4.4.1 - Size Selection

Parameter

Parameter Name	Default Value	Parameter Range	Description
FILTERING	false	{true, false}	switches filtering for fragment sizes on/off
SIZE_DISTRIBUTION			specifies the probability distribution over fragment sizes to be retained in the final set. The first possibility to provide a distribution is by a string that describes parameters of a normal distribution in the form N(mean, sd),e.g., N(800, 200); the alternative is to provide as parameter value the name of a file that contains an empirical distribution.
SIZE_SAMPLING	AC	{RJ,AC,MH}	identifies the method used for sub-sampling fragments: either rejection-sampling (RJ), a variant of rejection-sampling with a minimal rejection rate (so-called acceptance sampling, AC), or the well known Metropolis-Hastings algorithm (MH) can be selected

Algorithm

If a size filtering step is carried out (parameter FILTERING), each fragment gets first assigned a probability according to the provided distribution (SIZE_DISTRIBUTION); either normal distributions can be characterized by their characteristic attributes "mean" and "standard deviation", or empiric distribution can be provided in the form of a file. Subsequently, fragments are selected according to one of the following sub-sampling algorithms:

Rejection Sampling (RJ) - a Bernoulli trial is carried out directly against the probability assigned by the provided distribution, which then decides whether the fragment is retained or discarded. Fragment sizes obtained by this algorithm distribute as the distribution specified by SIZE_DISTRIBUTION.

Acceptance Sampling (AC) - a modification of rejection sampling which stretches the maximum value of the probabilities in SIZE_DISTRIBUTION to 1 before applying rejection filtering. The method shows a higher yield of retained fragments, and largely preserves the characteristics of the provided size distribution.

Metropolis Hastings (MH) - a Montecarlo Markov Chain algorithm is employed in the sub-sampling process. Iteratively applied assumptions on the posterior distribution may distort the shape of the distribution, especially if the initial fragment size distribution and the provided filter distribution (SIZE_DISTRIBUTION) differ significantly from each other. However--at these costs--the algorithm looses less fragments than RJ or AC.