Machine Learning-Based Identification of Therapeutic Peptides

Tushita T
Dept. of Computer Science & Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Bengaluru, India
bl.sc.u4aie24065@bl.students.amrita.edu

Viyaneeta Ramesh
Dept. of Computer Science & Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Bengaluru, India
bl.sc.u4aie24054@bl.students.amrita.edu

R Guruprasad Reddy
Dept. of Computer Science & Engineering
Amrita School of Computing
Amrita Vishwa Vidyapeetham, Bengaluru, India
bl.sc.u4aie24063@bl.students.amrita.edu

***Abstract*—Due to its high specificity and reduced cytotoxicity, therapeutic peptides have become an important group of biomolecules in modern pharmacological studies. Such peptides can only be detected by using precise computational methods to accelerate biomedical research. A dichotomous classification paradigm was developed in the current study which differentiated between therapeutic and non-therapeutic peptides. The Therapeutic Peptide Database (THPdb) was used as a source of a carefully curated dataset. The dataset balance was obtained by the use of synthetic decoy peptides, which were synthesized by stochastic sequence generation and permutation approaches. The issue of internal redundancy was addressed through global sequence alignment using the NeedlemanWunsch algorithm where the match, mismatch and gap cost parameters were set. Sequences with a similarity above a set similarity threshold were cut to eliminate model bias. The suitability of the filtered dataset was then established by maximum pairwise similarity tests. The resulting non-redundant collection provides a strong underpinning on which next generation machine-learning stratification can be carried out.**

I. Introduction

The therapeutic peptides are a rapidly growing category of pharmacological agents due to their high specificity in targets, excellent safety profile, as well as structural adaptability. In comparison to traditional small-molecule drug development, peptides exhibit better selectivity and less off-target toxicity, which makes them very promising targets in current drug development. Nevertheless, the identification and verification of therapeutic peptides that are time-consuming and resource-intensive methods are experimentally identified and validated; therefore, effective computational methods are required.

Classification based on machine learning has emerged as an effective approach that can be used to expedite peptide screening and discovery. Such models require the construction of their datasets with caution. The factors that a strong binary classification model needs are a balanced positive and negative sample, biologically significant negative examples, removal of repeats in sequences, and preprocessing procedures that are scientifically proved. In the absence of these precautions, there is a risk of overfitting and inflated estimates of performance of the models.

Redundancy is a vital issue in the biological sequence datasets. The similarity of peptides can also create systematic biases, as highly similar sequences are memorized by a model. Thus, the cancellation of redundancy by filtering of similarities based on alignment is a basic preprocessing phase before machine learning analysis. Here, we assemble a stringently selected binary set of therapeutic and non-therapeutic peptides, which is followed by global sequence alignment in order to achieve structural diversity and maximum reduction, followed by subsequent modeling.

II. Dataset Construction and Curation

A. Therapeutic Peptides

The Structurally Annotated Therapeutic Peptide Database (SATPdb) was used to identify therapeutic peptide sequences. SATPdb contains experimentally verified peptides that have been reported to have biological activity. Four therapeutically relevant functional categories were chosen to ensure biological diversity and clinical relevance. Namely, Antiviral peptides, Antifungal peptides, Antiparasitic peptides and Antihypertensive peptides. All sequences were downloaded and programmatically processed with Python in FASTA format. The original data comprised over 7,000 therapeutic sequences of peptides. Preprocessing measures such as elimination of misspelled or empty sequences, change of sequences in uppercase, eradication of non-standard symbols of amino acids, filtering of length (5-100 amino acids) and elimination of precise replicas have been used to guaranty the integrity of the data. Following the preprocessing, 6,121 high quality therapeutic peptide sequences were further analyzed.

B. Generation of Non-Therapeutic (Decoy) Sequences

Negative samples were synthetically prepared to build a supervised learning framework of detecting therapeutic peptides. The composition-preserving shuffling strategy was used to boost generation of decoy sequences, where each therapeutic peptide sequence was shuffled randomly to maintain composition of original amino acids and original sequence length. The methodology is such that the decoy sequences have the same physicochemical characteristics but contain biologically relevant sequence assembly. Consequently, this makes classification not dependent on the frequency of amino acids but rather on the model strength is enhanced. Equal number of decoy sequences (6,121) were generated so as to ensure strict class balance.

