Probabilistic and Machine Learning Models for the Protein Scaffold Gap Filling Problem

Kushal Badal[1], Letu Qingge[1], Xiaowen Liu[2], and Binhai Zhu[3(✉)]

[1] Department of Computer Science, North Carolina A&T State University, Greensboro, NC, USA
kbadal@aggies.ncat.edu, lqingge@ncat.edu

[2] John W. Deming Department of Medicine, Tulane University, New Orleans, LA, USA
xwliu@tulane.edu

[3] Gianforte School of Computing, Montana State University, Bozeman, MT, USA
bhz@montana.edu

Abstract. In de novo protein sequencing, we often could only obtain an incomplete protein sequence, namely scaffold, from top-down and bottom-up tandem mass spectrometry. While most sections of the proteins can be inferred from its homologous sequences, some specific section of proteins is always missing and it is hard to predict the missing amino acids in the gaps of the scaffold. Thus, we only focus on predicting the gaps based on a probabilistic algorithm and machine learning models instead predicting the complete protein sequence using generative AI models in this paper. We study two versions of the protein scaffold filling problem with known size gaps and known mass gaps. For the known size gaps version, we develop several machine learning models based on random forest, k-nearest neighbors, decision tree and fully connected neural network. For the known mass gap problem, we design a probabilistic algorithm to predict the missing amino acids in the gaps. The experimental results on both real and simulation data show that our proposed algorithms show promising results of 100% and close to 100% accuracy.

Keywords: Protein sequencing · Protein Scaffold filling · Machine learning · Probablistic model · Heuristic algorithms

1 Introduction

In the fields of proteomics, protein sequencing determines the amino acid code of a protein. Protein sequencing is a widely researched area, as it is beneficial for highlighting the structures and functions of proteins. With such information, researchers across a realm of fields in biology, chemistry, and medicine can identify and develop more effective solutions to long-standing problems, such as

W. Peng et al. (Eds.): ISBRA 2024, LNBI 14956, pp. 28–39, 2024.
https://doi.org/10.1007/978-981-97-5087-0_3

pharmaceutical drug development and understanding the role that proteins play in various diseases and conditions.

The advent of mass spectrometry marked a pivotal shift in protein sequencing technologies, offering substantial improvements over traditional methods like Edman degradation, which, despite its utility, was limited by low throughput and substantial sample requirements [4,8]. Mass spectrometry's sensitivity to attomole quantities of peptides represents a significant advancement, facilitating rapid and high-coverage data acquisition [4]. Early mass spectrometry-based strategies for peptide sequencing laid the groundwork for the sophisticated techniques in use today [2,12].

De novo sequencing and database searching are the two main methods commonly used in mass spectrometry protein sequencing [1]. With de novo protein sequencing, there are two mass spectrometry based methods, known as top-down and bottom-up approaches. Despite recent progress, most assembled proteins are still in an incomplete form with gaps in the scaffold [11]. While most sections of the proteins can be inferred from its homologous sequences, some specific section of proteins is always missing and hard to predict the missing amino acids in the gaps of the scaffold [11]. (Due to space constraint, we refer more background and references on *de novo* protein sequencing to [1,5,11].)

We study two versions of the protein scaffold gap filling problems (PSGF). In the first variant, we assume that the size of gaps (i.e., number of missing amino acids) within a protein scaffold is known. For this version, we develop several machine learning models with data pre-processing techniques to accurately predict the missing amino acids in the gaps. In the second variant, we handle the gap filling problem of known mass (of the gaps) but unknown gap size; to be precise, only the total mass of the missing amino acids required to fill the gaps is known. For this version, we design a probabilistic algorithm to fill the missing amino acids in the gaps.

