MSARC – user documentation

Michał Modzelewski   Wiktor Blaschke   Norbert Dojer

1 Getting started

MSARC requires a Python 2.7 interpreter, with the modules numpy, scipy and BioPython installed. The MSARC source code is provided as a gzip compressed tarball that needs to be extracted. This can be done with the following command under a UNIX style operating system:

```
$ tar -xzf msarc_x.x.tar.gz
```

where x.x should be replaced with actual version. The source code contains some Python extensions that need to be built using the included setup.py script.

```
$ python setup.py build_ext --inplace
```

The above command uses the system default compiler. Building on Windows has been tested using the MinGW compiler with the following commands at the command prompt:

```
>PATH=C:\Python27;C:\MinGW\bin;%PATH%
>python setup.py build_ext --inplace --compiler=mingw32
```

2 Invocation

```
msarc [-h] [-T TEMP] [-M] [-C REPS] [-R REPS] [-c CUT] [-d | -p]
      [-g SCORE] [-x SCORE] [-e SCORE | --no-end-gaps] [-m MATRIX]
      [-s SET] [-v]
      FILE
```

For example, if the msarc script is executable and Python 2.7 is the main Python interpreter installed on the system, MSARC may be invoked by the following command

```
$ ./msarc -s 160 test.fasta
```

Otherwise the script will need to be called through the interpreter

```
$ /usr/bin/python2.7 msarc -s 160 test.fasta
```

3 Options

-h, --help

Shows the help message and exits.

-T TEMP, --temperature TEMP

Sets the value of the thermodynamic temperature T, which defaults to $\frac{\log 10}{2}$.

-M, --multilevel

Enables the multilevel graph partitioning algorithm.

-w, --weighted-transformation

Enables weighting sequence pairs in the consistency transformation procedure (new in 1.1).

-C REPS, --consistency REPS

Sets the number of iterations of the consistency transformation to be performed, the default being 2.

-R REPS, --refinements REPS

Sets the number of iterations of horizontal refinement to be performed, the default being 100.

-c CUT, --cut CUT

Sets the cut-off value for posterior probabilities. A higher cut-off value increases the speed of the consistency transformation, and operations on sparse matrices, but may affect accuracy. This setting defaults to 0.01.

-d, --DNA

Forces input polymers to be treated as nucleic acids. By default, an attempt is made to determine the type of the input polymers automatically. This option is mutually exclusive with the following option.

-p, --protein

Forces input polymers will be treated as amino acids. By default, an attempt is made to determine the type of the input polymers automatically. This option is mutually exclusive with the previous option.

-g SCORE, --gap-open SCORE

Sets the gap opening penalty. By default, this is set based on the substitution matrix used.

-x SCORE, --gap-extend SCORE

Sets the gap extension penalty. By default, this is set based on the substitution matrix used.

-e SCORE, --end-gaps SCORE

Sets the penalty for terminal gaps, by default 0. This option is mutually exclusive with the following option.

--no-end-gaps

Turns off the special treatment of terminal gaps, causing terminal gaps to be scored just like internal gaps. This option is mutually exclusive with the previous option.

-m MATRIX, --matrix MATRIX

Selects the substitution matrix series to be used from among *blosum*, *gonnet* (the default) and *pam*.

-s SET, --set SET

Selects the substitution matrix set. The default value of -1 causes the set to be automatically selected based on the computed evolutionary distance between sequences. Accepted values are 30, 50, 62, and 80 for the *blosum* series; 40, 80, 120, 160, 250, 300, and 350 for the *gonnet* series; 20, 60, 120, and 350 for the *pam* series.

-P PROCESSES, --max-processes PROCESSES

Sets the maximum number of simultaneous processes (by default multiprocessing is off; new in 1.2).

-v, --verbose

Turns on verbose mode, which outputs progress to the terminal while aligning.

4 Example input

MSARC takes as input a file in *fasta* format.

```
>1aab_
GKGDPKKPRGKMSSYAFFVQTSREEHKKKHPDASVNFSEFSKKCSERWKT
MSAKEKGKFEDMAKADKARYEREMKTYIPPKGE
>1j46_A
MQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAE
KWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPK
>1k99_A
MKKLKKHPDFPKKPLTPYFRFFMEKRAKYAKLHPEMSNLDLTKILSKKYK
ELPEKKKMKYIQDFQREKQEFERNLARFREDHPDLIQNAKK
>2lef_A
MHIKKPLNAFMLYMKEMRANVVAESTLKESAAINQILGRRWHALSREEQA
KYYELARKERQLHMQLYPGWSARDNYGKKKKRKREK
```

