

Paper reference

Zhi,D., Raphael,B., Price,A., Tang,H. and Pevzner,P: **Identifying repeat domains in large genomes**. Genome Biology 2006, 7, R7..

Abstract

This paper presents a novel approach for the visualization and analysis of repeat domains (repeated subsequences amongst repeat families). Using the framework of A-Bruijn graphs, which generalize de Bruijn graphs to allow for the representation of imperfect similarities between sequences (as frequently encountered in DNA), the authors construct a repeat domain graph that represents of the structure of repeat families. In this graph, there are $2n$ sources and $2n$ sinks, and each path from a source to a sink represents a repeat sequence or its reverse complement. Sequence similar regions are represented in the graph by edges; repeat domains will be shown as edges with high multiplicity, since they are present in more than one family. The authors present a modification of their graph construction algorithm from their previous work in order to tackle tandem repeats and palindromic sequences which cause problems for the use of A-Bruijn graphs in repeat domain identification.

The repeat domain graph is constructed from a collection of repeat sequences, given as the input to the algorithm. In this paper, the authors constructed graphs for humans and *C. elegans* and analyzed the resulting graphs, revealing the mosaic structure of repeat families. The repeat domain graph for humans was found to contain many of the known domains and composite repeats (repeat families that contain at least 2 repeat domains with different biological origin) found in Repbase; in the *C. elegans* graph, the authors discover a set of new composite repeats and propose that the graph provides more descriptive annotations than those in the RECON library and also suggests annotations for families that remain unannotated.

Other than analyzing repeat domain graphs from individual organisms, the authors did a comparative analysis of the repeat domain graph generated from both *C. elegans* and *C. briggsae* by generating a combined repeat domain graph using the collection of repeat families from both organisms combined. Through this analysis, they discover conserved domains present in both species. Also, the tree structure present in their comparative graph induced a phylogenetic tree that was in extremely good agreement with one generated by CLUSTALW directly from the DNA sequences.

As a final application of their approach, the authors show how the repeat domain graph may be used to identify tandem repeats from repeat families generated by existing algorithms, and also how the graphs may be used to find domain recombinations or composite repeats; they present a known domain recombination known that was found in their graphs.

Discussion

The paper presents a framework for further analyzing repeat families once they are found, by organizing them according to commonly repeated subsequences (motifs). This helps biologists make sense of the vast amount of repeat family data that are generated as the graph may be more easily interpreted than a dot plot of the repeat sequences – common domains may be eyeballed by looking for high multiplicity edges.

Although this is not the first such algorithm that is able to identify repeat domains, the only other algorithm (also by the same group), RepeatGluer, is unable to directly identify domains from the entire genome sequence as it is too slow. Hence, this work represents the first work in efficiently identifying such domains by utilizing other repeat family finding algorithms in a preprocessing step to first find repeat families before operating on them.

Even though the authors present many biologically relevant results in their paper, these results were manually discovered from inspection of the domain graphs. Thus, perhaps the immediate utility of this algorithm will be the automated identification of tandem repeats and more importantly, in the more easily interpreted representation of the structure of repeat families, which will aid in annotation of repeat families and provide a conceptual framework for future study.