When a homologous reference protein is given, a number of useful polynomial-time algorithms based on local search and dynamic programming were developed in 2017 by Qingge et al. for the protein scaffold gap filling problem [7]. The running time of their developed algorithms to obtain optimal solutions is $O(n^{26})$, where n is the size of the reference protein (also the total length of the protein to be filled) [7]. When a reference protein is not given, different innovative approaches based on deep learning models such as a convolutional neural network and long short-term memory for the protein scaffold gap filling problem were designed [10]. Sturtz et al. also introduced a convolutional denoising autoencoder model, achieving remarkable accuracy in gap filling [9]. These methods mainly focus on the protein scaffold filling problem with known gap size. In this paper, we will also solve the gap filling problem with known mass gaps in the scaffold.

2 Methodology

Given an (unknown) target protein sequence T and a protein scaffold $S = (S_1, S_2, ..., S_m)$ of m contigs, with a gap composed of missing amino acids

between contigs S_i and S_{i+1}, the protein scaffold gap filling (PSGF) problem is to fill the missing amino acids in S to obtain S' such that the number of one-to-one matching amino acids between the filled sequence S' and target sequence T is maximized. (S' is used as the predicted protein sequence for T.)

It is important to note that the exact sizes and masses of these missing gaps may not be known. Our paper stands out due to its innovative approach to addressing protein sequencing challenge by employing different algorithms. Specifically, we introduce techniques for filling gaps in two different types of PSGF problems as illustrated in Fig. 1. The first PSGF problem is focused on the case with known gap size and the second on the case with known gap mass. Machine learning models, such as random forest, k-nearest neighbors, decision tree and fully connected neural network are used for the first version. For the second PSGF problem (with known gap mass) we design a new probabilistic algorithm (See Fig. 1). We provide a detailed description of our methods to tackle both version of the PSGF problem below.

Fig. 1. An illustration for the two PSGF problems.

2.1 Data Collection

Two types of protein scaffold datasets, alemtuzumab's light chain (MabCampath) [5] and antibody light chain proteoforms of Homo sapiens (P5A) [3] collected from Dupré et al. [3] are used to evaluate the performance of our proposed methods. The basic idea is that we will fill the gaps in the given scaffolds and compare the similarity between constructed protein sequence with the ground truth of the target protein sequences of alemtuzumab's light chain (MabCampath) [5] and the antibody light chain proteoforms of Homo sapiens proteoform (P5A) [3]. The real MabCampth scaffold 1 data [5] is generated by combining the bottom-up and top-down mass spectrometry approaches, which consists of five contigs and six gaps in the scaffold. We also generate simulated scaffold 2 data from MabCampth target sequence to further test the model performance. Additionally, we generate two sets of simulated scaffold 3 and 4 datasets from the proteoforms P5A by randomly introducing gaps in the target sequence. Figure 2 shows the features of the scaffold data, in which gap sizes are denoted by red dashed lines and gap mass are given by the mass in Dalton. Scaffolds 1 and 3 has fewer and shorter gaps than scaffolds 2 and 4. These scaffold data will be used to test models in different scenarios. Moreover, we retrieve 1000 homologous sequences for each scaffold sequence from NCBI's Protein Blast Server [6] as training data in our proposed machine learning models.

Fig. 2. Target and scaffold sequences of Mabcampath and P5A proteoform.

2.2 Data Preprocessing

For the PSGF problem with known gap size, we have the input-output samples in our training dataset by generating 11-mers starting from the first position of each homologous sequence and shift it to the next position until we reach to end of the sequence. Then we introduce gaps at the start and end position of each input 11-mer to simulate scenarios where certain amino acids are missing and the corresponding masked amino acids are output labels. For instance, in a 11-mer "DIQMTQSPSSL", gaps are added to create sequences "-IQMTQSPSSL" and "DIQMTQSPSS-". The corresponding output labels for these sequences would be "D" and "L" respectively. This technique of creating training data results in training the model twice (in forward direction and reverse direction) as in [10]. Additionally, these sequences undergo a label encoding transformation, converting each amino acid into a unique numeric value and making them compatible with machine learning algorithms. We further refine the feature space through feature engineering techniques, such as singular value decomposition (SVD) and row averaging. SVD is employed for dimensional reduction while preserving essential patterns in the data, and row averaging simplifies the sequences and calculates the average of numeric representations, aiding the models in detecting significant patterns within the protein sequences. For the PSGF problem with known mass, no specific data preprocessing is needed. We directly apply our proposed algorithm to fill the scaffolds using its homologous sequences (obtained from NCBI's server).

