Blob Metadata and Statistic for Two-Dimensional Chromatography (GCxGC and LCxLC)

GC Image, LLC. 03/2021

A blob is a set of samples (or pixels) identified with a single separated chemical. Each blob has its own distinct metadata and statistics. Blob metadata are informational fields set by the user, either interactively or automatically with blob templates. Blob statistics are numeric values computed from a two-dimensional chromatogram. The software records and reports various metadata and statistics about each blob. Configuration controls allow the user to select which metadata and statistics are included (or excluded) in blob reports.

Blob Metadata

The available metadata fields for each blob B are:

- *B.CompoundName* the name of the separated chemical compound
- B.ConstellationName the name of a constellation of chemical compounds
- B.GroupName the name of a group of chemical compounds
- B.Inclusion a flag indicating whether (or not) the blob is to be included in blob reports

The compound name allows individual descriptive names for each blob. Constellations and groups are constructs for defining named subsets of blobs.

The user has only indirect control over one other metadata field:

• B.InternalStandard — the associated internal standard

The user cannot directly set the value of B.InternalStandard, but can control it indirectly. The user can flag an included blob as an internal standard, in which case the value of B.InternalStandard is the blob itself. For included blobs that are not flagged as an internal standard, the initial value of B.InternalStandard is the nearest blob that is flagged as an internal standard. (If no blobs are flagged as an internal standard, then all internal standard associations are set to null or 0.) For individual included blobs that are not flagged as an internal standard, the user can manually cycle the value of B.InternalStandard through all of the possible internal standards. When internal standards are added or removed, all internal standard associations are recomputed.

Blob Statistics

The blob statistics can be defined mathematically using the following conventions. Let (x, y) designate a sample, where integer x indexes samples with respect to the first-column separation and integer y indexes samples with respect to the second-column separation. Let B be a set of samples defining a blob for a separated chemical. Let $(x, y) \in B$ have the value *True*, if the sample indexed by (x, y) is in the blob B, and have the value *False*, otherwise. Let p(x, y) be the value of the sample indexed by (x, y).

For many of the statistics, the units of (x, y) are configurable to either *Pixel* (i.e., column index and row index) or *Time* (with the first-column retention-time in minutes and the second-column retention-time in seconds). The equations in this white paper use the *Pixel* units of the raw data. The conversions from *Pixel* units (x_p, y_p) to *Time* units (x_t, y_t) are:

 $x_t = x_p \times modulation_cycle_time$ $y_t = y_p \times sampling_cycle_time$

The Reference List at the end of this whitepaper, indicates which statistics have configurable units.

Blob ID

Each blob has a unique numeric identification number:

B.BlobID

Blob IDs are positive integers numbered during blob detection according to the sorted order of the peak values. So, the blob with the largest peak value in a two-dimensional chromatogram has *B.BlobID* equal to 1; the blob with the second largest peak value has *B.BlobID* equal to 2; and so on. Note that the peak value after smoothing is used for ID numbering. If blobs are deleted, the blob ID values of remaining blobs are not changed. The blob ID of blobs added manually may not be indicative of peak order.

Peak

The peak of each blob is described by the indices of the sample with the largest value and the peak sample value:

$$(B.PeakI, B.PeakII) = (X, Y) \text{ s.t. } (\forall (x, y) \in B, p(X, Y) \ge p(x, y))$$
$$B.PeakValue = p(B.PeakI, B.PeakII)$$

The peak value is the actual data value (without smoothing) of the peak sample.

The interpolated peak in each dimension is estimated to sub-pixel precision by piecewise cubic convolution (PCC) of the row and column projections:

$$B.InterpolatedPeakI = PCC_peak(h_B(x))$$

$$B.InterpolatedPeakII = PCC_peak(v_B(y))$$

where PCC_peak interpolates a continuous function from the one-dimensional projection and determines the maximum value within one sample unit either side of the largest projected sample value. The projections are:

$$h_B(x) = \sum_{\substack{y|(x,y)\in B}} p(x,y)$$
$$v_B(y) = \sum_{\substack{x|(x,y)\in B}} p(x,y)$$

Note that the computations are more complicated if a blob crosses two thermal modulation cycles — starting near the end of one modulation cycle and ending after the start of the next modulation cycle. In the image, the blob appears to wrap around off the top to the bottom of the image. To compute the interpolated peaks of wrap-around blobs, the wrap-around samples are treated as if they extend the initial second-column separation of the blob (rather than occur in the subsequent second-column separation). This may result in a value of *B.InterpolatedPeakII* that is greater than the dimension of the second-column separation (in pixels or time), indicating that the blob-peak, second-column retention-time exceeded the second-column modulation period.