III. Binary Dataset Pre-redundancyt

A. Global Sequence Alignment

In order to remove similar sequences and to ensure that there is no bias in model estimation due to redundancy, global sequence alignment with the NeedlemanWunsch algorithm applied with PairwiseAligner as part of Biopython. The choice of global alignment rather than local alignment (i.e., BLAST) was based on the fact that peptide sequences in the current study are quite short (5-100 amino acids), and to perform redundancy analysis, full-length sequences should be compared. The local alignment algorithms can only identify high-scoring subsequences, which can be used to detect only part of the redundancy. A sequence identity global cut of 80 percent was used. Any pair of sequencing that showed similarity 0.80 was deemed as redundant and only a single representative sequence remained.

$Score = (Matches \times MatchScore) + (Mismatches \times MismatchPenalty) + (Gaps \times GapPenalty)$
where:

$$\text{MatchScore} = +1$$
$$\text{MismatchPenalty} = -1$$
$$\text{GapPenalty} = -2$$

B. Post-Filtering Statistics

Following the process of eliminating redundancy, the edited dataset retained 1000+ sequences. The distribution of classes was checked to ensure that the classes such as therapeutic and non-therapeutic were properly represented after filtering. This measure was critical to eliminating redundancy so that it did not cause unwanted imbalance in the classes.

$$Similarity = \frac{MaximumPossibleScore}{AlignmentScore}$$

where:

$$\text{Maximum Possible Score} = \min(L_1, L_2) \times \text{MatchScore}$$

A similarity threshold of 0.80 was used to determine redundancy.

C. Non-Redundancy Non-Quantitative Validation

In order to prove the efficiency of the redundancy elimination, similarity analysis was performed on a randomly selected set of sequences of the filtered dataset. The quantitative measurements that were calculated included: Mean pairwise global similarity, Median pairwise similarity, Optimal maximum similarity and Similarity Standard deviation. Additionally, a similarity matrix of two, a plot of a similarity distribution histogram and the similarity matrix was plotted using a heatmap.
The lack of pairs of sequences above the 80 percent similarity rate corroborates the effectiveness of removing sequences that are extremely similar.The general low average of similarity also proves the existence of high structural diversity in the data set.

D. Data Integrity

One of the most important steps in the peptide classification research is redundancy elimination since the highly similar sequences can make the model overfit and thus gave the impression that it is generalizing well. The similarity between repetitive sequence patterns can lead to an inflated classification accuracy, memorization, and train-test leakage, as well as lower biological diversity of the dataset when the redundant sequences are not eliminated. T he resulting curated dataset is of high biological diversity, with minimal structural redundancy, and greater generalization ability, accompanied by robust downstream machine learning performance, by filtering with global alignment followed by quantitative measures of redundancy, and by testing the validity of this approach with quantitative measures of redundancy. All these measures will ensure model predictions are based on realistic functional discrimination, and are not a simple memorisation of similar peptide sequences.

IV. Feature Extraction

The crucial step to successful supervised classification of therapeutic and non-therapeutic peptides is feature engineering. Since the dataset built contains composition-preserving shuffled decoy sequences the chosen features should be able to reflect both order of sequences and contextual dependencies as well as functional patterns on top of amino acid frequency. In the present study, this section explains the features of deep learning and biologically interpretable features used.

A. Contextual Sequence Embeddings

1) BERT Model Selection: The contextual sequence embeddings were obtained with the help of the pretrained ProtBERT model ("Rostlab/prot_bert"). ProtBERT is a protein-language model based on the transformer, which is trained on large-scale protein sequence databases with the masked language modeling objectives. Compared to general-purpose natural language BERT models, ProtBERT is domain-specific and needs to perform representation of amino acids. The model structure will be built with several self-attention layers, and the size of the hidden dimension will be 1024. This is the

hidden size that defines the dimension of the output feature representation of every token of the sequence.

