A Method for Querying Complete Genomes by Viewing Them as Structured Documents

Aaron J. Stokes
stokes@ics.es.osaka-u.ac.jp

Hideo Matsuda
matsuda@ics.es.osaka-u.ac.jp

Akihiro Hashimoto
hasimoto@ics.es.osaka-u.ac.jp

Department of Informatics and Mathematical Science
Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

1 Introduction

Recently, database tools for manipulating structured documents have attracted much attention. The application of database technology to querying structured documents makes it possible to retrieve such documents not only by content but also by structure using the structural components of the documents. In this paper, we present a method for querying information on complete genomes (i.e., whole DNA nucleotide sequences contained in organisms) that have been determined recently.

In our method, genomes, accompanied by the genetic information encoded on them, are modeled as structured documents: for example, genes are modeled as words, while sets of consecutive genes are modeled as sentences since many functionally-related genes tend to be neighbors on genomes. We propose a data model for representing the structure of a genome as a structured document.

2 Genome Language

For describing complete genomes, we designed a schema based on the ODMG standard. Our schema consists of three object-types, Genome, Strand and Gene, which describe biological entities; and one object-type, Region, which represents a region on a particular strand in a genome.

We introduce several notations based on *region algebras* [1]. For example, Fig. 1 illustrates some of the notations used to describe a circular genome. Note that `dist` represents the shortest of the two possible circumferential distances between the genes.

3 Experimental Results

Based on our query language, we have implemented a prototype system for querying complete genomes, using the Perl language. In our implementation, we extracted gene data on complete genomes from GenBank Release 104. We then computed the distance and similarity between all pairs of the extracted genes. Sequence similarity was computed using the FASTA program.

In our experiment, we performed several test queries. One of these queries demonstrates how our query language can be used to explore functionally-related genes on a single genome, using sequence similarity. The query can be expressed in natural language in the following way:

"Given two neighbor genes *phoP* and *phoQ* on the *Escherichia coli* genome, retrieve all pairs of two neighbor genes p_1 and p_2 on the same genome where p_1 and p_2 are paralogous to *phoP* and *phoQ*, respectively."

Here, we consider two genes to be paralogous if they are located on the same genome and their mutual similarity is at least 200.

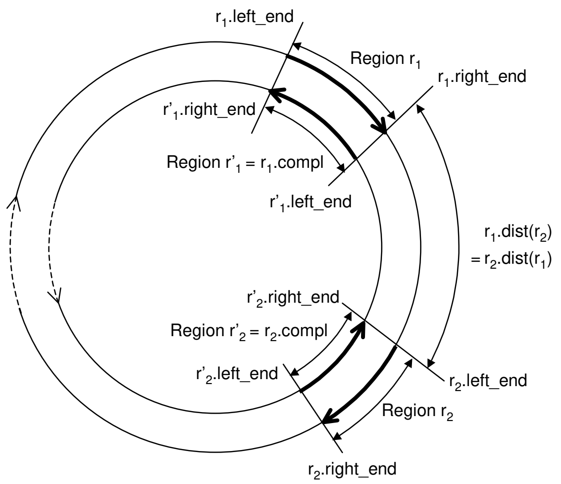

Figure 1: Notations used for describing objects on circular genomes.

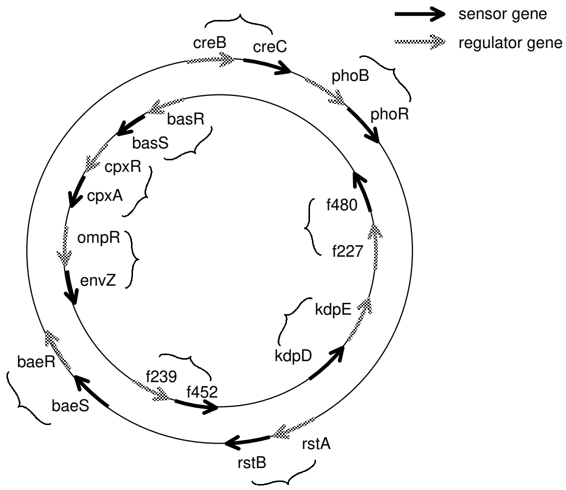

Figure 2: Results for the test query.

In our query language, the query can be expressed as follows.

```
{  Gene g1, Gene g2, Gene p1, Gene p2 :  p1.name, p2.name ;
   g1.name = "phoP" & g2.name = "phoQ" &
   g1.genome.name = "Escherichia coli" & g2.genome.name = "Escherichia coli" &
   g1.region.dist(g2.region) < 100 & p1.genome = g1.genome & p2.genome = g2.genome &
   g1.similarity(p1) >= 200 & g2.similarity(p2) >= 200 & p1.region.dist(p2.region) < 100
}
```

As shown in Fig. 2, we obtained 10 pairs of neighbor genes as a result of the test query. The fact that these gene pairs are paralogous suggests that they may be functionally-related. We confirmed this in biological literature [2], where the same 10 pairs we obtained are reported to be functionally-related as cognate sensor/regulator pairs.

Our system was thus able to obtain a result that accurately suggests a functional relationship between gene pairs. Further, the fact that the genome involved was circular posed no obstacle to obtaining this result. The computation time required was only 2.5 seconds.

This and other results show that our method is effective in retrieving information on complete genomes by modeling them as structured documents. Our current research topics include the further enhancement of the modeling capability of the language.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Science', from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Consens, M.P. and Milo, T., Algebras for Querying Text Regions, *ACM Symposium on Principles of Database Systems (PODS'95)*, 11–22, 1995.

[2] Mizuno, T., Compilation of All Genes Encoding Two-component Phosphotransfer Signal Transducers in the Genome of *Esherichia coli*, *DNA Research*, 4:161–168, 1997.