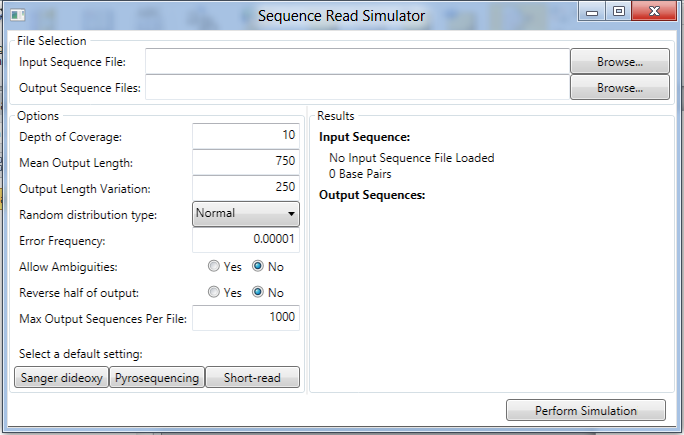
# Read Simulator

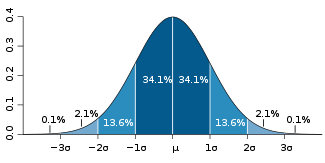
Read Simulator is an application for generating artificial reads from a given reference sequence mimicking sequencing machine outputs. This application can be used for generating test data for sequence assembly algorithms.

On running the ReadSimulator application, the following **Sequence Read Simulator** window is displayed:



The window is divided into four areas plus the **Perform Simulation** button:

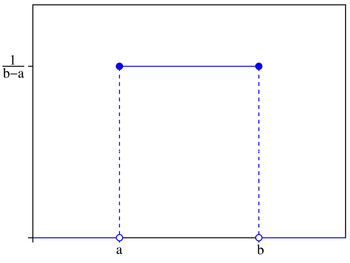
1. **File Selection** 
   1. **Sequence Input File**: The input sequence file name in any supported format.
   2. **Output Sequence Files**: The output sequence file name. This is where the generated reads will be written. If multiple output files are needed, the file names will be appended with a number indicating the order in which they were produced. The output files will be written in the FASTA format.
2. **Options**
   1. **Depth of Coverage:** The number of times on average that each base in the input sequence is represented across all output sequences.
   2. **Mean Output Length:** The mean length of the output sequences (values are limited to a range of 30 to 2000).
   3. **Output Length Variation:** For uniform distributions this represents the range in which the output length may vary from the mean length. For normal distributions this represents the standard deviation from the average.
   4. **Random Distribution Type:** The read lengths follow random distributions.
      1. **Normal Distribution:** Normal distribution is an absolutely continuous probability distribution.



**Fig 1: Normal distribution**

About 68% of values drawn from a normal distribution are within one standard deviation *σ* > 0 away from the mean *μ*; about 95% of the values are within two standard deviations and about 99.7% lie within three standard deviations.

* + 1. **Uniform Distribution:** The continuous uniform distribution is a family of probability distributions such that for each member of the family, all intervals of the same length on the distribution's support are equally probable. The support is defined by the two parameters, *a* and *b*, which are its minimum and maximum values. The distribution is often abbreviated *U*(*a*,*b*).



**Fig 2: Continuous Uniform Distribution where a = µ - σ and b = µ + σ**

* 1. **Error Frequency:** The likelihood that any given sequence base will be incorrectly copied (values range from 0.00001 to 1.0.). These errors are added to mimic SNP’s (Single Nucleotide Polymorphisms) read errors.
  2. **Allow Ambiguity:** Allow use of ambiguous characters when applying errors to base copying.
  3. **Reverse half of output:** Half the reads will be represented as a reverse complement.
  4. **Max Output Sequences Per File**: The maximum number of sequences to store in a single output file. When the output file reaches this maximum number, a new output file will be started for the remaining sequence outputs.

1. **Select Default Setting:** These platform settings will update the input options to closely mimic the specified sequencing platforms.
   1. **Sanger dideoxy:** The key principle of the Sanger method is the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.
   2. **Pyrosequencing:** This method is based on the "sequencing by synthesis” methodology. It relies on the detection of pyrophosphate release on nucleotide incorporation.
   3. **Short-Read:** Short reads are generated by next-generation sequencing platforms.
2. **Results**: Presents information on the loaded input sequence and the resultant output sequences.
   1. **Input Sequence**: The input sequence that was loaded and the number of base pairs.
   2. **Output Sequences**: The number of sequences that were generated, the number of files parsed, notification when the simulation is complete and the output format.

**Perform Simulation Button:** Selecting this option will start the read simulation which generates the simulated reads from the Sequence Input File.