2) Interpretation of the Model Output: Given a peptide input of length L, the output dimension of ProtBERT is:

$$(L, 1024)$$

where, L is the length of sequence and 1024 is the hidden dimension of the transformer. Mean pooling was used in the residue dimension to give a fixed length representation that is not dependent on the length of the sequence. In this way, every peptide sequence was coded representationally into a 1024-dimensional embedding vector.

$$E = [e1, e2, ..., e1024]$$

In the case of the last curated dataset where there were 10,789 sequences, the output feature matrix was of the dimension (10789,1024). These 1024 features are learned latent embedding dimensions, as a result of large-scale self-supervised training. They lack explicit biological names but represent contextual associations among amino acids, motif integrity, evolutionary hints and high-order functional semantics.

3) Dimensionality Reduction: In order to avoid over-fitting and decrease the complexity of the computations, the Principal Component Analysis (PCA) of the embedding space with 1024 dimensions was selected. The leading principal elements were used to hold most of the variance and at the same time reduce features dimensionality. This move enhanced model generalization and computation in downstream classification.

B. Statistical Features Statistical features implemented using sequences

Besides profound contextual embeddings, biologically understandable sequence features were also taken into account to increase discriminatory power.

1) Amino Acid Composition (AAC): The composition of amino acids was calculated as normalized frequency of each of the 20 standard amino acids:

$$AAC(i) = \frac{Countof AminoAcid(i)}{SequenceLength}$$

AAC records patterns of residue distribution in the world. Therapeutic peptides may have certain compositional biases, such as positive charge enrichment of Lysine (K) and Arginine (R). Despite the decoy sequences maintaining amino acid structure, AAC offers a baseline biochemical structure of comparative modeling.

2) Peptide Length: The length of peptides was added as a scalar property. Therapeutic peptides are usually within a certain range of size that affects stability and activity of the presented structure. Length also adds to functional characterization as well as complements contextual embeddings.

C. Physicochemical Features

Physicochemical properties were also introduced to offer biologically significant descriptors in line with the peptide activity.

1) Net Charge: The net charge was determined on the basis of the difference between positively charged residues (K, R) and negatively charged residues (D, E). A good number of therapeutic peptides are cationic especially antimicrobial peptides. The electrostatic interactions are highly important in attachment to membranes and specificity of the targets.

2) Hydrophobicity: Standard amino acid hydrophobicity scales were used to calculate hydrophobicity. Amphipathic, i.e., hydrophobic and positively charged patches are common in membrane-active therapeutic peptides. Hydrophobicity hence helps in differentiating functional peptides and shuffled decoys.

3) Isoelectric Point (pI): Isoelectric point based on theory was calculated to describe the behavior of peptide solubility and charge under physiological conditions. Functional activity and biological interaction mechanisms are related to variations in pI.

D. Conceptual Framework of Feature Retrieval

Input was given in the form of non-redundant peptide sequences as a result of filtering based on global alignment. Then the tokenization of sequences was done by splitting amino acids into tokens. ProtBERT was used to produce embeddings on a contextual basis at the residue level. Mean pooling was done on the results to get fixed length 1024-dimensional sequence embeddings. Lastly, dimensionality reduction was done using PCA and deterministic analytical formulas were taken to calculate statistical and physicochemical descriptors. All features had been joined together to constitute the final feature matrix on which supervised binary classification was done.

E. Relatability of the Selected Features to the Research Problem

The strategy of dataset composition consisted of composition-preserving decoys that were shuffled. Thus, amino acid frequency is not enough to base discrimination on. Contextual features and order-dependent features appear necessary.
The chosen groups of features serve the following purposes:

- Transformer embeddings represent order within sequences, organization of motifs and dependencies between contextual residues.
- The world biochemical bias is captured in the amino acid composition.
- Length connects constraints of structural feasibility.
- Hydrophobicity and net charge entrap membrane interaction potential and electrostatic properties.

Collectively, the above features constitute a biologically and computationally powerful model of differentiating between therapeutic and non-therapeutic shuffled peptides.

V. Results

Dataset Filtering Results

Original number of sequences was 12,242 sequences, of which 6,121 were therapeutic peptides and 6,121 decoy peptides. Finally, 1000+ sequences non-redundant dataset size. Following global alignment based redundancy elimination with an 80percent similarity cutoff, the dataset was further narrowed down.

