 2** and **3**); for example, selecting higher SVM threshold leads to greater stringency of prediction of cell-penetrating capacity.

One of the important features of this module is that it allows users to design novel analogs of existing CPP. In wet lab, after getting a suitable CPP lead, one uses to generate all possible analogs of the lead in order to have better and efficient CPPs. CellPPD

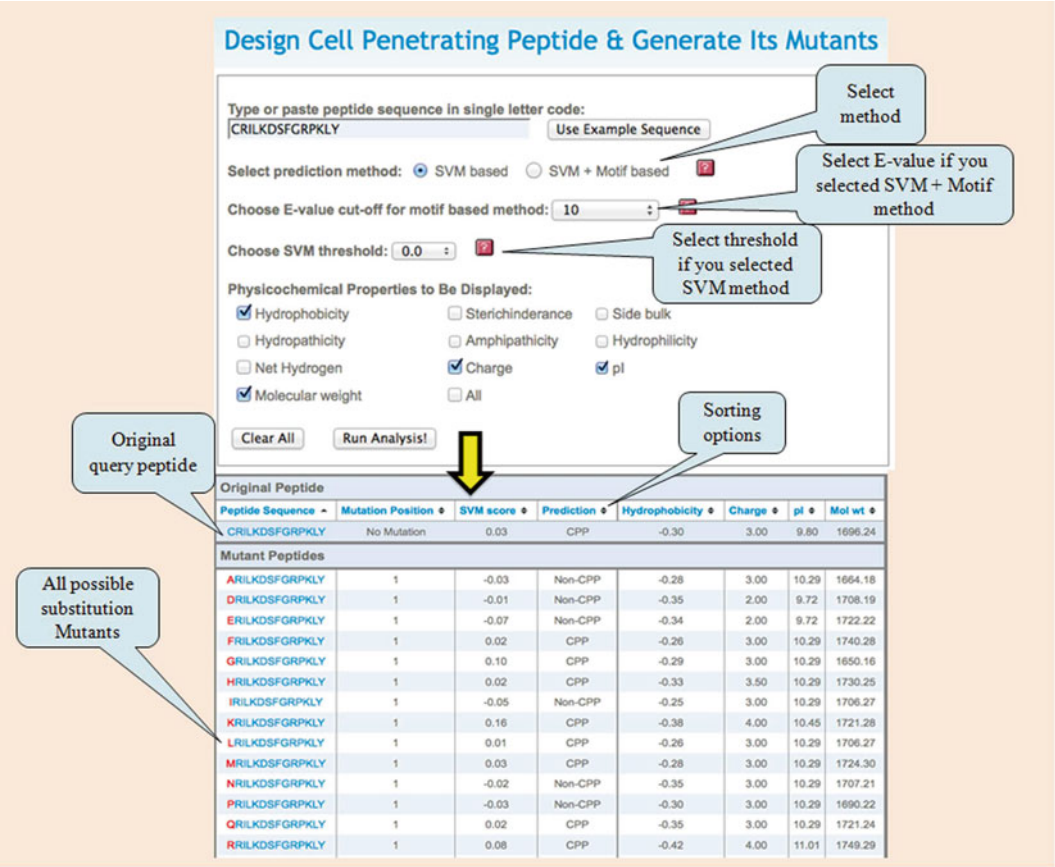

Fig. 2 Schematic representation of "Design Peptide" module and its output

works in the same manner and after submission of a query peptide, first it will predict whether the query peptide is CPP or not. In addition, server also generates all possible substitution mutants (mutated residues are depicted in red color) of the query peptide along with their SVM score, prediction status, and physicochemical properties, like hydrophobicity, charge, and molecular weight. User can sort any of these properties and select the best analog as CPP candidates. All these mutants are clickable and user can select any of the analogs and further generate all possible mutants of selected analog by clicking on that sequence (Fig. 3). In this manner, user can design novel CPP analogs with desired physicochemical properties.