Note that the point indicated by the pair of values (B.InterpolatedPeakI, B.InterpolatedPeakII) is not the same as the peak of the interpolated two-dimensional function. The interpolated peaks are computed in a separable manner, not by two-dimensional interpolation.

Bounding Boxes

The bounding box is the smallest rectangular region oriented with the axes that contains all samples in the blob:

$$\begin{array}{rcl} B.StartI &=& X \text{ s.t. } (\exists (X,y) \in B) \text{ and } (\forall (x,y) \in B, x \geq X) \\ B.EndI &=& X \text{ s.t. } (\exists (X,y) \in B) \text{ and } (\forall (x,y) \in B, x \leq X) \\ B.StartII &=& Y \text{ s.t. } (\exists (x,Y) \in B) \text{ and } (\forall (x,y) \in B, y \geq Y) \\ B.EndII &=& Y \text{ s.t. } (\exists (x,Y) \in B) \text{ and } (\forall (x,y) \in B, y \leq Y) \end{array}$$

Note that the mathematical definitions of *B.StartII* and *B.EndII* are more complicated for wraparound blobs. To compute the start and end of wrap-around blobs, the wrap-around samples are treated as if they extend the initial second-column separation of the blob (rather than occur in the subsequent second-column separation). So, the index *B.EndII* may be reported as a value that is greater than the dimension of the second-column separation (in pixels or time), indicating that the blob started near the end of one modulation cycle and ended after the start of the next modulation cycle.

Additional bounding boxes can be defined for the samples with values that are at least w% of the peak value:

$$B.StartI(w) = X \text{ s.t. } (\exists (X, y) \in B, p(X, y) \ge B.PeakValue \times w/100) \text{ and} \\ (\forall (x, y) \in B, (x < X) \to (p(x, y) < B.PeakValue \times w/100))$$

The user can configure the value of w to record a parametric bounding box. The default value of the parametric bounding box is w = 10, giving the bounding box for values 10% of peak value. Note that B.StartI = B.StartI(0), B.EndI = B.EndI(0), B.StartII = B.StartII(0), and B.EndII = B.EndII(0). As for the full bounding box, the values of these boxes for wrap-around is computed as if the wrap-around samples extend the initial second-column separation (rather than occur in the subsequent second-column separation).

The bounding box for w = 50 is commonly used and so is always available (in addition to the full bounding box and the parametric bounding box):

B.StartI(50)B.EndI(50)B.StartII(50)B.EndII(50)

The dimensions of the bounding boxes are available. The equations for reporting in *Pixel* units are:

$$\begin{array}{rcl} B.SizeI &=& B.EndI-B.StartI+1\\ B.SizeII &=& B.EndII-B.StartII+1\\ B.SizeI(w) &=& B.EndI(w)-B.StartI(w)+1\\ B.SizeII(w) &=& B.EndII(w)-B.StartII(w)+1\\ B.SizeI(50) &=& B.EndI(50)-B.StartI(50)+1\\ B.SizeII(50) &=& B.EndII(50)-B.StartII(50)+1\\ \end{array}$$

The value of w is the same as used for the parametric bounding box.

Symmetry

Simple measures of the symmetry of the bounding boxes with respect to the peak indices are available:

$$B.SymmetryI = \frac{B.EndI - B.PeakI + 0.5}{B.PeakI - B.StartI + 0.5}$$
$$B.SymmetryII = \frac{B.EndII - B.PeakII + 0.5}{B.PeakII - B.StartII + 0.5}$$

B.SymmetryI(w)	=	$\frac{B.EndI(w) - B.PeakI + 0.5}{B.PeakI - B.StartI(w) + 0.5}$
		B.EndII(w) - B.PeakII + 0.5
B.Symmetry II(w)	=	$\frac{B.PeakII - B.StartII(w) + 0.5}{B.PeakII - B.StartII(w) + 0.5}$
B.SymmetryI(50)	=	B.EndI(50) - B.PeakI + 0.5
		$\overline{B.PeakI - B.StartI(50) + 0.5}$
B.Symmetry II(50)	=	B.EndII(50) - B.PeakII + 0.5
		$\overline{B.PeakII - B.StartII(50) + 0.5}$

For wrap-around blobs, the computation for symmetry treats the wrap-around samples as if they extend the initial second-column separation of the blob (rather than occur in the subsequent second-column separation). In this, the peak location values are adjusted accordingly.

