03a-Batch-Alignment

March 26, 2025

1 Test 03: Performing Batch Alignment

1.0.1 Overview

This notebook demonstrates batch alignment of a series of sets of sequences from BALIS-2 using MAli v1.31.

The same set of inputs included with Test 01 are being used in this test.

Expected runtime: ~60 seconds or less

1.0.2 Context

This notebook is intended to test the following requirements of MAli:

Requirement 3.4 - Supports multiple bioinformatics file formats for outputting alignments. - The alignments produced in this notebook are outputted in the ClustalW file format rather than FASTA (used previously in tests 01 & 02)

Requirement 4.3 - Can specify a randomness seed to support reproduction of results under the same settings. - A randomness seed & number of iterations is specified for exact reproduction of results

Requirement 6.1 - Supports batch alignment of a series of sets of sequences from a directory. - This notebook demonstrates the batch alignment of a series of 6 sets of 6 sequences.

Requirement 6.2 - Interface displays progress on the current alignment task - in terms of time or iterations. - In this directory, a console command can be run to view progress on individual alignments within a batch alignment task. - Run CLI: `MAli-v1.31\MAli.exe -input data\input -output data\output -batch -iterations 100 -seed 25032025 -format clustalw -debug`

1.0.3 Installing Prerequisites

[1]:
```
!pip install biopython
```

```
Requirement already satisfied: biopython in
c:\users\pdmoo\appdata\local\programs\python\python310\lib\site-packages (1.85)
Requirement already satisfied: numpy in
c:\users\pdmoo\appdata\local\programs\python\python310\lib\site-packages (from
biopython) (1.26.2)
```

Imports

[2]:
```
import os
import shutil
import subprocess
import time
from presentation_helper import PresentationHelper
import random
```

[3]:
```
presenter = PresentationHelper()
```

Reproducibility Parameters

[4]:
```
SEED_VALUE = 25032025
NUM_ITERATIONS = 100
```

MAli v1.31

[5]:
```
ALIGNER_NAME = "MAli-v1.31"
ALIGNER_PATH = "MAli-v1.31/MAli.exe"
INPUT_FOLDER = "data/input"
OUTPUT_FOLDER = "data/output"
```

[6]:
```
# creating empty output folder
if os.path.exists(OUTPUT_FOLDER):
    shutil.rmtree(OUTPUT_FOLDER)
os.makedirs(OUTPUT_FOLDER)
```

Performing Batch Alignment

[7]:
```
SECONDS_OF_COMPUTATION_PER_TESTCASE = 2
ALIGNMENT_COMMAND = f"{ALIGNER_PATH} -input {INPUT_FOLDER} -output␣
 ↪{OUTPUT_FOLDER} -batch -iterations {NUM_ITERATIONS} -seed {SEED_VALUE}␣
 ↪-format clustalw"
print(f"CLI command to be run: '{ALIGNMENT_COMMAND}'")
```

```
CLI command to be run: 'MAli-v1.31/MAli.exe -input data/input -output
data/output -batch -iterations 100 -seed 25032025 -format clustalw'
```

[8]:
```
subprocess.run(ALIGNMENT_COMMAND)
```

[8]:
```
CompletedProcess(args='MAli-v1.31/MAli.exe -input data/input -output data/output
-batch -iterations 100 -seed 25032025 -format clustalw', returncode=0)
```

[9]:
```
ALIGNMENTS_PRODUCED = os.listdir(OUTPUT_FOLDER)
COUNT_PRODUCED = len(ALIGNMENTS_PRODUCED)

print(f"Performed batch alignment: produced {COUNT_PRODUCED} alignments")
print(f"Filenames: {ALIGNMENTS_PRODUCED}")
```

```
Performed batch alignment: produced 6 alignments
Filenames: ['BB20016.aln', 'BB20018.aln', 'BB20020.aln', 'BB20036.aln',
'BB20039.aln', 'BB40044.aln']
```

Comparing Alignments Produced by MAli to Expected States

[10]:
```
ALIGNMENT_FILEPATHS = []
for ALIGNMENT_FILENAME in ALIGNMENTS_PRODUCED:
    ALIGNMENT_FILEPATH = f"{OUTPUT_FOLDER}/{ALIGNMENT_FILENAME}"
    ALIGNMENT_FILEPATHS.append(ALIGNMENT_FILEPATH)
```