Reduction Percentage

$$Reduction = \frac{12242 - N}{12242} \times 100$$

This is equivalent to a redundancy cut of R meaning that one eliminates the most similar entries in the dataset in terms of peptide sequences without losing the diversity of the dataset.

Maximum Pairwise Similarity After Filtering

$$S < 0.80$$

The dataset satisfies the non-redundancy criterion. This observation is further supported by the similarity distribution in pairs as shown in Figure 1, which shows that there is no significant density after the similarity threshold given.

Moreover, the representation of the similarity matrix in Figure 2 demonstrates that there is a diffuse similarity pattern with no major clusters, thus proving the existence of a low level of structural redundancy.

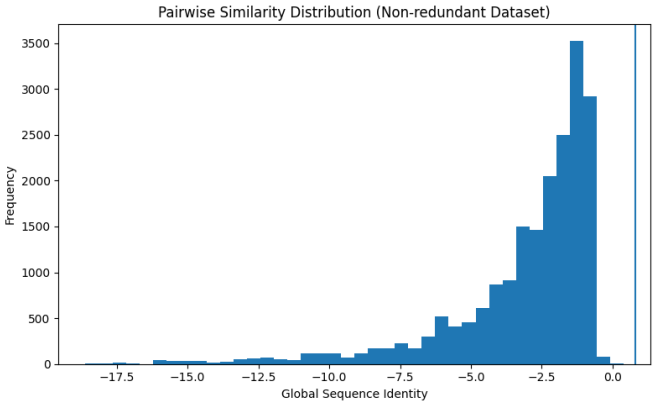

Fig. 1. Pairwise Similarity Distribution (Non-redundant Dataset)

Final Class Distribution Final class counts after redundancy removal:

- 6121 Decoys
- 6121 Therapeutic Peptides

$$\text{Ratio} = 1:1$$

The dataset remains balanced and suitable for binary classification.

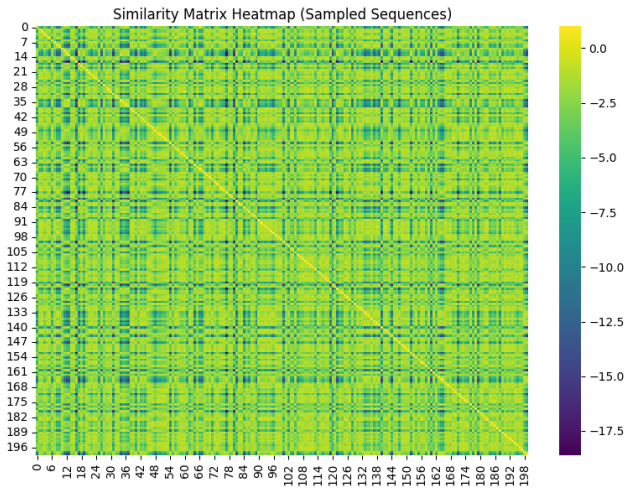

Fig. 2. Similarity Matrix Heatmap (Sampled Sequences)

A. Interpretation of Results

A relatively moderate percentage of the reduction indicates that redundancy filtering eliminated very similar sequences and retained a sizeable amount of the biologically important information. The lack of pairs of sequences that are greater than 80 percent similar proves the fact that the structural redundancy was successfully removed. As a result, the resulting dataset is non-redundant and heterogeneous and it can be used in powerful supervised machine learning tasks.

B. Contextual Embedding

Physicochemical properties were also introduced to offer biologically significant descriptors in line with the peptide activity.

1) Embedding Dimensionality: An embedded model based on contextual sequences on the curated peptides dataset was generated using the pretrained ProtBERT model. Position mean pooling of the residues was then carried out as well as representation of each peptide sequence by a fixed length (1024) embedding vector. The feature matrix obtained in the instance of the whole dataset was of the dimension (10789,1024) and informed that the projection of the 10,789 peptide sequences took place in 1024 dimensional embedding space.

2) Dimensionality Reduction: The dimensions of the feature were reduced with Principal Component Analysis (PCA), and so as to render the embedding matrix computationally tractable. The dimension was reduced and the majority of variance in data was not lost. The compressed feature representation applied the identical amount of samples but minimized the embedding dimension, hence being suitable to downstream supervised classification.

