

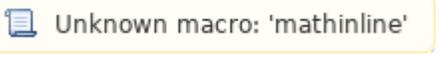
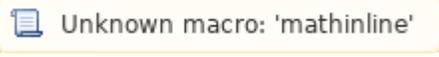
## 4.3.1 - RNA Hydrolysis

### Parameters

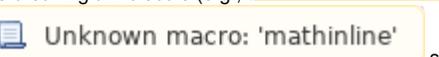
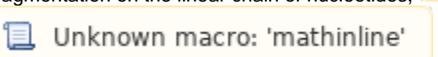
Parameter Name	Variable	Default Value	Parameter Range	Description
FRAG_UR_DO		1	>0	minimum length of fragment
FRAG_UR_DE LTA		NaN <sup>1</sup>		geometry of the fragment: 2=surface-diameter, 3=volume; if not explicitly specified (NaN), depends logarithmically on length
FRAG_UR_ETA		NaN <sup>1</sup>		intensity of fragmentation, breaks per unit length; if not explicitly specified (NaN), is determined by the correlation of the number of fragments and the expected fragment size, if size selection is used

<sup>1</sup> NaN stands for "Not a Number" and marks the uninitialized state of a parameter

### Algorithm

Frequencies  of fragment sizes  produced by a uniform random fragmentation process have demonstrated to fall along Weibull distributions , if the fragmentation thermodynamics depends on the molecule size:



Scale parameter  represents the intensity of fragmentation (i.e., breaks per unit length), and—as a determinant of the mean expected fragment size—is assumed to be constant across molecules of different lengths for fragmentation protocols where the number of produced fragments depends on the molecule length. Shape parameter  reflects the geometric relation in which random fragmentation is breaking a molecule (e.g.,  corresponds to uniform fragmentation on the linear chain of nucleotides,  splits uniformly the surface, and  the volume, etc.).

The Flux Simulator uses a 3-step algorithm to tokenize a molecule; first, geometry  $\frac{L}{h}$  and the number  $n$  of fragments that are obtained from the molecule are determined. We found empirically that parameter

$d$  depends logarithmically on  $L$ , the length of the molecule that is fragmented

$n$ . The number of fragments produced from a specific RNA molecule is determined by

$n$ , where  $\frac{L}{h}$  is the expectancy of the most abundant fragment size, computed from  $h$  and the gamma-function  $\Gamma$  of

$n$ :

$$n = \frac{L}{h} \Gamma\left(\frac{L}{h}\right)$$

Second,  $n$  breakpoints are sampled uniformly from the interval  $[0,1]$ , resulting in relative length fractions

$\frac{L_i}{L}$  for all  $i$  fragments. Third, relative fragment sizes

$\frac{L_i}{L}$  are transformed from unit space to sizes  $L_i$  that follow

a Weibull distribution of shape  $\frac{L}{h}$  by:

$$L_i = L \left(\frac{L_i}{L}\right)^{\frac{L}{h}}$$

where  $\frac{L}{h}$  is a constant of the transformation to ensure that the sizes of the

$n$  fragments sum up exactly to the given molecule length

$n$ . Latter transformation produces a slightly distorted Weibull distribution for the sizes

$L_i$ , however the deviation is sufficiently small to be neglected in our applications.