3.3 Virtual Screening of CPPs

Another relevant module of CellPPD is "Multiple Peptides," which enables the users to predict large numbers of peptides at a time (Fig. 4). User can virtually screen libraries of a large number of peptides in order to identify novel efficient CPP candidates. It is just an extended version of "Design Peptide" module, which was

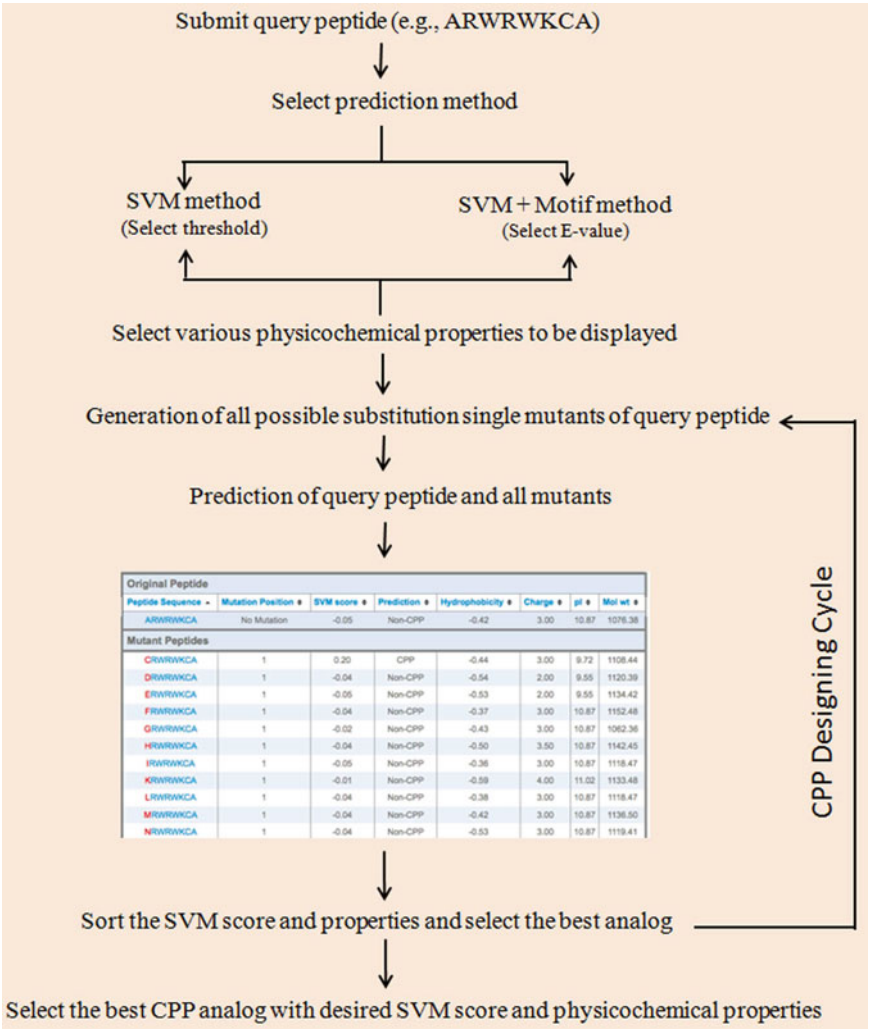

Fig. 3 Flow chart showing functioning of "Design Peptide" module for designing of CPP analogs (Color figure online)

built to save time and effort, if users have large numbers of peptides to predict. Interface of this module asks the users to select the method of prediction, i.e., "SVM" and "SVM + Motif" based. Then users have to define the *E*-value (*see* **Note 3**) if they are using "SVM + Motif"-based method as shown in Fig. 4.

Next is to select the SVM threshold; higher threshold leads to higher stringency (*see* **Note 2**). Users can also select the physicochemical properties, which they wish to display in outputs. Result table shows the prediction result and physicochemical properties, which users have selected (Fig. 4). In the result table, users can further click the individual peptide, which was submitted with multiple peptides earlier to generate its mutants. This facility of "Multiple Peptides" is homologous to "Design Peptide" module.

