Manuals how to use ASR-CC-ranking.py

[Summary]

ASR-CC-ranking.py ranked the sequences based on the correlation coefficients calculated by analyzing “the number of occurrences of 20 amino acids for the template sequence and the sequence” and “Mean difference for number of AAs per thousands in mesophiles and thermophiles (Gregory A.C. Singer and Donal A. Hickey, Gene 317, 2003, 39-47)”. The detail was reported in the research paper in H. Arazeki et al,.

[Required calculation environment]

Linux (CentOS 6, 7)

Python >3.7, BioPython, Numpy, Scikit-learn

[Required input data]

・Sequence data in Fasta format (Multiple number of sequences are available)

・Template protein sequence data (only one sequence, fasta format)

[Construction of running environment]

1. Install Biopython, Numpy and Scikit-learn from Conda.
2. Run Python by interactive mode and enter the following commands one after another. Make sure there are no errors.

```
from Bio import AlignIO
import numpy as np
import os,sys,re,random,shutil,subprocess
from sklearn.linear_model import LinearRegression
```

3. Save the script directly under the analysis directory or in the script storage area.

[Preparation of input data]

0. Make the directory for analysis. In the directory, sequence data containing multiple number of sequences that would be ranked by the analysis and template sequence data were saved.
1. Make sure that there is one blank line between arrays. Make sure it is in the following format.

>PgiDAPDH
MTDDKKIRAAIVGYGNIGRYALQALREAPDFEIAGIVRRNPAEVPFELQPFRVVSDIEQLESV
DVALVCSPSREVERTALEILKKGICTADSFDIHDGILALRRSLGDAAGKSGAAAVIASGWDP
GSDSVVRTLMQAIVPKGITYTNFGPGMSMGHTVAVKAIDGVKAALSMTIPLGTGVHRRMV
YVELLPGHNLEEVSAAIKADEYFVHDETHVIQVDEVDALIDMGHGVRMVRKGVSGSTQNQ
RMSFDMEINNPALTGQVLVCAARAAMRQQPGAYTLQEIPVIDLLPGDREQWIGKLC

>CbDAPDH
MAIRVGILGYGNLGRGVECAVKHNPDMELKAVFTRRNPDSLSILTEGAKVCRAEDVLSMKD
QIDVMILCGGSATDLPGQTPEMAAHFNVIDSFDTHANIPRHFEAVDKAAKESGHVGIISVG
WDPGMFSLNRLYANAILPGGSDYTFWGKGVSQGHSDAIRRIKGVKDARQYTIPVEAALTAV
RSKKAPELTTRDKHTRECFVVAEEGADLKAIEEAIVTMPNYFADYDTTVHFISQEELMRDH
AGIPHGGFVIRTGSTGWNDENGHVIEYSLKLDSNPEFTASVIASYARAAYRLSREGQSGCK
TVFDIAPAYLSAADGAELRKHLL

>CtDAPDH
MNSKIRIGIVGYGNIGKGVEKAIKQNDDMELEAIFTRRDINKVDSNNSKLVHISRLELYKDTV
DVMILCGGSATDLVEQGPMIASQFNTVDSFDNHGRIPQHFERMDEISKKAGNISLISTGWD
PGLFSLNRLLGESILPKGKTHTFWGKGVSLGHSDAIRRVQGVKNGIQYIIPIKGALDKARSG
EQCDFTTREKHEMVCYVVPEENADLKKIEQDIKTMPDYFADYNTTVHFITEEELKLNHAGL
SNGGFVIRSGNTQGGAKQVMEFNLNLESSAEFTSSVLVAYSRAIYKLSKEGKKGAVTVLDI
PFSYLSPKTPEELRKELL

>PasDAPDH
MQLRSTTLYPLLLLLLLPLSGWAQEDHIDTTQAVQILQQAEQRGEARYGVSVWRIDEDKPL
LDYRSRERFTPASVTKIFSSATALIALGADYQFPTEIGYRGDITNDGVLKGDLIIVGHGDPSL
ESKHYPRRKGIFYEQVYLALQQAGIRQIRGRIIVDASAYCDEGYLDVWPREDWGRRYAPAV
YGVNLCDNIMQVGISAQEVAKGAKAPTFLHPSTPGHAWQMDIQLVKRGRLLAISADRNSR
TTRRLSGRLVRGSSKRQVIACDLSNPAMALALQLAEHLQQRGIELTDCQSVAYYDKSAPAL
TTLLDIYLSPHLSELIRTCNYHSVNLYAEALLRSIGNRFGSVQQGGCISTSEALRQEMNYWR
ETCSLSANELELYDGSGLSPRSKLSPYALTAALRQVYRLPLPLSDPFILSLPQVGREGTVRK
LLSASQLTAYFKSGSIRGVQNYAGYVSYNGHTYCVSLLANDMRHRGTTRRTMTQVLEALF
PNSPTTRASNP

1. Input the following command through the terminal.

sudo python ASR-CC-ranking.py -STP < Template protein sequence data> -ASRSEQ

Sequence data to be analyzed> -OUTPUT <The name of output file>

[Analysis of Output file]

0. The output data is stored in the specified path. The files in this section should be viewed sequentially.
1. Open the output file specified by -OUTPUT with a text editor.

