Automatic identification of Transposable Elements

ID - 142

Del Fabbro Cristian[1,2], Cattonaro Federica[2], Policriti Alberto[1,2], Morgante Michele[3,2]
[1]Department of Mathematics and Computer Science, University of Udine, Udine
[2]Istituto di Genomica Applicata, Udine
[3]The Department of Agriculture and Environmental Sciences, University of Udine, Udine

Motivation

Our knowledge of the structure and components of genomes is rapidly progressing in pace with their sequencing. The emerging data show that a significant portion of genomes is composed of transposable elements (TEs) The proper annotation of repeated sequences and transposable elements (SINEs, LINEs, DNA-retrotransposons, Helitrons, etc. see [1]) in an assembled genome, is both important and computationally hard.

Typically researchers in this area use a combination of visual tools - most often Dotter[2] and variants of it - and alignment tools - typically members of the Blast[3] suite - against know proteins databases in order to combine information on repeated sequences and coding.

[1] 'A unified classification system for plant transposable elements' Thomas Wicker, Francois Sabot, Jeffrey Bennetzen, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, Alan Schulman - in preparation [2] 'A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis' Erik L.L. Sonnhammer and Richard Durbin, Gene 167:GC1-10 (1995) [3] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) 'Basic local alignment search tool.' J. Mol. Biol. 215:403-410

Methods

Results

Our goal is to develop a tool for automatic annotation of repeated sequences, integrated within a system capable to search and classify coding regions characterizing specific TEs.

We have developed a prototype taking as input information on repeated sequences in gff format (provided as output by most practical tools for this kind of search like, for example, RepeatMasker[3]). The information we need is only the relative position in a genomic target sequence (e.g. a coordinate in a contig) and the name, of an aligned sequence from a data set S of generic repeated sequences (this information can also be computed de novo by software tools like, for example, ReAS[4]).

Using this information we build an simplified representation of our target string (e.g. our contig) written in an extended alphabet A. One letter of A will encode a sequence in S. Consequently, a word in A encodes a sequence of overlapping repeating fragments. Two A-words at distance d are separated by d occurrences of a letter N not in A. Our contig becomes a sequence of words in A*, with words separated by a stretches of Ns. Both visual and computational analysis of the simplified representations of our contigs (i.e. the encoded versions) can be carried out using Dotter. In order to do this in a practical manner we can first re-encode A into an isomorph version of it in which every character corresponds uniquely to a word in the amminoacids alphabet: ARNDCQEGHILKMFPSTWYVBZX. A letter in A becomes a unique sequence of amminoacids and, using an ad hoc cost matrix, we use Dotter for search, visualization, and analysis of a patterns of repeated requences within contigs.

The next step is to analyze the output of Dotter using information about structured sequences corresponding proteins commonly occurring in TEs. This step is performed using a characterization of patterns typical of families of transposable elements by structured motifs and SmartFinder[5], a tool for structured motif search expressly developed for TE search.

[3] Smit AFA, Hubley R. & Green P. - 'RepeatMasker' at http://www.repeatmasker.org [4] ReAS: Recovery of Ancestral Sequences for Transposable Elements from the Unassembled Reads of a Whole Genome Shotgun' Ruiqiang L., Jia Y., Songgang L., Jing W., Yujun H., Chen Y., Jian W., Huanming Y., Jun Y., Gane K. W., and Jun W.

[5] 'Structured Motifs Search' M. Morgante, A. Policriti, N. Vitacolonna, A. Zuccolo - Journal of Computational Biology, Vol. 12, no. 8, October 2005

Email: delfabbro@dimi.uniud.it