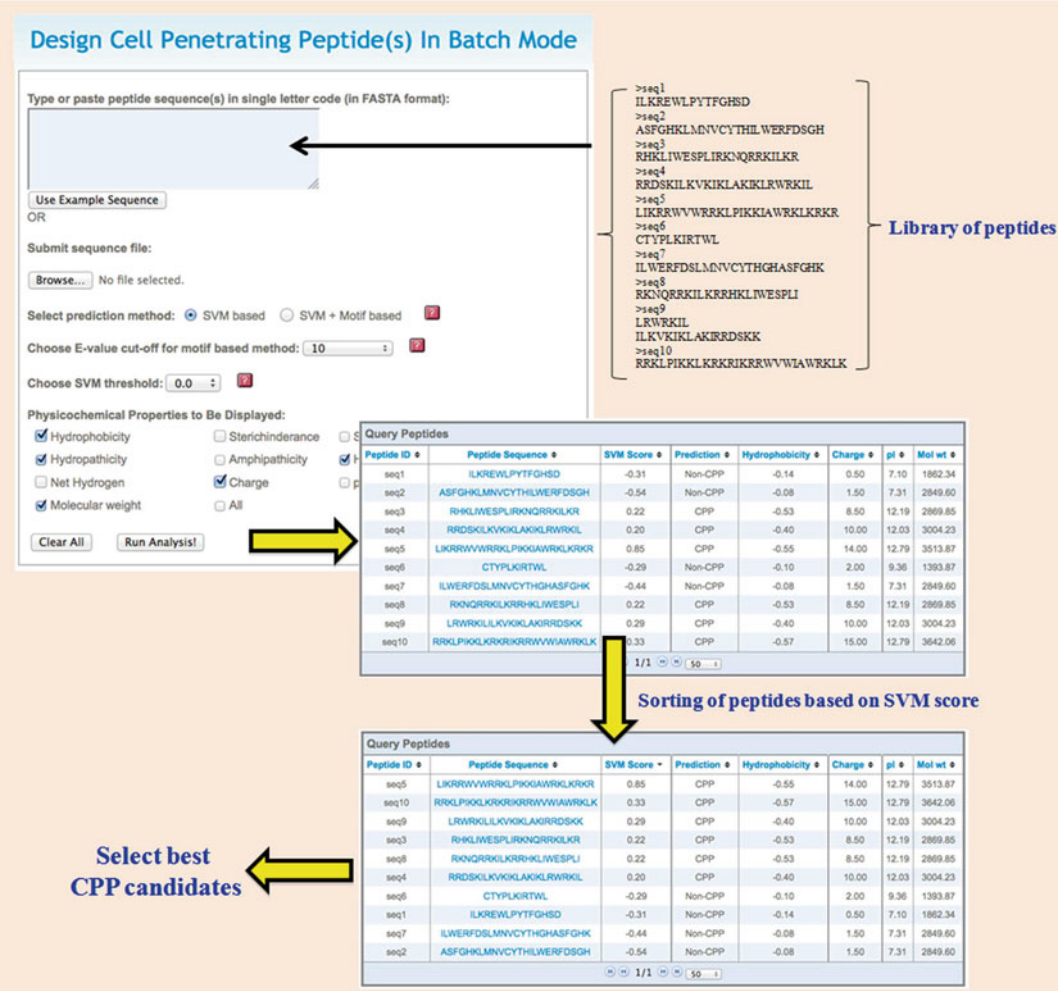

Fig. 4 Schematic representation of working flows of "Multiple Peptides" module of CellPPD

3.4 Protein Scanning

Since many of the CPPs are derived from natural proteins, user may be interested to identify novel CPPs that are derived from natural proteins. To identify those sections of a protein of interest, it has to be scanned throughout its length and resulting fragments can be identified as CPPs/non-CPPs based upon SVM score. For a protein of length l, a total number of overlapping fragments (n) is given by the equation $n = l - w + 1$, where w is the size of each fragment.