3) Implication on Classification: The fact that the non-therapeutic sequences were produced by composition maintaining shuffling means that amino acid frequencies are not enough in discrimination. The transformers as the form of

embedding incorporate the sequence order and the contextual association of the residues that are necessary to differentiate functional peptides and shuffled peptides. The ensuing embedding space consequently provides an adequate representation when it comes to features of binary therapeutic peptides classification.

VI. Discussion

The process of redundancy elimination enhanced the structural integrity and diversity of the dataset to a large extent since it eliminated highly similar peptide sequences that would otherwise be biased in machine learning models. The choice of global sequence alignment over the local one with the Needleman-Wunsch algorithm was not accidental as the evaluation of redundancy in short peptides needs the entire sequence to be compared and not just the motif. The 80 percent global similarity threshold was employed to make sure that only homologous sequences were removed and functional diversity was not destroyed. The effectiveness of the redundancy removal was also validated using the quantitative validation in terms of the pairwise similarity analysis, where none of the sequence pairs would surpass the predetermined threshold. The mean similarity is low, and there are no clusters of high similarity in the similarity matrix, which even more corroborates that structural repetitive entries are absent in the dataset.

The resulting dataset is:

- Balanced
- Non-redundant
- Scientifically validated
- Appropriate in the case of supervised learning

This preprocessing pipeline is important in facilitating a strong base further feature extraction and classification modeling.

In addition, the dataset composition, controlled by redundancy and balance, improves the quality of downstream learning algorithms, and makes sure that the facts obtained are an accurate representation of actual functional variation and not circular sequence artefacts. Structural diversity maintenance following filtering increases the number of representational space to model with, and the ability to generalize better to unseen peptide sequences. The dataset, in turn, is curated, and therefore, forms a statistically and biologically sound base of further feature extraction and supervised classification.

Novelty

The study is a systematically constructed a large-scale non-redundant binary dataset and validated the use of therapeutic peptide classification with experimentally validated sequences with composition-preserving decoy generation. In comparison to the traditional research which has utilized only available curated datasets or naive sequence clustering algorithms, the work combines therapeutic peptides of SATPdb verified experimentally. Preserving shuffled decoy generation composition-wise. Redundancy removal on a strict 80 percent similarity global alignment.This quantitative method of validation compares and contrasts similarities between pairs of data in the form of a data matrix and the output is visualized in a matrix format.

Detection of redundancy based on full-length global alignment, and statistical confirmation of non-redundancy, provides integrity of datasets not just by removing duplicates or using heuristic clustering methods. The redundancy filtering was also done in conjunction between therapeutic and decoy classes and the leakage of cross-class similarity is reduced, and robustness to supervised learning is enhanced..

Conclusion

The paper introduces a preprocessing pipeline framework to build a quality binary dataset of therapeutic peptides prediction. Therapeutic peptides that were experimentally validated were experimentally combined with composition conserving decoys to form a balanced dataset used in supervised learning. To remove redundancy, global sequence alignment was undertaken basing on a classical scoring scheme and threshold of 80 percent. The quantitative similarity confirmation was used to verify the lack of very homologous sequences, which contributed to the increased reliability of the dataset and minimized bias.

The resulting data set is not redundant, biologically varied, and statistically confirmed, which serves as a good base in future machine learning operations. The future work will be aiming at the developments of higher features of the work, optimizing the model, and eventually extending the framework to the multi-class classification of certain therapeutic peptide subtypes.

References

[1] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” J. Mol. Biol., vol. 48, no. 3, pp. 443–453, Mar. 1970, doi: 10.1016/0022-2836(70)90057-4.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” J. Mol. Biol., vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[3] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” Bioinformatics, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.

[4] K.-C. Chou, “Prediction of protein cellular attributes using pseudo-amino acid composition,” Proteins: Struct., Funct., Genet., vol. 43, no. 3, pp. 246–255, May 2001, doi: 10.1002/prot.1035.

[5] D. Veltri, U. Kamath, and A. Shehu, “Deep learning improves antimicrobial peptide recognition,” Bioinformatics, vol. 34, no. 16, pp. 2740–2747, Aug. 2018, doi: 10.1093/bioinformatics/bty179.