BdClust

Clustering of gene expression profiles or samples by Ben-Dor algorithm.
Algorithm description

The program allows clustering genes by their expression profile similarity. The purpose of the analysis is to select groups of genes that have common patterns of expression in different experiments, e.g. high expression in cancer tissues and low expression in normal tissues. These patterns of co-expression are usually treated as co-regulation. The similarity of the expressions patterns may not be limited by simple rules and can be described by similarity (or distance) Measures. There are several measures of expression profile similarity between two genes:

(1) *Euclidean distance.* This is the geometric distance in the multidimensional space. It is computed as: $d_{ij} = [\Sigma_k (x_{ik} - x_{ik})^2]^S$, where x_i, x_j are two expression profiles for genes i,*j* , *k* is the index of experiment (field), x_{ik} is the expression value of gene *i* in the experiment *k*.

(2) *Squared Euclidean distance.* The squared Euclidean distance can be implemented in order to place progressively greater weight on objects that are further apart. The squared Euclidian distance is computed as: $d_{ij} = \Sigma_k (x_{ik} - x_{ik})^2$ (see explanation above). The Euclidian and squared Euclidian distances are computed from raw data (non-standardized), therefore they may be affected by differences in scale among the expression values in different experiments.

(3) *Manhattan distance.* This distance is the average absolute difference for the set of experiments calculated by the formula $d_{ij} = \Sigma_k |x_{ik} - x_{ik}|$. In most cases, this distance measure yields results similar to the simple Euclidean distance, for this measure, the effect of single large differences is dampened (since they are not squared).

(4) *Chebychev distance.* This distance is computed as $d_{ij} = \max_k | x_{ik} - x_{ik} |$. The measure is useful when one wants to define two objects as "different" if they are different on any one of the experiments.

In SelTag all distance measures (1-3) are normalized to the number of fields involved in calculation. This is useful when take into account expression data with missing values.

Other measures involve correlation coefficient r_{ij} between two expression profiles of genes *i* and *j*.

(5) 1-r_{ij}; This measure keep close profiles with positive correlation coefficients and is useful when one wants to detect co-regulated genes.

(6) 1-|r_{ij}|; This measure keep close profiles with higher absolute value of correlation coefficients.

(7) 1+|r_{ij}|; This measure keep close profiles with negative value of correlation coefficients (anti-correlated).

Three types of correlation are possible for correlation distance option:
Pearson's *r* - Pearson's correlation coefficient. The Pearson product moment correlation coefficient between expression profiles *i* and *j* is calculated as follows:

$$r_{ij} = \frac{\sum_k (y_{ki} - \bar{y}_i)(y_{kj} - \bar{y}_j)}{\left(\sum_k (y_{ki} - \bar{y}_i)^2 \sum_k (y_{kj} - \bar{y}_j)^2\right)^{1/2}} \quad ,$$

where y_{ki} is the expression level of gene *i* in the experiment *k*; $\bar{y}_i$ is the mean expression level of the gene *i*. Positive correlation implies that the expression levels of genes *i*,*j* are related positively, the higher expression of gene *i*, the higher expression of gene *j*. Negative correlation means that the expression levels of genes *i*,*j* are related negatively, the higher expression of gene *i*, the lower expression of gene *j*. If the r_{ij} is close to zero, two expression profiles are unrelated.
Spearman *r* - Spearman's correlation coefficient.
This correlation coefficient is computed for ranks. Let R_{ki} is the rank of the expression level in the experiment *k* of gene *i* (relatively to other experiments), R_{kj} is the rank of the expression level in the experiment *k* of gene *j*. Then Spearman's correlation coefficient is calculated by the formula

$$r_{ij} = \frac{\sum_k (R_{ki} - \bar{R}_i)(R_{kj} - \bar{R}_j)}{(\sum_k (R_{ki} - \bar{R}_i)^2 \sum_k (y_{kj} - \bar{R}_j)^2)^{1/2}}$$

Kendall's τ - Kendall's *tau* correlation coefficient.
To calculate Kendall's τ K for data points (y_{ki}; y_{kj}) 2K(K − 1) pairs considered (without self-pairing, the points in either order count as one pair). Pairs in which $y_{ki} > y_{mi}$ and $y_{kj} > y_{mj}$ or $y_{ki} < y_{mi}$ and $y_{kj} < y_{mj}$ are called concordant pairs (agreement between ranks), pairs with rank disagreement are called discordant pairs. In general, τ is calculated as

τ = ([number of concordant] – [number of discordant]) / total number of pairs

Clustering algorithm

The program implements Cluster Affinity Search Technique (CAST), proposed by Ben-Dor et al [Ben-Dor A., Shamir R., Yakhini Z. (1999) *J. Comput. Biol*. 6, 281–297].

A common shortcoming of hierarchical clustering techniques, such as single-linkage, complete-linkage, group-average, and centroid, is due to their "greedy" nature, once a decision to join two elements in one cluster is made, it cannot be undone. The CAST algorithm use the "affinity" values to perform "cleaning" step while making clusters by removing low-affinity elements of the cluster. The affinity in the CAST algorithm is the average similarity between gene expression profile and gene profiles already included to the cluster. The threshold for affinity is user-defined.

Example of output data

```
status=Correlation matrix calculation...
status=CAST clustering...
status=done [0.0 sec]
Number of gene clusters obtained 4.
Cluster Sizes and Scores:
Cluster 1      2      1.7469
Cluster 2      10     1.6321
Cluster 3      7      1.7248
Cluster 4      4      1.6679
List of selected genes, their cluster indices and scores :
No     DataIndex     Name    Cluster Score
1      1      GEN30482       2      1.6892
2      2      GEN03437       2      1.6962
3      3      GEN03687       2      1.6649
4      4      GEN24649       2      1.6463
```

Some lines starting from "`status=`" are just output the status of the calculation and can be ignored. Then the result cluster information is output: number of clusters, their list with cluster scores. Then list of selected genes with their cluster indices and scores is printed out.