Alignement d'une paire de séquences

Alignements globaux (Needleman-Wunsch) versus locaux (Smith-Waterman)

- Alignement global
  - Approprié, par exemple, pour les protéines homologues qui sont conservées sur toute leur longueur.
  - L'alignement final inclut obligatoirement les deux séquences complètes.

```
LQGPSKTGKGS-SRSWDN
|----|-|||---|--|-
LN-ITKAGKGAIMRLGDA
```

  - Algorithme: ***Needleman-Wunsch*** (1970).
  - Outil web EMBOSS : ***needle*** (nucleic acids (nucleic acids or proteins).
- Alignement local
  - Approprié, par exemple, pour les protéines qui partagent un domaine commun, restreint à un segment de chaque séquence.

```
LQGPSSKTGKGS-SSRIWDN
      |-|||
LN-ITKKAGKGAIMRLGDA
```

  - L'alignement final est restreint aux segments conservés.

```
KTGKG
|-|||
KAGKG
```

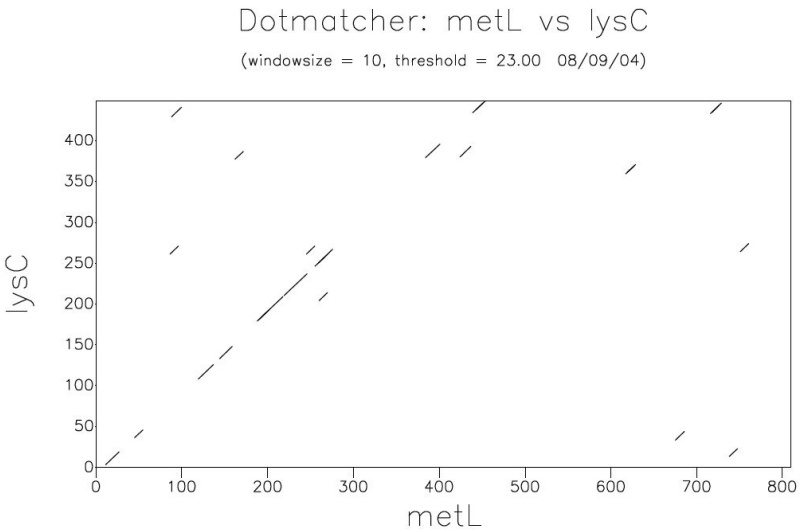
  - Algorithme: ***Smith-Waterman*** (1981).
  - Outil Web EMBOSS : ***water*** (nucleic acids(nucleic acids or proteins)

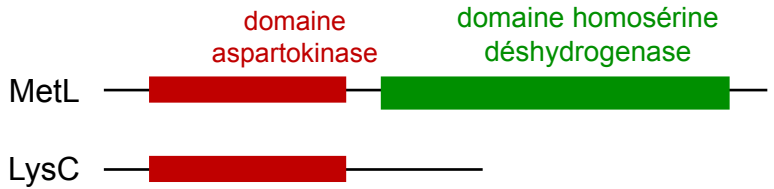
Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443-53.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol 147, 195-7.

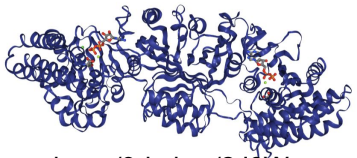
Aspartokinases: dot plot avec matrice de substitution (BLOSUM62)

- Avec le logiciel *dotmatcher*, une matrice de substitution est utilisée pour assigner un score à chaque paire de résidus. Les segments de lignes indiquent des régions de correspondance entre séquences.
- Ceci révèle la similarité entre les domaines aspartokinase de LysC (l'ensemble de la séquence) et de MetL (positions 1 à ~450).
- La région de similarité ne recouvre que la partie N-terminale (gauche) de MetL, car il s'agit d'une enzyme bi-fonctionnelle. La région C-terminale de MedL contient un domaine homosérine déshydrogenase qui est absent de la protéine LysC.
- Sur base de ce dessin, on comprend qu'un alignement local sera plus pertinent qu'un alignement global car il révèlera le domaine que ces protéines ont en commun.





