

LCRep

Program for mapping low complexity regions in nucleotide sequences.

Search for the low complexity regions is performed with using Shannon's information measure. Shannon's information is defined as follows:

$$H = - \sum_{i=1}^k P(a_i) \log_2 P(a_i)$$

where: $\{a_1, \dots, a_k\}$ is the alphabet of the size k , and $P(a_i)$ is a fractional composition of a_i

The search is carried out as follows. For each position i of the sequence S calculation of the Shannon's information $H(i, l)$ is performed in the window of size l within the range $[l_{begin}, l_{end}]$. If $H(i, l)$ turns out below prespecified threshold $H_{thr}(l)$ then fragment $[i, i+l]$ is declared low complex. Intersection of all such fragments at the end of calculation gives a map of low complexity regions of the sequence S .