MAS5Baseline

Comparison of the Affymetrix gene expression row data to the baseline data by MAS 5.0 algorithm.

Data specification

The input for MAS5Baseline is the set of expression row data in Affymetrix CEL data format, corresponding CDF file and file with list of CEL files to be processed and their short description (this file is provided by user). The CEL file stores the results of the intensity calculations on the pixel values on the chip. The CDF file describes the layout for an Affymetrix GeneChip array. The output is SelTag data file with gene expression data. The baseline experiment name should be provided by user.

Algorithm description

The purpose of the algorithm is to perform noise correction and data normalization for each experiment and to estimate the change of the gene expression signal relatively to the baseline experiment signal. The method is known as MAS 5.0 statistical algorithm implemented in the Affymetrix Microarray Suite version 5.0. The algorithm details are described in the Affymetrix documentation at http://www.affymetrix.com/support/technical/technotesmain.affx ("Statistical Algorithms Description Document", Affymetrix, 2002; "Statistical Algorithms Reference Guide", Affymetrix, 2001).

The algorithm contains of several steps.

1. Background noise correction for baseline and experiment
2. Change of the expression value (signal change) calculation between experiment and baseline
3. Estimation of the signal change value statistical significance (change detection p-values)
4. Estimation of the of the signal change (change detection call)

Background noise correction. At the first step the chip area is divided into K squared zones of the same size (default number of zones is 16). Then the 2% probes with the lowest intensity define the background intensity for each zone. The background noise level for each *k*-th zone bZ^k is the calculated as the average for those lowest intensity probes. The background noise level *b(x,y)* for each probe at the chip location *x,y* is calculated as weighted sum of zone background values

$$b(x,y) = \frac{1}{\sum_{k=1}^{K} w_k(x,y)} \sum_{k=1}^{K} w_k(x,y) bZ_k$$

,

where weights wk(x,y) are calculated as follows:

$$w_k(x,y) = \frac{1}{d_k^2(x,y) + smooth}$$

,

where $d_k(x,y)$ is the distance from the point *x,y* to the center of the *k*-th zone, *smooth* - is the smoothing parameter (by default is 100).

The noise correction procedure is as follows. First, standard deviations of the 2% probes with the lowest intensity nZ_k are calculated for each zone. For each probe the noise intensity *n(x,y)* is is estimated by above formulas (substitute *n(x,y)* for *b(x,y)* and nZ_k for bZ_k in the formulas above). Then the probe intensity corrected for noise is calculated from actual probe intensity *I(x,y)* as follows:

A(x,y)=max(*I'(x,y)-b(x,y),NoiseFrac***n(x,y)*),

where *I'(x,y)*=max(*I(x,y)*,0.5), *NoiseFrac* is the fraction of noise and is set to 0.5 as in MAS 5.0 algorithm description.

Expression value (signal) calculation. After background subtraction from each probe intensity value, the signal values for the probesets are calculated. The calculation uses "ideal mismatch" technique that allows to

process probe pairs for which the mismatch (MM) signal is greater than the match (PM) signal (see details in the Affymetrix documentation). When the ideal mismatch is calculated for each probe pair *j* of the each probeset *i*, the probe value PV_{ij} is calculated: $PV_{ij} = \log_2(\max(PM_{ij}-IM_{ij}, 2^{-20}))$. The signal log value ($SLV_i$) for the probeset *i* is calculated as the one-step biweight estimate for the corresponding probeset SLVs. Then the algorithm scales all the probesets to target scale value *Sc* (default is 500) estimating the scale factor *sf*

$$sf = \frac{Sc}{TrimMean\left(2^{SignalLogValue_i}, 0.02, 0.98\right)}$$

and using normalization factor *nf*:

$$nf = \frac{TrimMean\left(SPVb_{i,}\ 0.02, 0.98\right)}{TrimMean\left(SPVe_{i,}\ 0.02, 0.98\right)}$$

where $SPVb_i$ is the baseline signal, $SPVe_i$ is the experiment signal, the scaled probe intensity values are calculated as $SPV_{ij}=PV_{ij}+ log_2(nf+sf)$. The *TrimMean* function calculates the mean value of the data without highest 2% and lowest 2% values. The probe log ratio *PLR* is calculated for probe pair *j* in probeset *i* on both the baseline *b* and experiment *e* arrays $PLR_{ij}={}_eSPV_{ij}-{}_bSPV_{ij}$. Having the probe log ratios *PLR* the *SignalLogRatio* is calculated using the biweight algorithm. *SignalLogRatio* is the reported value for this algorithm.

Estimation of the signal statistical significance (detection p-values). To estimate the significance of the change of the expression signal between experiment and baseline two additional sets of values for each probeset are calculated:

$$q_i = PM_i - MM_i, \ (i = 1, \ldots, n)$$

and

$$q_i = PM_i - MM_i, \ (i = 1, \ldots, n)$$

They are used to estimate two balancing factors:

$$nf = \frac{sfE}{sfB}$$

as the ratio of scaling factors of the of the *q* values for experiment *sfE* and baseline sfB data. The second balancing factor

$$nf_2 = \frac{sf_2E}{sf_2B}$$

is calculated as the ratio of scaling factors of the of the *z* values for experiment sf_2E and baseline sf_2B data. The balancing factor range is extended by using three balancing factors for the *q* values

$$f[0] = nf * d \qquad f[1] = nf \qquad f[2] = \frac{nf}{d}$$

and for *z* values

$$z_i = PM_i - b_i, \ (i = 1, \ldots, n)$$

where *d* is perturbation parameter and is set by default to 1.1.

