

From Satellite DNA to HTML Table Understanding

Wen-Lian Hsu

Institute of Information Science

Academia Sinica

Taipei, Taiwan, ROC

Abstract

A Satellite DNA is a DNA sequence that repeats itself one after another (called tandem repeats) for a number of times. Satellite DNA composes 10% of the genome and is useful for gene mapping. In this paper we study minisatellites, which are blocks of tandem repeats 20 to 70 in length which usually have a total length of a few thousand base pairs. Given a minisatellite DNA, we want to determine its repeat pattern and the number of repeats. However, each repeat might contain a few errors caused by deletion, insertion or substitution. The similarity among these repeat sequences is usually very high. If we cut out these repeat sequences and make an alignment, normally ninety-five percent of each column contains the same base pair.

We developed a tandem repeat algorithm to cut a given minisatellite DNA into segments so that their alignments closely satisfy the above similarity measure. Our method uses a PAT-tree of the DNA sequence to aid the analysis of the alignment. Probability is used to determine the confidence of a repeat segment being part of the consensus of the alignment. Overlapping relationships of these repeating segments are exploited to weed out local noises.

Our algorithm for tandem repeat can also be applied to solve a difficult problem in HTML document analysis, namely, understanding a table in an HTML document. Many HTML tables are irregular in that they are not ordinary matrices with attributes listed only in the first row and the first column. Sometimes, each entry of a row or a column could be subdivided into smaller inner tables, making the whole analysis much harder. Through natural language analysis, we first determine the properties of the entries of the table such as proper names, ranks and etc. Then, substitute the entries by symbols representing their properties. Apply the tandem-repeat algorithm to the above sequence to discover the repeat patterns and cut-off locations. These repeat sequences would give us the boundary of blocks of information that are closely related.