BB20020 Expected Alignment

- SEED_VALUE = 25032025
- ITERATIONS = 100

```
1mrj_           --D-----VSFRLSGATSSSYGVFISNLRK-ALPNERKLY-DIP-LL-RSSLPGSQ-RYALIHLTNYADETISVAIDV
1apg_A          IFPKQYPIINFTTAGATVQSYTNFIRAVRG-RLTTGADVRHEIPVLPNRVGLPINQ-RFILVELSNHAELSVTLALDV
1abr_A          --EDR-P-IKFSTEGATSQSYKQFIEALRE-RLRGGL-I-HDIPVLPDPTTLQERN-RYITVELSNSDTESIEVGIDV
1qi7_A          V--T--S-ITLDLVNPTAGQYSSFVDKIRN-NVKDPNLKYGGTDIAVIGPPSKEK-F--LRINFQ-SSRGTVSLGLKR
1apa_           I--N--T-ITFDVGNATINKYATFMKSIHN-QAKDPTLKCYGIPMLPNTN-LTPK-Y--LLVTLQDSSLKTITLMLKR
1dm0_A          ---KEFTLD-FSTAKTYVDS-LNVIRSAIGTPLQT-I-SSGGTSLLMIDDNLFAVDVRGIDPEEGRFNNL-R-LIVER

1mrj_           -----N-EASATEAAKYVFKDAMRKVTLPYS-G-NYERLQTAAGKIR---ENIPLGLPA-LDSAITTLFYYNANSA---
1apg_A          HPD--N-QEDA-EAITHLFTDVQNRYTFAFG-G-NYDRLEQLAGNLR---ENIELGNGP-LEEAISALYYYST-GGTQ
1abr_A          LRD--A-PSSA-SDYLFTGTD-QHSLPF-YG-T--YGDLERWAHQSR---QQIPLGLQA-LTHGIS-F-FRS--GGND
1qi7_A          FK--SEITSAE-L--TALFPEATTANQKALEYTEDYQSIEKNAQITQGDKSRKELGLGIDLLLTFMEAVNKKAR-V-V
1apa_           FKDISNTTERN-DVMTTLCPNPSSRVGKNINYDSSYPALEKKVGRPR---SQVQLGIQI-LNSGIGKIYGVDSF-T-E
1dm0_A          A-DF---SHVTFP-GT---TAV----TLSGD-S-SYTTLQRVAGISR---TGMQINRHS-LTTSYLDLMSH---SGTS

1mrj_           YKFIEQQIGKRVDK-TFL-PSLAIISLENSWSALSKQI-QIA-STN-NGQFESP-VVL---INAQNQRVTITNVDAGV
1apg_A          FQYIEGEMRTRIRYNRRSAPDPSVITLENSWGRLSTAI-Q-ES--N-QGAFASP-IQL-QRR-NGSKFS-VYDVSILI
1abr_A          FRYISNRVRVSIQTGTAFQPDAAMISLENNWDNLSRGV-Q-ES--V-QDTFPNQ-VTLTNIR-NEPVI--VDSLSHPT
1qi7_A          FRYIQNLVTK--NFPNKFDSDNKVIQFEVSWRKISTAIYG-D-A-K-NGVFNKD-YDF-GF---G-K---VRQVKD-L
1apa_           FKYIENQVKT--NFNRAFYPNAKVLNLEESWGKISTAI-H-N-A-K-NGALTSP-LEL-KNA-NGSKWI-VLRVDDIE
1dm0_A          FRQIQRGFRTTLDSYVMTAEDVDL-TL-N-WGRLSSVL-P-D--YHGQDSVRVGRISF-GSI-NAILGS-VALILNCF

1mrj_           A-------
1apg_A          S--Q---F
1abr_A          ---N----
1qi7_A          ---P---K
1apa_           ---A---T
1dm0_A          STLGAILM
```