Structure 3D du domaine aspartokinase



http://www.rcsb.org/3d-view/2J0W

Alignement global : exemple (Needleman-Wunsch)

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 867
# Identity:     254/867 (29.3%)
# Similarity:   423/867 (48.8%)
# Gaps:         104/867 (12.0%)
# Score: 929.0

metL              1 MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDM-MVVSA     49
                      .::.||||:|:|:.:.:||||.|:..::::..: .|:||
thrA              1           MRVLKFGGTSVANAERFLRVADILESNARQGQVATVLSA     39

metL             50 AGSTTNQLINWLK----------LSQTDRLSAHQVQQTLRRYQCDLISG     88
                    ....||.|:.:::          :|..:|:.|            :|::|
thrA             40 PAKITNHLVAMIEKTISGQDALPNISDAERIFA----------ELLTG     77

metL             89 LLPAEEADSL--ISAFV-SDLERLAALLDSGIN------DAVYAEVVGHG    129
                    |..|:....|  :..|| .:.::::.:| .||:      |::.|.::.:|
thrA             78 LAAAQPGFPLAQLKTFVDQEFAQIKHVL-HGISLLGQCPDSINAALICRG    126

metL            130 EVWSARLMSAVLNQQGLPAAWLD-AREFLRAERAAQPQVD--EGLSYPLL    176
                    |..|..:|:.||..:|.....:| ..:.|.....:..||  |.......
thrA            127 EKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAA    176

metL            177 QQLLVQHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRV    226
                    .:::...|   .::::.||.:.|..||.|:|||||||||..:.|.......
thrA            177 SRIPADH---MVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCC    223

metL            227 TIWSDVAGVYSADPRKVKDACLLPLLRLDEASELARLAAPVLHARTLQPV    276
                    .||:||.|||:.|||:|.||.||..:...||.||:..|.|||.||:.|:
```

- Alignement des protéines metL et thrA d'E.coli avec l'algorithme de Needleman-Wunsch.
- Barres verticales « | »
  - **Identité**: les deux résidus alignés sont identiques.
- Doubles points « : »
  - **Substitution « conservative »**
  - Les deux résidus alignés sont différents mais ***similaires*** (la paire de résidus a un score positif dans la matrice de substitution utilisée (ici, BLOSUM62). Voir plus loin pour comprendre ces matrices.
- Points « . »
  - **Substitution non-conservative**
  - Cette paire de résidus (distincts) a un score négatif dans la matrice de substitution.
- Espace: « »
  - **Gap**: les résidus d'une des deux séquences ne correspondent à aucun résidu sur l'autre.
  - Le gap peut provenir soit d'une délétion, soit d'une insertion, on parle donc d'***indel,*** pour désigner l' événement évolutif d'où provient ce gap.

Needleman-Wunsch with partial similarities

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 854
# Identity:     136/854 (15.9%)
# Similarity:   209/854 (24.5%)
# Gaps:         449/854 (52.6%)
# Score: 351.0
metL                1 MSVIAQAGAKGRQLHKFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAA     50
                      ||.|.        :.|||||:|:||.....|.|.|:...:.. .::|:||:
lysC                1 MSEIV--------VSKFGGTSVADFDAMNRSADIVLSDANV-RLVVLSAS     41

metL               51 GSTTNQLINWLK-LSQTDRLSAHQVQQTLRRYQCDLISGL----LPAEEA     95
                      ...||.|:...: |...:|.   :....:|..|..::..|    :..||.
lysC               42 AGITNLLVALAEGLEPGERF---EKLDAIRNIQFAILERLRYPNVIREEI     88

metL               96 DSLISAFVSDLERLAALLDSGINDAVYAEVVGHGEVWSARLMSAVLNQQG    145
                      :.|:.. ::.|...|||..|   .|:..|:|.|||:.|..|...:|.::.
lysC               89 ERLLEN-ITVLAEAAALATS---PALTDELVSHGELMSTLLFVEILRERD    134

metL              146 LPAAWLDAREFLRA-ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-GF    193
                      :.|.|.|.|:.:|. :|..:.:.|......|....|:....:.||:| ||
lysC              135 VQAQWFDVRKVMRTNDRFGRAEPDIAALAELAALQLLPRLNEGLVITQGF    184

metL              194 ISRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV    243
                      |...|.|.|..|||.||||:|..:.....|||.||:||.|:|:.|||.|
lysC              185 IGSENKGRTTTLGRGGSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVV    234

metL              244 KDACLLPLLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ    293
```

- Alignment of *E.coli* lysC and metL proteins with Needleman-Wunsch algorithm.
- metL contains two domains: aspartokinase and homoserine dehydrogenase.
- LysC only contains the aspartokinase domains.
- With Smith-Waterman, the %similarity is calculated over the whole length of the alignment (854aa), which gives 24.5%.
- Actually, most of the alignment length is in the terminal gap (the homoserine dehydrogenase domain of metL).
- This percentage is lower than the usual threshold for considering two proteins as homolog.

