SIRAC : Supervised Identification of Region of Aberration in aCGH datasets

For a detailed description of the algorithm see the paper:

Lai, H.M. Horlings, M.J. van de Vijver, E.H. van Beers, P.M. Nederlof, L.F.A. Wessels, and M.J.T. Reinders. Sirac: Supervised identification of regions of aberration in acgh datasets. BMC Bioinformatics, 2007.

In the following, a step by step description of the algorithm scripts and execution is given.

Files description

The .zip files contains:

a) the directory Sirac_matlabcode where the main file Sirac.m is given
b) the subdirectory Functions, where all functions called by Sirac are provided
c) the subdirectory prtools4.1, with the software package PRTools required by the SAM routines
d) the subdirectory Exampledata which contains:
   1. exampleAmplifier: acgh data with the inputs required by Sirac; the dataset is a two class version of the dataset provided by: *"J. Fridlyand, A.M. Snijders, and B. Ylstra et al. Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer, 6(96), 20,* at http:// www.biomedcentral.com/1471-2407/6/96
   2. chromStartEnd.txt, text file with the start and end of each chromosome (from Ensembl build 32)
   3. cytoBand.csv, text file with the start and end of the cytoband (from Ensembl build 32)

Required INPUT

✓ acgh = struct array with the acgh data, the following fields are required:
- .data => matrix NxP with N samples and P DNA-probes
- .ch => vector 1xP with the chromosome in which the P probes are located
- .kbmidpos => vector 1xP with the midposition of the probes
- .kbcumdis => vector 1xP with the cumulative position along the genome of the probes
- .name => name of the data(to be used to save the results
- .cyto => string 1xP with the cytoband location of the genes (optional, if not given the cytoband are determined using the location and the cytoBand.csv file)

✓ l = vector Nx1 with the labels of the N samples (must be a two class problem)
✓ dok = delta parameter for the SAM analysis (if not known,use [] for default delta=[0.1:0.01:1]);
✓ fdrex = desired false discovery rate for the SAM analysis (default 0.05)
✓ utili = struct array with strings for plots title and legend, the following fields are required
- .class1 => string with the definition of class1 (default 'class A')
- .class2 => string with the definition of class1 (default 'class B')
- .title1 => default: 'class A/class B: median of the two classes'
- .title2 => default: 'class A/class B: relevant regions'
- .name => default: 'yourDataset';
- .savedir=> (optional) string with the directory where to save the results, if provided the results are saved

✓ param = struct array with parameter for the analysis
- .SAMiter => number of iteration for the SAM analysis (default: 1000);
- .WINDOW => length of half window (default [250, 500,1000:1000:12000])
- .th = 0.05 => threshold for the Bonferroni correction
- .numth => minimum number of scales in which the location has to be significant (default num=2)
-

✓ plotyes = binary variable, if plotyes =1 then the matlab figures with the results of the algorithm are shown (default plotyes = 0)
✓ mrna = (optional) mrna dataset with the same structure as the acgh dataset (if provided the genes in the identified aberrated regions are determined)

Algorithm execution

Unzip the SIRAC.zip and add SIRAC routines in your matlab path:

```
addpath SIRAC_matlabcode
addpath SIRAC_matlabcode/functions
addpath SIRAC_matlabcode/ExampleDATA
addpath SIRAC_matlabcode/prtools
```

Load the data

```
load ExampleDATA/exampleAmplifier
```

this will load the variable: acgh (the example dataset), param and utili

Execution 1st step: SAM analysis

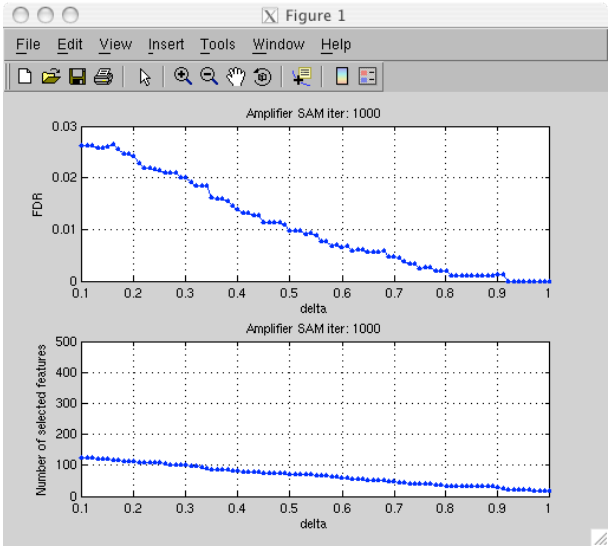
```
[res]=Sirac (acgh,acgh.l,[],0.005,utili,param,1)
```

The 1st step of the execution is the SAM analysis, the goal is to evaluate which DNA-probes significantly discriminate the two classes. If the parameter delta is unknown ([] in the input variables), a search is performed in order to select a value of delta that provides the desired FDR (0.005 in the example).

```
>fdr=0.0052  fdrex=0.0050 d= 0.70 n_called=   45
>Do you want to go further with this value of d (YES => 1)?
```

the above message is print on the matlab screen, it gives the delta value that provide the FDR closest to the desired one (fdrex), and the number of DNA-probes that are judged significant by the SAM analysis (n_called). (note that the values of the fdr and n_called may change for different execution of the algorithm, since the SAM analysis uses a randomize procedure to estimate the parameters)

The following plot (Figure 1) is also provided. It shows the FDR (upper plot) and the corresponding number of DNA-probes selected as relevant (lower plot) for different values of delta.
The higher the value of delta the stricter the FDR and the smaller the number of probes selected.
The plot is meant to assist the user in the selection of a different delta values.



