SelCorr

The program select most correlated genes for specified gene set.

Algorithm

The ***SelTag:SelCorr*** program allows selecting genes which have expression profiles highly correlated to the profile of the user-defined gene(s).
User should provide list of fields to calculate correlation.
Three types of correlation are possible:
Pearson's *r* - Pearson's correlation coefficient. The Pearson product moment correlation coefficient between expression profiles *i* and *j* is calculated as follows:

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

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

For the specified gene user can select other genes that have correlation coefficient between target gene expression profile greater than threshold. There are several threshold types: "Best N" - select N most correlated genes from set; "Best %" - select a fraction (in %) of most correlated genes from set; "Value" - select the genes with the absolute correlation value equal or higher than the threshold; "All" - select all genes from list.

If a number of genes are selected in target list, several options exist how to treat the correlation of profile with this groups of profiles: "Max. correlation value to select" - when comparing genes, the key parameter is the maximum coefficient of correlation of a gene from Set 1 with genes from Set 2; "Aver. correlation value to select" - when comparing genes from Set 1, the key parameter is the average coefficient of the correlation of a gene from Set 1 with genes from Set 2; "Corr. for aver. field values to select" - when comparing genes from Set 1, the key parameter is the coefficient of correlation of a gene from Set 2 with an average profile of genes from Set 2. This means that the program creates an "imaginary" average gene from Set 2 and uses this average value to calculate the correlation coefficient.
Example of the output data

```
status=Correlation matrix for cards...

status=Correlation matrix calculation...

status=done [0.0 sec]

List of selected genes [30 total]:

1     6718  X54232
```

```
2      4575  R81175
3      7132  X79981
4      5493  T78432
5      3454  R06627
6      5166  T59895
7      6042  U14394
8      6690  X52947
```

Some lines starting from "`status=`" just output the status of the calculation and can be ignored. Then the result information (with the number of selected genes) is output. Then list of selected genes with their indices in data file and gene names are printed out.