A Structured Genome Document Database System for Making High-Level Queries on Diverse Genome Data

Aaron Stokes[1] stokes@ics.es.osaka-u.ac.jp
Hideo Matsuda[1,2] matsuda@ics.es.osaka-u.ac.jp
Akihiro Hashimoto[2] hasimoto@ics.es.osaka-u.ac.jp

[1] CREST, JST (Japan Science and Technology)
[2] Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan

Keywords: structured document, XML, genome database, pathway analysis, query language

1 Introduction

With recent advances in biotechnology, tremendous amounts of sequence, gene, protein, pathway, and other genome data are being accumulated. By using computer analyses to identify correspondences and correlations among such data, researchers hope to elucidate the functions of putative genes, infer networks of gene interactions, and generally gain a better understanding of the genome. However, performing such analyses on genome data is considered difficult, because diverse genome data are scattered across many highly heterogeneous databases, and because existing database systems lack the facilities to expose and analyze functional relationships among the data.

2 Method and Results

In our approach, we focus on developing an effective method for the *representation* of genome data that is conducive to making high-level analyses. The result, GXML (Genome-oriented eXtensible Markup Language), is an application of the eXtensible Markup Language (XML)[1] that describes how to represent genome data (the entire DNA sequence of an organism, together with all its genes and associated information that constitute its genome) in self-descriptive *genome documents* [2]. The representation is more effective than traditional data models because genomic relationships are expressed more directly, data is more readily accessible, and the self-descriptiveness of the documents helps cope with diversity and heterogeneity among the data. Fig. 1 shows an excerpt from a sample GXML document.

We have defined a query language called GQL (Genome-oriented Query Language) that is equipped with powerful functions that expose the biological relationships represented in GXML documents. These functions can be used as building blocks for higher level queries. GQL also exposes pathway relationships and performs pathway computation based on a bidirectional minimum-weight search algorithm (Fig. 2).

We propose a new type of genome database system, based on GXML and GQL, that can be applied to high-level genome analyses that are difficult to perform using existing databases and tools. For example, suppose we want to know whether some genes that are neighbors on a genome also encode for proteins that are consecutive components of a pathway: "Find all pairs of neighbor genes g_1 and g_2 on the *Escherichia coli* genome and a pathway *pw*, where g_1 and g_2 encode enzymes that are consecutive components of *pw*." Fig. 3 shows this query expressed in GQL. Results for the *tryptophan biosynthesis* pathway are depicted graphically in Fig. 4. Some interesting gene correspondences could be identified, including sub-unit and multi-functional relationships. The time required for execution was only 12.8 seconds (PII-400MHz processor) using a prototype implementation.

```xml
<?xml version="1.0" ?>
<!DOCTYPE gxml SYSTEM "gxml.dtd" >

<gxml>
   <genome>
      <gid>Escherichia coli K-12...</gid>
      <whose>E. coli Genome Project</whose>
      <date>98Nov18</date>
      <contig>
         <cid>c000</cid>
         <dna>ATGCGAGTGTTGAAGTTCGGCGG...</dna>
      </contig>
      ...
      <feature type="orf">
         <fid>b1263</fid>
         <alias>trpD</alias>
         <location>
            <cid>c000</cid>
            <start>1317813</start>
            <end>1319408</end>
            <strand>-</strand>
         </location>
         <dna>ATGGCTGACATTCTGC...</dna>
         <prot>MADILLLDNIDSF...</prot>
      </feature>
      ...
      <pw>
         <pid>00401</pid>
         <pwname>tryptophan biosynthesis</pwname>
         <pwrole>
           <rid>4.1.3.27</rid>
           <fid>b1263</fid>
         </pwrole>
         ...
      </pw>
      ...
      <role>
         <rid>4.1.3.27</rid>
         <rdescription> ... </rdescription>
         <fid>b1263</fid>
         ...
         <substrate>Pyrophosphate</substrate>
         <product>Anthranilate</product>
         ...
      </role>
   </genome>
</gxml>
```

Figure 1: Example GXML document.

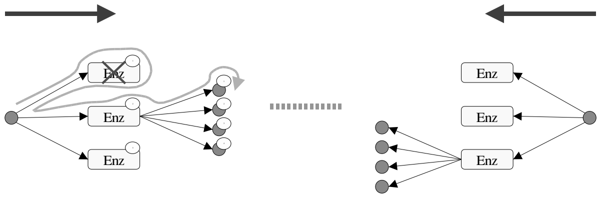

Figure 2: Pathway reconstruction algorithm.

```
WHERE
  <pw>
    <pwname>$pwname</>
  </> AS $pw IN "ecoli.gxml"
ORDER $pw BY $pwname
CONSTRUCT
  <neighborsbypathway>
    $pw
    WHERE
      <feature><alias>$a1</></> AS $f1
      <feature><alias>$a2</></> AS $f2
        IN "ecoli.gxml",
      neighbors($f1,$f2),
      upstream($f1,$f2),
      pwneighbors($f1,$f2,$pw)
    ORDER $f1, $f2 BY $a1, $a2
    CONSTRUCT
      <neighborpair>
        $f1
        $f2
      </>
  </>
```

Figure 3: Example GQL query.

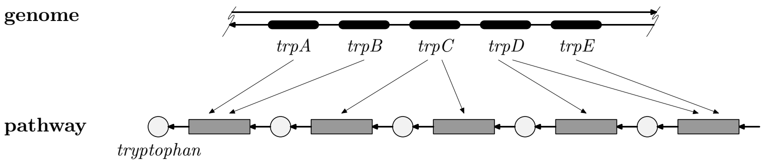

Figure 4: Results obtained from example query.

We are confident that the proposed GXML/GQL genome database will prove an invaluable tool for helping researchers identify complex genome interactions on many levels.

References

[1] Bray, T., Paoli, J. and Sperberg-McQueen, C.M. ed., Extensible Markup Language (XML) 1.0, *W3C Recommendation 10-Feb-98*, available at `http://www.w3.org/TR/REC-xml`, 1998.

[2] Stokes, A.J., Matsuda, H. and Hashimoto, A., Making high-level queries on diverse genome data: a structured genome document database system based on GXML and GQL, *Genome Informatics 1999* (ed. Asai, K. *et al.*), Universal Academy Press, Tokyo, 176–185, 1999.