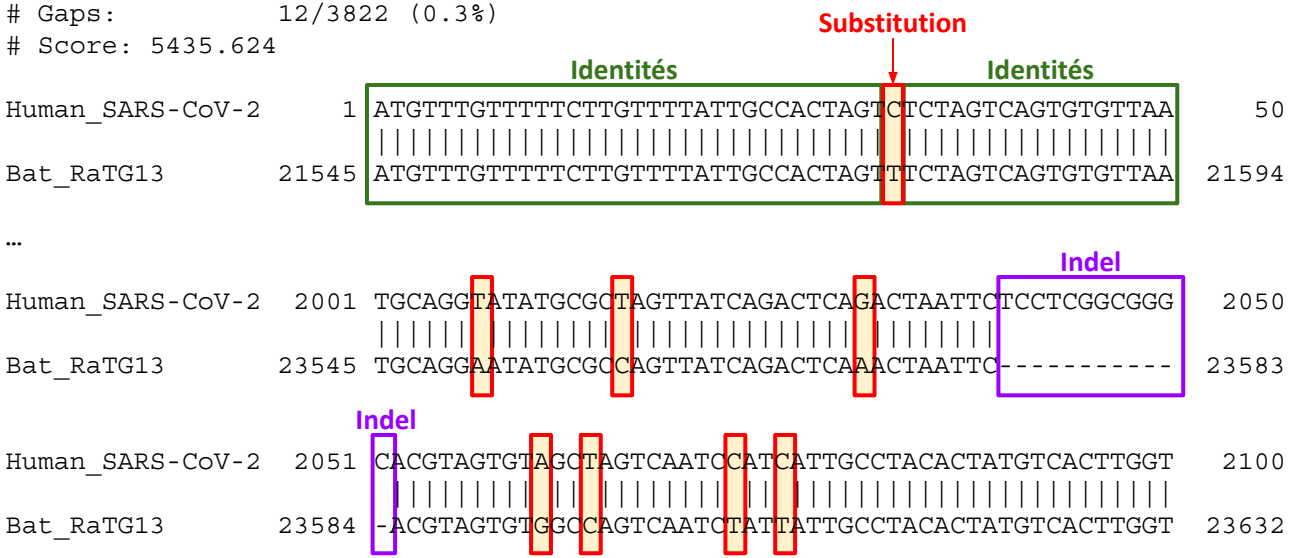
Smith-Waterman with partial similarities

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 482
# Identity:     133/482 (27.6%)
# Similarity:   205/482 (42.5%)
# Gaps:          85/482 (17.6%)
# Score: 353.5
metL              16 KFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINWLK-LS     64
                     ||||:|:||.....|.|.|:...:. .::|:||:...||.|:...: |.
lysC               8 KFGGTSVADFDAMNRSADIVLSDANV-RLVVLSASAGITNLLVALAEGLE     56

metL              65 QTDRLSAHQVQQTLRRYQCDLISGL----LPAEEADSLISAFVSDLERLA    110
                     ..:|.   :....:|..|..:.:..|    :..||.:.|:.. ::.|...|
lysC              57 PGERF---EKLDAIRNIQFAILERLRYPNVIREEIERLLEN-ITVLAEAA    102

metL             111 ALLDSGINDAVYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRA-    159
                     ||..|   .|:..|:|.|||:.|..|...:|.:::.:.|.|.|.|:.:|.
lysC             103 ALATS---PALTDELVSHGELMSTLLFVEILRERDVQAQWFDVRKVMRTN    149

metL             160 ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-GFISRNNAGETVLLGRN    208
                     :|..:.:.|......|....|:....:.||:| |||...|.|.|..|||.
lysC             150 DRFGRAEPDIAALAELAALQLLPRLNEGLVITQGFIGSENKGRTTTLGRG    199

metL             209 GSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLLPLLRLDEAS    258
                     ||||:|..:.....|||.||:||.|:|:.|||.|..|..:.:...||:
lysC             200 GSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVVSAAKRIDEIAFAEAA    249

metL             259 ELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRI---------E    299
```

- Alignment of *E.coli* lysC and metL proteins with Smith-Waterman algorithm.
- The alignment is almost identical to the one reported by Needleman-Wunsch, but the score is now considered on the aligned segments only (482 aa).
- On this region, there is 42.5% of similarity.

Alignement de séquences – Gènes S de SARS-CoV-2 et RaTG13

```
# Aligned_sequences: 2
# 1: Human_SARS-CoV-2_BetaCoV/Wuhan/IPBCAMS-WH-01/2019
# 2: Bat_RaTG13
#
# Length: 3822
# Identity:    3549/3822 (92.9%)
# Similarity:    NA/3822 (NA%)
# Gaps:           12/3822 (0.3%)
# Score: 5435.624
```



Note

- "**Indel**" signifie "Insertion ou délétion"
- Sur base de ce résultat, la différence observée peut provenir soit d'une insertion chez un ancêtre de SARS-CoV-2, soit d'une délétion chez un ancêtre de RaTG13

Des insertions bizarres?

- Figure from Pradhan et al (2020), initially published on bioRxiv and retracted.
- The “multiple alignment” is actually a pairwise alignment + a consensus.
- The gaps obtained from a multiple alignment overlap with these ones, but they start and end at different positions.
- It is precisely because they did not do a multiple alignment that they did not realize that 3 of these insertions were not unique to SARS-CoV-2.

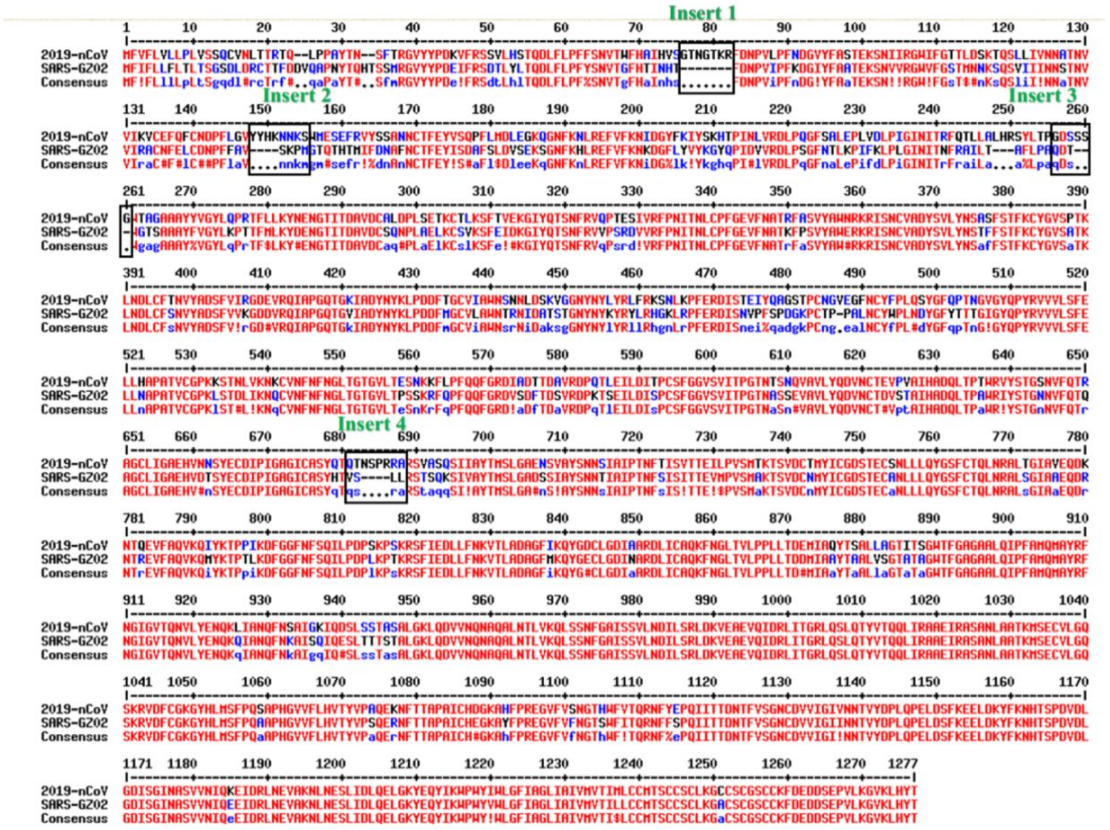


Figure 2: Multiple sequence alignment between spike proteins of 2019-nCoV and SARS. The sequences of spike proteins of 2019-nCoV (Wuhan-HU-1, Accession NC_045512) and of SARS CoV (GZ02, Accession AY390556) were aligned using MultiAlin software. The sites of difference are highlighted in boxes.