2.3 The Proposed Models for the PSGF Problem with Known Gap Size

In this subsection, we develop machine learning models, such as decision tree, KNN, random forest and combination of these models with row average and singular value decomposition (SVD) techniques to solve the PSGF problem with known gap size. Due to space constraint, we leave out all the methods related to k-nearest neighbor and random forest to the full version.

The use of singular value decomposition (SVD) and row average in the context of processing and analyzing protein scaffold gap filling is motivated by their ability to simplify complex protein sequence data. By applying SVD, the dimensionality of the sequence can be reduced while maintaining its primary structural and functional properties. This reduction technique improves the models' capacity to correctly predict the missing amino acids in the scaffold with less time and resources. The row average approach simplifies complex protein sequences by reducing the complexity of protein sequences to a single numerical number that represents the average of the encoded values of the amino acids in a kmer. This simplicity becomes particularly advantageous when they are used as a preprocessing step for machine learning algorithms, such as decision trees or random forest or KNN, allowing for quick and efficient analysis.

Decision Tree. In this subsection, we employ a decision tree algorithm for the protein scaffold gap filling problem. The initial step involves preprocessing protein sequences into 11-mers which includes gap ("-") at the start or end positions, which will be used as input data and the corresponding amino acid at the gap position will be the output label. Algorithm 1 illustrates the proposed decision tree algorithm designed to solve protein scaffold gap filling problem. In Algorithm 1 there are terms as Gini impurity, features available for splitting and stopping criteria. Gini impurity is a way to measure how mixed up or "impure" a group of items is in decision trees. In PSGF, Gini impurity measures how mixed the target y labels (missing amino acids) are among a set of items (11-mer sequences) at a specific node in the decision tree. The best feature (position in the 11-mer) and its value are selected by decision trees using Gini impurity at each node to separate the dataset. In order to create child nodes that are more "pure" in terms of their target y labels, the algorithm looks for splits that will produce the largest decrease in Gini impurity. In decision trees, stopping criteria are guidelines or conditions that determine when the algorithm should stop dividing the nodes further. Feature availability refers to whether a feature can still be used for making further splits in a decision tree. Feature availability refers to whether a feature (every position in 11-mers) can still be used for making further splits in a decision tree. If all features have been visited or if the potential splits do not reduce Gini impurity, then no more features are available for splitting at that node. Figure 3 shows simple decision tree illustration for smaller 11-mers dataset.

We have also combined row average with decision tree provides a novel method (denoted as **Row Average + Decision Tree**) for prediction of

Algorithm 1: Decision Tree Algorithm for Protein Scaffold Gap Filling

Input: Training dataset of numerically encoded 11-mers with gaps ("-") and corresponding output labels y for each sample, which will be used to fill gaps;
Step 1: Initialize the Tree
Root node has a set of samples S including all 11-mers and corresponding output labels y; Calculate the initial Gini impurity $Gini(S)$ using the formula $Gini(S) = 1 - \sum(p_i)^2$, where p_i is the proportion of items labeled with the i^{th} y labels in the set S
Step 2: Build the Decision Tree
while *features available for splitting and the stopping criteria (Gini impurity = 0) not met* **do**
    for *each feature f at each position in the 11-mers* **do**
        Identify all unique values V at feature position f and sort them
        Calculate midpoints between consecutive values in V to determine potential thresholds
        for *each potential threshold t calculated from midpoints* **do**
            Partition S into subsets S_{left} and S_{right} based on t
            S_{left} contains all 11-mers with feature f value $\leq t$
            S_{right} contains all 11-mers with feature f value $> t$
            Calculate Gini impurity for S_{left} and S_{right}
        end
        Compute weighted Gini impurity for each split using the formula:

