

## Image Quantitation and Flagging for CodeLink™ Bioarrays



The CodeLink™ Bioarray System includes software for quantitative image data extraction. This application note provides detailed descriptions of the image quantitation and spot flagging methodologies, and the data export fields employed by the CodeLink Expression Analysis software (versions 3 through 5).

### Introduction

The last step of a laboratory bioarray processing workflow consists of scanning the bioarray to generate a Tagged Image File Format (TIFF) image of each bioarray run. For robust gridding, we recommend that the image borders surrounding the arrays be nearly equivalent in size, that is, the arrays should be centered within the image in both the x and y directions. As a general guideline, set the borders on all sides of the array to approximately 1.0-2.0 mm. The TIFF images are then analyzed to quantitate spot intensities. CodeLink Expression Analysis software provides automated spot finding and data extraction for all CodeLink Bioarrays, allowing either individual or batch processing. After a TIFF image is submitted, the software will automatically find all spots within the image and extract the corresponding signal intensity. Additionally, the software will use the bioarray's unique serial number to specify the gene identification and description for each spot within the bioarray. The CodeLink Expression Analysis software evaluates the quality of each spot, and displays the results both numerically and graphically. This application note describes the image quantitation attributes and quality flag terms.

### Image Quantitation: Description of Background and Signal Calculations

#### *Spot measurements calculated by the software and reported*

- Spot mean intensity
- Spot median intensity
- Spot standard deviation of intensity
- Spot area
- Spot diameter
- Spot total intensity
- Spot density
- Spot position
- Raw intensity
- Signal strength
- Spot background mean
- Spot background median
- Spot background standard deviation

**Note:** There is a buffer around the spot. The buffer is typically a few pixels and is used to prevent inclusion of any signal bleeding into the computation of the background signal.

**Note:** Many of the plots provided in the analysis portions of the software display a general noise level. This is calculated as the average local background of the entire bioarray.

#### **Spot Intensity Calculation**

To accurately quantitate bioarray spot values, the spot itself must first be physically defined, and a representative background area selected. Based upon the CodeLink™ bioarray product design, a particular spot diameter is expected per product (*specDiameter*). In addition, empirical testing of each product has determined a set of

expected variations of the spot size (defining *spExtend* and *spReduce*). The actual spot radius is determined for each spot on the bioarray through data extraction, and spot signal area – the number of pixels per spot (Fig. 1, region A) – is determined through automated computations, as described below. Based on these, the basic spot mean signal intensity can then be calculated. The spot mean is the sum of all pixel intensity values in the spot area divided by the number of pixels.

To account for the contribution of background to the spot signal intensity, a local background region is defined as a concentric ring around the spot of several pixels in width (Fig. 1, Region C). To ensure that no spot signal bleeds into the defined background area, a concentric buffer area of several pixels separates the spot signal area from the background area (Fig. 1, region B). The raw intensity is then defined as the mean of the pixel intensities within the area of interest (spot mean: Fig. 1, region A) minus local background (Fig. 1, region C).

### **Spot Area Calculation**

Number of pixels in the local region (Fig. 1, region A), includes:

- 1) All pixels inside a circle  $R \leq (\text{spotRad} - \text{spotReduce})$
- 2) All pixels from  $(\text{spotRad} - \text{spotReduce}) \leq R$  area with  $\text{pixelSignal} - \text{bkgrMedian}/\text{bkStdDev} > 1.5$
- 3) Pixels from  $R \leq (\text{spotRad} + \text{spotExtend})$  area with  $\text{pixelSignal} - \text{bkgrMedian}/\text{bkStdDev} > 5$

where

R = distance from spot center

spotRad = spot radius found by Data Extraction

spotExtend = *spExtend* x *specDiameter*

spotReduce = *spReduce* x *specDiameter*

*spExtend* = 10% (Config file)

*spReduce* = 10% (Config file)

*SpecDiameter* = diameter from product information

### **Spot Diameter Calculation**

Computation:

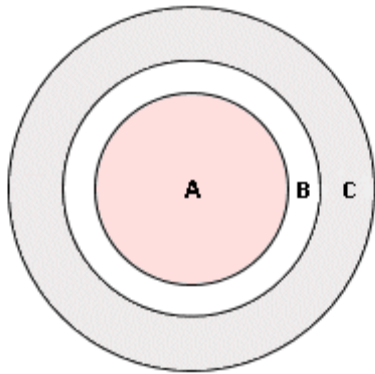
$$\text{SQRT}(\text{spArea})/3.14$$

$$\text{minDiam} \times \text{specDiameter} \leq \text{maxDiam} \times \text{specDiameter}$$

where

$$\text{minDiam} = 70\%, \text{maxDiam} = 110\% \text{ (from configuration file)}$$

$$\text{specDiameter} = \text{diameter from product information}$$



**Fig. 1.** Delineation of the spot regions in the calculation of various spot values. The pixels the software uses when computing the local background mean, median, and standard deviation. In the diagram above, **A** indicates the spot signal area, **B** the background buffer for which pixels are excluded from the background computations, and **C** the background area.