5 Example output

If verbose output is requested with the appropriate option, MSARC prints progress information during the alignment process. Once the process is complete, the alignment is output to the screen in *msf* format.

```
$ ./msarc -v -M -R 2 -g -22 -x -1 -s 160 tests/BB11001.tfa
reading sequences ... done
  1aab_
    GKGDPKKPRGKMSSYAFFVQTSREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKFEDMAKADKARYEREMKT
    YIPPKGE
  1j46_A
    MQDRVKRPMNAFIVWSRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRP
    RRKAKMLPK
  1k99_A
    MKKLKKHPDFPKKPLTPYFRFFMEKRAKYAKLHPEMSNLDLTKILSKKYKELPEKKKMKYIQDFQREKQEFERNLA
    RFREDHPDLIQNAKK
  2lef_A
    MHIKKPLNAFMLYMKEMRANVVAESTLKESAAINQILGRRWHALSREEQAKYYELARKERQLHMQLYPGWSARDNY
    GKKKKRKREK

calculating pairwise probabilities ...
  1aab_ with 1j46_A ... done
    matrix: gonnet 160
    gap penalties: -22.000000 (open), -1.000000 (extend), 0.000000 (terminal)
  1aab_ with 1k99_A ... done
    matrix: gonnet 160
    gap penalties: -22.000000 (open), -1.000000 (extend), 0.000000 (terminal)
  1aab_ with 2lef_A ... done
    matrix: gonnet 160
    gap penalties: -22.000000 (open), -1.000000 (extend), 0.000000 (terminal)
  1j46_A with 1k99_A ... done
    matrix: gonnet 160
    gap penalties: -22.000000 (open), -1.000000 (extend), 0.000000 (terminal)
  1j46_A with 2lef_A ... done
    matrix: gonnet 160
    gap penalties: -22.000000 (open), -1.000000 (extend), 0.000000 (terminal)
  1k99_A with 2lef_A ... done
    matrix: gonnet 160
    gap penalties: -22.000000 (open), -1.000000 (extend), 0.000000 (terminal)

performing consistency transformation ...
  iteration 1 ... done
  iteration 2 ... done
```

```
partitioning graph ............ done

realigning sequences ...
  1aab_ with 1j46_A, 1k99_A, 2lef_A ... done
  1aab_, 1j46_A, 2lef_A with 1k99_A ... done
  1aab_, 1j46_A, 1k99_A with 2lef_A ... done
  1aab_, 1k99_A, 2lef_A with 1j46_A ... done

refining alignment ...
  1aab_, 1k99_A with 1j46_A, 2lef_A ... done
  1j46_A, 2lef_A with 1aab_, 1k99_A ... done

!!AA_MULTIPLE_ALIGNMENT 1.0
PileUp of: @tests/BB11001.tfa

 <stdout>  MSF: 107  Type: P  September 25, 2012 00:27  Check: 6956 ..

 Name: 1aab_            Len:   107  Check: 6605  Weight:  1.00
 Name: 1j46_A           Len:   107  Check:  349  Weight:  1.00
 Name: 1k99_A           Len:   107  Check:  683  Weight:  1.00
 Name: 2lef_A           Len:   107  Check: 9319  Weight:  1.00

//

        1                                                  50
1aab_   ~~~GKGDPKK PRGKMSSYAF FVQTSREEHK KKHPDASVNF SEFSKKCSER
1j46_A  ~~~~~~MQDR VKRPMNAFIV WSRDQRRKMA LENPR..MRN SEISKQLGYQ
1k99_A  MKKLKKHPDF PKKPLTPYFR FFMEKRAKYA KLHPE..MSN LDLTKILSKK
2lef_A  ~~~~~~~~MH IKKPLNAFML YMKEMRANVV AESTL..KES AAINQILGRR

        51                                                100
1aab_   WKTMSAKEKG KFEDMAKADK ARYEREMKTY IPPKGE~~~~ ~~~~~~~~~~
1j46_A  WKMLTEAEKW PFFQEAQKLQ AMHR...... .EKYPNYKYR P.......RR
1k99_A  YKELPEKKKM KYIQDFQREK QEFERNLARF REDHPDLIQN A.......KK
2lef_A  WHALSREEQA KYYELARKER QLHM...... .QLYPGWSAR DNYGKKKKRK

        101
1aab_   ~~~~~~~
1j46_A  KAKMLPK
1k99_A  ~~~~~~~
2lef_A  REK~~~~
```

The alignment in *msf* format is also saved to a file with the same name as the input file and the .msf file extension.

6 Additional programs

batch-msarc

Similar to the main msarc program, but uses multiple processes to align multiple sets of sequences simultaneously. Invoked with the same arguments as the msarc program.

compare

Used to compare different alignments of the same sequences with regard to the internal scoring function. Alignments are searched for within the current directory and sub-directories. Invoked with the same arguments as the msarc program, to set the parameters of the scoring function.

balitest

Scores alignments against BAliBASE reference alignments. Takes a list of reference files as arguments, and searches for corresponding alignments in the current directory.

balitest-compare

Compares different alignments against BAliBASE reference alignments. Takes a list of reference files as arguments, and searches for all alignments within the current directory and sub-directories.