Actual Alignment

[11]:
```
presenter.present_aligned_clustalw(ALIGNMENT_FILEPATHS[2])
```

```
Displaying ClustalW format alignment from data/output/BB20020.aln:

1mrj_           --D-----VSFRLSGATSSSYGVFISNLRK-ALPNERKLY-DIP-LL-RSSLPGSQ-
RYALIHLTNYADETISVAIDVTNVYIMGY---RA--GDTSYFF
1apg_A          IFPKQYPIINFTTAGATVQSYTNFIRAVRG-RLTTGADVRHEIPVLPNRVGLPINQ-
RFILVELSNHAELSVTLALDVTNAYVVGY---R--AGNSAYFF
```

```
1abr_A          --EDR-P-IKFSTEGATSQSYKQFIEALRE-RLRGGL-I-HDIPVLPDPTTLQERN-
RYITVELSNSDTESIEVGIDVTNAYVVAY---RA--GTQSYF-
1qi7_A          V--T--S-ITLDLVNPTAGQYSSFVDKIRN-NVKDPNLKYGGTDIAVIGPPSKEK-F--
LRINFQ-SSRGTVSLGLKRDNLYVVAYLAMDNTNVNRAYY-
1apa_           I--N--T-ITFDVGNATINKYATFMKSIHN-QAKDPTLKCYGIPMLPNTN-LTPK-Y--
LLVTLQDSSLKTITLMLKRNNLYVMGY-A-DTYNGKCRYHI
1dm0_A          ---KEFTLD-FSTAKTYVDS-LNVIRSAIGTPLQT-I-
SSGGTSLLMIDDNLFAVDVRGIDPEEGRFNNL-R-LIVERNNLYVTGFVN-RT--NNVFYRF

1mrj_           -----N-EASATEAAKYVFKDAMRKVTLPYS-G-NYERLQTAAGKIR---ENIPLGLPA-
LDSAITTLFYYNANSA--------ASALMVLIQSTSEAAR
1apg_A          HPD--N-QEDA-EAITHLFTDVQNRYTFAFG-G-NYDRLEQLAGNLR---ENIELGNGP-
LEEAISALYYYST-GGTQL-P-TLARSFIICIQMISEAAR
1abr_A          LRD--A-PSSA-SDYLFTGTD-QHSLPF-YG-T--YGDLERWAHQSR---QQIPLGLQA-
LTHGIS-F-FRS--GGNDN-E-EKARTLIVIIQMVAEAAR
1qi7_A          FK--SEITSAE-L--
TALFPEATTANQKALEYTEDYQSIEKNAQITQGDKSRKELGLGIDLLLTFMEAVNKKAR-V-V---
KNEARFLLIAIQMTAEVAR
1apa_           FKDISNTTERN-DVMTTLCPNPSSRVGKNINYDSSYPALEKKVGRPR---SQVQLGIQI-
LNSGIGKIYGVDSF-T-E---KTEAEFLLVAIQMVSEAAR
1dm0_A          A-DF---SHVTFP-GT---TAV----TLSGD-S-SYTTLQRVAGISR---TGMQINRHS-
LTTSYLDLMSH---SGTSLTQ-SVARAMLRFVTVTAEALR

1mrj_           YKFIEQQIGKRVDK-TFL-PSLAIISLENSWSALSKQI-QIA-STN-NGQFESP-VVL---
INAQNQRVTITNVDAGVVTSNIA-LLLN-RN----N--M
1apg_A          FQYIEGEMRTRIRYNRRSAPDPSVITLENSWGRLSTAI-Q-ES--N-QGAFASP-IQL-QRR-
NGSKFS-VYDVSILIPII--A-LMVY-RCA-P-PP-S
1abr_A          FRYISNRVRVSIQTGTAFQPDAAMISLENNWDNLSRGV-Q-ES--V-QDTFPNQ-VTLTNIR-
NEPVI--VDSLSHPTVAVL-A-LMLF-VCN-P--P--
1qi7_A          FRYIQNLVTK--NFPNKFDSDNKVIQFEVSWRKISTAIYG-D-A-K-NGVFNKD-YDF-GF---
G-K---VRQVKD-L-QM--G-LLMY-L-G---K---
1apa_           FKYIENQVKT--NFNRAFYPNAKVLNLEESWGKISTAI-H-N-A-K-NGALTSP-LEL-KNA-
NGSKWI-VLRVDDIEPDV--G-LLKY-VNG-TCQ---
1dm0_A          FRQIQRGFRTTLDSYVMTAEDVDL-TL-N-WGRLSSVL-P-D--YHGQDSVRVGRISF-GSI-
NAILGS-VALILNCFPSMCPADGRVRGITHNKILWDS

1mrj_           A-------
1apg_A          S--Q---F
1abr_A          ---N----
1qi7_A          ---P---K
1apa_           ---A---T
1dm0_A          STLGAILM
```