### ***Background Intensity Calculation***

To calculate the background area, the software determines the width of region C. If two spots are very close together, such that the diameter of the spot is less than one-half the spot pitch (the distance between two adjacent spots), the software divides the spot pitch by 2 and uses the result for the width. The software then uses the resulting C width to compute the area of C. The local background is then determined as the median signal of the pixels within the background ring C (Fig. 1).

It is important to understand the magnitude of the spot signal relative to its local background. The signal strength, given by the following calculation, is a metric that allows this type of comparative assessment.

### ***Signal Strength Calculation***

The signal strength is calculated as follows:

$$\text{signal strength} = 100\% \times (\text{signal mean} - \text{local background median}) / \text{local background standard deviation}$$

**Note:** When the signal strength is below 1.0, the spot is flagged as "L" (i.e. below noise).

### **Description of Quality Flag Calculations**

CodeLink™ Expression Analysis software also assesses the quality of each spot through pre-defined metrics and assigns quality flags as tools to allow the researcher to utilize only the highest quality microarray data. These quality flags vary in their definitions and thresholds for assignment. The definitions for each flag are given below.

### **Quality Control Flag Definitions**

Below is a description of the quality control flags assigned by the CodeLink™ software.

**The CodeLink™ software excludes the following flagged spots in the analysis:**

**M:** The spot is identified to be defective through image inspection at manufacturing, and is flagged in this category to indicate exclusion from further analysis. No expression value is provided for this flag.

**X:** The spot was manually excluded from the analysis (by user).

The following spot signal values are considered questionable and may be included in an analysis at the user's discretion:

**C:** The spot has a high level of background contamination. Its background is above the global background population.

**I:** The spot has an irregular shape.

**L:** The spot has a signal that is below local background noise.

**S:** The spot has a high number of saturated pixels.

*Note: Any spot can have a combination of flags assigned to it, such as "CI," "LC," or any other flag combination.*

**Computation of Flags**

**Irregular spot (I):** The spot has an irregular shape. Below are the calculation details.

The flag is switched on when:

$$QVI > shlrreg$$

$$QVI = 100\% \times (Dmax - Dmin)/Dmax$$

where

$$Shlrreg = 30\%$$

The shape irregularity flag is applied only on spots with signal strength > 4.

**Signal below noise (L):** The spot has a signal that is below local background noise.

The flag is switched on when:

$$Signal\ mean < local\ background\ median + 1.5 \times local\ background\ standard\ deviation$$

**Background contamination (C):** The spot has a high level of background contamination. Its background is above the global background population. Below are the calculation details.

The flag is switched on when:

$$100\% \times (local\ Bkgr\ Mean - local\ Bkgr\ Median)/local\ Bkgr\ Median > bkCont$$

where **bkCont** = 10% (Config file)

or

$$globalBkgrMedian - localBkgrMedian > 7 \times globalBkgrMedianStdDev$$

where

*globalBkgrMedian* = median of all *localBkgrMedian* values calculated for the whole image

*backMedianStdDev* = standard deviation of all *localBkgrMedian* values calculated for entire image

**Saturation (S):** The spot has a high number of saturated pixels.

The flag is switched on:

If % [nSaturated/spotArea] > ***saturThreshold***

Pixel is saturated if pixel intensity > ***saturLevel***

where

**nSaturated** refers to number of pixels saturated. The default level is 20%.

### Data Fields in Text Export from CodeLink™ Software

All data derived from CodeLink™ Expression Analysis software can be exported into a standard text output file. This data will include information on the physical spot characteristics, such as size and array position, the spot probe identity and annotation information, as well as spot intensity and background data. This output will also include any quality flags assigned to each spot. Table 1 lists the parameters exported to the text file.

