

dataset_1: sequências maiores

a)

seq_1: Bacillus anthracis
seq_2: Clostridium tetani 12124569
seq_3: Escherichia coli
seq_4: Mycobacterium tuberculosis
seq_5: Pseudomonas aeruginosa
seq_6: Salmonella enterica
seq_7: Staphylococcus aureus
seq_8: Streptococcus pneumoniae
seq_9: Treponema pallidum subsp. pertenue
seq_10: Vibrio cholerae

Para a questão b), usou-se um tamanho de 10.000 nucleotídeos a partir do início.

b)

seq_1 vs seq_2: -1303

```
O score do alinhamento é: -1303.0
A identidade é: 38.99972916854744%
```

seq_1 vs seq_3: -1580

```
O score do alinhamento é: -1580.0
A identidade é: 37.94287780187997%
```

seq_1 vs seq_4: -1833

```
O score do alinhamento é: -1833.0
A identidade é: 37.47744496571635%
```

seq_1 vs seq_5: -1700

```
O score do alinhamento é: -1700.0
A identidade é: 37.932910244786946%
```

seq_1 vs seq_6: -1712

```
O score do alinhamento é: -1712.0
A identidade é: 37.48865080806247%
```

seq_1 vs seq_7: 1254

```
O score do alinhamento é: 1254.0
A identidade é: 53.442622950819676%
```

seq_1 vs seq_8: -732

```
O score do alinhamento é: -732.0
A identidade é: 42.11618257261411%
```

seq_1 vs seq_9: -1391

```
O score do alinhamento é: -1391.0  
A identidade é: 38.56020233041279%
```

seq_1 vs seq_10: -1470

```
O score do alinhamento é: -1470.0  
A identidade é: 38.62218605912666%
```

seq_2 vs seq_3: -1733

```
O score do alinhamento é: -1733.0  
A identidade é: 37.254723837209305%
```

seq_2 vs seq_4: -2139

```
O score do alinhamento é: -2139.0  
A identidade é: 35.60399636693914%
```

seq_2 vs seq_5: -2125

```
O score do alinhamento é: -2125.0  
A identidade é: 35.852484331001904%
```

seq_2 vs seq_6: -1913

```
O score do alinhamento é: -1913.0  
A identidade é: 36.440984793627806%
```

seq_2 vs seq_7: -1214

```
O score do alinhamento é: -1214.0
```

seq_2 vs seq_8: -1377

```
O score do alinhamento é: -1377.0
```

seq_2 vs seq_9: -1534

```
O score do alinhamento é: -1534.0
```

seq_2 vs seq_10: -1748

```
O score do alinhamento é: -1748.0
```

seq_3 vs seq_4: -1464

```
O score do alinhamento é: -1464.0  
A identidade é: 38.42200180342651%
```

seq_3 vs seq_5: -1399

```
O score do alinhamento é: -1399.0  
A identidade é: 38.556831539922236%
```

