README: implementation for the "average over repeated measurements", "weighted similarity", "FITSS" approaches

The byte code files (.class files from java) for performing hierarchical agglomerative algorithms (avg_weighted_hie.tar.gz), k-means (avg_weighted_kmeans.tar.gz) and FITSS (fitss.tar.gz) are available. Theoretically, these byte code files should be platform independent, but we make absolutely no guarantee that they will run properly on your system.

To uncompress the archives:

- On linux:
  tar –xzvf avg_weighted_hie.tar.gz
  tar –xzvf avg_weighted_kmeans.tar.gz
  tar –xzvf fitss.tar.gz
- On windows: winzip can extract these files. There are probably other windows applications that can do it as well.

Format of input files:

Tab-delimited text files:

- The first row is a header row.
- Column 1: gene identifiers (eg. Gene names or ORF)
- Column 2: documentation or other gene names
- Column 3 onwards: expression values for repeated measurements of the same experiments are in consecutive columns
- Example input file: test_4rep_err.txt

Format of output files:

- When clustering genes: a tab-delimited file "Out<algorithm>_<number of clusters>.txt" with the following headings
  GeneID<tab>ORF<tab>GeneName<tab>Cluster#(number of clusters<newline>
  Eg. gene_OutAvgLinkCorrErr_2.txt
- When clustering experiments: a tab-delimited file "Out<algorithm>_<number of clusters>.txt" with the following headings
  ExptID<tab>Sample<tab> Cluster#(number of clusters<newline>
  Eg. expt_OutAvgLinkCorrErr_3.txt
- GeneID and ExptID are unique identifiers for genes and experiments. They follow the order of the genes or experiments appearing in the input file. For example, the first gene/experiment in the file is 0, and the second gene/experiment is 1 etc.
- Genes or experiments with the same cluster number belong to the same cluster. For example, gene #0 and gene #1 were assigned to the same cluster in gene_OutAvgLinkCorrErr_2.txt.

How to run the executables from "avg_weighted_avg"?

- These bytecode files produce clustering results from **hierarchical agglomerative** clustering algorithms using array data with repeated measurements, using either the "average over repeated measurements" approach or "variability-weighted similarity" approach.
- General format:

java hieclust -r <# genes> -c <# experiments> -NoC_range <range of number of clusters> -step <step size of number of clusters> -sim <similarity measure> -err <average over repeated measurements or variability weighted> -rep <# repeated measurements> -errOp <sd or cv or in> -obj <gene or expt> -alg <algorithm> inputfilename

- To see all these options:
  java hieclust –
- Interpretation of options:
  - -sim <similarity measure>: specifies the pairwise similarity measure where <similarity measure> is either **corr** (correlation coefficient) or **dist** (Euclidean distance)
  - -err <average over repeated measurements or variability weighted>:
    - **-err 0** means that the average over repeated measurements approach will be used
    - **-err 1** means that the variability weighted approach will be used
  - -errOp <sd or cv or in>: ignored if –err 0 is used.
    - **-errOp sd** means that standard deviation will be used to compute variability in variability-weighted approach.
    - **-errOp cv** means that coefficient of variation will be used to compute variability in variability-weighted approach.
    - **-errOp in** means that error estimates are given for each array measurements in the input file. In this case, # of repeated measurements is set to 2, where 2 values are given for each experiment in consecutive columns such that

<measured value for expt 1><tab><error for expt1><tab><measured value for expt 2><tab><error for expt2><tab> etc…

  - -obj <gene or expt>:
    - **-obj gene** means that genes are clustered (default, if unspecified)
    - **-obj expt** means that experiments are clustered
  - -alg <algorithm>, where <algorithm> can be:
    - **-alg avg** : average linkage (default, if unspecified)
    - **-alg centroid**: centroid linkage
    - **-alg complete**: complete linkage
    - **-alg single**: single linkage

- Examples:
  - To cluster the five genes in the input file "test_4rep_err.txt" to produce 2 to 3 clusters using **average-link with the SD-weighted correlation**:
    java hieclust -r 5 -c 6 -NoC_range 2 3 -step 1 -sim corr -err 1 -rep 4 -errOp sd -obj gene -alg avg test_4rep_err.txt

This command should produce clustering results "OutAvgLinkCorrErr_2.txt" and "OutAvgLinkCorrErr_3.txt".

- To cluster the 6 experiments in the input file "test_4rep_err.txt" to produce 2 to 3 clusters using **complete-link with Euclidean distance and average over repeated measurements**:

java hieclust -r 5 -c 6 -NoC_range 2 3 -step 1 –sim dist -err 0 -rep 4 -obj expt -alg complete test_4rep_err.txt

This command should produce clustering results "OutCompleteLinkDist_2.txt" and "OutCompleteLinkDist_3.txt".

How to run the executables from "avg_weighted_kmeans"?

