Geometric DNA Alignment with Range Trees

D. Wagner[1]

Abstract: DNA Alignment often involves the search for a short query gene within a much longer sequence of DNA. If mutations are a possibility, then it becomes important to search for highly similar but inexact matches. Matches are typically scored as a function of the number of substitutions, insertions, and deletions that exist between the query gene and the sequence at the chosen location within the DNA.

The optimal match location can be found with a dynamic programming algorithm, such as Needleman-Wunsch or Smith-Waterman, however this takes time proportional to the product of the lengths of the DNA sequence and the query gene. The complete genome of a species can be billions of characters in length, so this method may be too slow for some applications.

Rapid searching can be achieved by building a suffix tree during preprocessing of the larger sequence. The matching locations of a query string can then be found by traversing the suffix tree. This method enables searching in time prortional to the length of the query string, however it reports only exact matches.

We introduce a new method for DNA alignment that maintains speeds similar to those of the suffix tree, while allowing inexact matches. By searching for fragments of a query gene, it is possible to find inexact matches. However, short query fragments could be found in numerous locations, resulting in billions of matches.

During preprocessing, we consider a DNA sequence to be a one dimentional geometric object, and thereby construct a set of range trees over the DNA. Short gene fragments can then be located with a range search of the DNA. Even if there are billions of matches, a range search allows us to report the existence of any number of matches in logarithmic time.

[1] Department of Electronic Engineering, Hanyang University
222 Wangsimni-ro, Seongdong-gu, Seoul, 04763, Korea
dwagnkorea@gmail.com