This measure of symmetry is determined by only the extrema samples of the blobs. Moments account for all samples of the blobs and may provide better measures of blob symmetry.

Unweighted Moments

The unweighted moments are useful in describing the shape of the area occupied by the blob (but not the volume defined by its values). A generic unweighted moment is defined as:

$$u_{a,b} = \sum_{(x,y)\in B} x^a y^b$$

For wrap-around blobs, the computation of unweighted moments treats the wrap-around samples as if they extend the initial second-column separation of the blob (rather than occur in the subsequent second-column separation).

The zero-order unweighted moment is the area (*i.e.*, the number of samples in the blob):

$$B.Area = u_{0,0}$$

The first-order unweighted moments, normalized by the area, are the averages of the indices:

$$B.MiddleI = u_{1,0}/B.Area$$

 $B.MiddleII = u_{0,1}/B.Area$

The second-order unweighted moments, normalized by the area and the average indices, are useful for measuring the spread of the samples:

$$\begin{split} B.SpreadI &= \sqrt{\sum_{(x,y)\in B} (x-B.MiddleI)^2/B.Area} \\ &= \sqrt{u_{2,0}/B.Area - (B.MiddleI)^2} \\ B.SpreadII &= \sqrt{\sum_{(x,y)\in B} (y-B.MiddleII)^2/B.Area} \\ &= \sqrt{u_{0,2}/B.Area - (B.MiddleII)^2} \end{split}$$

Spread is used as the denominator of other statistics (e.g., shape), so a value of zero is problematic. To address this problem, the minimum width of each blob is taken to be one sample and the spread is set to a minimum $\frac{1}{12}$ if the computed value is less than that value.

Two parametric unweighted moments, normalized by the area, the average indices, and the spread, also are provided to further describe the shape:

$$B.ShapeA = \sum_{(x,y)\in B} ((x - B.MiddleI)/B.SpreadI)^{a_{u1}} ((y - B.MiddleII)/B.SpreadII)^{a_{u2}}/B.Area$$

$$B.ShapeB = \sum_{(x,y)\in B} ((x - B.MiddleI)/B.SpreadI)^{b_{u1}} ((y - B.MiddleII)/B.SpreadII)^{b_{u2}}/B.Area$$

Because the sample values are important for statistical descriptions of the blob, weighted moments (described below) may be more useful statistics.

Weighted Moments

The weighted moments are useful in describing the shape of the blob volume. A generic weighted moment is defined as:

$$m_{a,b} = \sum_{(x,y)\in B} x^a y^b p(x,y)$$

For wrap-around blobs, the computation of weighted moments treats the wrap-around samples as if they extend the initial second-column separation of the blob (rather than occur in the subsequent second-column separation).

The zero-order weighted moment is the volume (*i.e.*, the sum of the sample values in the blob):

$$B.Volume = m_{0,0}$$

The software also reports the percent response, which is the volume of the blob divided by the sum of the volumes of all blobs:

$$B.PercentResponse = B.Volume / \sum_{i} B_{i}.Volume$$

The first-order weighted moments, normalized by the volume, index the center-of-gravity of the blob:

$$B.CenterI = m_{1,0}/B.Volume$$

 $B.CenterII = m_{0,1}/B.Volume$

The second-order weighted moments are useful for measuring the variance of the blob values. The weighted variance and covariance are normalized by the volume and center-of-gravity:

$$B.VarianceI = \sum_{(x,y)\in B} (x - B.CenterI)^2 p(x,y) / B.Volume$$

$$= m_{2,0}/B.Volume - (B.CenterI)^{2}$$

$$B.VarianceII = \sum_{(x,y)\in B} (y - B.CenterII)^{2}p(x,y)/B.Volume$$

$$= m_{0,2}/B.Volume - (B.CenterII)^{2}$$

$$B.Covariance = \sum_{(x,y)\in B} (x - B.CenterI)(y - B.CenterII)p(x,y)/B.Volume$$

$$= m_{1,1}/B.Volume - (B.CenterI)(B.CenterII)$$

Variance is used as the denominator of other statistics (e.g., correlation), so a value of zero is problematic. To address this problem, the minimum width of each blob is taken to be one sample and the variance is set to a minimum $\frac{1}{12}$ if the computed value is less than that value.

