TandemRep-P

Program for mapping the Tandem Repeats Regions in protein sequences.

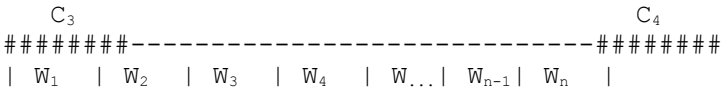
TandemRep mapping is performed by searching regions with uniform dinucleotide composition. The searching is initiated for the regions flanked by short ideal repeated elements.

Tandem searching algorithm consists of the following stages:

1) Find a pair of l-plets C_1 and C_2 with a distance between C_1 and C_2 not exceeding predefined N. The region between and including C_1 и C_2 will be denoted as R_1 with the length L_1. If C_1 and C_2 overlap then tandem unit size can be found trivially, jump to p.5.

2) Implying that C_1 and C_2 flanks do not contain insertions/deletions, extend synchronously C_1 and C_2 allowing 1 mismatch per several matches. Extended C_1 and C_2 we will denote as C_3 and C_4. After this operation the region will be denoted as R_2 with the length L_2 (>= L_1). If extension performed without mismatches and C_3 and C_4 overlap then we have ideal tandem which unit size again can be found trivially, followed by jump to p.5. If extension performed with mismatches and C_3 and C_4 overlap then we have almost ideal tandem which unit size can be found according p.4 (jump to p.4). Proceed if C_3 and C_4 do not overlap.

3) Now region R_2 looks as follows



For the region R_2 perform the following test. Divide region into set of windows W_1, …, W_n, each of size U. Consequently compare mono- (or di-) plet composition of the windows W_1 and W_i. If the difference in such composition between W_1 and some window W_i exceeds predefined threshold then stop. Test is not passed, jump to the p.1 to consider the next pair of l-plets. If the difference is low for all windows W_2, …, W_n then the test is passed and at least fragment R_2 could be declared tandem region.

Since we don't know the size of the window at which test described above could be passed, the test is performed for the window sizes U = 2, …, $L_2/2$.

Remember the lowest U at which the test is passed. Denote it U_1.

3a) Since uniform mono- (or di-) plet composition does not guarantee homology in windows W_1 and W_i, at this step the identity calculated by cycled Smith-Waterman algorithm is used for the additional filtering. If such an identity does not exceed predefined threshold then calculation is stopped for the C_1 and C_2 pair.

4) Calculate more precisely unit size U_{opt} of the tandem using two small windows synchronously sliding at the distance U one from another, U changes from U_1 to $L_2/2$.

5) Using U_{opt} calculated at the previous step find precise margins of the tandem using again two small synchronously sliding windows.

Such a procedure is carried out for all pairs C_1 and C_2 possible in the sequence. The final map of the tandems is an interception of tandems found for all l-plet pairs.