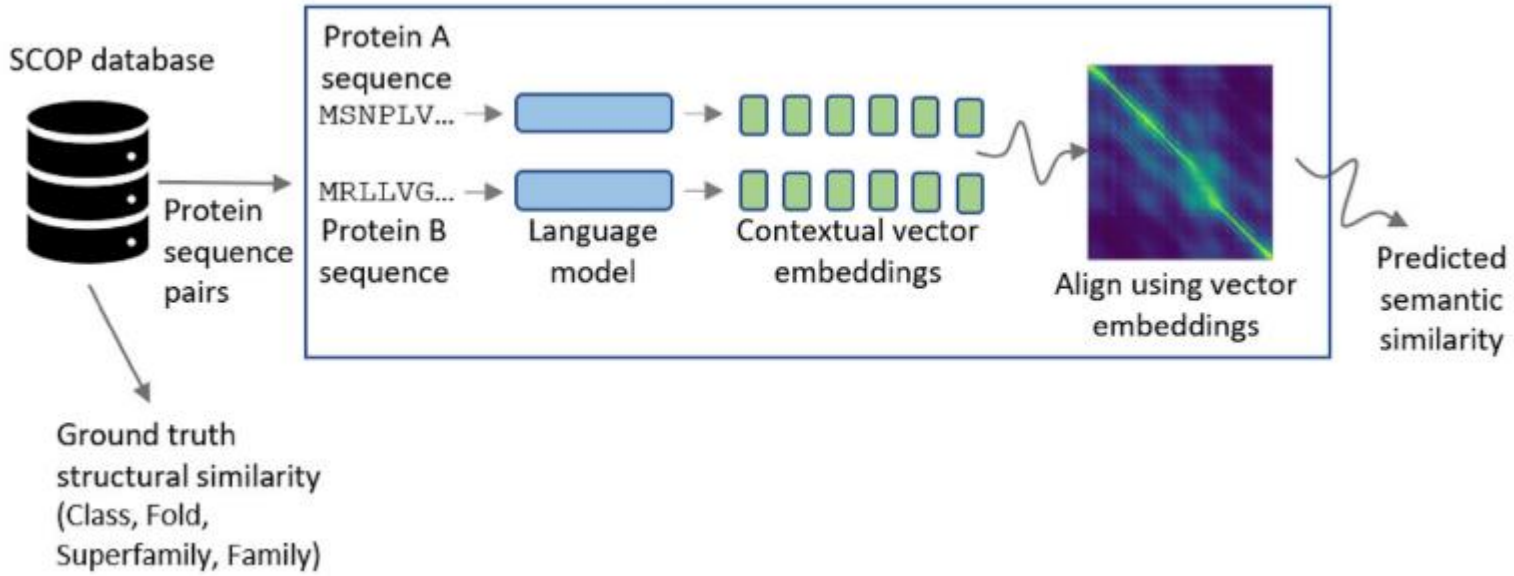
Production of an alignment program based on embedding by dynamic programming

Clémence Lauden

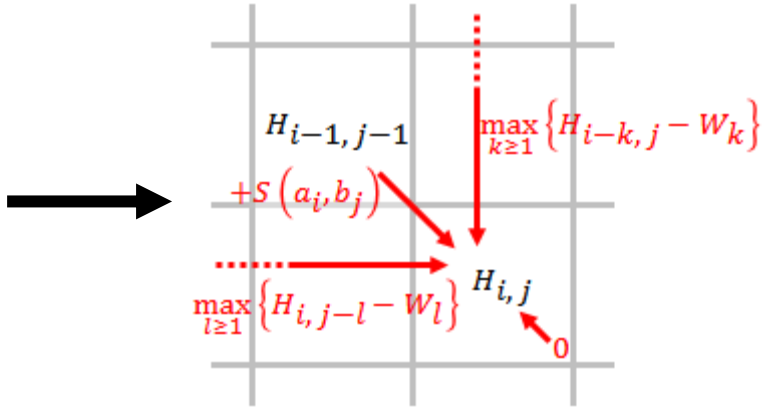
Embedding



Embedding production (From Bepler et al., 2021)

