04a-Iterative-Refinement

March 26, 2025

1 Test 04: Iterative Refinement

1.0.1 Overview

This notebook demonstrates iterative refinement of an alignment produced by a progressive alignment algorithm in ClustalW2.

Expected runtime: ~30 seconds or less

1.0.2 Context

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

Requirement 3.3 - Can load an existing alignment state from an appropriate bioinformatics file format, for iterative refinement. - In this notebook, an initial alignment is produced in the ClustalW format using ClustalW2, and then refined by MAli.

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
```

ClustalW2

[3]:

```
ALIGNER_NAME = "ClustalW2"
ALIGNER_PATH = "ClustalW2/clustalw2.exe"
ALIGNER_OUTPUT_FOLDER = "data/w2_output"
```

MAli v1.31

```
[4]: REFINER_NAME = "MAli-v1.31"
     REFINER_PATH = "MAli-v1.31/MAli.exe"
     REFINER_OUTPUT_FOLDER = "data/refined_output"
```

```
[5]: # creating empty output folders
     for OUTPUT_FOLDER in [ALIGNER_OUTPUT_FOLDER, REFINER_OUTPUT_FOLDER]:
         if os.path.exists(OUTPUT_FOLDER):
             shutil.rmtree(OUTPUT_FOLDER)
         os.makedirs(OUTPUT_FOLDER)
```

Testcase The BB20016 testcase from BAliBASE has been chosen as it contains 6 biological sequences and has a structural reference available.

All testcases from BALIS-2 (subset of BAliBASE used for development) containing 6 sequences have been included in /data

```
[6]: TESTCASE_NAME = "BB20018"
     INPUT_FILEPATH = f"data/input/{TESTCASE_NAME}"
     ALIGNED_OUTPUT_FILEPATH = f"{ALIGNER_OUTPUT_FOLDER}/{TESTCASE_NAME}"
     REFINED_OUTPUT_FILEPATH = f"{REFINER_OUTPUT_FOLDER}/{TESTCASE_NAME}"
```

Viewing Testcase

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

```
[8]: presenter.present_unaligned_fasta(INPUT_FILEPATH)
```

```
Displaying Sequences from data/input/BB20018:

>1ldg_
APKAKIVLVGSGMIGGVMATLIVQKNLGDVVLFDIVKNMPHGKALDTSHTNVMSNCKVSGSNTYDDLAGSDVVIVTAGFT
KEWNRLDLLPLNNKIMIEIGGHIKKNCAFIIVVTNPVDVMVQLLHQHSGVPKNKIIGLGGVLDTSRLKYYISQKLNVCPR
DVNAHIVGAHGNKMVLLKRYITVEFINNKLISDAELEAIFDRTVNTALEIVNLHASPYVAPAAAIIEMAESYLKDLKKVL
ICSTLLEGQYGHSDIFGGTPVVLGANGVEQVIELQLNSEEKAKFDEAIAETKRMKALA

>1lld_A
PTKLAVIGAGAVGSTLAFAAAQRGIAREIVLEDIAKERVEAEVLDMQHGSSFYPTVSIDGSDDPEICRDADMVVITAGPR
QKPGQSRLELVGATVNILKAIMPNLVKVAPNAIYMLITNPVDIATHVAQKLTGLPENQIFGSGTNLDSARLRFLIAQQTG
VNVKNVHAYIAGEHGDSEVPLWESATIGGVPMSDWTPLPGHDPLDADKREEIHQEVKNAAYKIINGKGATNYAIGMSGVD
IIEAVLHDTNRILPVSSMLKDFHGISDICMSVPTLLNRQGVNNTINTPVSDKELAALKRSAETLKETAAQFGF

>1i0z_A
ATLKEKLIAPVAEEEATVPNNKITVVGVGQVGMACAISILGKSLADELALVDVLEDKLKGEMMDLQHGSLFLQTPKIVAD
KDYSVTANSKIVVVTAGVRQQEGESRLNLVQRNVNVFKFIIPQIVKYSPDCIIIVVSNPVDILTYVTWKLSGLPKHRVIG
SGCNLDSARFRYLMAEKLGIHPSSCHGWILGEHGDSSVAVWSGVNVAGVSLQELNPEMGTDNDSENWKEVHKMVVESAYE
VIKLKGYTNWAIGLSVADLIESMLKNLSRIHPVSTMVKGMYGIENEVFLSLPCILNARGLTSVINQKLKDDEVAQLKKSA
DTLWDIQKDLKD