**Table 1. Description of the Data Headers Within the Standard Text Output File**

Parameter Name	Description
Idx	Counter (or index) that tracks the number of records (rows) that were exported
Array	Array position in the bioarray. The array index starts at the top left corner of the image and increments from the left to the right, and then continues in the subsequent row.
Sample_name	Name of the biological sample that is profiled by the bioarray, and was input by the user during scanning
Accn	Accession number, the gene information from one of the four NCBI database fields
Probe_name	Name of the probe
Annotation_PIN	Probe ID, which is a unique identifier for the probe sequence in the CodeLink™ WEBB database
Annotation_NCBI_Acc	Accession number
Annotation_LocusLink	Locus Link ID
Annotation_OGS	Official gene symbol
Annotation_UniGene	Unigene ID
Annotation_ENSEMBL	Ensembl ID
Annotation_NCBI_NID	GI number of the nucleotide sequence
Probe_type	Probe type (see below)
Feature_ID	Unique identifier for a spot within an array
Raw_intensity	Spot density, which is the difference between the spot mean and the

	local background median
Normalized_intensity	Normalized intensity value (the raw intensity divided by the normalization factor)
Quality_flag	Metrics the software uses to indicate the quality of a spot (see previous section)
Signal_strength	Estimated signal above its local background noise level. The ratio is expressed as the spot mean divided by the local background standard deviation
Logical_raw	Raw number of the bioarray where the probe is physically located
Logical_col	Column number of the bioarray where the probe is physically located
Center_X	X coordinate, in pixels, of the physical position of the spot center in the image. The origin is at the top right corner of the image; therefore the X coordinate increases as you move to the left from the origin.
Center_Y	Y coordinate, in pixels, of the physical position of the spot center in the image. The origin is at the top right corner of the image; therefore the Y coordinate increases as you move downward from the origin.
Spot_mean	Sum of all pixel values in the spot area divided by the number of pixels
Spot_median	Median pixel intensity computed over the spot area
Spot_stdev	Standard deviation of the pixel intensity in the spot
Spot_area	Number of pixels in the spot
Spot_diameter	Spot diameter (in pixels) that the software calculates during data extraction
Spot_noise_level	Local background noise level, given by the median signal of the pixels within the background ring C (Fig. 1)
Bdgd_mean	Estimate of the background mean that the software computes, using the pixel intensities in the background area surrounding the spot
Bkgd_median	Estimate of the background median the software computes, using the pixel intensities inside the background area surrounding the spot
Bkgd_stdev	Estimate of the background standard deviation that the software computes, using the pixel intensities inside the background area surrounding the spot
Bkgd_area	Number of pixels that the software includes to compute the estimated local background mean, median, and standard deviation
Annotation_Molecular Function	Gene Ontology Consortium GO term ID for molecular function
Annotation_Biological Process	Gene Ontology Consortium GO term ID for biological process
Annotation_Cellular Component	Gene Ontology Consortium GO term ID for cellular component
Annotation_Cytoband	Cytological band position
Annotation_HS_Homology	NCBI UniGene cluster number of the homologous Homo sapiens gene, if available
Annotation_MM_Homology	NCBI UniGene cluster number of the homologous Mus musculus gene, if available
Annotation_RN Homology	NCBI UniGene cluster number of the homologous Rattus norvegicus gene, if available
Annotation_Analogous_CodeLink	Analogous CodeLink™ probe for homologous genes in humans
Annotation_Legacy_Probe_Name	Probe name used in previous version of the software
Description	Description of the probe

## Probe Type Definitions

Each probe or spot on the array is assigned one of seven different probe types, based on the use of the probe or spot in the microarray experiment. Each of the seven probe type definitions is described below, and probe definition assignments can be found in the standard text output file.

**D:** Discovery probes – probes corresponding to the genes of interest for a particular species

**P:** Positive controls – probes designed against *E. coli* transcripts that can be ‘spiked’ during the assay at either the total RNA or cRNA level. These controls can be used for quality control by evaluating sensitivity and dynamic range of the platform.

**N:** Negative controls – probes designed against *E. coli* genes that are used for evaluating the degree of nonspecific assay background

**F:** Fiducial spots – labeled probes that emit a high level of fluorescence which are only used during gridding

**B:** Blank spots – area where a spot is expected, but no probe has been deposited

**M:** Mismatch spots – probes used internally throughout development to determine specificity of platform

**O:** Other – probes that do not meet above categories and are not used for data analysis; probes used by development

## Ordering Information

CodeLink™ Expression Analysis v5.0 software

**310035**

Applied Microarrays, Inc. (“AMI”) reserves the right, subject to any regulatory approval if required, to make changes in specifications and features shown herein, or discontinue the product described at any time without notice or obligation. Contact your AMI Representative for the most current information. © 2007 Applied Microarrays, Inc. – All rights reserved. The AMI logo and CodeLink™ are trademarks of Applied Microarrays, Inc.