Space-Gene : Microbial Gene Prediction System Based on Linux Clustering

Jong-won Chang[1] brutus@bric.postech.ac.kr
Chungoo Park[1] madreach@bric.postech.ac.kr
Dong Soo Jung[1] viroid97@bric.postech.ac.kr
Mi-hwa Kim[2] bfpark@posdata.co.kr
Jae-woo Kim[2] jawookim@posdata.co.kr
Seung-sik Yoo[2] ssyoo@posdata.co.kr
Hong Gil Nam[1] nam@postech.ac.kr

[1] Biological Research Information Center, POSTECH, Pohang, 790-784, Korea
[2] Linux Parallel Processing Computing Team, POSDATA, Seongnam-Si, 463-775, Korea

Keywords: space-gene, prokaryotes, Gene prediction, hidden Markov model, parallel processing

1 Introduction

With the increasingly popularity of genome sequencing, the large-scale identification and functional analysis of the genes that underlie life phenomena have become the principle research subjects in this field, which in turn has stimulated the development of computer applications for gene prediction based on pattern recognition [1, 2] and neural networks. At present, there are still some missing genes that couldn't be found using the existing gene prediction methods. And many annotated genes are having to be re-annotated, because of uncertainty about their identity. As for the genes registered in the Genbank database so far, of the 148 species for which the full genome sequencing data has been published, 129 entries belong to the prokaryotes. This implies that the prokaryotes are more suitable than the eukaryotes for research using comparative genomics, because their simple gene structure makes gene prediction easier. Since many of the existing gene prediction tools only have an accuracy of 80% to 90%, the development of gene prediction systems with more accurate precision is urgently required. With this in mind, many scientists have tried to upgrade the existing gene prediction programs [1] or to enhance their accuracy through the integration of new features and enhancements [3].

In this paper, to enhance the accuracy of gene prediction, we propose a scheme that merges the *ab-initio* method with the homology-based one. While the latter identifies each gene by taking advantage of the known information for previously identified genes, the former makes use of predefined gene features. Also, the proposed scheme adopts parallel processing to guarantee optimum system performance, in the face of the crucial drawback of the homology-based method, i.e. the bottleneck that inevitably occurs due to the large amount of sequence information that has to be processed.

2 System Implementation and Experiment

The proposed scheme, which we refer to as Space-Gene [4], takes advantage of the Glimmer2.0, Blastn, Blastx and Hammer methods. Glimmer2.0 performs gene prediction using the interpolated Markov model which is based on the *ab-initio* method. Blastn and Blastx are the most widely used programs among the homology based methods. Finally, Hammer is a technique which is employed to search for a particular motif during the BLAST search, which enables the selection of genes having a smaller homology. The flowchart for gene prediction and the configuration of our system are shown in Figures 1 and 2, respectively. In our method, the problem causing the bottleneck is overcome through the implementation of parallel processing techniques, allowing for the rapid analysis of a large amount of sequence information, as depicted in Figure 2. As a result, we can obtain an increase in performance which is linearly proportional to the number of nodes, as shown in Figure 3. When the performance of the Space-Gene algorithm is compared with that of Glimmer2.0, which is known to be extremely efficient, it is found to represent a superior upgrade path than Glimmer2.0, due to its superior system of merging sequences (Table 1).

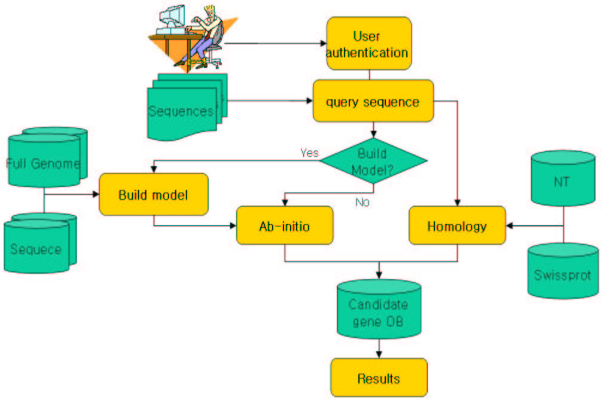

Figure 1: Space-Gene Flowchart.

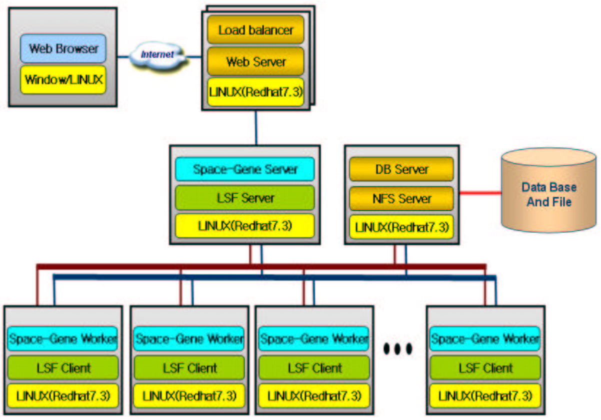

Figure 2: Space-Gene H/W Configuration.

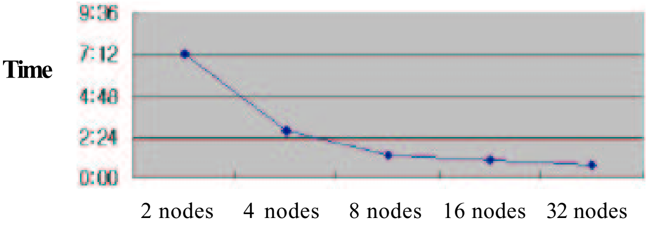

Figure 3: Performance vs. the number of processing nodes (*E.coli* k-12 full genoem)

Table 1: Performance Comparison of Space-Gene with Glimmer 2.0.

Strain	Genbank	Glimmer	Space-Gene	Remark
E.coli k-12	4279	4167(97.30%)	4216(98.50%)	Merged with *B.halodurans*
S.typhi	4365	4300(97.80%)	4340(98.70%)	Merged with *B.subtilis*

References

[1] Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L., Improved microbial gene identification with GLIMMER, *Nucleic Acids Research*, 27(23):4636–4641, 1999.

[2] Lukashin, A.V. and Borodovsky, M., GeneMark.hmm: new solutions for gene finding, *Nucleic Acids Res.*, 26(4):1107–1115, 1998.

[3] Murakami, K. and Takagi, T., Gene recognition by combination of several gene-finding programs, *Bioinformatics*, 14(8):665–675, 1998.

[4] `http://218.155.24.139/gene/html/main.htm`