$$Weighted\_Gini_{f,t} = \left(\frac{|S_{left}|}{|S|}\right) \times Gini(S_{left}) + \left(\frac{|S_{right}|}{|S|}\right) \times Gini(S_{right})$$

        where $|S_{left}|$ and $|S_{right}|$ are the counts of unique output y labels in each subset, and $|S|$ is the total count of 11-mers in set S
    end
    Choose the feature f and threshold t combination with the lowest $Weighted\_Gini_{f,t}$ for splitting
    If multiple thresholds yield the same lowest $Weighted\_Gini_{f,t}$, select one randomly
    Split S into S_{left} and S_{right} using the chosen threshold t at feature f
    Create child nodes for S_{left} and S_{right}, assigning the corresponding subset to each
    Assign a class (y label) to each node based on the majority class of the subset at that node
    Recursively apply the above steps to each child node, treating each as a new root
end
Step 3: Predict Missing Amino Acid for New 11-mers with gaps
for *each 11-mer* **do**
    node $\leftarrow$ root of the tree
    while *node is not leaf* **do**
        $v \leftarrow$ value at *node*'s feature in 11-mer
        node $\leftarrow v \leq$ *node*'s threshold ? left child : right child
    end
    Output *node*'s class
end

missing amino acids for PSGF. An additional efficient method for predicting missing amino acids in protein sequences is to combine decision tree modeling and SVD in the PSGF (denoted as **SVD + Decision Tree**). Again, due to space constraint we will cover the details in the full version.

Algorithm 2: Algorithm for the PSGF Problem with Known Gap Masses

Set-up and the Goal:
Assume $P(B) > 0.5$ (say 0.9, which means that the contigs are more similar to ground truth) and $P(\overline{B}) < 0.5$ (say 0.1, which means that the contigs are less similar to ground truth), the algorithm considers $P(\overline{B})$ a small probability.
The goal is to maximize $P(A)$, ensure $P(A \cap B)$ is large and $P(A \cap \overline{B})$ is small.
Compute $P(A \cap B)$ and $P(A \cap \overline{B})$:
Generate all possible sequences of amino acids of the mass Δ_i and pick one at a time, say t_i.
Break $S_i t_i S_{i+1}$ into peptides of different lengths and compute the number of these peptides appearing in the raw peptide data from the input sequence, say $\alpha(t_i)$.
Among all sampled t_i, select the one with the $\max(\alpha(t_i))$. (In practice, record the largest 10, say.)
For $P(A \cap \overline{B})$, peptides are formed by the left end of S_{i+1}, t_i, and the right end of S_i.
Similarly, to compute $P(A \cap \overline{B})$, select t_i with the $\min(\beta(t_i))$. (In practice, record the smallest 10, say).
Compute $P(A)$:
If $\alpha(t_i)$ and $\beta(t_i)$ are recorded, select the t_i with $\max\{\alpha(t_i) - \beta(t_i)\}$ to maximize $P(A)$.

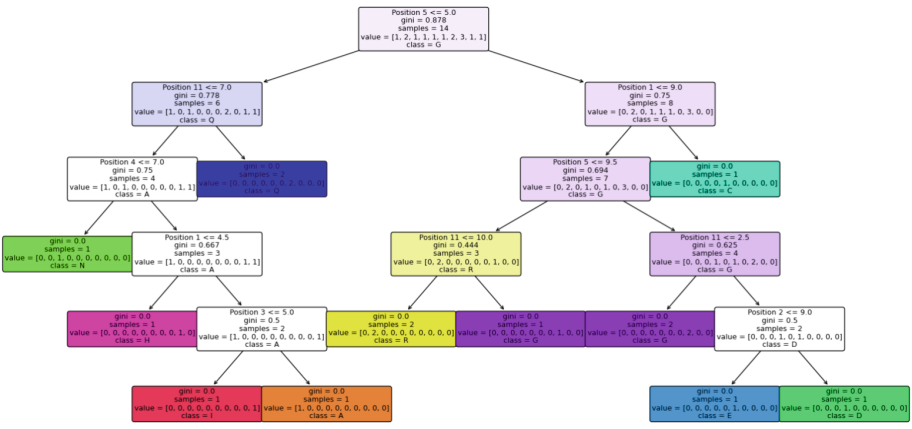

Fig. 3. Simple Decision Tree Illustration Example for PSGF.

