MSACompro, Release 1.1.0

MSACompro is a novel and practical multiple protein sequence alignment algorithm based on secondary structure, solvent accessibility, and contact map information.

This software is developed by Xin Deng, a PhD student in computer science department in University of Missouri-Columbia.

MSACompro1.1.0 is an updated version of MSACompro1.0.1. It can automatically predict secondary structure, solvent accessibility, as well as contact map information when running MSACompro1.1.0.

Users just need to run perl script auto_run_msacompro.pl.

```
*********************************************************
*            installation                     *
*********************************************************
```

To install MSACompro, users need to fulfill the following steps:

1. download pspro package, and install it according to its instruction. pspro is a bond software package of SSpro, ACCpro and NNcon.

You may have to set the installation dir(unzip dir), nr_database_dir,

big_database_dir, nr database name, big database name in configure.pl. Then go to the directory of pspro, and run ./configure.pl.

2. download MSACompro package and unzip it.

3. go into the ./MSACompro directory, and run make.

```
*********************************************************
*   How to use MSACompro1.1.0                 *
*********************************************************
```

There are three ways to use MSACompro1.1.0

I. Automatically use MSACompro1.1.0(default parameters) by running auto_run_msacompro.pl, secondary structure, solvent accessibility and contact map information is automatically predicted.

go to ./script directory, run ./auto_run_msacompro.pl arg1 arg2 arg3

arg1: path for pspro bin directory

arg2: given file of target protein sequences

arg3: resulting multiple sequence alignment file for target proteins

eg: ./auto_run_msacompro.pl
/storage/shared/pspro2/bin/ ../test/BB40004.fasta ../test/BB40004.msa

If users want to set the values of the parameters, then can change line 72 in auto_run_msacompro.pl.

eg. $msacommand = "../MSACompro/msacompro -w1 0.4 -w2 0.5 -wc 0.9 -cm ".$cmdir." -pspro ".$pspro."predict_ssa.sh ".$infile." >".$outfile;

can be changed to $msacommand = "../MSACompro/msacompro -w1 0.3 -w2 0.2 -wc 0.5 -cm ".$cmdir." -pspro ".$pspro."predict_ssa.sh ".$infile." >".$outfile;

II. users can prepare the data step by step, and then To implement multiple protein sequence alignment by MSACompro, following process need to be done:

1). prepare the secondary structure and solvent accessibility files for the target group of proteins by pspro 2.0.

go to ./script directory, run ./predict_ssa.pl arg1 arg2 arg3

arg1: the path for shell program predict_ssa.sh in pspro

arg2: given file of target protein sequences

arg3: directory containing all the predicted secondary structure/solvent accessibility files for all the proteins in target file

eg. ./predict_ssa.pl /storage/shared/pspro2/bin/predict_ssa.sh
/storage/homes/xd9d3/MSA_copy/MSA_project/data/BB1.fasta
/storage/homes/xd9d3/MSA_copy/MSA_project/data/BB1ssa/

2). predict the contact map files for the target group of proteins by pspro 2.0.

still go to ./script directory, run ./predict_cm.pl arg1 arg2 arg3

arg1: the path for shell program predict_ss_sa_cm_simple.sh in pspro

arg2: given file of target protein sequences arg3: directory containing all the predicted contact map files for all the proteins in target file

eg. ./predict_cm.pl /storage/shared/pspro2/bin/predict_ss_sa_cm_simple.sh /storage/homes/xd9d3/MSA_copy/MSA_project/data/BB1.fasta /storage/homes/xd9d3/MSA_copy/MSA_project/data/BB/

Each contact map file under the directory is supposed to be named the same as the head name of each protein sequence in the target file with a suffix _fasta.cm8a.

However, in some special cases such as some proteins in OXBENCH database, whose head names end with .1 or .2.

Eg. 1bnec-1-AUTO.2. After using pspro, the name of predicted contact map file changes to be 1bnec-1-AUTO_2_fasta.cm8a, which should be 1bnec-1-AUTO.2_fasta.cm8a.

Then users need do post-processing for the names of the predicted contact map files by running ./cm_name_trim.pl arg1

arg1: given directory containing the prediction files for the target group of sequences

3). start multiple protein sequence alignment by MSACompro.