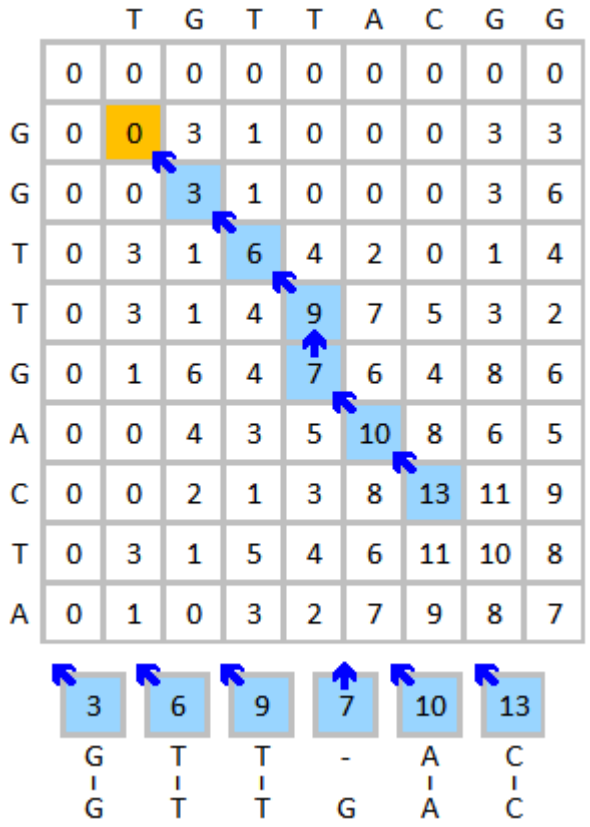
Alignment algorithm : Matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								



		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

Alignment methods : Backtracking



		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

3	6	9	7	10	13
G	T	T	-	A	C
I	I	I		I	I
G	T	T	G	A	C

Alignment methods : difference

Matrix

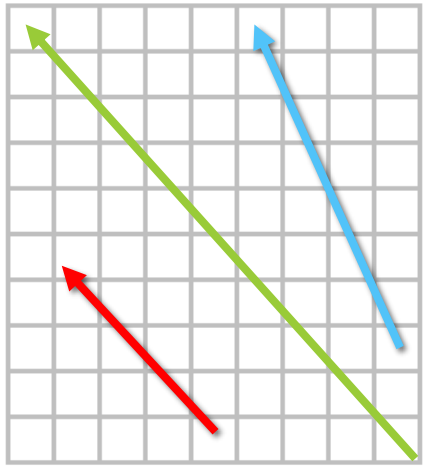
		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Scoring matrix initialization for local alignment

		G	C	A	T	G	C	G
	0	−1	−2	−3	−4	−5	−6	−7
G	−1							
A	−2							
T	−3							
T	−4							
A	−5							
C	−6							
A	−7							

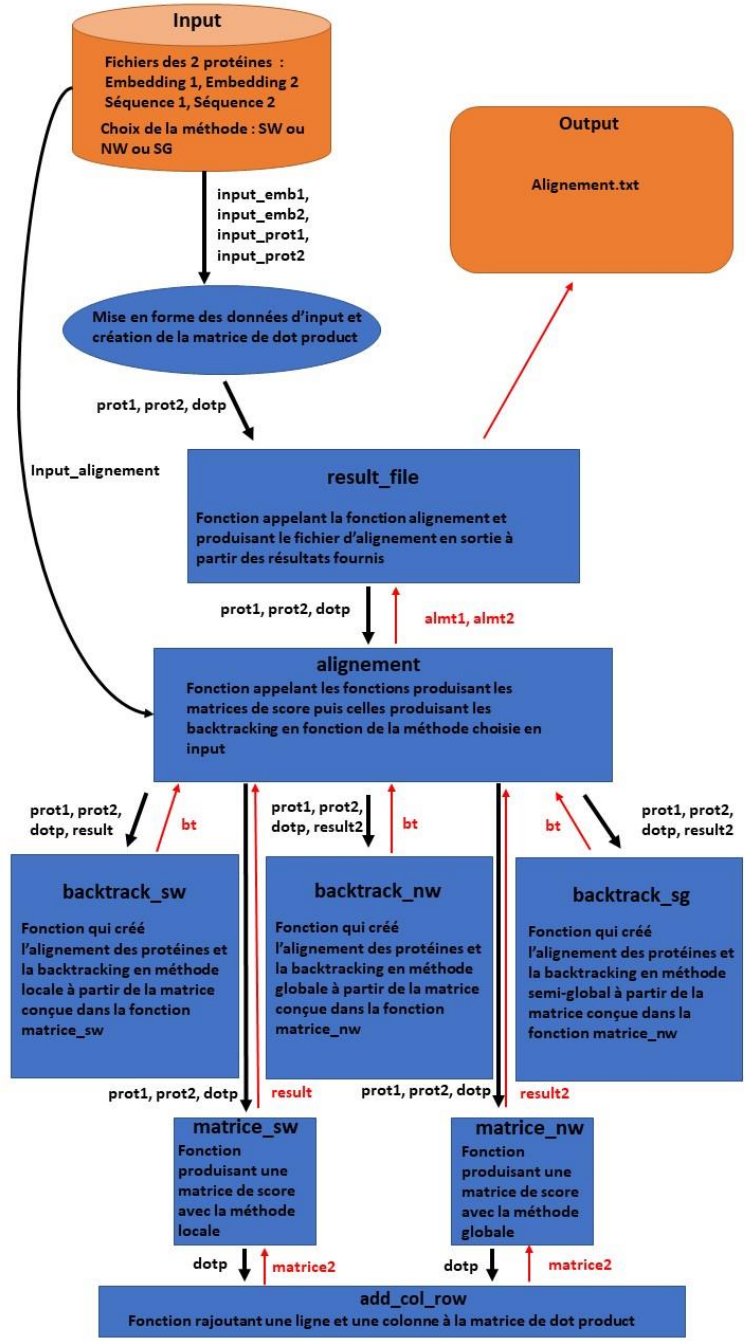
Scoring matrix initialization for global alignment