The weighted standard deviations and correlation are defined from the variances and covariance:

$$B.DeviationI = \sqrt{B.VarianceI}$$

$$B.DeviationII = \sqrt{B.VarianceII}$$

$$B.Correlation = B.Covariance/(B.DeviationI \times B.DeviationII)$$

The orientation, inertia, and eccentricity are defined from the second-order moments:

$$B.Orientation = \tan^{-1} \left(2(B.Covariance)/(B.VarianceI - B.VarianceII) \right) / 2$$

$$B.Inertia = B.VarianceI + B.VarianceII$$

$$B.Eccentricity = (B.VarianceI - B.VarianceII)^2 + 4(B.Covariance)$$

Note that inertia and eccentricity are the second-order moments (for continuous functions) that are invariant with respect to affine transforms.

The number of theoretical plates are defined from the center and variances:

$$B.PlatesI = (B.CenterI)^2/B.VarianceI$$

 $B.PlatesII = (B.CenterII)^2/B.VarianceII$

Skewness is a measure asymmetry based on the third-order moments, normalized by the volume, centerof-gravity, and standard deviation:

$$B.SkewnessI = \sum_{(x,y)\in B} ((x - B.CenterI)/B.DeviationI)^3 p(x,y)/B.Volume$$

$$= \left(\frac{m_{3,0}}{B.Volume} - 3(B.CenterI)m_{2,0}/B.Volume + 2(B.CenterI)^3 \right)$$

$$/(B.DeviationI)^3$$

$$B.SkewnessII = \sum_{(x,y)\in B} ((y - B.CenterII)/B.DeviationII)^3 p(x,y)/B.Volume$$

$$= \left(\frac{m_{0,3}}{B.Volume} - 3(B.CenterII)m_{0,2}/B.Volume + 2(B.CenterII)^3 \right)$$

$$/(B.DeviationII)^3$$

Kurtosis is a measure of flatness based on the fourth-order moments, normalized by the volume, centerof-gravity, and standard deviation:

$$B.KurtosisI = \sum_{(x,y)\in B} ((x - B.CenterI)/B.DeviationI)^4 p(x,y)/B.Volume = (m_{4,0}/B.Volume - 4(B.CenterI)m_{3,0}/B.Volume + 6(B.CenterI)^2m_{2,0}/B.Volume - 3(B.CenterI)^4) /(B.DeviationI)^4 B.KurtosisII = \sum_{(x,y)\in B} ((y - B.CenterII)/B.DeviationII)^4 p(x,y)/B.Volume = (m_{0,4}/B.Volume - 4(B.CenterII)m_{0,3}/B.Volume + 6(B.CenterII)^2m_{0,2}/B.Volume - 3(B.CenterII)^4) /(B.DeviationI)^4$$

Two parametric weighted moments, normalized by the volume, center-of-gravity, and standard deviation, also are provided to further describe the blob volume:

$$B.WeightA = \sum_{(x,y)\in B} ((x - B.CenterI)/B.DeviationI)^{a_{m1}}((y - B.CenterII)/B.DeviationII)^{a_{m2}} \times p(x,y)/B.Volume$$

$$B.WeightB = \sum_{(x,y)\in B} ((x - B.CenterI)/B.DeviationI)^{b_{m1}}((y - B.CenterII)/B.DeviationII)^{b_{m2}} \times p(x,y)/B.Volume$$

Percent Response and Volume Ratio with Internal Standard

The percent response is the ratio of the blob volume to the sum of the volumes of all blobs as a percent:

$$B.PercentResponse = 100 \times B.Volume / \sum_{B_n} B_n.Volume$$

The volume ratio with the internal standard is reported as:

$$B.VolumeRatio = B.Volume/B.InternalStandard.Volume$$

where B.InternalStandard is the internal standard associated with blob B.