2.4 A Probabilistic Algorithm for the PSGF Problem with Known Gap Mass

In this section, we consider another variant of the protein scaffold gap filling problem where the mass of missing amino acids in the gaps are known, while their precise sizes of how many amino acids missing are unknown. Given the scaffold $S = (S_1, S_2, \cdots, S_m)$, to fill the gap between S_i and S_{i+1} with protein sequence T_i of the right total mass Δ_i, our idea is roughly sample through all peptides with the total mass Δ_i (and we will show how to avoid a brute-force method).

Define the Event Space

Let A denote the event that the gap between contigs S_i and S_{i+1} is filled with some protein sequence T_i of the desired mass Δ_i. Let B denote the event that S_i and S_{i+1} are the correct contigs (i.e., with no error in them). The event space can be further illustrated as following

$$\underbrace{S_i \quad \overbrace{T_i}^{\text{A}} \quad S_{i+1}}_{\text{The gap between contigs is filled by } T_i \text{ with mass } \Delta_i}$$

Probability Computation:

The probability of A is calculated as:

$$P(A) = P(A|B) + P(A|\overline{B}).$$

And by the formula of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(A|\overline{B}) = \frac{P(A \cap \overline{B})}{P(\overline{B})}.$$

Algorithm 2 shows the procedure to solve the PSGF problem with known gap masses.

The main challenge of this algorithm is to generate all possible sequences of amino acids of certain mass Δ_i, even though we know the mass of each of the amino acids. To tackle this challenge we first find the minimum and maximum possible length of a sequence with the targeted mass, i.e., min_length and max_length, and then generate all sequences with a length from min_length to max_length. For a large mass max_length can be large too (and it can take $O(n^{20})$ time to generate them, where n = max_length). Hence, to make this algorithm more feasible, we generate the possible sequences of amino acids of a total mass of Δ_i using homologous sequences generated from NCBI's server for each scaffold.

We generate possible sequences of amino acids of a target mass Δ_i using the concept of kmers. For each length in [min_length, max_length] we generate kmers using homologous sequences and choose the kmers of the target mass Δ_i as a candidate for T_i. For instance, consider a certain mass 300 Daltons for

which the minimum and maximum possible length of a sequence with this mass can be 3 and 6 respectively. We can then generate kmers of length 3,4,5 and 6 respectively, using homologous sequences like "DIQMTQSPSSLSASVI...", etc. We select the kmers "DIQ", "SIS", "SGTD", "NRGEC", "IISSCT", etc., which has a total target mass $\Delta_i = 300$. Generating t_i this way makes it feasible to fill the gap with a large mass.

Additionally, when we construct the peptides from the left of S_i and the right of S_{i+1}, we simply select up to 5 amino acids from them. In other words, in $S_i t_i S_{i+1}$ and when computing $P(A \cap \overline{B})$, we only consider up to 5 amino acids on the right (resp. left) end of S_i (resp. S_{i+1}).

3 Experimental Results

We test our proposed protein scaffold gap filling problem with known size on Scaffold 1 and Scaffold 3, and test the proposed protein scaffold gap filling problem with known mass on Scaffold 2 and Scaffold 4, which have been introduced in Sect. 2.1.

3.1 Results for PSGF with Known Gap Size Using ML Models

To increase the proposed machine learning models prediction accuracy, we utilize data preprocessing techniques including row average and singular value decomposition (SVD) to extract the important features from input data. We evaluate the model performance by comparing the similarity between the filled gaps in the scaffold with the corresponding positions of target sequence. The gap filling accuracy is a fraction of the number of one-to-one matching amino acids in the gaps with the corresponding position at the target sequence with the length of amino acids in the gaps of the scaffold. Each model is used to predict the

Table 1. The Model Training and Validation Accuracy on Mabcampath Data.