There are some important parameters:

-w1 W1_SUB

users define their own weight w1 for the substitution matrix, range is 0<=w1<=1, users must set both w1 and w2

-w2 W2_SS

users define their own weight w2 for the secondary structure similarity matrix, range is 0<=w2<=1-w1, users must set both w1 and w2, if not, w1 and w2 are set as 0.4 and 0.5 respectively by default

-wc WCM

users define their own weight wc for the contact map score in calculating pairwise distance function, range is 0<=wc<=1, and wopt = 1-wc, wopt is the weight for optimal global alignment score in pairwise distance function. If users set -cm but didn't set -wc, then wc in our program is 0.9 by default

-ss ss_folder

users can provide the secondary structure information of each sequence by themselves. Just give the folder which contains separate secondary structure information file for each sequence of the target input file each secondary structure file should be exactly named by the sequence, eg. 1aqt.

In each file, the first line contains name of the sequence, the second contains the sequence, the third contains the corresponding secondary structure information of all the amino acids, eg. ECCCCHHHH........

-sa sa_folder

users can provide the solvent accessibility information of each sequence by themselves. Just give the folder which contains separate secondary structure information file for each sequence of the target input file

each solvent accessibility file should be exactly named by the sequence name, eg. 1aqt.

In each file, the first line contains name of the sequence, the second contains the sequence, the third contains the corresponding solvent accessibility information of all the amino acids, eg. eeebbbe........

-ssa ssa_folder

users can provide the folder which contains the file combining both the secondary structure and solvent accessibility information for each sequence of the target input file

For example, the folder format can be /data/ssa_predict/BAliBASE3.0/RV11/BB11001/,each file in this folder should be exactly named by the sequence name,eg. 1aqt.

In each file, the first line contains name of the sequence, the second contains the sequence, the third contains the corresponding secondary structure information of all the amino acids,eg. ECCCCHHHH........,the fourth contains the corresponding solvent accessibility information of all the amino acids, eg. eeebbbe........

Noted that if you choose -ss or -sa, then don't choose -ssa!

-cm cmdir

users can provide the contact map information of each sequence by themselves. Just give the folder which contains separate contact map information file for each sequence of the target input file

For example, the folder format can be /data/cm_predict/BAliBASE3.0/RV11/BB11001/,each file in this folder should be named as the sequence nme with suffix name "cm8a", eg.1aqt.cm8a

The fifth line is followed by n * n contact probability matrix. n is the length of the sequence.

III. Users don't need use pspro to provide secondary structure and solvent accessibility files in advance, MSACompro can also automatically generate secondary structure/solvent accessibility information when running, once the users use command:

-pspro pspro_path

users can provide the path for the shell program predict_ssa.sh under pspro software folder. for example, -pspro /data/pspro2/bin/predict_ssa.sh

For the usage of other parameters:

use "./msacompro -help" or "./msacompro -?" to get command line options

use "./msacompro infile >outfile" to get the multiple alginment in FASTA format.

****

*   Test MSACompro   *

****

go to ./MSACompro sub-directory

For case BB40004.fasta,

1) just use secondary structure and solvent accessibility information,

run: ./msacompro -w1 0.4 -w2 0.5 -ssa ../test/ssa_predict/BB40004/ ../test/BB40004.fasta >../test/BB40004_test_1.msa

compare: diff ../test/BB40004_test_1.msa ../test/BB40004_1.msa

If the same, that means you succeeded to install MSACompro

2) use combined information of secondary structure, solvent accessibility, and solvent accessibility

run: ./msacompro -w1 0.4 -w2 0.5 -wc 0.3 -ssa ../test/ssa_predict/BB40004/ -cm ../test/cm_predict/BB40004/ ../test/BB40004.fasta >../test/BB40004_test_2.msa

compare: diff ../test/BB40004_test_2.msa ../test/BB40004_2.msa

For case _____9,

1) just use secondary structure and solvent accessibility information,

run: ./msacompro -w1 0.5 -w2 0.3 -ssa ../test/ssa_predict/_____9/ ../test/_____9 >../test/_____9_test_1.msa compare: diff ../test/_____9_test_1.msa ../test/_____9_1.msa

2) use combined information of secondary structure, solvent accessibility, and solvent accessibility

run: ./msacompro -w1 0.5 -w2 0.3 -wc 0.5 -ssa ../test/ssa_predict/_____9/ -cm ../test/cm_predict/_____9/ ../test/_____9 >../test/_____9_test_2.msa

compare: diff ../test/_____9_test_2.msa ../test/_____9_2.msa