Noise and Error

The sample noise for the blob is described by the estimated noise standard deviation σ_n at the peak sample and the signal-to-noise ratio (SNR) is reported as the ratio of the peak value to the noise:

$$B.Noise = \sigma_n(B.PeakI, B.PeakII)$$

$$B.SNR = B.PeakValue/B.Noise$$

If sample noise is uncorrelated, then the standard error of the volume estimate is the product of the noise standard deviation and square-root of the number of samples. Both the volume standard-error and the ratio of the volume to the volume standard-error are computed:

$$B.Error = B.Noise\sqrt{B.Area}$$

 $B.VNR = B.Volume/B.Error$

Column-Dependent Statistics

Several statistics are derived from the following column characteristics:

- C.DiameterI and C.DiameterII diameters of columns in cm
- C.LengthI and C.LengthII lengths of columns in cm
- C.FlowRateI and C.FlowRateII[DimensionI] flow rates of columns in ml/min
- C.VoidVolumeI and C.VoidVolumeII[DimensionI] void volume in ml
- *C.VoidTimeI* and *C.VoidTimeII*[*DimensionI*] void time or dead time, the retention time of unretained compounds

where DimensionI and DimensionII are the dimensions of a two-dimensional chromatogram.

The adjusted retention time indices are:

B.AdjustedTimeI = (B.PeakI - C.VoidTimeI)B.AdjustedTimeII = (B.PeakII - C.VoidTimeII[B.PeakI])

The capacity factors or capacity ratios are:

B.CapacityFactorI = B.AdjustedTimeI/C.VoidTimeI B.CapacityFactorII = B.AdjustedTimeII/C.VoidTimeII[B.PeakI]

The plate height is:

$$B.HETPI = C.LengthI/B.PlatesI$$

 $B.HETPII = C.LengthII/B.PlatesII$

Nearest Blob

Several statistics are defined relative to the nearest neighboring blob, including distance to the nearest peak, selectivity with the nearest peak, and resolution with the nearest blob:

B.NearestBlob	=	B_n s.t. $(B_n \neq B)$ and $\forall B_j \neq B$,
		$\sqrt{(B.PeakI - B_n.PeakI)^2 + B.PeakII - B_n.PeakII)^2}$
		$\leq \sqrt{(B.PeakI - B_j.PeakI)^2 + (B_0.PeakII - B_j.PeakII)^2}$
B.Separation	=	$\sqrt{(B.PeakI-B.NearestBlob.PeakI)^2 + (B.PeakII-B.NearestBlob.PeakII)^2}$
B.SelectivityI	=	Larger(B.AdjustedTimeI, B.NearestBlob.AdjustedTimeI)
		/Smaller(B.AdjustedTimeI, B.NearestBlob.AdjustedTimeI)
B.Selectivity II	=	Larger (B. Adjusted Time II, B. Nearest Blob. Adjusted Time II)
		/Smaller(B.AdjustedTimeII, B.NearestBlob.AdjustedTimeII)
B.ResolutionI	=	B.PeakI - B.NearestBlob.PeakI
		/(B.DeviationI + B.NearestBlob.DeviationI)
B.Resolution II	=	B.PeakII - B.NearestBlob.PeakII
		/(B.Deviation II + B.NearestBlob.Deviation II)
B.Resolution	=	$B.Separation/\sqrt{B.Inertia}$

Reference List

B.AdjustedTimeI	Difference between peak and void time *†
B.AdjustedTimeII	Difference between peak and void time *†
B.Area	Number of samples in blob
B.BlobID	Unique identification number of the blob
B.CapacityFactorI	Ratio of adjusted time to void time
B.CapacityFactorII	Ratio of adjusted time to void time
B.CenterI	Center-of-gravity ^{*†}
B.CenterII	Center-of-gravity ^{*†}
B.CompoundName	Name of the separated chemical compound
B. Constellation Name	Name of a constellation of chemical compounds
B.Correlation	Weighted correlation coefficient
B.Covariance	Weighted covariance
B.Deviation I	Square-root of weighted variance *
B.Deviation II	Square-root of weighted variance *
B.Eccentricity	Weighted second-order moment related to eccentricity
B.EndI	End of bounding box *†
B.EndI(50)	End of bounding box *†
B.EndI(w)	End of bounding box *†
B.EndII	End of bounding box *†