BB40044 Expected Alignment

- SEED_VALUE = 25032025
- ITERATIONS = 100

```
1a8l_           MGLISDADKK-VIKEEFF-SKMV-----NPVKLIVFVRKDHCQYCDQLKQLVQELSELTDKLSYEIVDFDTP-EGKELA
2trc_P          EGQATHTGPKGVINDWRKFKLESEDGDSIPPSKKEILRQMSSPQSRDDKDSKERXSRKXSIQEYELIHQDKEDEG-CLR
1mek_           ----------------------------------------------------------------------------
1erv_           ----------------------------------------------------------------------------
1r26_A          ----------------------------------------------------------------------------
1thx_           ----------------------------------------------------------------------------

1a8l_           VRYFGLPAG-HEFAAFLEDIVDVSREETNLMDETKQAIRNIDQD-VRILV-FVTPTCPYCPLAVRMAHKFAIEN-TKAG
2trc_P          PRY-GFVY-ELETGEQFLETIEKEQKVTTIVVNIYEDGVRGCDA-LNSSLECLA--AEY-P-XVKFCKIRASNTGAGDR
1mek_           ------------------DAPEEEDHVLVLRKSNFAEAL-AAHKYLLV-E-FYAPWCGHCKALAPEYAKAAGKLK--AE
1erv_           ------------------MVKQIE------SKTAFQEALDAAGDKLVV-VDFSATWCGPCKMIKPFFHSLS------EK
1r26_A          --------------------PSVVD-VYSV-EQ-FRNI--MSEDILTVAW-FTAVWCGPCKTIERPMEKIAY------K
1thx_           ----------------------SKGVITITDAEFESEVL-KAEQPVLV-Y-FWASWCGPCQLMSPLINLAANTYS--DR

1a8l_           NVMAVPKIVIQVNGEDRVE-FEGAYPEKMFL-EKLLSALS-------
2trc_P          IS----------VAEQFAEDFFAADVE-SFLNEYGLLPER-------
1mek_           GVRGYPTIKFFRNGDTASP--KEYTAGREAD-DI-VNWLKKRTGPAA
1erv_           EVKSMPTFQFFKKGQKVGE----F-SG--ANK----EKLEATINELV
1r26_A          RVLQLPTFIIARSGKMLGH-VI----G--AN-PG-MLRQKLRDIIKD
1thx_           KVEGVPALRLVK-GEQILD----STEGVISK-DK-LLSFLD-TH-LN
```

Actual Alignment

[12]:
```
presenter.present_aligned_clustalw(ALIGNMENT_FILEPATHS[5])
```

```
Displaying ClustalW format alignment from data/output/BB40044.aln:

1a8l_           MGLISDADKK-VIKEEFF-SKMV-----
NPVKLIVFVRKDHCQYCDQLKQLVQELSELTDKLSYEIVDFDTP-EGKELAKRYRIDRAPATTITQDGKDFG
2trc_P
EGQATHTGPKGVINDWRKFKLESEDGDSIPPSKKEILRQMSSPQSRDDKDSKERXSRKXSIQEYELIHQDKEDEG-
CLRK-YR--RQCXQDXHQKL-SFG
1mek_           ------------------------------------------------------------
------------------------------------
1erv_           ------------------------------------------------------------
------------------------------------
1r26_A          ------------------------------------------------------------
------------------------------------
1thx_           ------------------------------------------------------------
------------------------------------

1a8l_           VRYFGLPAG-HEFAAFLEDIVDVSREETNLMDETKQAIRNIDQD-VRILV-
FVTPTCPYCPLAVRMAHKFAIEN-TKAGKGKILGDMVEAIEYPEWADQY
2trc_P          PRY-GFVY-ELETGEQFLETIEKEQKVTTIVVNIYEDGVRGCDA-LNSSLECLA--AEY-P-
XVKFCKIRASNTGAGDRFSSDVLPTLLVYKGGELISNF
1mek_           ------------------DAPEEEDHVLVLRKSNFAEAL-AAHKYLLV-E-
FYAPWCGHCKALAPEYAKAAGKLK--AEGSEIRLAKVDATEESDLAQQY
1erv_           ------------------MVKQIE------SKTAFQEALDAAGDKLVV-
VDFSATWCGPCKMIKPFFHSLS------EKYSNVIFLEVDVDDCQDVASEC
```

```
1r26_A          --------------------PSVVD-VYSV-EQ-FRNI--MSEDILTVAW-
FTAVWCGPCKTIERPMEKIAY------EFPTVKFAKVDADNNSEIVSKC
1thx_           ----------------------SKGVITITDAEFESEVL-KAEQPVLV-Y-
FWASWCGPCQLMSPLINLAANTYS--DRLKVVKLE-IDPNPTT-V-KKY

1a8l_           NVMAVPKIVIQVNGEDRVE-FEGAYPEKMFL-EKLLSALS-------
2trc_P          IS----------VAEQFAEDFFAADVE-SFLNEYGLLPER-------
1mek_           GVRGYPTIKFFRNGDTASP--KEYTAGREAD-DI-VNWLKKRTGPAA
1erv_           EVKSMPTFQFFKKGQKVGE----F-SG--ANK----EKLEATINELV
1r26_A          RVLQLPTFIIARSGKMLGH-VI----G--AN-PG-MLRQKLRDIIKD
1thx_           KVEGVPALRLVK-GEQILD----STEGVISK-DK-LLSFLD-TH-LN
```

[ ]:

[ ]: