Design of an embedding alignment program by dynamic programming

Jean Delhomme, Tatiana Galochkina, Jean-Christophe Gelly

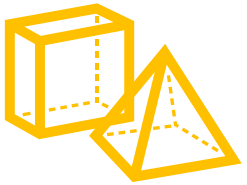
16/09/2022

Université Paris Cité

Introduction (1)

Sequence alignment is used for :

Evolutionary links identification



Structure prediction

function prediction

Traditional methods are looking for sequence homology, residue per residue

CPTLIVMVGLPARGKTYISKKLTRYLNFIGVPTREFNVGQYRRDMVKTYKSFEFFLPDNEEGLKIR

TNTTRERRAMIFNTGEQNGYKTFFVESICVDPEVIAANIVQVKLGSPDYVNRDSDEATEDESYKLN

Introduction

New approach : deep learning and vector embeddings

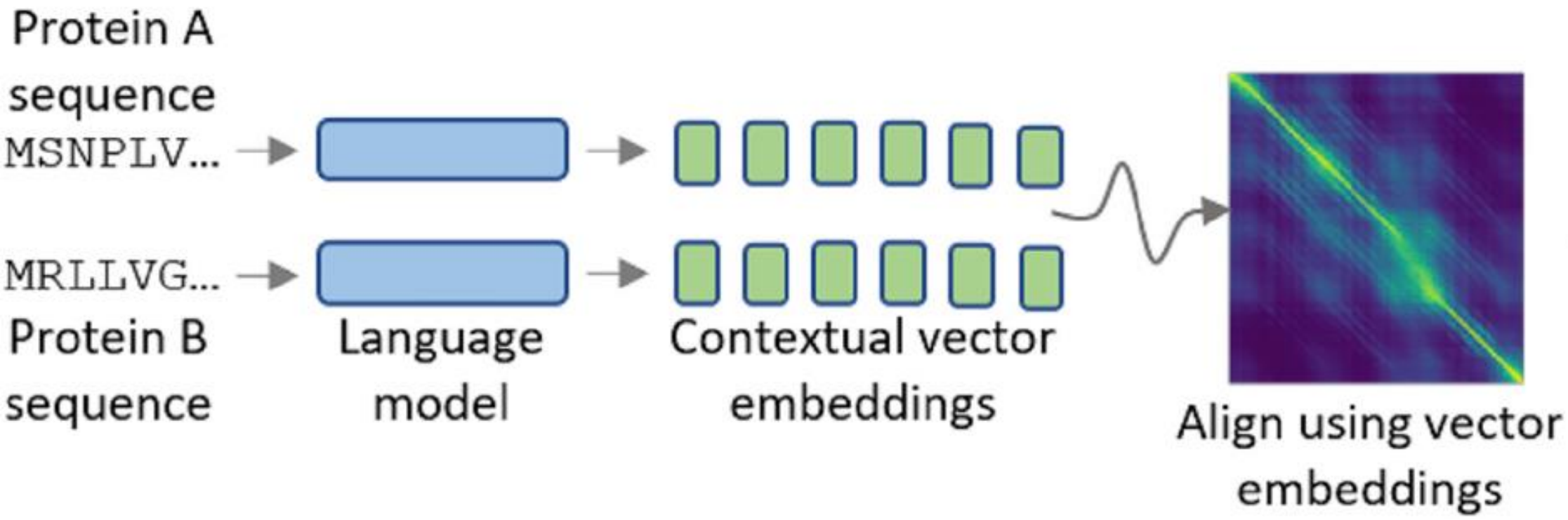


Figure 1 : the use of embeddings for sequence alignment. [2]

Table of contents

Table of contents

1. The algorithm

Environment : `python 3.10.4  numpy 1.23.1`

Functionalities :

- Vector embeddings
- Dynamic programming
- 3 alignment methods : Global [3], local [4] and semi-global [5]
- Linear ~~and affine~~ gaps

Input : 1 fasta file and 1 embedding file (T5 ProtTrans) for each protein.

Output : aligned sequence in a separate txt file.

