CALIFORNIA STATE UNIVERSITY
FULLERTON

Evolution of **Optimal Global Alignment** Methods in Bioinformatics

Hardik Bhawsar

885191064

Abstract

Optimal global alignment is a fundamental problem in bioinformatics, aiming to identify the best alignment between two biological sequences while maximizing the number of matches and minimizing gaps or mismatches. Optimal global alignment is a crucial task in bioinformatics, enabling the identification of similarities and evolutionary relationships between biological sequences. This survey paper explores the evolution of optimal global alignment methods, tracing their development from seminal works to more recent advancements. We discuss the Needleman-Wunsch algorithm, Gotoh's algorithm, the Smith-Waterman algorithm, profile hidden Markov models (HMMs), and the application of deep learning techniques in computational biology. This survey paper provides a comprehensive overview of the evolution of optimal global alignment methods and their impact on bioinformatics.

Introduction

Optimal global alignment plays a critical role in bioinformatics, facilitating the identification of similarities and evolutionary relationships between biological sequences. This survey paper presents a historical overview of the evolution of optimal global alignment methods. The invention of optimal global alignment can be attributed to Needleman and Wunsch, who introduced the concept in their seminal paper in 1970.

Needleman and Wunsch proposed the Needleman-Wunsch algorithm in 1970. Their algorithm introduced the concept of dynamic programming to align amino acid sequences by considering all possible alignments and finding the optimal one based on a scoring scheme. The algorithm considers all possible alignments and constructs a scoring matrix to identify the best alignment. The matrix is filled iteratively, considering the scores of three possible operations: match, mismatch, and gap. By backtracking through the matrix, the optimal alignment is determined. The Needleman-Wunsch algorithm laid the foundation for subsequent developments in optimal global alignment methods.

In 1982, Gotoh introduced an improved algorithm for matching biological sequences. Gotoh's algorithm enhanced the Needleman-Wunsch algorithm by incorporating affine gap penalties. By assigning separate penalties for gap opening and extension, the algorithm better accounted for the characteristics of biological sequences. This modification significantly improved the accuracy of optimal global alignment, especially when dealing with sequences containing gaps. The Smith-Waterman algorithm, proposed in 1981, introduced the concept of local alignment that can be adapted for global alignment. The algorithm determines the optimal global alignment by finding local alignments and extending them. This approach is beneficial when aligning sequences with regions of high similarity amid unrelated sections. The Smith-Waterman algorithm provides a more flexible and versatile approach to optimal global alignment.

In 1998, Eddy introduced profile hidden Markov models (HMMs) to capture evolutionary information and improve alignment accuracy. HMMs combine sequence profiles and probabilistic models to represent conserved patterns in sequences. By modeling sequence similarity and variation, HMMs provide more sophisticated scoring schemes for optimal global alignment. This approach enables the detection of remote homologs and improves the alignment of distantly related sequences. More recently, deep learning techniques have been applied to computational biology, including optimal global alignment. Angermueller and Stegle demonstrated the potential of deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in improving alignment accuracy. These models learn complex features

directly from input sequences, enhancing the capability to capture intricate relationships and patterns. Deep learning approaches show promise in further advancing optimal global alignment methods.

S. No.	Title	Year	Analyzing	Drawbacks/Possible Improvement
1.	A general method applicable to the search for similarities in the amino acid sequence of two proteins.	1970	• The algorithm introduced the concept of dynamic programming to align amino acid sequences by considering all possible alignments and finding the optimal one based on a scoring scheme. • The matrix is filled iteratively, considering the scores of three possible operations: match, mismatch, and gap. • By backtracking through the matrix, the optimal alignment is determined.	The Needleman-Wunsch algorithm has a quadratic time complexity, making it inefficient for aligning large-scale sequence databases. To address this limitation, researchers explored optimization techniques to reduce time and space complexity. One notable improvement is the introduction of the Hirschberg algorithm, which employs a divide-and-conquer strategy to achieve linear space complexity without sacrificing alignment accuracy.
2.	An improved algorithm for matching biological sequences	1982	• Gotoh's research paper presented an improved algorithm that addressed the limitations of the Needleman-Wunsch algorithm. One major enhancement was the incorporation of affine gap penalties, which introduced separate penalties for gap openings and gap extension. This modification better reflected the biological reality of sequence gaps and improved alignment accuracy. • Gotoh's algorithm extended the concept of dynamic programming to handle affine gap penalties. It introduced an additional matrix to track the	Gotoh's algorithm had a profound impact on bioinformatics applications that rely on optimal global alignment. It facilitated more accurate sequence comparison, aiding in the identification of homologous sequences, conserved motifs, and functional domains. The improved alignment accuracy also contributed to more reliable phylogenetic tree construction, protein structure prediction, and sequence annotation. Subsequent research built upon Gotoh's work led to further algorithm enhancements and adaptations. Researchers explored