seq_3 vs seq_6: -1374

```
O score do alinhamento é: -1374.0
```

seq_3 vs seq_7: -1607

```
O score do alinhamento é: -1607.0
```

seq_3 vs seq_8: -1636

```
O score do alinhamento é: -1636.0
```

seq_3 vs seq_9: -1550

```
O score do alinhamento é: -1550.0
```

seq_3 vs seq_10: -1444

```
O score do alinhamento é: -1444.0
```

seq_4 vs seq_5: -950

```
O score do alinhamento é: -950.0
```

seq_4 vs seq_6: -1247

```
O score do alinhamento é: -1247.0
```

seq_4 vs seq_7: -1542

```
O score do alinhamento é: -1542.0
```

seq_4 vs seq_8: -1929

```
O score do alinhamento é: -1929.0
```

seq_4 vs seq_9: -1589

```
O score do alinhamento é: -1589.0
```

seq_4 vs seq_10: -1475

```
O score do alinhamento é: -1475.0
```

seq_5 vs seq_6: -1234

```
O score do alinhamento é: -1234.0
```

seq_5 vs seq_7: -1987

O score do alinhamento é: -1987.0

seq_5 vs seq_8: -2153

O score do alinhamento é: -2153.0

seq_5 vs seq_9: -1636

O score do alinhamento é: -1636.0

seq_5 vs seq_10: -1497

O score do alinhamento é: -1497.0

seq_6 vs seq_7: -1733

O score do alinhamento é: -1733.0

seq_6 vs seq_8: -1811

O score do alinhamento é: -1811.0

seq_6 vs seq_9: -1502

O score do alinhamento é: -1502.0

seq_6 vs seq_10: -1433

O score do alinhamento é: -1433.0

seq_7 vs seq_8: -948

O score do alinhamento é: -948.0

seq_7 vs seq_9: -1397

O score do alinhamento é: -1397.0

seq_7 vs seq_10: -1483

O score do alinhamento é: -1483.0

seq_8 vs seq_9: -1411

O score do alinhamento é: -1411.0

seq_8 vs seq_10: -1537

O score do alinhamento é: -1537.0

seq_9 vs seq_10: -1504

O score do alinhamento é: -1504.0

dataset_2: sequências menores

a)

seq_1: Streptococcus agalactiae
seq_2: Neisseria gonorrhoeae
seq_3: Mycobacterium tuberculosis variant bovis
seq_4: Staphylococcus aureus
seq_5: Treponema pallidum
seq_6: Bacillus anthracis
seq_7: Yersinia pestis
seq_8: Acinetobacter baumannii ATCC 17978
seq_9: Pseudomonas aeruginosa
seq_10: Helicobacter pylori

b)

seq_1 vs seq_2: -26

AAACACC-T--CCAG---TCAT AAT ATT CGT AAACCAAT CAAAAACTCATGTTTAAATCAATAAAAAATAC-TTAG-
CCCTGCGCGATTTCGGAGCTAGACCGTGCGTAATATAAA----GCCGGCCCGCGA---TGTATTT-GCCGTGGC

O score do alinhamento é: -26.0
A identidade é: 27.848101265822784%

seq_1 vs seq_3: -15

AAACA-CCTCC-A-G-TCAT AAT -ATT CGT AAACCAAT CAAAAACTCATG --TTTTAAAT-CAATAAAAAATAC-TTAG-
TCACAC-GCCCCACGACGGAGCTGGGGCACCCAGC-ATTCACT --GCTTA-CC- ACTACGATCTCGCTCAAG

O score do alinhamento é: -15.0
A identidade é: 34.61538461538461%

seq_1 vs seq_4: -17

AAACACCTCCA-GTCAT AAT A-TTCGTAAACCAAT CAAAAACTCATG --TTTTAAAT-CAATAAAAAATAC-TAG-
TTATTTT-GAAAGTAAATGGCAT-CATTATTA--TTAAATGGTA-TAGGTTCTGGTACTGGTAT --GCTTC

O score do alinhamento é: -17.0
A identidade é: 36.36363636363637%

seq_1 vs seq_5: -17

AAACACCTCCA-GTCAT AAT ATT CGT AAACCAAT CAAAAACTCATG --TTTTAAAT-CAATAAAAAATAC-TAG-
TGATATAC-TTC-AAGC-AATCAC --GCAGCCCACCACTTTT-CCCGGGAGCAAGACATCTTCCA-GGAAAC

O score do alinhamento é: -17.0
A identidade é: 35.064935064935064%

seq_1 vs seq_6: -13

AAACACCTCCA-GTCAT AAT ATT CGT AAACCAAT CAAAAACTCATG --TTTTAAAT-CAATAAAAAATAC-TAG-
CTTGC-ATAA-GTA-CAACACCGCAT AAT AAT ACCCGA-CATAACTAATTCCTCATGGGAGTTTTATG

O score do alinhamento é: -13.0
A identidade é: 36.486486486486484%

seq_1 vs seq_7: -25

AAACACCTCCA-GTCAT AAT ATT CGT AAACCAAT CAAAAACTCATG --TTTTAAAT-CAATAAAAAATAC-TAG-
GGTTGACAGGAATA-CAATACTGCCGC-AACAGTGTGAAGCTGCCGCCCTTTGGCGTCAAA---GC-A--

O score do alinhamento é: -25.0
A identidade é: 32.91139240506329%

seq_1 vs seq_8: -20

AAACACCTCCA-GTCAT AAT ATT CGT AAACCAAT CAAAAACTCATG --TTTTAAAT-CAATAAAAAATAC-TAG-
AGAGACTAATATGTTTTTCAACTTTTCA-T-CCTGAAATATTTATCATTAAA-GGGT

O score do alinhamento é: -20.0
A identidade é: 32.432432432432435%

seq_1 vs seq_9: -29

```
AAACA--CCT-CCAGTCATAATATT CGTAAACCAATC - AAAAACTCATGTTTAATCAATAAAAATACTTAG -  
TCGCCGAAGTCAGCGGT --CGCCCTGCC-AGCGCCA---CGGCCGTACGCC-TGGAACGCCGTGCCAGCC
```

O score do alinhamento é: -29.0
A identidade é: 28.57142857142857%

seq_1 vs seq_10: -11

```
AAACACCTCCAGTCATAAT-ATT CGT-AAAACAATCAAAACTCATGTTT -AAA-TCAATAAAAAATACTTA-G--  
CACTGAATCAAT-CCTTCTTAAC-TTTA-GGA--TCACTTATTA-TGGGGCTAGGATCAATAAGGCTAT-CAAGCA
```

O score do alinhamento é: -11.0
A identidade é: 38.961038961038966%

seq_2 vs seq_3: -19

```
CCCC-TG-CGGATTTC-GGAGT-CAGACCGTGC GT AAT AT AAAACGCCGCC-GCCGATGTATTTGCCGTGCC  
TCACCAAGCCCCACGACGGAGCTG-TGGGCACCCAGCATTCACTGCTTACACTACGATCTCGCTACG---
```

O score do alinhamento é: -19.0
A identidade é: 36.0%

seq_2 vs seq_4: -27

```
CCCCCTGCCGA-TTTTCGGAGTCAGACCGTGC GT AAT AT AAAACCG-CCGG-CCCGCCGATGTATTTGCCGTGCC  
TTATTTG-AAAAT-AACATGGCATCATTATTAAA---TGGTATAGGTTCTGTTACTGGTATGCTCA
```

O score do alinhamento é: -27.0
A identidade é: 28.0000000000004%

seq_2 vs seq_5: -23

```
CCCCCTGCCGA-TTTTCGGAGT-CAGACCGTG-CGTATAA-TAAAACGCCGCCGCCGATGTATTTGCCGTGCC  
TGATATAC-TT-CAAGCACATCA--C-GCAGCCCACCCACTTTCCC-G-CGGAC-GAACACATCTCCCAGGGAAAC
```

O score do alinhamento é: -23.0
A identidade é: 31.645569620253166%

seq_2 vs seq_6: -24

```
CCCCCTGCCGA-TTTTCGGAGT-GGAGTCAGACCGTGC GT AAT AT AAAA- -CGCCGGCCGCCGATGTATTTGCCGTGCC  
CTT--GCATAAAGTACAACACCGCAT-AAAATATACCGACATAACTAATTCTTCATTGGGA-TTTTTATG---
```

O score do alinhamento é: -24.0
A identidade é: 34.61538461538461%

seq_2 vs seq_7: -15

```
CCCCCTGCCG---ATTTCGGAGT-CAGACCGTGC GT AAT AT AAAACCG-CCGGCCGCCGATGTATTTGCCGTGCC  
GGTT-GACAGCGAATAACATCG-CC-GCCAACAGTGTGAAGCTGCCGCC-G---TTT-CTTGGCGTCAAAGCA
```