	Train Acc.	Validation Acc.
KNN	94.37	92.84
Decision Tree	97.52	94.46
Random Forest	97.52	95.60
SVD + KNN	94.43	93.01
SVD + Decision Tree	97.52	92.40
SVD + Random Forest	97.52	94.16
Row Average + KNN	95.74	94.50
Row Average + Decision Tree	97.51	96.92
Row Average + Random Forest	97.51	97.11
Fully Connected NN	94.92	92.83

missing amino acids in the gaps of the scaffold one after another. The training and validation accuracy of these models on Mabcampath data are illustrated in Table 1.

To demonstrate the performance of our models on different scaffolds with larger gaps, we run the models on Scaffold 1–4 data. All models show promising performance on filling the gaps of these scaffolds. Table 4 shows the gap filling accuracy of the proposed models on MabCampath and P5A protein scaffolds. The prediction results show the proposed machine learning models can fill the gaps of the protein scaffold accurately. Our proposed models also achieve 100% gap filling accuracy compared with CNN-LSTM model developed in [10] on the real MabCampth data. Figure 4 shows the 100% gap filling prediction accuracy results on the Scaffold 1 and 3, which have smaller size of gaps. For the scaffolds 2 and 4 having the larger size of gaps, our proposed machine learning models also achieve higher prediction accuracy up to 94.73% and 100% respectively. Table 4 summarizes all models prediction accuracy on Scaffold 1–4 datasets.

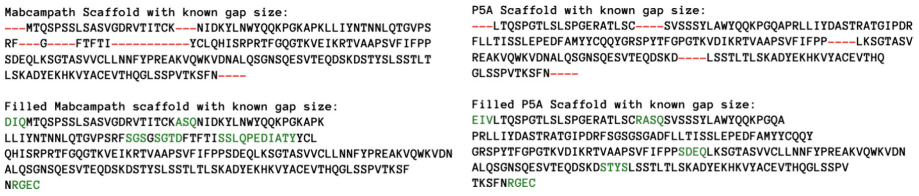

Fig. 4. Filled Mabcampath scaffold 1 and P5A proteoform scaffold 3 for all the models.

Table 2. Gap filling accuracy on MabCampath and P5A scaffold.

	Gap Filling Acc.			
	Mab		P5A	
	Scaffold 1	Scaffold 2	Scaffold 3	Scaffold 4
KNN	100.0	94.73	100.0	97.73
SVD + KNN	100.0	94.73	100.0	94.73
Row Average + KNN	100.0	94.73	100.0	94.73
Decision Tree	100.0	93.42	100.0	100.0
SVD + Decision Tree	100.0	93.42	100.0	94.73
Row Average + Decision Tree	100.0	93.42	100.0	100.0
Random Forest	100.0	93.42	100.0	100.0
SVD + Random Forest	100.0	93.42	100.0	97.36
Row Average + Random Forest	100.0	93.42	100.0	100.0
Fully Connected Neural Network	100.0	92.10	100.0	97.36

3.2 Results for PSGF with Known Gap Masses Using the Probabilistic Algorithm

We design and implement the Algorithm 2 to fill the gaps in the protein scaffold gap filling problem. To fill the gaps, we search the maximum value of $\alpha(t_i)$ and minimum value of $\beta(t_i)$ from all sample t_i. We test our algorithm on all the generated scaffolds 1–4 shown in Fig. 2. Table 3 shows $\max(\alpha(t_i))$ and $\min(\beta(t_i)))$ values of each gap for Mabcampath scaffold. We achieve 100% gap filling accuracy on all the scaffolds data. It demonstrates that our designed probabilistic algorithm can fill missing amino acids in gaps accurately in the scenario of known gap mass with unknown size of the gap. Table 3 and 4 show the computed amino acids to fill Mabcampath and P5A scaffolds.

Table 3. Gap filling on MabCampath scaffold 1 with known mass.