- These bytecode files produce clustering results from **kmeans** using array data with repeated measurements, using either the "average over repeated measurements" approach or "variability-weighted similarity" approach.
- General format:

java kmeansclust -r <# genes> -c <# experiments> -NoC <number of clusters> -sim <similarity measure> -err <average over repeated measurements or variability weighted> -rep <# repeated measurements> -errOp <sd or cv or in> -obj <gene or expt> -init <initialization> inputfilename

- To see all these options:
  java kmeansclust –
- Interpretation of options:
  - -sim <similarity measure>: specifies the pairwise similarity measure where <similarity measure> is either **corr** (correlation coefficient) or **dist** (Euclidean distance)
  - -err <average over repeated measurements or variability weighted>:
    - **-err 0** means that the average over repeated measurements approach will be used
    - **-err 1** means that the variability weighted approach will be used
  - -errOp <sd or cv or in>: ignored if –err 0 is used.
    - **-errOp sd** means that standard deviation will be used to compute variability in variability-weighted approach.
    - **-errOp cv** means that coefficient of variation will be used to compute variability in variability-weighted approach.
    - **-errOp in** means that error estimates are given for each array measurements in the input file. In this case, # of repeated measurements is set to 2, where 2 values are given for each experiment in consecutive columns such that

<measured value for expt 1><tab><error for expt1><tab><measured value for expt 2><tab><error for expt2><tab> etc…

  - -obj <gene or expt>:
    - **-obj gene** means that genes are clustered (default, if unspecified)
    - **-obj expt** means that experiments are clustered
  - -init <initialization>, where <initialization> can be:

- **-init random**: initialize k-means with randomly chosen K objects as initial centroids, where K = number of clusters
- **-init avglink**: initialize k-means with centroids computed from results from average linkage hierarchical clustering algorithm (this option is slower than the other two options)
- **-init <clustering result>**: initialize k-means with centroids computed from a given clustering result in the same output format

- Examples:
  - To cluster the five genes in the input file "test_4rep_err.txt" to produce 3 clusters using **random initializations with the SD-weighted correlation**:
    java kmeansclust -r 5 -c 6 -NoC 3 -sim corr -err 1 -rep 4 -errOp sd -obj gene –init random test_4rep_err.txt
    This command should produce clustering result "OutKmeansCorrErr_3.txt".
  - To cluster the 6 experiments in the input file "test_4rep_err.txt" to produce 3 clusters using **average linkage hierarchical algorithm to compute initial centroids with Euclidean distance and average over repeated measurements**:
    java kmeansclust -r 5 -c 6 -NoC 3 –sim dist -err 0 -rep 4 -obj expt –init avglink test_4rep_err.txt
    This command should produce clustering results "OutAvgLinkDist_3.txt" and "OutKmeansDist_3.txt".
  - To cluster the five genes in the input file "test_4rep_err.txt" to produce 2 clusters using **initializations using file "gene_OutAvgLinkCorrErr_2.txt" with the SD-weighted correlation**:
    java kmeansclust -r 5 -c 6 -NoC 2 -sim corr -err 1 -rep 4 -errOp sd -obj gene –init gene_OutAvgLinkCorrErr_2.txt test_4rep_err.txt
    This command should produce clustering result "OutKmeansCorrErr_2.txt" .

How to run the executables from "fitss"?

- These bytecode files produce clustering results from **hierarchical agglomerative** clustering algorithms using array data with repeated measurements, using the FITSS approach. The FITSS (Force Into The Same Subtree) approach initializes hierarchical agglomerative algorithms by forcing expression values from repeated measurements into the same subtrees.
- General format:

java hieclustRep -r <# genes> -c <# experiments> -NoC_range <range of number of clusters> -step <step size of number of clusters> -sim <similarity measure> -rep <# repeated measurements> -alg <algorithm> inputfilename

- To see all these options:
  java hieclustRep –
- Limitation: can ONLY cluster genes, not experiments
- Interpretation of options:
  - -sim <similarity measure>: specifies the pairwise similarity measure where <similarity measure> is either **corr** (correlation coefficient) or **dist** (Euclidean distance)

- -alg <algorithm>, where <algorithm> can be:
  - **-alg avg** : average linkage (default, if unspecified)
  - **-alg centroid**: centroid linkage
  - **-alg complete**: complete linkage
  - **-alg single**: single linkage

- Examples:
  - To cluster the five genes in the input file "test_4rep_err.txt" to produce 2 to 3 clusters using **average-link, correlation and FITSS**:

    java hieclustRep -r 5 -c 6 -NoC_range 2 3 -step 1 -sim corr -rep 4 -alg avg test_4rep_err.txt

    This command should produce clustering results "OutAvgLinkCorrRep_2.txt" and "OutAvgLinkCorrRep_3.txt".