End of bounding box $*^{\dagger}$ B.EndII(50)B.EndII(w)End of bounding box $*^{\dagger}$ B.Error Standard error of volume B.GroupName Name of a group of chemical compounds B.HETPIRatio of column length to number of plates **B.HETPII** Ratio of column length to number of plates **B.Inclusion** Flag indicating inclusion for reporting Weighted second-order moment related to inertia B.Inertia B.InternalStandard Associated internal standard B.InterpolatedPeakIInterpolated peak of projected values *[†] Interpolated peak of projected values *[†] B.InterpolatedPeakIIB.KurtosisI Weighted fourth-order moment related to flatness B.KurtosisII Weighted fourth-order moment related to flatness B.MiddleIIndex average *[†] Index average *† B.MiddleII B.NearestBlobBlob ID of nearest peak B.NoiseSample noise at peak sample **B**.Orientation Weighted second-order moment related to orientation B.PeakIIndex of largest valued sample *[†] B.PeakIIIndex of largest valued sample *[†] B.PeakValue Largest value Ratio of volume to sum of all blob volumes as percent B.PercentResponse B.PlatesIRatio of center squared to variance B.PlatesIIRatio of center squared to variance **B.**Resolution Ratio of separation from nearest peak to inertia B.ResolutionIRatio of separation from nearest peak to variance B.Resolution IIRatio separation from nearest peak to variance B.PercentResponse Ratio of volume to sum of volumes of all blobs B.SNRRatio of peak value to sample noise B.SelectivityIRatio of adjusted times with nearest peak B.SelectivityIIRatio of adjusted times with nearest peak B.SeparationDistance to nearest peak $B.ShapeA(a_{u1}, a_{u2})$ Parametric unweighted moment $B.ShapeB(b_{u1}, b_{u2})$ Parametric unweighted moment B.SizeISize of bounding box * B.SizeI(50)Size of bounding box for values 50% of peak * B.SizeI(w)Size of bounding box for values w% of peak * B.SizeIISize of bounding box * Size of bounding box for values 50% of peak * B.SizeII(50)B.SizeII(w)Size of bounding box for values w% of peak * B.SkewnessI Weighted third-order moment related to asymmetry B.SkewnessII Weighted third-order moment related to asymmetry B.SpreadIIndex standard deviation *

B.SpreadII	Index standard deviation *
B.StartI	Start of bounding box ^{*†}
B.StartI(50)	Start of bounding box for values 50% of peak *†
B.StartI(w)	Start of bounding box for values w% of peak *†
B.StartII	Start of bounding box * [†]
B.StartII(50)	Start of bounding box for values 50% of peak *†
B.StartII(w)	Start of bounding box for values w% of peak *†
B.SymmetryI	Symmetry ratio of bounding box
B.SymmetryI(50)	Symmetry ratio of bounding box for values 50% of peak
B.SymmetryI(w)	Symmetry ratio of bounding box for values w% of peak
B. Symmetry II	Symmetry ratio of bounding box
B.Symmetry II(50)	Symmetry ratio of bounding box for values 50% of peak
B.Symmetry II(w)	Symmetry ratio of bounding box for values w% of peak
B.VNR	Ratio of volume to standard error of volume
B.VarianceI	Weighted variance *
B.Variance II	Weighted variance *
B.Volume	Sum of sample values
B.VolumeRatio	Volume ratio with the internal standard
$B.WeightA(a_{m1}, a_{m2})$	Parametric weighted moment
$B.WeightB(b_{m1}, b_{m2})$	Parametric weighted moment

* These statistics can be reported in *Pixel* or *Time* units as configured by the user.

[†] These statistics are changed by phase shift.

Programming Notes

The software also computes and records a bit mask for the samples in the bounding box, specifying inclusion of samples in the blob. This structure cannot be accessed directly by the user. Only a small subset of the statistics (including those required to efficiently recompute the other blob statistics) are stored in the GC Image (.gci) file. Some statistics, called primary statistics, are computed during the scan of the image (either during blob detection or during file loading) and stored in the runtime data structure. Others statistics, called secondary statistics, can be computed from the primary features and are not stored in the runtime data structure.