Gap Mass	Predicted combination	$\text{Max}(\alpha(t_i))$ and $\text{Min}(\beta(t_i))$
356 Dalton	DIQ	[572, 3]
286 Dalton	ASQ	[3, 9]
231 Dalton	SGS	[734, 31]
360 Dalton	SGTD	[70, 24]
1204 Dalton	SSLQPEDIATY	[8, 2]
445 Dalton	RGEC	[886, 0]

Table 4. Gap filling on P5A scaffold 3 with known mass.

Gap Mass	Predicted combination	$\text{Max}(\alpha(t_i))$ and $\text{Min}(\beta(t_i))$
341 Dalton	EIV	[292, 0]
442 Dalton	RASQ	[238, 200]
459 Dalton	SDEQ	[991, 0]
438 Dalton	STYS	[990, 3]
445 Dalton	RGEC	[944, 0]

4 Conclusions

In this paper, we consider two versions of the protein scaffold gap filling (PSGF) problem. For PSGF with known gap size, we propose several machine learning models combined with data preprocessing engineering techniques, such as decision tree, KNN, random forest and also develop fully connected neural network

to fill the gaps iteratively until all the gaps are filled. For the PSGF problem with known gap mass, we design a probabilistic model to fill the missing amino acids in the gaps. The experimental results on different scenario of scaffolds show that our proposed algorithms achieve promising results on both real and simulation datasets, in fact, with 100% or close to 100% accuracy in general. Moreover, the proposed algorithms are simple, yet effective to solve the protein scaffold gap filling problem.

Acknowledgements. This work is supported by the NSF of the United States under Award 2307571, 2307572 and 2307573. We also thank anonymous reviewers for their insightful comments and inputs.

References

1. Aebersold, R., Mann, M.: Mass spectrometry-based proteomics. Nature **422**(6928), 198–207 (2003)
2. Bricas, E., Van Heijenoort, J., Barber, M., Wolstenholme, W., Das, B., Lederer, E.: Determination of amino acid sequences in oligopeptides by mass spectrometry. IV. Synthetic n-acyl oligopeptide methyl esters. Biochemistry **4**(10), 2254–2260 (1965)
3. Dupré, M., et al.: De novo sequencing of antibody light chain proteoforms from patients with multiple myeloma. Anal. Chem. **93**(30), 10627–10634 (2021). pMID: 34292722. https://doi.org/10.1021/acs.analchem.1c01955
4. Kinter, M., Sherman, N.E.: Protein Sequencing and Identification Using Tandem Mass Spectrometry. Wiley, Hoboken (2005)
5. Liu, X., et al.: De novo protein sequencing by combining top-down and bottom-up tandem mass spectra. J. Proteome Res. **13**(7), 3241–3248 (2014)
6. National Center for Biotechnology Information: Blast (2023). https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins
7. Qingge, L., Liu, X., Zhong, F., Zhu, B.: Filling a protein scaffold with a reference. IEEE Trans. Nanobiosci. **16**(2), 123–130 (2017)
8. Standing, K.G.: Peptide and protein de novo sequencing by mass spectrometry. Curr. Opin. Struct. Biol. **13**(5), 595–601 (2003)
9. Sturtz, J., Annan, R., Zhu, B., Liu, X., Qingge, L.: A convolutional denoising autoencoder for protein scaffold filling. In: Guo, X., Mangul, S., Patterson, M., Zelikovsky, A. (eds.) Bioinformatics Research and Applications, ISBRA 2023. LNCS, vol. 14248, pp. 518–529. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-7074-2_42
10. Sturtz, J., Zhu, B., Liu, X., Fu, X., Yuan, X., Qingge, L.: Deep learning approaches for the protein scaffold filling problem. In: 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1055–1061. IEEE (2022)
11. Tran, N.H., Rahman, M.Z., He, L., Xin, L., Shan, B., Li, M.: Complete de novo assembly of monoclonal antibody sequences. Sci. Rep. **6**(1), 1–10 (2016)
12. Wulfson, N., et al.: Mass spectrometric determination of the amino (hydroxy) acid sequence in peptides and depsipeptides. Tetrahedron Lett. **6**(32), 2805–2812 (1965)