				optimal alignment of gap extensions, allowing for the computation of alignment scores that considered both gap opening and extension penalties. This extension enhanced the alignment accuracy by capturing the finer details of gap placement.	variations of affine gap penalties, including non-linear gap penalties and gap penalties dependent on sequence context. Additionally, the algorithm's efficiency was improved through optimization techniques, enabling the alignment of larger datasets.
3.	Identification of common molecular subsequences	1981	• The Smith-Waterman algorithm addressed the limitations of global alignment algorithms by introducing a dynamic programming approach that allowed for local alignment. The algorithm computed a score matrix by considering the similarity of subsequence pairs and their neighboring regions. It then utilized backtracking to determine the optimal local alignment, thereby enabling the identification of common molecular subsequences. • The Smith-Waterman algorithm employed a scoring system that considered the similarity between residues in the sequences being aligned. It also introduced gap penalties to account for gaps in the alignment. The flexibility of the gap penalties allowed for the identification of local similarities, as gaps could be introduced without incurring significant penalties.	The Smith-Waterman algorithm has a quadratic time complexity, making it computationally demanding for large-scale sequence alignment. As the algorithm considers all possible alignments in the dynamic programming matrix, the time required increases with the length of the sequences being aligned. This can be a limiting factor when dealing with extensive genomic or proteomic datasets. Due to its quadratic time complexity and memory requirements, the Smith-Waterman algorithm may not scale well for aligning multiple sequences simultaneously. When dealing with multiple sequence alignments or comparing a large number of sequences, the algorithm's computational and memory limitations become more pronounced.
4.	Profile hidden Markov models	1998	• Eddy's research paper detailed the construction of profile HMMs from a multiple sequence alignment. The process involves training an	Profile HMMs require significant computational resources and time for training and learning the model parameters. Constructing a profile

			HMM on the aligned sequences, learning the emission and transition probabilities, and creating a probabilistic model that represents the conserved regions, insertions, and deletions in the family of sequences. • Profile HMMs enable the alignment of a query sequence against a profile representing a sequence family. The alignment is performed using the Viterbi algorithm, which identifies the most probable alignment path. The scoring of the alignment provides a measure of similarity or homology between the query sequence and the profile HMM.	HMM from a multiple sequence alignment involves estimating emission and transition probabilities, which can be computationally intensive for large datasets. The complexity of the training process can limit the scalability and practicality of profile HMMs in certain applications. The accuracy of profile HMMs heavily relies on the quality of the initial multiple sequence alignments used for training. Inaccurate or poorly aligned sequences in the training set can result in suboptimal profile HMMs. Thus, the performance of profile HMMs is influenced by the quality of the underlying alignment data.
5.	Deep learning for computational biology	2013	• Angermueller and Stegle's research paper outlined the process of training deep learning models for global alignment. This involved the preparation of training datasets, including pairs of aligned sequences, and the design and training of CNN or RNN architectures to learn the alignment patterns and similarity measures. The models were trained using gradient-based optimization algorithms to minimize alignment errors. • Deep learning models for alignment leverage the learned representations and patterns to align sequences. CNN-based models excel in capturing local	The paper does not focus on proposing a specific alignment algorithm but rather explores the potential of deep learning techniques for global alignment. As a result, the paper lacks a detailed analysis and evaluation of a specific alignment method based on deep learning. The paper does not mention the datasets available for training and assessing deep learning models. Lack of data availability hinders the reproducibility of the experiments and limits the ability of other researchers to validate and build upon the proposed methods. The paper lacks a comprehensive evaluation of the proposed deep-learning models for global alignment. While the authors

				sequence features, while RNN-based models can model long-range dependencies. Alignment with deep learning models involves feeding the sequences into the trained model and obtaining the aligned outputs based on learned representations and alignment rules.	mention the competitive performance of their models, there is little discussion on the specific evaluation metrics used, such as alignment accuracy, sensitivity, specificity, or computational efficiency.

Conclusion

The evolution of optimal global alignment methods in bioinformatics has witnessed significant advancements. From the pioneering work of Needleman and Wunsch in 1970, who introduced the Needleman-Wunsch algorithm as a general method for sequence alignment, to Gotoh's 1982 improved algorithm that incorporated affine gap penalties, researchers have continuously strived to enhance alignment accuracy and efficiency. The Smith-Waterman algorithm, proposed in 1981, introduced the concept of local alignment that can be adapted for global alignment, allowing for more flexible and versatile alignment strategies. In 1998, Eddy's introduction of profile hidden Markov models (HMMs) revolutionized alignment methods by capturing evolutionary information and improving alignment accuracy. More recently, the application of deep learning techniques by Angermueller and Stegle in 2013 has shown promising results in improving alignment accuracy through the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Overall, the evolution of optimal global alignment methods has been driven by a combination of algorithmic improvements, such as affine gap penalties and the concept of local alignment, as well as the integration of probabilistic models and deep learning techniques. These advancements have greatly enhanced the accuracy, efficiency, and flexibility of optimal global alignment, enabling researchers to align larger and more diverse biological sequences.

References

1. Needleman, S.B., & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443-453.
2. Gotoh, O. (1982). An improved algorithm for matching biological sequences. Journal of Molecular Biology, 162(3), 705-708.
3. Smith, T.F., & Waterman, M.S. (1981). Identification of common molecular subsequences. Journal of Molecular Biology, 147(1), 195-197.
4. Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics, 14(9), 755-763.
5. Angermueller, C., & Stegle, O. (2013). Deep learning for computational biology. Molecular Systems Biology, 9(1), 1-13.