Table of contents

2. Results

Objectives :

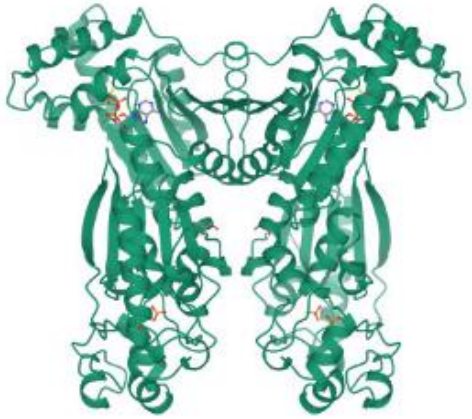
- The algorithm must work
- The results must be coherent

3 tests :

- Best case : one protein on itself
- Worst case : two very different proteins
- In between : two not so different proteins

2. Results : best case

6PF2K



6PF2K

Global alignment of 6PF2K_1bif and 6PF2K_1bif

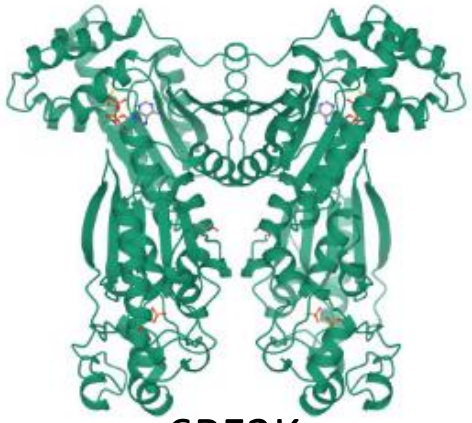
TM score : 1,00

Alignment_score = 6595.005491138418

```
6PF2K_1bif
CPTLIVMVGLPARGKTYISKKLTRYLNFIGVPTREFNVGQYRRDMVKTYKSFEFFLPDNEEGL
CPTLIVMVGLPARGKTYISKKLTRYLNFIGVPTREFNVGQYRRDMVKTYKSFEFFLPDNEEGL
6PF2K_1bif
6PF2K_1bif
KIRKQCALAALNDVRKFLSEEGGHVAVFDATNTTRERRAMIFNFGEQNGYKTFFVESICVDPE
KIRKQCALAALNDVRKFLSEEGGHVAVFDATNTTRERRAMIFNFGEQNGYKTFFVESICVDPE
6PF2K_1bif
6PF2K_1bif
VIAANIVQVKLGSPDYVNRDSDEATEDFMRRIECYENSYESLDEEQDRDLSYIKIMDVGQSYV
VIAANIVQVKLGSPDYVNRDSDEATEDFMRRIECYENSYESLDEEQDRDLSYIKIMDVGQSYV
6PF2K_1bif
6PF2K_1bif
VNRVADHIQSRIVYYLMNIHVTPR
VNRVADHIQSRIVYYLMNIHVTPR
6PF2K_1bif
```

Figure 2 : alignment of 6PF2K against itself. (6)

2. Results : worst case



6PF2K

7kD_DNA_binding

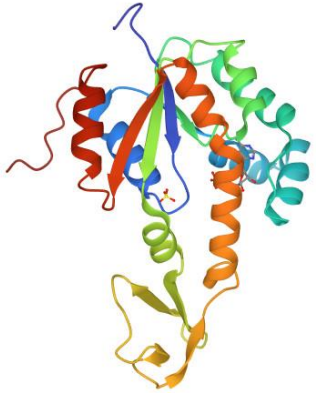
```
Global alignment of 6PF2K_1bif and 7kD_DNA_binding_1azpa
Alignment_score = 229.09141635142177
6PF2K_1bif
CPTLIV--MVGL-PA--RGKTYISKKLTRYLNFIGVPTREFNVGQ-YRRDM-VKTYKSFEFFLPDNEE
MVKV--KF--K-Y--KGEE-------------------KEVDTSKI-KKVWR----------------
7kD_DNA_binding_1azpa
6PF2K_1bif
GLKIRKQCALAALNDVRKFLSEEGGHVAVFDATNTTRER-RAM-I-FNF-G--E-QN-GYKTFFVEIC
--------------------------------------V-G-K-MVSF--T-YD-D--NGK-TG-----
7kD_DNA_binding_1azpa
6PF2K_1bif
ICVDPEVIAANIVQVKLG--SPDYVNRDSDEATE-DFMRRIECYENSYESLDEEQDR—DLSYIDLSYI
-----------------RGAV-----SEKDAPKE-LLDMLARAER------------EK----------
7kD_DNA_binding_1azpa
6PF2K_1bif
IMDVGQSYVVNRVADHIQSRIVYYLMNIHVTPR
-----------------------------K
7kD_DNA_binding_1azpa
```

TM score : 0,16

Figure 3 : alignment of 6PF2K against 7kD_DNA_binding. [6][7]

2. Results : in between

6PF2K



adk

Alignment of 6PF2K_1bif and adk_2ak3a

TM score : 0,62

Alignment_score = 2326.0764011769847

Global	Local	Glocal
6PF2K_1bif HV---------TPR -LPQRSQETSVTP- adk_2ak3a	6PF2K_1bif HV---------TPR -LPQRSQETSVTP- adk_2ak3a 6PF2K_1bif HV---------TP -LPQRSQETSVTP adk_2ak3a	6PF2K_1bif HV---------TPR -LPQRSQETSVTP- adk_2ak3a 6PF2K_1bif HV---------TP -LPQRSQETSVTP adk_2ak3a

Figure 4 : alignment of 6PF2K against adk. [6][8]

Table of contents

3. Discussion

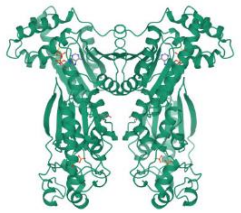
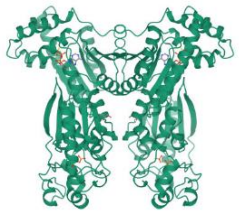
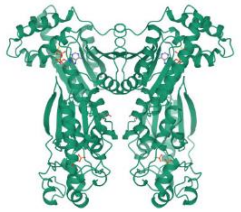
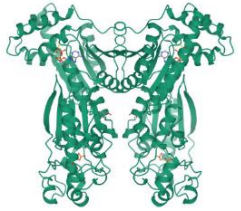
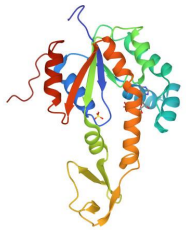
Protein 1	Protein 2	Alignment score	TM score
		6595	1,00
		229	0,16
		2326	0,62

Table 1 : result summary. [6][7][8]

Conclusion

Functioning program

Embedding is a plus

Affine gap

argparse

OOP

Python

Github

sys.argv

numpy

markdown

programming practices

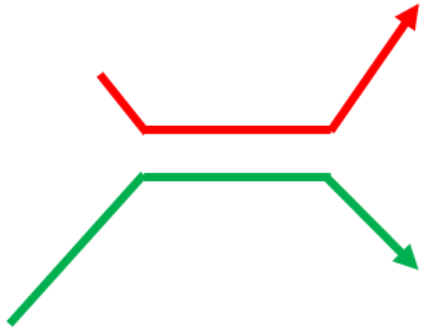
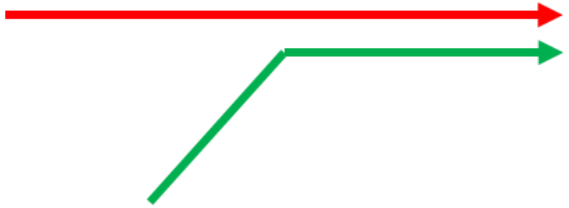
References

1. S. Altschul et al. (2017) Handbook of Discrete and Combinatorial Mathematics. 2nd edition, Chapter 20.1 Sequence
2. T. Bepler et al. (2021) Learning the protein language: Evolution, structure, and function.
3. Needleman, Saul B. & Wunsch, Christian D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins.
4. Smith, Temple F. & Waterman, Michael S. (1981) Identification of Common Molecular Subsequences.
5. A Elnaggar et al. (2020) ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing
6. Page pdb de la 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase bifunctional enzyme complexed with atp-g-s and phosphate : https://www.rcsb.org/structure/1bif
7. Page pdb de la hyperthermophile chromosomal protein sac7d bound with kinked dna duplex : https://www.rcsb.org/structure/1azp
8. Page pdb de 2AK3 : https://www.rcsb.org/structure/2ak3

Thank you for your listening

QUESTIONS

Alignment methods

	Global	Local	Glocal
Reference Querry			
$H_{i,j} =$	$max\begin{cases} H_{i-1,j-1} + S_{i,j} \\ H_{i-1,j} + gap \\ H_{i,j-1} + gap \end{cases}$	$max\begin{cases} H_{i-1,j-1} + S_{i,j} \\ H_{i-1,j} + gap \\ H_{i,j-1} + gap \\ 0 \end{cases}$	$max\begin{cases} H_{i-1,j-1} + S_{i,j} \\ H_{i-1,j} + gap \\ H_{i,j-1} + gap \end{cases}$
Starting point	Bottom-right corner	Anywhere	Right column
Ending point	Top-left corner	Anywhere	Left column

Program structure