O score do alinhamento é: -15.0
A identidade é: 38.46153846153847%

seq_2 vs seq_8: -20

```
CCCCCTGCCGA-TTTTCGGAGT--CAGACCGTGC GT AAT AT AAAACGCCGCCGCCGATGTATTTGCCGTGCC  
AGAGAC--TAATAT-G-TTTTTCACACTTTCT-GCTTTTTACATCC-TG-AAATATA-TTATCATTTAAAGGGGT
```

O score do alinhamento é: -20.0
A identidade é: 30.76923076923077%

seq_2 vs seq_9: -10

```
CCCCCTGCCGA-TTTTCGGAGT-AQ-TCAGACCGTGC GT AAT AT AAAACGCCGCCGCCGATGTATTTGCCGTGCC  
TGGCC---GAAGCTCCAGGCCGTG-CCCT-GCCCAAGGCCACC-GCCCGTACGCCCT-GGAACGCCGTGCCAGCC
```

O score do alinhamento é: -10.0
A identidade é: 37.66233766233766%

seq_2 vs seq_10: -23

```
CCCCCTGCCGA-TTTTCGGAGT-CAGACCGTGC GT -AAT -AT AAAACGCCGCCGCCGATGTATTTGCCGTGCC  
CACT-GAATCAATCCTTCACTTACTTAGGATCACTTA-TTATGGGGCTAGGATC-AATAAGGCTATC--AAG-CA
```

O score do alinhamento é: -23.0
A identidade é: 30.263157894736842%

seq_3 vs seq_4: -22

```
TCACCAACGCCGCCACGA-CGGA-GCT-CGTTGGGCCACCGATTCACTGCTTACCA-CTACGATCTCGCTACG  
TTATTTGAAAGTAACATGGCATCATTATTAAATGGTATAGGTTCTGTTACTGGTAT-G-CTTC-A
```

O score do alinhamento é: -22.0
A identidade é: 31.08108108108108%

seq_3 vs seq_5: -11

TCACCACGCCCGACGGAGCTCGTGGGCA-CCCAGATT-CACTGCTTACCACTACGATCTCGC-TG--ACG-TGATA-TACTTCAAGCACA--TCACGCAGCCAAACAC-TTTTCCCG-CGGACGAAGACA-TCTTCCCAGGCAAAC

O score do alinhamento é: -11.0
A identidade é: 39.473684210526315%

seq_3 vs seq_6: -16

T-CACCA--CGCCCCCA-GCAC-GGAGCTGTGGGACCCAGCATTCACTGCTTACCACTACG--ATCTCGCTCAG-CTTGC-ATAAGTAC--AACAA-CGCATAAAATAA-CCCGACATAAA-CTAATTCCTCATGGAGTTTTA-TG

O score do alinhamento é: -16.0
A identidade é: 35.064935064935064%

seq_3 vs seq_7: -14

TCACACG-CCCCCACGAGCTCGTGGGACCCAGCATTAC-TCTTACCACTACGATCTCGCTCA-C-GGGTTGACAGGAATACAATA-C-TCGCAGCCAAAC-A-TGTGAAGCTGCCGCCGTTCTGGGT-CAAAGCA

O score do alinhamento é: -14.0
A identidade é: 37.333333333333336%

seq_3 vs seq_8: -32

TCACCA-CGCCCCACGAGCTCGTGGGACCCAGC-ATTCACGTCTCGCTTACCACTACGATCTCGCTCAG-AGAGACTAATATGTTTACAC-TTTCTGCTTTTACATCC-TGAAATATA-TTATCATTAAAGGGT