If the user would like to change the delta, he/she should type any key different then 1 on the command line. The new value of delta will be then requested by Sirac.

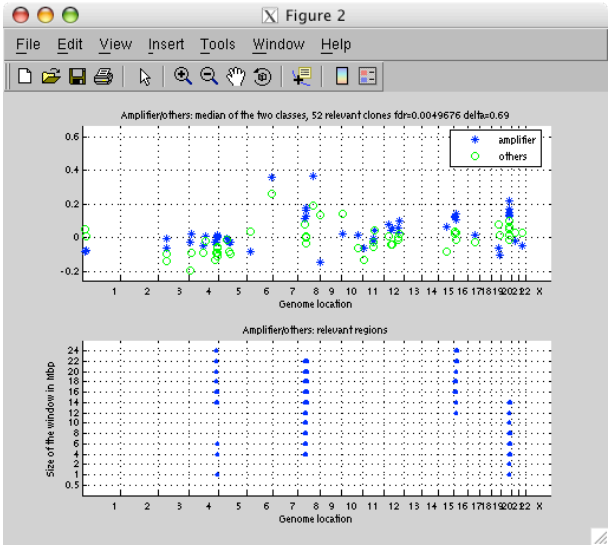
The search is repeated until the user is happy with the result, and types 1 on the command line when asked for confirming the delta value.

Execution 2nd step

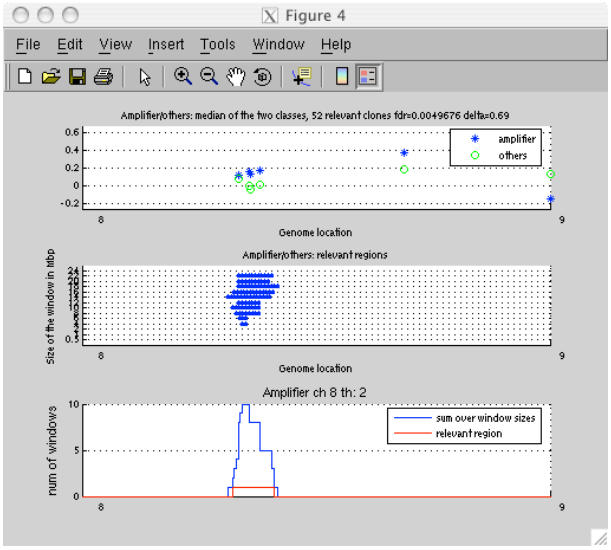
For each window scales (in param.WINDOW) a window is slid over the genome. The enrichment of relevant probes, within each window, is determined using the hypergeometric test. Significant windows at different scales are selected.

The results of the scale search are provided as:

- Summary plot (Figure 2). In the upper plot, each relevant DNA-probe selected by the SAM analysis is depicted with two different symbols: the mean of the "Amplifier" class and the mean of the remaining samples (class "Others"). In the lower plot the regions judged significant are depicted for the different window scales used.



- Detailed plots for each chromosome containing genome regions judged significantly enriched at some window scales. An example is given in Figure 4, which zooms on Chromosome 8. The first two plots are similar to the plots in Figure 2, only focusing on the specific chromosome. The third plot identifies the number of scales in which a genome region was judged significantly enriched of relevant probes. The red line shows the region judged significant in at least s different scales (here s=2, i.e. param.numth = 2).



Execution 3rd step

If the expression measurements are provided (in a struct format with the same fields required in the acgh variable), then Sirac searches for the genes that are present in the identified regions and provide the user with a list in a .txt file for each chromosome. The .txt file contains not only the gene names list, but also several information that can be used for further prioritization of the gene list. The information fields are listed below. For eample the F column provides the p-value of the t-test of the genes, which indicated how well the genes selected are able to distinguish the classes of interest. Column G provides the values of the correlation between the genes and the closest relevant DNA-probes, which indicate if the aberration of the DNA do affect the expression of the genes in the same region.

A. Gene name
B. cytoband
C. kb middle position
D. median of the gene for samples in class A
E. median of the gene for samples in class B
F. p-value of the t-test gene
G. correlation between the gene and the closest relevant DNA-probe
H. p-value of the correlation with the closest relevant DNA-probe
I. F*H
J. identifier (stored in the .gnames field) of the closest DNA-probe
K. kb middle DNA-probe position
L. distance gene/DNA-probe (in kb)
M. p-value t-test of the DNA-probe (in J)
N. number of windows where the DNA-probe (in J) is included in a region significantly aberrated
O. median of the DNA-probe in the samples of class A
P. median of the DNA-probe in the samples of class B