To capitalize this concept, "Protein Scanning" module has been integrated into "CellPPD" web server (Fig. 5). Here, the user can give protein sequence in plain one-letter amino acid format and can select the window length of fragments in which the protein has to be fragmented in sliding window fashion. Output can be obtained in tabular or graphical format based upon the selected model ("SVM" based or "SVM + Motif" based). Tabular output contains all the fragments of a query protein along with

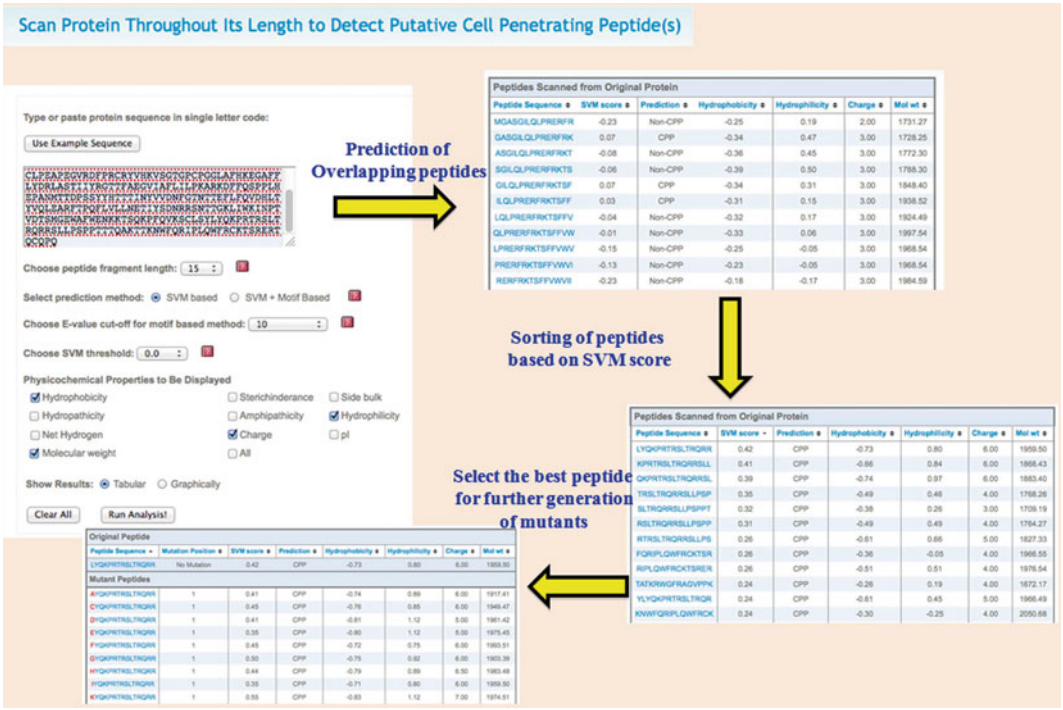

Fig. 5 Schematic representation of functioning of "Protein Scanning" module

their prediction score and prediction status, which is further accompanied by the user-selected physicochemical properties (hydrophobicity, hydrophilicity, molecular weight, etc.). These fragments are further clickable to generate single-point mutants/analogs along with their prediction scores. Parent peptide fragment is displayed in the top panel and respective analogs are in the lower panel of the sortable table. In this manner, user can reiterate and can select the best possible fragments and their analogs. Graphical output is more intuitive as it makes a clear picture of the trend of SVM scores and physicochemical properties of all generated fragments. Plots can be seen by clicking on the range of fragments given at the bottom, where plot's *y*-axis is the SVM score/property value and *x*-axis represents the fragments generated. Graphs are also supplemented with the tabular output.

3.4.1 *Example*

For a given Ebola protein, **Q66811**, if a user wants to see the CPP-rich regions of size 15 in above protein, then "Protein Scanning" module is useful. This query protein has a length of 365 amino acids and generates a total of 351 fragments (size = 15). If the user selects other default parameters and SVM-based model, then he or she gets the tabular output as shown in Fig. 5. Further output can be sorted based upon SVM score to get most potent CPP in the query protein and latter can be subject to generate single-point mutation analog generation (Fig. 5). Also to see the