Backtracking



: Global
: Local
: Semi-global

Program pipeline


Input
Fichiers des 2 protéines : Embedding 1, Embedding 2 Séquence 1, Séquence 2
Choix de la méthode : SW ou NW ou SG
Output
Alignement.txt
input_emb1, input_emb2, input_prot1, input_prot2
Mise en forme des données d'input et création de la matrice de dot product
prot1, prot2, dotp
Input_alignement
result_file
Fonction appelant la fonction alignement et produisant le fichier d'alignement en sortie à partir des résultats fournis
prot1, prot2, dotp
almt1, almt2
alignement
Fonction appelant les fonctions produisant les matrices de score puis celles produisant les backtracking en fonction de la méthode choisie en input
prot1, prot2, dotp, result
bt
prot1, prot2, dotp, result2
bt
bt
prot1, prot2, dotp, result2
backtrack_sw
Fonction qui créé l'alignement des protéines et la backtracking en méthode locale à partir de la matrice conçue dans la fonction matrice_sw
backtrack_nw
Fonction qui créé l'alignement des protéines et la backtracking en méthode globale à partir de la matrice conçue dans la fonction matrice_nw
backtrack_sg
Fonction qui créé l'alignement des protéines et la backtracking en méthode semi-global à partir de la matrice conçue dans la fonction matrice_nw
prot1, prot2, dotp
result
prot1, prot2, dotp
result2
matrice_sw
Fonction produisant une matrice de score avec la méthode locale
matrice_nw
Fonction produisant une matrice de score avec la méthode globale
dotp
matrice2
dotp
matrice2
add_col_row
Fonction rajoutant une ligne et une colonne à la matrice de dot product

Example

- Input

  6PF2K_1bif.t5emb

  SKI_1shka.t5emb

  6PF2K_1BIF.fasta

  SKI_1SHKA.fasta

  SW

- Output

```
IRRMFDETAEDSDRNV-DPSGLKVQVINAAIVEPDVCISEVFFTKYGN-EGFNFIMARRERTTNTADFVAVHGGEESLFKRVDNLL--AALACQKRIKLGEENNDPLFFEFSKYTKVMDRR-QGVNFERTPVGIFNLYRTLKKSIYTKGRAPLGVMVILTPC
APLRMTQMLECVIAAPPPQTADVVYHAVDQYLAERERLVAEMEEAIQLLRLALEEAPAFLYVVTGHARMFQRNQELLVMGGGTAV-VV--PTAVAQLAESERR-FGPWGEAAVVDAVTMGSSTHQMFIDTDVFEYGLARALERGVTTKGCGRAGVMFIPETM
```

Comparison of the 3 alignment methods

Local

```
PDNEEGLKIRKQCALAALNDV-KFLSEEGGHVAVFDATNTTTRERRAMIFNFGEQNGYKTFFVESSICVDPEVIAANIVQVKLGSPDYVNRDSDEATEEDFMRRIECYENSYESLDEEQDRDLSYIKIMDVGQSYVVNRVADHIQSRIVYYLM-IHVTP
MTEPIFMVGARGCGKTTVGREELARALGYEFVDTDIFMQHT-GMTVADVVAAEGWPGFRRRESEA-QAVATPNRVVATGGGMVLLEQNRQFMRAHGTV-YLFAPAEELALRLQIAEEMEAVLREREALYQDVAHYVVDATQPPAAIVCELMQTTMRLPA
```