If the algorithm settings indicate a user defined balancing factor and the factor is not equal to 1 then, *nf* = *nf2* = *user defined normalization factor*·*sfE* /*sfB*, where *sfE* is the experiment *sf* and *sfB* is the baseline *sf* as described in the ***Expression value (signal) calculation*** section.

The critical *p*-value is estimated for all three *f*[*k*] (*k*=0,1,2) parameters and are designated below as

p[0],*p*[1],*p*[2] correspondingly. These values are used to estimate the signal *p*-value for the signal change:

p=max(*p*[0],*p*[1],*p*[2]) if *p*[0] < 0.5, *p*[1] < 0.5 and *p*[2] < 0.5
p=min(*p*[0],*p*[1],*p*[2]) if *p*[0] > 0.5, *p*[1] > 0.5 and *p*[2] > 0.5
p=0.5 otherwise.

Estimation of the presence/absence of the signal (detection call). The algorithm report several types of detection calls in the output file: increase (I - is the designation of the detection call in the SelTag file), marginally increase but not increase (i), decrease (D), marginally decrease but not decrease (d), no change / unchanged (U). The definition of the detection change is dependent on several parameters: γ_1High, γ_1Low, γ_2High, γ_2Low, yielding two parameters γ_1 as linear interpolation of γ_1High and γ_1Low (if γ_1High = γ_1Low, then γ_1= γ_1High = γ_1Low), and 2 as linear interpolation of γ_2High and γ_2Low (if γ_2High = γ_2Low, then γ_2= γ_2High = γ_2Low).

The rule for the detection change is as follows:

$$\text{increase} \quad \begin{cases} p[0] < \gamma_1 \\ p[1] < \gamma_1 \\ p[2] < \gamma_1 \end{cases}$$

$$\text{marginally increase but not increase} \quad \begin{cases} p[0] < \gamma_2 \\ p[1] < \gamma_2 \\ p[2] < \gamma_2 \end{cases}$$

$$\text{decrease} \quad \begin{cases} p[0] > 1 - \gamma_1 \\ p[1] > 1 - \gamma_1 \\ p[2] > 1 - \gamma_1 \end{cases}$$

$$\text{marginally decrease but not decrease} \quad \begin{cases} p[0] > 1 - \gamma_2 \\ p[1] > 1 - \gamma_2 \\ p[2] > 1 - \gamma_2 \end{cases}$$

The MAS 5.0 default values for the gamma parameters are: γ_1High=0.0025, γ_1Low=0.0025; γ_2High=0.003, γ_2Low=0.003 (for 16-20 probe pairs).

Example of experiment list file

```
GSM42890        DEHP_48hr_Veh1  DEHP 48hr Veh1
GSM42891        DEHP_48hr_Veh2  DEHP 48hr Veh2
GSM42892        DEHP_48hr_Veh3  DEHP 48hr Veh3
GSM42893        DEHP_48hr_Veh4  DEHP 48hr Veh4
GSM42894        DEHP_48hr_Veh5  DEHP 48hr Veh5
```

This file contains three columns separated by symbol. First column is the experiment data name (the corresponding CEL file should start from this name and have extension *.cel, for example GSM42890.cel). Second column is the name of the variable in the output SelTag file, corresponding to this experiment (see below example of SelTag output file). This column should not contain spaces. Third column is the extended description of the experiment that will appear at the SelTag file header section.

Example of output data

```
#HEADER
Multiple chip data analysis by Affymetrix MAS5.0 algoritm [comparison with baseline].
ChipName=RG_U34A.
    BaselineDataFilename=GSM42895.cel.cel
    BaselineDataHeader=Baseline experiment
```

```
    BaselineDataScalingFactor=3.0104
    BaselineDataNormalizationFactor=1.0000
    BaselineDataSignalTrimmedMean=500.0000

  1 ExperimentDataFilename=GSM42907.cel
  1 DataHeader=VPA_48hr_Ve        VPA 48hr Veh POOLED
  1 DataScalingFactor=2.3930
  1 DataNormalizationFactor=1.0000
  1 DataSignalTrimmedMean=500.0000

  2 ExperimentDataFilename=GSM42913.cel
  2 DataHeader=DEHP_48hr_t        DEHP 48hr treated POOLED
  2 DataScalingFactor=2.6396
  2 DataNormalizationFactor=1.0000
  2 DataSignalTrimmedMean=500.0000

MAS5 algorithm parameters:
BF=2.0000
NZ=2.0000
Bsmooth=100.0000
Alpha1=0.0400
Alpha2=0.0600
Gamma1H=0.0025
Gamma1L=0.0025
Gamma2H=0.0030
Gamma2L=0.0030
Perturbation=1.1000
Tau=0.0150
TGT=500.0000
#ENDHEADER
ProbesetName     STRING
VPA_48hr_Ve_SignalLogRatio        FVALUE
VPA_48hr_Ve_Change       WORD
VPA_48hr_Ve_Change_p     FVALUE
DEHP_48hr_t_SignalLogRatio        FVALUE
DEHP_48hr_t_Change       WORD
DEHP_48hr_t_Change_p     FVALUE
END
DATA
AFFX-MurIL2_at   -0.0952  U       0.32868  -0.3230  U       0.28164
AFFX-MurIL10_at  0.5692   U       0.12112  0.3852   U       0.66645
AFFX-MurIL4_at   -0.1952  U       0.16996  -0.3095  U       0.30476
AFFX-MurFAS_at   -1.3517  U       0.49464  -0.2080  U       0.04914
AFFX-BioB-5_at   -0.7911  D       0.99998  0.0126   U       0.79768
AFFX-BioB-M_at   -0.7021  D       1.00000  -0.2708  D       0.99997
AFFX-BioB-3_at   -0.5249  D       0.99998  -0.4171  D       0.99987
```