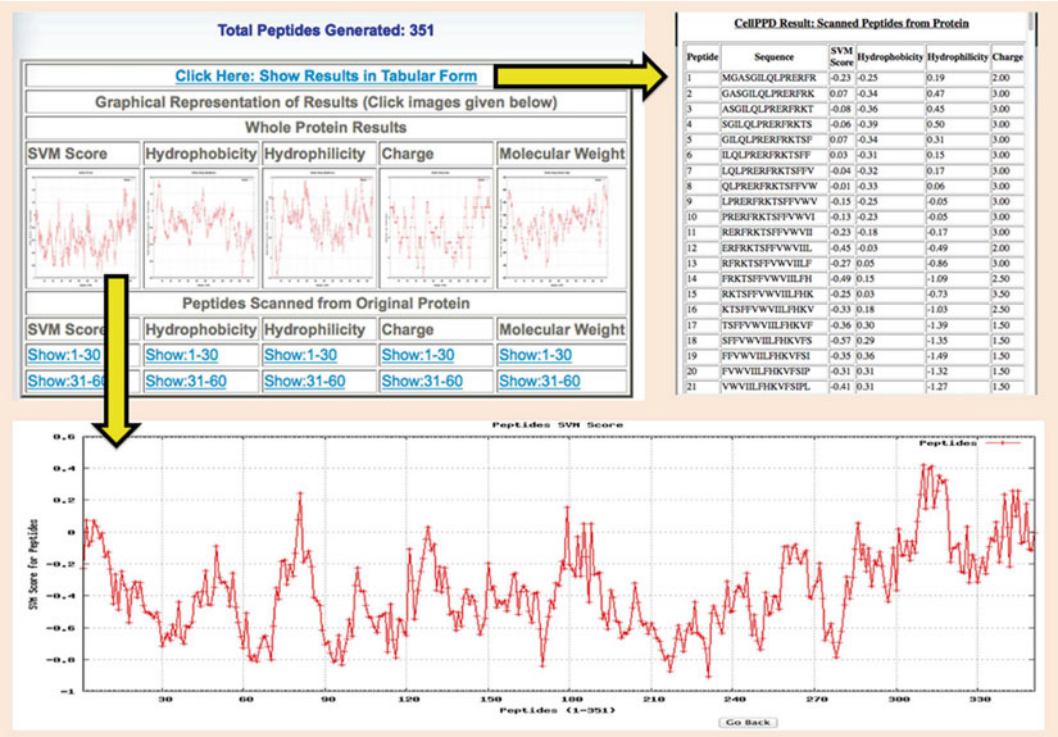

Fig. 6 Tabular and graphical output of "Protein Scanning"

plots of fragments having different SVM scores and physicochemical property values, user has to select graphical output (Fig. 6).

3.5 Motif Scanning

During preliminary analysis, it was observed that CPPs have significant sequence motifs, which are common to most of the CPPs. We have deduced a list of 120 significant motifs ranging from 4 to 30 in length using MEME suite. Whole list of CPP motifs is available at crdd.osdd.net/raghava/cellppd/. These 120 CPP motifs were further used for scanning of unknown proteins for any putative CPP activity, which leads to the concept of "Motif Scanning" module in CellPPD. In this module, user can submit any unknown protein to scan its sequence for the presence of known CPP motifs (Fig. 7). Along with the query sequence(s), user has to select few more options available at the CellPPD interface as shown in Fig. 7. These options are crucial for the scanning of significant motif in query protein sequence. User has to choose the *E*-value cutoff for motif scanning. This parameter is required by the MAST module of MEME suite, which CellPPD uses at the backend to scan motifs in a protein sequence. Users also have to define the length of the putative motifs they are looking for. As a case study, we used 43 protein sequences from five Ebola virus strains and scanned CPP motifs of different lengths in these protein sequences (Table 1).

