Visualysis: A Tool for Biological Sequence Analysis

SungHwan Kim[1] shkim@bric.postech.ac.kr
YangSuk Kim[1] biostone@bric.postech.ac.kr
TaeJin Ahn[1] tj@bric.postech.ac.kr
HongGil Nam[2] hgn@bric.postech.ac.kr
ByungJun Han[3] bjhan@ce.knu.ac.kr
SamMyo Kim[3] kims@bh.knu.ac.kr

[1] Biological Research Information Center, Pohang University, San 37, Hyoja-Dong, Nam-Gu, Pohang, Kyungpook, Korea
[2] Department of Computer Engineering, Kyungpook National University, 1370, SanKyuk-Dong, Buk-Gu, TaeGu, Korea

Keywords: suffix tree, visualization, repeat sequence

1 Introduction

Genomes contain much biological information for the living organism in addition to protein-coding genes. However, because of its length and complexity, it is hard to identify the characteristics of the genome and its derivatives, such as RNAs, proteins, ESTs, restriction fragments, etc., The underlying complexity does not allow current computing technology a direct analytic approach. It requires good huristics and good interactive, sophisticated tools.

We have developed a software tool, called VISUALYSIS, for visualizing (in 2D and 3D) the genomic context and analysis based on the suffix tree, which is a data structure constructed with all suffices of given sequences [3]. It is well known in computer science that with this data structure, which can be constructed in linear time, many interesting problems concerning string analysis and comparisons can be efficiently solved. The algorithms for these problems are directly applicable to biological sequence problems such as, among others, mapping and sequencing, finding tandem repeats, recognizing DNA contamination, constructing phylogenetic trees, finding maximum parsimony [1].

The main objective for the development of VISUALYSIS is to visualize possible hidden structural information that cannot be accessed through the conventional algorithms. We believe that typical structural properties of a suffix tree, such as bushiness, branch distribution, isolated branches, etc., may lead to some interesting biological information. Highlighted branches or trunks matched with a set of given sequences would help the user evaluating the output of an algorithm. The 3D visualization (in color) let the user peer into the hidden structure of a tree (see Fig. 2), which is difficult to view in 2D, especially when the tree is bushy.

2 Method and Results

VISUALYSIS has been implemented on Pentium PC in Visual C++ 6.0 based on the Ukkonnen's algorithm for constructing suffix tree [2]. (So the complexities for the running time and space are both linear in the size of the input string.) Along the visualization it shows the statistics on the number of nodes, the number of edges and their average length, and the branching density (see Fig. 1). It allows tree traversal showing the sequence along the traversed path, or tree walking along a given sequence. We are going to implement all the known algorithms that run on suffix tree with proper visualization for their output. Current version provides exact matching and tandem repeat algorithms, which show profiles with information such as the locations, repeat patterns, and the number of repeats (see Fig. 2).

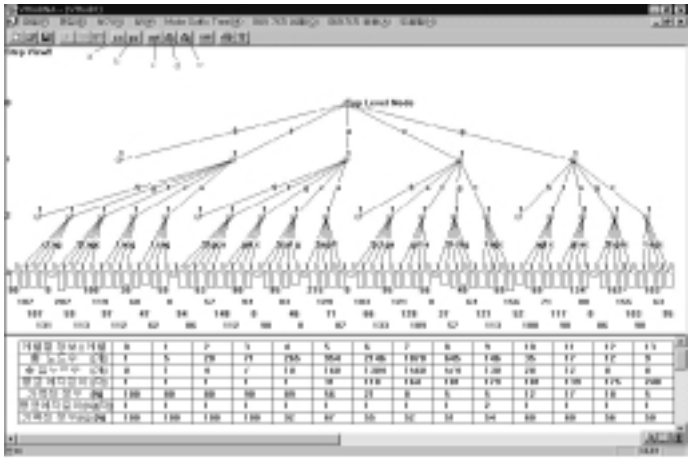

Figure 1: Visualizing a suffix tree and its structural information.

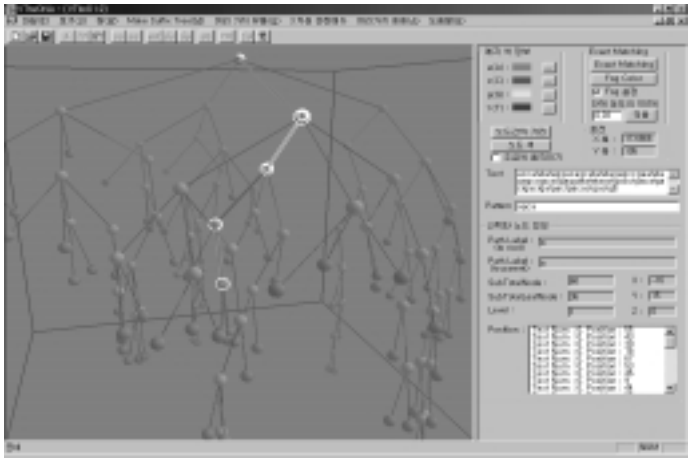

Figure 2: 3D visualization and exact matching.

References

[1] Gusfield, D., *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1997.

[2] Ukkonen, E., On-line construction of suffix-tree *Algorithmica*, 14: 249–260, 1995.

[3] Weiner, P., Linear pattern matching algorithms, *Proc. 14th IEEE Symp. on Switching and Automata Theory*, 1–11, 1973.