----------(Start) The basic information used in the calculations can be found here----------

#The correlation of the prameters between standard and target protein.

#Mean difference for num. of AAs per thousands in mesophiles and thermophiles (thermophiles-mesophiles)

#standard_param = [A, C, E, H, I, K, Q, T, V, W, Y] of which p-value were <0.05.

#ref. Gregory A.C. Singer and Donal A. Hickey, Gene 317, 2003, 39-47

---------(End) -----

----------(From) The sequence data of the template protein is output.----------

#STP name is as follows:

>CgDAADH_STP

MTNIRVAIVGYGNLGRSVEKLIAKQPDMDLVGIFSRRATLDTKTPVFDVADVDKHADDVDVL
FLCMGSATDIPEQAPKFAQFACTVDTYDNHRDIPRHRQVMNEAATAAGNVALVSTGWDPG
MFSINRVYAAAVLAEHQQHTFWGPGLSQGHSDALRRIPGVQKAVQYTLPSEDALEKARRG
EAGDLTGKQTHKRQCFVVADAADHERIENDIRTMPDYFVGYEVEVNFIDEATFDSEHTGM
PHGGHVITTGDTGGFNHTVEYILKLDRNPDFTASSQIAFGRAAHRMKQQGQSGAFTVLEV
APYLLSPENLDDLIARDV

----------(End) ----------

----------(From) The calculated correlation coefficients are output. The larger this value, the closer the distribution of amino acid residues to thermophiles (that are relative to the template protein)----------

#Rank 1: CC_value was 0.9052431083047059

>UtDAPDH

MSKIRIGIVGYGNLGRGVEAAIQQNPDMELVAVFTRRDPKTVAVKSNVKVLHVDDAQSYKD
EIDVMILCGGSATDLPEQGPYFAQYFNTIDSFDTHARIPDYFDAVNAAAEQSGKVAIISVGW
DPGLFSLNRLLGEVVLPVGNTYTFWGKGVSQGHSDAIRRIQGVKNAVQYTIPIDEAVNRVR
SGENPELSTREKHARECFVVLEEGADPAKVEHEIKTMPNYFDEYDTTVHFISEEELKQNHS
GMPHGGFVIRSGKSDEGHKQIIEFSLNLESNPMFTSSALVAYARAAYRLSQNGDKGAKTVF

DIPFGLLSPKSPEDLRKELLTR

#Rank 2: CC_value was 0.8088190661754731
>CtDAPDH
MNSKIRIGIVGYGNIGKGVEKAIKQNDDMELEAIFTRRDINKVDSNNSKLVHISRLELYKDTV
DVMILCGGSATDLVEQGPMIASQFNTVDSFDNHGRIPQHFERMDEISKKAGNISLISTGWD
PGLFSLNRLLGESILPKGKTHTFWGKGVSLGHSDAIRRVQGVKNGIQYIIPIKGALDKARSG
EQCDFTTREKHEMVCYVVPEENADLKKIEQDIKTMPDYFADYNTTVHFITEEELKLNHAGL
SNGGFVIRSGNTQGGAKQVMEFNLNLESSAEFTSSVLVAYSRAIYKLSKEGKKGAVTVLDI
PFSYLSPKTPEELRKELL

#Rank 3: CC_value was 0.6727124275558394
>CbDAPDH
MAIRVGILGYGNLGRGVECAVKHNPDMELKAVFTRRNPDSLSILTEGAKVCRAEDVLSMKD
QIDVMILCGGSATDLPGQTPEMAAHFNVIDSFDTHANIPRHFEAVDKAAKESGHVGIISVG
WDPGMFSLNRLYANAILPGGSDYTFWGKGVSQGHSDAIRRIKGVKDARQYTIPVEAALTAV
RSKKAPELTTRDKHTRECFVVAEEGADLKAIEEAIVTMPNYFADYDTTVHFISQEELMRDH
AGIPHGGFVIRTGSTGWNDENGHVIEYSLKLDSNPEFTASVIASYARAAYRLSREGQSGCK
TVFDIAPAYLSAADGAELRKHLL
----------(End)---------