O score do alinhamento é: -32.0
A identidade é: 24.324324324324326%

seq_3 vs seq_9: -6

TCACCA-CGCCCCACGAGCTCGTGGGACCCAGC-AGCATTCACTGCTTACCACTACGATCTCGCTCAG-TCGCGGAAGTCC-AGCCCGTGCCTT-GCCAGCCACCGCCCGT-ACCGCCCTGGAAAGGCC-TG-CCCAGCC

O score do alinhamento é: -6.0
A identidade é: 39.473684210526315%

seq_3 vs seq_10: -19

TCACCA-CGCCCCACGAGCTCGTGGGACCCAGCATTAC-TG-CTTACCACTACGATCTCGCTCAG-C-ACTGATCAATCC-TTCTTAACTTAGGATCACTTATTATGGGGTAGGATC-AATAAAAGGCTATCAAGCA

O score do alinhamento é: -19.0
A identidade é: 32.432432432432435%

seq_4 vs seq_5: -23

TTATTTT---GAAAGTAA-C-AATGGCATCATTATTATT-AAATGGTATAGGTTT-TTGGTTACTGGTATGCTTC-TGATATACCAA-GCACATCA-CGC-AGCCCAACCACTTTCCCGGGACGAAGACATCTCCAGGCAAAC---

O score do alinhamento é: -23.0
A identidade é: 33.76623376623377%

seq_4 vs seq_6: -10

TT-ATTTGAAAGTAA-C-AATGGCATCATTATTAAATGGTATAGGTTT-CTTGGTTACTGGTATGCTTC-CA-CTTGAT-AA-G-TACAACACCG-CATAAA-TAATA-CCCGACATAACTAACTCCTT-CATTGGAG-TTTTTATG

O score do alinhamento é: -10.0
A identidade é: 36.708860759493675%

seq_4 vs seq_7: -17

TT--A-TTTGAAAGTAAACATGCCATCA-TTATTATTAAATGGTATAGGTTTCTTGTGTTACTGGTATGCTTC-GGTTGACAGGAATACAATACTGCCGCCAACAGTGTGAAGCTGCCGCCGTTCTT-GCGTCAAAG-CA--

O score do alinhamento é: -17.0
A identidade é: 34.66666666666667%

seq_4 vs seq_8: -23

TTATTTGAAAG-TAACATGGCATCATT-ATTATT---AAA-TGGTA---TAGGTTTCTTGTGTTACTGGTATGCTTC-AAGAGACTAATATGTTTTT-CACAC-TTTCTGCTTTTACATCTGAAATATT-ATCATTAAAGGGT-----

O score do alinhamento é: -23.0
A identidade é: 34.17215189873415%

seq_4 vs seq_9: -32

TTATTTGAAAGTAAACATGCCATCA-TTATTATTAAATGGTATAGGTTTCTTGTGTTACTGGTATGCTTC-A-TCGCGG-AACGTC-AGG--CCGTCGCCCTGCCAGCG-CCACCGCCCGT-ACCGCCCTGGAAAGCCCTGCCAGCC

O score do alinhamento é: -32.0
A identidade é: 26.923076923076923%

seq_4 vs seq_10: -13

TTATTTGAAAGAACATGGCATATTATTAAATGGTATGGTTCTGTGTTA--CTGGT-ATGCTTCACACT-G-AATCAATCCTTAACTTAGATCACTTATTATGGGCTAGGAT-CAATAAGGCTATCAAGCA

O score do alinhamento é: -13.0
A identidade é: 39.189189189189186%

seq_5 vs seq_6: -17

TGATATACCTCAAGCACATCAG-CAGCCCAACCACCTTTCCGC-GGAC-GAAGACATC-TT-CCCAGGCAAAC-C-CTT-GCAT-AA-GTACAACCGCATAAAAA-TAAT---ACCGGACATAACTAATTCTCATGGGAGTTTTATG

O score do alinhamento é: -17.0
A identidade é: 32.05128205128205%

seq_5 vs seq_7: -6

TGATATACCTCAAGCACATCAC-GCAGCCCAACCACCTTTCCGC-GGACGAAGACATCTCCAGGCAAAC-C-GTTGA-CAGCGAATACAATACTGCCGCCAA-CA-GTGTGAAG-CTGCCGCCGTTCTGGCGTCAAAGCA

O score do alinhamento é: -6.0
A identidade é: 43.24324324324324%

seq_5 vs seq_8: -18

TGATAT-ACCTCAAGCACATCAGCAGCCCAACCACCTTTCCGC-GGACGAAGACATCTTCCA-GGCAAACAGAGACTAATA---TGTTTTTACACTTCTGCTTTTAC-ATCTGAAATATTATCATTAAAGGGT--

O score do alinhamento é: -18.0
A identidade é: 36.0%

seq_5 vs seq_9: -4

TGATATA-CTTCAAGCACATCAC-GCAGCCCAACCACCTTTCCGC--GGACGAAGACATC-TTCCCAGGCAAAC-TGCCGAACGTCAGGCCGTGCCAGGCCACCGGCCGTAACGCCCTGAAACGCCCTGCCAGCC---

O score do alinhamento é: -4.0
A identidade é: 47.2972972972973%

seq_5 vs seq_10: -16

TGATATACCTCAAGCACATCA---CGCA-GCCCAACCACCTTTCCGC-GGACGAAGACATC-AT-CTTCCCAGGCAAAC-CACTGAAT-C-AATCC-TTCTTAACCTTAGA-TC-ACTTATTATGGGC-TAGGATCAATAAGGCTATCAAGCA

O score do alinhamento é: -16.0
A identidade é: 35.064935064935064%

seq_6 vs seq_7: -18

CTT-GC-ATAA-GTACAACACC-GCATAAAATAACCGACATAACTAAATTCTT---CATTGGGAGTTTTATG-GGTTGACAGCGAATACAATACTGCCGCCACAGTGTG-AA-GC-TGCCGCCGTTCTGGCG-TCAAAGCA-

O score do alinhamento é: -18.0
A identidade é: 33.7662376623377%

seq_6 vs seq_8: -26

CTTGCATAACTAACACCGCATAAAAATAACCGACATAACTAAATTCTTATTGGGAGTTTTATG---GGGAGTTTTA-TG---AGAGACTAAT-ATGTTTTTACACTTCTGCTTTT-AC-A-T-CCTGAAATATTATCATT-AAAGGGT

O score do alinhamento é: -26.0
A identidade é: 31.16883116883117%

seq_6 vs seq_9: -24

CTTGCATAA-G-TACAAACCCGATAAAATAACCGACATAACTAAATTCTTATTGGGAGTTTTATG--T-TGCCGAACGTCAGGCC-GTCCGCCAGCG-CCACCGGCCGTAACGCCCTGG-AACGCCCTGCCAGCC

O score do alinhamento é: -24.0
A identidade é: 31.08108108108108%

seq_6 vs seq_10: -23

CTT-GCATAGTAC---AA---CACCG-CATAAAAATAACCGACATAACTAAATTCTTATTGGGAGTTTTATG---CACTGAATCAATCTCTTAACCTTAGGATCACTTATTATGGGCT-AGGA-T-CAAT-AAAGGCTAT----CAAGCA

O score do alinhamento é: -23.0
A identidade é: 30.37974683544304%

seq_7 vs seq_8: -26

GGTTGACAGCGAATACAATACTC-GCCGCCAAC-AGTGT-GAAGCTGCCGCCGTTCTGGCGT-CAAAQCA-AGAG-ACTAA---TATGTTTTTACACTTCTGCTTTACATCTGAAATATTATC-ATTAAAGGGT

O score do alinhamento é: -26.0
A identidade é: 30.263157894736842%

seq_7 vs seq_9: -15

```
GG-TTGACAGGGATACAATACT-CGCCG--CCAACAGTGTGAAG-CTG-CC--GCCCG-TTTCTTT-GGCCTAAAGCA  
TCGCGAA-C-GT--CCAGGCGTCGCCCTGCCA--GCCAACGCCCGTACCGCCTGGAACGCCG-CCC-A--GCC
```

O score do alinhamento é: -15.0
A identidade é: 33.75%

seq_7 vs seq_10: -15

```
GGTTGACAGCGAATACAATACT-CGCCGCCAA--C-AGTGTGAAGCTGCCGCCGTTCTTTGG---CGT-CAAAGCA  
CACTGAA-TCAA-TCTT-C-TAAC-TTAGGACTT-A-TTATGGGCTAGGATCAATAAGGCTATCAAG-CA
```

O score do alinhamento é: -15.0
A identidade é: 33.333333333333%

seq_8 vs seq_9: -34

```
AG-AG-AC-TAATATGTTTTACACTTTCTGCTTTTACAT--CCTGAAATATTATCATTTAAAGGGT  
TCGCCGAAACGTCCAGGCGT--CGCCCTGCCAGGCCACGCCCGTACCGCCTGGAACGCCGCCA-G--CC
```

O score do alinhamento é: -34.0
A identidade é: 25.333333333336%

seq_8 vs seq_10: -14

```
A-GAGACTAATATGTTTTACACTTTCTGCTTTTACA-TCCGAAATA-TA--TTA-TCACTT-AAAGGGT  
CACTGAATCAA-TCTTCTAAC--TTTAGG-ATCACTTATTATGGGCTAGGATCAAT-AAAGGCTATCAAG--CA
```

O score do alinhamento é: -14.0
A identidade é: 36.36363636363637%

seq_9 vs seq_10: -21

```
TGCGCGAACGTCCAGGCCGTGCC-TGCCCA-GGCCA-CCG---CCCGTACGCCCTGGAACG-CCTGCCAGGCC  
C-ACTGAAT--CAATCC-TCTTAACCTTAGGACTCTTATTATGGGCT-AGGATC-AATAAGGC-TATCAAGCA
```

O score do alinhamento é: -21.0
A identidade é: 31.16883116883117%

c) Em anexo, .csv dentro de cada pasta especificada por dataset.

d) programa_UPGMA.py e árvores derivadas da execução em anexo em cada pasta especificada por dataset.

Observação: o alinhamento foi realizado com os pesos:

Match = 1

Mismatch = -1

Gap = -2

É possível observar que, para o dataset_1, nem todas as imagens identificam a identidade e parte da sequência. Isso ocorre porque antes eu estava utilizando o código mais completo de alinhamento que fiz para o Trabalho I, mas, visando otimizar a velocidade de execução, encurrei o código para relatar apenas o score. O alinhador.py anexado é esta versão simplificada.

Também, é possível verificar que existem dois programas UPGMA, sendo que a diferença reside na função min/max_valor. Isto ocorre porque, como estamos falando de score, quanto maior, mais próximo. Porém, por receio por conta dos métodos geralmente buscarem um menor valor, fiz desta forma. Acredito que o correto seja a busca pelo maior valor, mas, por todavia, fiz das duas formas.

DESCRIÇÃO DO CÓDIGO

Bibliotecas utilizadas:

```
import csv
from dataclasses import replace
import numpy as np
```

Função min_matriz: encontra o menor valor dentro de uma matriz e retorna as coordenadas em que este valor se encontra dentro da matriz.

Existe uma função análoga que procura o maior valor dentro da matriz.

```
5  # Menor valor: localiza, dentro de uma matriz, o menor valor
6  def min_matriz(matriz):
7      # Inicializa o menor valor como infinito (um numero muito grande)
8      min_val = float("inf")
9      x, y = -1, -1
10
11     # Procura o menor valor
12     for i in range(len(matriz)):
13         for j in range(len(matriz[i])):
14             if matriz[i][j] < min_val:
15                 min_val = matriz[i][j]
16                 x, y = i, j
17
18     # Retorna as coordenadas do menor valor dentro da matriz
19     return x, y
20
```

Função Junta_Seq: junta o nome das colunas e salva no local em que estava a coluna “a”, dando origem a um modelo NEWICK de árvore. Ainda, remove a identificação da coluna “b”.

```
21  # Junta a identificacao das colunas das sequencias mais proximas
22  def Junta_Seq(nomes, a, b):
23      # Organiza a ordem das sequencias pela identificacao
24      if b < a:
25          a, b = b, a
26
27      # Junta as colunas na primeira coluna
28      nomes[a] = "(" + nomes[a] + "," + nomes[b] + ")"
29
30      # Remove a coluna extra
31      del nomes[b]
```

Função Junta_coluna: recalcula os valores da nova coluna “(a, b)”, atribuindo à ex-coluna “a”. Ainda, deleta a coluna “b”.

```

52
33 # Junta colunas com as sequencias mais proximas
34 def Junta_coluna(matriz, a, b):
35     # Organiza/ordena pelo indice
36     if b < a:
37         a, b = b, a
38
39     # Recalcula e salva numa lista os valores da nova coluna
40     linha = []
41     for i in range(0, a):
42         linha.append((matriz[a][i] + matriz[b][i])/2)
43     matriz[a] = linha
44     for i in range(a+1, b):
45         matriz[i][a] = (matriz[i][a] + matriz[b][i])/2
46     for i in range(b+1, len(matriz)):
47         matriz[i][a] = (matriz[i][a] + matriz[i][b])/2
48     # Remove a coluna redundante
49     del matriz[i][b]
50
51     # Remove a linha redundante
52     del matriz[b]
53

```

Função ini_nomes: gera um vetor com as identificações das colunas a partir de um range que pode ser alterado dentro da main.

```

54 # Inicializa os nomes de identificação das sequências
55 def ini_nomes(inicio, fim):
56     nomes = []
57     for i in range(ord(inicio), ord(fim)+1):
58         nomes.append(chr(i))
59     return nomes
60

```

Função UPGMA: integra todas as demais funções num sistema cíclico até que o vetor gerado pela ini_nomes tenha tamanho igual a 1. Retorna uma string de uma árvore no modelo NEWICK.

```

61 # Compilação das funções
62 def UPGMA(matriz_dist, nomes):
63     while len(nomes) > 1:
64         x, y = min_matriz(matriz_dist)
65         Junta_coluna(matriz_dist, x, y)
66         Junta_Seq(nomes, x, y)
67
68     # Retorna o Newick
69     return nomes[0]
70

```

Main

Passo uma matriz

Inicilizo o vetor com os nomes das colunas (identificação das seq)

Executo a UPGMA e salvo numa variável

Com o uso das funções do Biopython, desenho a árvore de forma gráfica.

```
71 # CÓDIGO
72
73 matriz_dist = [
74     [],
75     [-1303],
76     [-1580, -1733],
77     [-1833, -2139, -1464],
78     [-1700, -2125, -1399, -950],
79     [-1712, -1913, -1374, -1247, -1234],
80     [1254, -1214, -1607, -1542, -1987, -1733],
81     [-732, -1377, -1636, -1929, -2153, -1811, -948],
82     [-1391, -1534, -1550, -1589, -1636, -1502, -1397, -1411],
83     [-1470, -1748, -1444, -1475, -1497, -1433, -1483, -1537, -1504]
84 ] # Obtido através do alinhamento no programa do Tabalho I
85
86 NOMES = ini_nomes("A", "J") # De A até H
87 print(NOMES)
88 newwick_final = UPGMA(matriz_dist, NOMES)
89
90 from Bio import Phylo
91 from Bio.Phylo.PhyloXML import Phylogeny
92 from io import StringIO
93
94 handle = StringIO(newwick_final)
95 tree = Phylo.read(handle, "newick")
96 Phylo.draw(tree)
```