LadderFinder: a tool for allelic ladders

ID - 141

Casagrande Alberto[1], Lain Orietta[2], Policriti Alberto[1,3], Testolin Raffaele[1,2]
[1]Istituto di Genomica Applicata, Udine
[2]Dipartimento di Scienze Agrarie ed Ambientali, Università di Udine, Udine
[3]Dipartimento di Matematica ed Informatica, Università di Udine, Udine

Motivation

Microsatellites, also known as simple sequence repeats (SSR), consist of contiguous repeats of short DNA fragments of at most 10bp. Since the lengths of their alleles are high variable, SSR were used to type different organisms from a molecular point of view. As a matter of fact, we can distinguish two strains of the same species simply by searching for different alleles lengths in a set of microsatellite loci. Unfortunately, allele lengths are measured by PCR and returned as peaks. Thus, to classify more than two strains and avoid errors during the interpretation of PCR output, it is important to prevent the use of microsatellite loci having common allele lengths over species. For such a reason, given a SSR locus, it may be useful to discover the set of disjoint allele lengths which maximize the number of identified strains. Such set is known as 'allelic ladder'.

Methods

Results

In this paper, we present a software tool, called LadderFinder, designed to find allelic ladders. LadderFinder takes in input a table whose rows and columns denote strains and SSR loci, respectively. Each cell of the input table reports the pair of allelic lengths of the strain locus corresponding to the cell position. The program returns an allelic ladder for each microsatellite locus.

To find the allelic ladders of each columns, LadderFinder codes the problem into a graph problem called the 'maximal matching problem'. First of all, LadderFinder builds a graph for each columns of the input table. In such graphs, nodes stand for allelic lengths and edges denote pairs of length. Hence, the built graphs contain an edge connecting two nodes if and only if there exists a cell in the corresponding input column containing the pair of lengths represented by such two nodes. Then, the program computes the maximal matching of each graph, that is the maximal set of edges which do not share any node. Since each edges corresponds to pairs of allelic lengths, the maximal matching denotes the maximal set of alleles whose lengths are not repeated and, thus, to the allelic ladder. To compute the maximal matching, we implemented an optimized version of the Edmondss non-bipartite matching algorithm. Such algorithm reduces the original problem to a maximal matching computation over a bipartite graph by collapsing odd length cycles.

Once the bipartite graph is obtained, the algorithm two-colors its dual graph and deduces the maximal matching from such a coloring. Then, it extend the computed matching to the original graph by coloring edges that do not introduce conflicts. The overall time complexity of such algorithm is O(|N|^4), where N is the set of nodes in the original graph, but it is known that can be decreased to O(|N|^3) by careful use of modern data structures.

Availability: http://www.appliedgenomics.org/LadderFinder

Image: http://www.appliedgenomics.org/LadderFinder/img/main.png

Email: casagrande@appliedgenomics.org