# Fitting and Local Aligner with Multiple Solutions

**Payal Mantri, Sindhu Inuganti**
**NetID:** payaljm2, vsi3

**February 27, 2024**

---

## Introduction

Alignment algorithms are used to find similarity between biological sequences, such as DNA and proteins. Proteins consist of amino acid chains, and aligning the amino acid sequences of related proteins may show which regions have functional or structural importance. Gene sequences are aligned to find regions of similarity, which may help to indicate if the genes are related. Sequence alignments are also used in the field of phylogenetics to construct and interpret phylogenetic trees.
One of the commonly used algorithms for sequence alignment is the Needleman-Wuncsh algorithm[4]. This global alignment algorithm is used to compare sequences in cases where we have reason to believe that the sequences are related along their entire length. However, local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm[5] is a general local alignment method also based on dynamic programming.
Sequence alignments help microbiologists to find gene sequences of high similarity and thus help in the identification of the species.
As a part of this project, we plan to extend the functionality of existing Needleman-Wunsch aligner[git] to support fitting alignments and local alignments.

## Motivation

While learning sequence alignments in class, we realized that the web tools[**?** ] available online for fitting and local alignment give only one solution. However, there can be multiple solutions for the same problem, using fitting and local alignment algorithm.
Thus, we decided to extend the functionality of the existing aligner to support multiple solutions for fitting and local alignments. This will also help the new students to visualize all the possible alignments and choose the best one for further analysis. It can also be useful to programmers debugging alignment algorithms, since it allows you to see both the aligned sequences and the dynamic programming matrix (including trace-back directions) for an alignment.

## Datasets

We plan to use the following datasets for testing and evaluating our aligner:

- **HOMSTRAD**: HOMologous STRucture Alignment Database [3] is a curated database of structure-based alignments for homologous protein families.
- **Pfam** : Pfam[2] is a database of protein families, each represented by multiple sequence alignments and Hidden Markov Models (HMMs).

## Plan of Work: Algorithms and Evaluation

We will be using the existing implementation of Needleman Wunsh Algorithm(Global Alignment Algorithm). The aligner is a web tool implemented in javascript and is available at http://experiments.mostafa.io/public/needleman-wunsch/index.html[git]
We plan to use this as base code . Initial step would be to implement the fitting and local alignment to get one solution. Later it could be generalized to generate all possible optimal alignments. The main challenge we see is that for long sequences, precomputing all local alignment might take indefinitely long time. Hence we plan to just show the number of solutions and click on solution number will generate the alignment on the fly.
The implementation will be evaluated by using sequences from the above mentioned datasets and comparing the results with the existing one solution aligner [**?** ].

**Project Timeline**

| | |
|---|---|
| Project Proposal | 11/10/2022 |
| Understanding the base code and data collection | 11/13/2022 |
| Design the UI requirements | 11/15/2022 |
| Fitting Alignment Implementation | 11/20/2022 |
| Local Alignment Implementation | 11/24/2022 |
| Testing and Evaluation | 11/27/2022 |
| Presentation | 11/30/2022 |
| Report Writing and Submission | 12/14/2022 |

## References


[git] GitHub - drdrsh/Needleman-Wunsch: Interactive Visualization of Needleman-Wunsch Algorithm in Javascript — github.com. https://github.com/drdrsh/Needleman-Wunsch. [Accessed 10-Nov-2022].

[2] Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Orengo, C. A., Pandurangan, A. P., Rivoire, C., Sigrist, C. J. A., Sillitoe, I., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., Bateman, A., and Finn, R. D. (2020). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1):D344–D354.

[3] Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998). Homstrad: A database of protein structure alignments for homologous families. *Protein Science*, 7.

[4] Needleman, S. B. and 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*, 48:443–453.

[5] Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197.