>1ez4_A
```

```
SMPNHQKVVLVGDGAVGSSYAFAMAQQGIAEEFVIVDVVKDRTKGDALDLEDAQAFTAPKKIYSGEYSDCKDADLVVITA
GALVNKNLNILSSIVKPVVDSGFDGIFLVAANPVDILTYATWKFSGFPKERVIGSGTSLDSSRLRVALGKQFNVDPRSVD
AYIMGEHGDSEFAAYSTATIGTRPVRDVAKEQGVSDDDLAKLEDGVRNKAYDIINLKGATFYGIGTALMRISKAILRDEN
AVLPVGAYMDGQYGLNDIYIGTPAIIGGTGLKQIIESPLSADELKKMQDSAATLKKVLNDGLAELEN

>1guy_A
MRKKISIIGAGFVGSTTAHWLAAKELGDIVLLDIVEGVPQGKALDLYEASPIEGFDVRVTGTNNYADTANSDVIVVTSGA
LIKVNADITRACISQAAPLSPNAVIIMVNNPLDAMTYLAAEVSGFPKERVIGQAGVLDAARYRTFIAMEAGVSVEDVQAM
LMGGHGDEMVPLPRFSTISGIPVSEFIAPDRLAQIVERTRKGGGEIVNLLKTGSAYYAPAAATAQMVEAVLKDKKRVMPV
AAYLTGQYGLNDIYFGVPVILGAGGVEKILELPLNEEEMALLNASAKAVRATLDTL

>1b8p_A
KTPMRVAVTGAAGQICYSLLFRIANGDMLGKDQPVILQLLEIPNEKAQKALQGVMMEIDDCAFPLLAGMTAHADPMTAFK
DADVALLVGARPRGPGMERKDLLEANAQIFTVQGKAIDAVASRNIKVLVVGNPANTNAYIAMKSAPSLPAKNFTAMLRLD
HNRALSQIAAKTGKPVSSIEKLFVWGNHSPTMYADYRYAQIDGASVKDMINDDAWNRDTFLPTVGKRGAAIIDARGVSSA
ASAANAAIDHIHDWVLGTAGKWTTMGIPSDGSYGIPEGVIFGFPVTTENGEYKIVQGLSIDAFSQERINVTLNELLEEQN
GVQHLLG
```

Initial Alignment with ClustalW2 Here, -OUTPUT=CLUSTAL is specified such that ClustalW2 will output a ClustalW format alignment.

[9]:
```python
ALIGNMENT_COMMAND = f"{ALIGNER_PATH} -INFILE={INPUT_FILEPATH}␣
 ↪-OUTFILE={ALIGNED_OUTPUT_FILEPATH} -OUTPUT=CLUSTAL -ALIGN"
print(f"CLI command to be run: '{ALIGNMENT_COMMAND}'")
```

```
CLI command to be run: 'ClustalW2/clustalw2.exe -INFILE=data/input/BB20018
-OUTFILE=data/w2_output/BB20018 -OUTPUT=CLUSTAL -ALIGN'
```

[10]:
```python
subprocess.run(ALIGNMENT_COMMAND)
print(f"Performed alignment of {TESTCASE_NAME} with ClustalW2")
```

```
Performed alignment of BB20018 with ClustalW2
```

Performing Refinement with MAli Here, MAli is tasked with accepting a ClustalW format alignment as input. This will be a starting point for iterative refinement.

[11]:
```python
REFINEMENT_COMMAND = f"{REFINER_PATH} -input {ALIGNED_OUTPUT_FILEPATH} -output␣
 ↪{REFINED_OUTPUT_FILEPATH} -refine"
print(f"CLI command to be run: '{REFINEMENT_COMMAND}'")
```

```
CLI command to be run: 'MAli-v1.31/MAli.exe -input data/w2_output/BB20018
-output data/refined_output/BB20018 -refine'
```

[12]:
```python
subprocess.run(REFINEMENT_COMMAND)
print(f"Performed refinement of {TESTCASE_NAME} with MAli")
```

```
Performed refinement of BB20018 with MAli
```

Viewing Refined Alignment Produced by MAli

[13]:
```
UNREFINED_ALIGNMENT_FILEPATH = ALIGNED_OUTPUT_FILEPATH
REFINED_ALIGNMENT_FILEPATH = REFINED_OUTPUT_FILEPATH + ".faa"
presenter.present_interleaved_aligned_fasta(REFINED_ALIGNMENT_FILEPATH)
```

```
Displaying interleaved alignment from 'data/refined_output/BB20018.faa:

1ldg_           -----------------APKAKIVLVGSG-MIGG-----VMATLIVQKNLG-DVVLFDIV
1guy_A          ------------------MRKKISIIGAG-FVGS-----TTAHWLAAKELG-DIVLLDIV
1lld_A          -------------------PTKLAVIGAG-AVGSTLAFAAAQ-----RGIAREIVLEDIA
1ez4_A          ---------------SMPNHQKVVLVGDG-AVGSSYAFAMAQ-----QGIAEEFVIVDVV
1i0z_A          ATLKEKLIAPVAEEEATVPNNKITVVGVG-QVGM-----ACAISILGKSLADELALVDVL
1b8p_A          -----------------KTPMRVAVTGAAGQICYSLLFRIANGDMLGKDQPVILQLLEIP