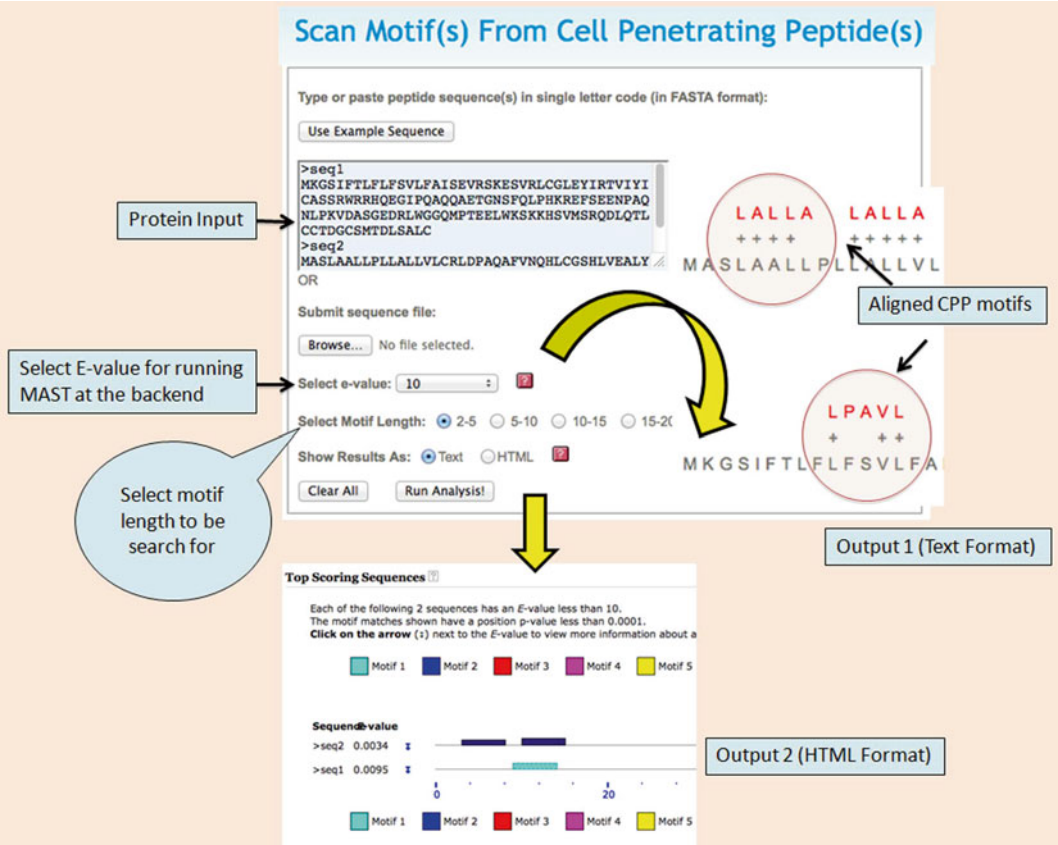

Fig. 7 Schematic representation of "Motif Scanning" module

Table 1
Ten most prominent CPP motifs (of different length) available in the Ebola protein sequences

S. no.	Motif sequence	*E*-value
1	YSPTT	1.70*E*–12
2	GNGYC	1.80*E*–07
3	MIIYR	3.50*E*–05
4	LALLA	2.00*E*–05
5	VALLPAVLLA	4.30*E*–07
6	KKTTTKPTKK	7.50*E*–06
7	ALWVTLWRDV	3.40*E*–05
8	GRQLRIAGKRLEGRS	1.10*E*–07
9	DSDCPGACICNGNGY	9.80*E*–05
10	GLWRALWRLLESLWWLLWEV	9.70*E*–06

4 Notes

1. All modules of CellPPD provide mainly two types of prediction models: (1) SVM based and (2) SVM + Motif based.
2. In the SVM-based model, the prediction status is assessed using the threshold of prediction score ranging nearly from −1 to +1. Default threshold is the one at which learning of a model is the best. If a user wants to shift from default threshold, then sensitivity and specificity also change accordingly. If a user wishes to get more accurate true positives rate (i.e., sensitivity) and false positive rate (i.e., specificity), then the user can move threshold towards +1 and −1, respectively.
3. In the case of motif-based models, users have to select *E*-values. Lower the *E*-value selected, the higher the stringency for that motif to be found in that sequence. But in that case, coverage of the samples decreases. So to compensate less coverage from motif-based approach, SVM and motif (hybrid) approach gives better performance than SVM-based approach, where the lower coverage is compensated by SVM performance and additional motif information attributes higher confidence.

5 Limitations and Future Prospects

Though CPPs are highly efficient and versatile delivery vehicles, their physicochemical properties often restrict their progression from bench to bedside. One of the major limitations is their low stability. Another limitation in CPP-based drug delivery is that most of the internalized CPP cargoes are entrapped in endosomes following endocytosis and the bioavailability is, therefore, severely reduced.

In order to overcome these limitations, novel CPPs with modified and nonnatural amino acids have been developed over the last decade. The in silico methods, which have been developed to date, can predict cell-penetrating potential of peptides consisting of only L-amino acids. Despite the enormous information on modified CPPs, no in silico method has been developed so far, which can predict peptides having modified or nonnatural amino acids. Also, there is a need to develop in silico tools, which can predict, as well as design CPPs with enhanced abilities to promote endosomal escape. Computational methods predicting intracellular localization of CPPs will also be useful for the scientific community, and thus provide momentum to peptide-based drug delivery.

References

1. Tyagi A, Tuknait A, Anand P et al (2014) CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res 43:D837–43
2. Kumar R, Chaudhary K, Sharma M et al (2015) AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides. Nucleic Acids Res 43:D956–D962
3. Gautam A, Chaudhary K, Singh S et al (2014) Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. Nucleic Acids Res 42:D444–D449
4. Mehta D, Anand P, Kumar V et al. (2014) ParaPep: a web resource for experimentally validated antiparasitic peptide sequences and their structures. Database (Oxford) 2014
5. Van Dorpe S, Bronselaer A, Nielandt J et al (2012) Brainpeps: the blood-brain barrier peptide database. Brain Struct Funct 217: 687–718
6. Wynendaele E, Bronselaer A, Nielandt J et al (2013) Quorumpeps database: chemical space, microbial origin and functionality of quorum sensing peptides. Nucleic Acids Res 41: D655–D659
7. Waghu FH, Gopi L, Barai RS et al (2014) CAMP: collection of sequences and structures of antimicrobial peptides. Nucleic Acids Res 42:D1154–D1158
8. Vlieghe P, Lisowski V, Martinez J et al (2010) Synthetic therapeutic peptides: science and market. Drug Discov Today 15:40–56
9. Craik DJ, Fairlie DP, Liras S et al (2013) The future of peptide-based drugs. Chem Biol Drug Des 81:136–147
10. Kaspar AA, Reichert JM (2013) Future directions for peptide therapeutics development. Drug Discov Today 18:807–817
11. Cerrato CP, Lehto T, Langel U (2014) Peptide-based vectors: recent developments. Biomol Concepts 5:479–488
12. Milletti F (2012) Cell-penetrating peptides: classes, origin, and current landscape. Drug Discov Today 17:850–860
13. Copolovici DM, Langel K, Eriste E et al (2014) Cell-penetrating peptides: design, synthesis, and applications. ACS Nano 8:1972–1994
14. Gautam A, Singh H, Tyagi A et al. (2012) CPPsite: a curated database of cell penetrating peptides. Database (Oxford) bas015
15. Hällbrink M, Kilk K, Elmquist A et al (2005) Prediction of cell-penetrating peptides. Int J Pept Res Ther 11:249–259
16. Hansen M, Kilk K, Langel U (2008) Predicting cell-penetrating peptides. Adv Drug Deliv Rev 60:572–579
17. Dobchev DA, Mager I, Tulp I et al (2010) Prediction of cell-penetrating peptides using artificial neural networks. Curr Comput Aided Drug Des 6:79–89
18. Lindgren M, Langel U (2010) Classes and prediction of cell-penetrating peptides. Methods Mol Biol 683:3–19
19. Sanders WS, Johnston CI, Bridges SM et al (2011) Prediction of cell penetrating peptides by support vector machines. PLoS Comput Biol 7:e1002101
20. Gautam A, Chaudhary K, Kumar R et al (2013) In silico approaches for designing highly effective cell penetrating peptides. J Transl Med 11:74
21. Holton TA, Pollastri G, Shields DC et al (2013) CPPpred: prediction of cell penetrating peptides. Bioinformatics 29:3094–3096