Global

```
PDNEEGLKIRKQCALAALNDV-KFLSEEGGHVAVFDATNTTTRERRAMIFNFGEQNGYKTFFVESSICVDPEVIAANIVQVKLGSPDYVNRDSDEATEEDFMRRIECYENSYESLDEEQDRDLSYIKIMDVGQSYVVNRVADHIQSRIVYYLM-IHVTPR
MTEPIFMVGARGCGKTTVGREELARALGYEFVDTDIFMQHT-GMTVADVVAAEGWPGFRRRESEA-QAVATPNRVVATGGGMVLLEQNRQFMRAHGTV-YLFAPAEELALRLQIAEEMEAVLREREALYQDVAHYVVDATQPPAAIVCELMQTTMRLPAA
```

Gloloc

```
CPTLIVMVGLPARGKTYISKKLTRYLNFIGVPTREFNVGQYRRDMVKTYKSFEFFPDNEEGLKIRKQCALAALNDV-KFLSEEGGHVAVFDATNTTTRERRAMIFNFGEQNGYKTFFVESSICVDPEVIAANIVQVKLGSPDYVNRDSDEATEEDFMRRIECYENSYESLDEEQDRDLSYIKIMDVGQSYVVNRVADHIQSRIVYYLM-IHVTPR
AEELALRLQIAEEMEAVLREREALYQDVAHYVVDATQPPAAIVCELMQTMRLPAAMTEPIFMVGARGCGKTTVGREELARALGYEFVDTDIFMQHT-GMTVADVVAAEGWPGFRRRESEA-QAVATPNRVVATGGGMVLLEQNRQFMRAHGTV-YLFAPAEELALRLQIAEEMEAVLREREALYQDVAHYVVDATQPPAAIVCELMQTTMRLPAA
```

- Global alignment is too short
- Gap underestimation
- Local and Global alignments are very similar

Comparaison with Blastp alignment

Score	Expect	Method	Identities	Positives	Gaps
25.0 bits(53)	4e-04	Compositional matrix adjust.	23/84(27%)	36/84(42%)	6/84(7%)

```
Query  5   IVMVGLPARGKTYISKKLTRYLNFIGVPTREFNVGQYRRDMVKTYKSFEFFLPDNEEGLK  64
           I MVG    GKT + ++L R L +  V T  F   Q+   M       +    +   G +
Sbjct  5   IFMVGARGCGKTTVGRELARALGYEFVDTDIFM--QHTSGMTVA----DVVAAEGWPGFR  58

Query  65  IRKQCALAALNDVRKFLSEEGGHV  88
            R+  AL A+    + ++  GG V
Sbjct  59  RRESEALQAVATPNRVVATGGGMV  82
```

Local alignment with Blastp

```
PDNEEGLKIRKQCALAALNDV-KFLSEEGGHVAVFDATNTTTRERRAMIFNFGEQNGYKTFFVESSICVDPEVIAANIVQVKLGSPDYVNRDSDEATEEDFMRRIECYENSYESLDEEQDRDLSYIKIMDVGQSYVVNRVADHIQSRIVYYLM-IHVTP
MTEPIFMVGARGCGKTTVGREELARALGYEFVDTDIFMQHT-GMTVADVVAAEGWPGFRRRESEA-QAVATPNRVVATGGGMVLLEQNRQFMRAHGTV-YLFAPAEELALRLQIAEEMEAVLREREALYQDVAHYVVDATQPPAAIVCELMQTTMRLPA
```

Local alignment with the program

Discussion

Program improvement

- Result
- Ergonomic

Using embedding for protein alignment allows more sensitive homology search but not always optimal

Bibliographie

Bepler T, Berger B. Learning the protein language: Evolution, structure, and function. Cell Syst. 2021 Jun 16;12(6):654-669.e3. doi: 10.1016/j.cels.2021.05.017. PMID: 34139171; PMCID: PMC8238390

Smith, Temple F. & Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences" (PDF). Journal of Molecular Biology. 147 (1): 195–197. CiteSeerX 10.1.1.63.2897. doi:10.1016/0022-2836(81)90087-5. PMID 7265238

Needleman, Saul B. & Wunsch, Christian 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): 443–53. doi:10.1016/0022-2836(70)90057-4. PMID 5420325.

Konstantin Schütze, Michael Heinzinger, Martin Steinegger, Burkhard Rost, Nearest neighbor search on embeddings rapidly identifies distant protein relations, bioRxiv. Preprint. 2022.09.04.506527; doi: https://doi.org/10.1101/2022.09.04.506527