1ldg_           KNMPH---GKALDTSHTNVMS--NCKVSGSNTYDDLAGSDVVIVTAG--FTKEWNRLDLL
1guy_A          EGVPQ---GKALDLYEASPIEGFDVRVTGTNNYADTANSDVIVVTSG--ALIKVN-ADIT
1lld_A          KERVE---AEVLDMQHGSSF-YPTVSIDGSDDPEICRDADMVVITAGPRQKPGQSRLELV
1ez4_A          KDRTK---GDALDLEDAQAFTA-PKKIY-SGEYSDCKDADLVVITAG----------ALV
1i0z_A          EDKLK---GEMMDLQHGSLF-LQTPKIVADKDYSVTANSKIVVVTAGVRQQEGESRLNLV
1b8p_A          NEKAQKALQGVMMEIDDCAFPLLAGMTAHADPMTAFKDADVALLVGARPRGPGMERKDLL

1ldg_           PLNNKIMIEIGGHIKKNC--AFIIVVT-NPVDVMVQLLHQHSGVPKNKIIGLGGVLDTSR
1guy_A          ----RACISQAAPLSPN---AVIIMVN-NPLDAMTYLAAEVSGFPKERVIGQAGVLDAAR
1lld_A          GATVNILKAIMPNLVKVAPNAIYMLIT-NPVDIATHVAQKLTGLPENQIFGSGTNLDSAR
1ez4_A          NKNLNILSSIVKPVVDSGFDGIFLVAA-NPVDILTYATWKFSGFPKERVIGSGTSLDSSR
1i0z_A          QRNVNVFKFIIPQIVKYSPDCIIIVVS-NPVDILTYVTWKLSGLPKHRVIGSGCNLDSAR
1b8p_A          EANAQIFTVQGKAIDAVASRNIKVLVVGNPANTNAYIAMKSAPSLPAKNFTAMLRLDHNR

1ldg_           LKYYISQKLNVCPRDVN-AHIVGAHGNKMVLLKRYITVEFINN---------KLISDAEL
1guy_A          YRTFIAMEAGVSVEDVQ-AMLMGGHGDEMVPLPRFSTISGIPVS--------EFIAPDRL
1lld_A          LRFLIAQQTGVNVKNVH-AYIAGEHGDSEVPLWESATIGGVPMSDWTPLPGHDPLDADKR
1ez4_A          LRVALGKQFNVDPRSVD-AYIMGEHGDSEFAAYSTATIGTRPVRDVAKEQG---VSDDDL
1i0z_A          FRYLMAEKLGIHPSSCH-GWILGEHGDSSVAVWSGVNVAGVSLQELNPEMGTD-NDSENW
1b8p_A          ALSQIAAKTGKPVSSIEKLFVWGNHSPTMYADYRYAQIDGASVKDMIN------DDAWNR

1ldg_           EAIFDRTVNTALEIVNLHAS--PYVAPAAAIIEMAESYLKDLKKVLICSTLLEGQYGHS-
1guy_A          AQIVERTRKGGGEIVNLLKTGSAYYAPAAATAQMVEAVLKDKKRVMPVAAYLTGQYGLN-
1lld_A          EEIHQEVKNAAYKIINGKGA--TNYAIGMSGVDIIEAVLHDTNRILPVSSMLKDFHGIS-
1ez4_A          AKLEDGVRNKAYDIINLKGA--TFYGIGTALMRISKAILRDENAVLPVGAYMDGQYGLN-
1i0z_A          KEVHKMVVESAYEVIKLKGY--TNWAIGLSVADLIESMLKNLSRIHPVSTMVKGMYGIEN
1b8p_A          DTFLPTVGKRGAAIIDARGVSSAASAANAAIDHIHDWVLGTAGKWTTMGIPSDGSYGIPE

1ldg_           DIFGGTPVVLGANGVEQVIELQLNSEEKAKFDEAIAETKRMKALA------
1guy_A          DIYFGVPVILGAGGVEKILELPLNEEEMALLNASAKAVRATLDTL------
1lld_A          DICMSVPTLLNRQGVNNTINTPVSDKELAALKRSAETLKETA----AQFGF
1ez4_A          DIYIGTPAIIGGTGLKQIIESPLSADELKKMQDSAATLKKVLNDGLAELEN
1i0z_A          EVFLSLPCILNARGLTSVINQKLKDDEVAQLKKSADTLWDIQ----KDLKD
1b8p_A          GVIFGFPVTTENGEYKIVQGLSIDAFSQERINVTLNELLEEQNGVQHLLG-
```

5

[ ]: