Index-Driven XML Data Integration to Support Functional Genomics

Ela Hunt[1], Evangelos Pafilis[1, 2], John N. Wilson[2], Barry A. Gusterson[1], Torsten Stein[1]
[1]University of Glasgow, [2]University of Strathclyde
Glasgow, Scotland, UK
ela@dcs.gla.ac.uk, jnw@cis.strath.ac.uk
http://www.xtect.cis.strath.ac.uk/

UNIVERSITY of GLASGOW

THE UNIVERSITY OF STRATHCLYDE IN GLASGOW

BIOINFORMATICS RESEARCH CENTER

Introduction

- XTeCT[1] (XML Technologies for Cancer Treatment) aims to automate biological data integration
- Many biological databases offer a download in XML[2]
- We index these large XML datasets, maintaining structural and content information
- The index is materialised in a relational database that supports data mining operations
- Our hypothesis is that if a significant number of leaf values of the XML trees can be reached by a certain set of paths then these paths may semantically match

Objectives

Interpretation of large scale experiments in **Functional Genomics** requires access to high volumes of information. The task of characterising the top 100 hits of a microarray experiment comprises a manual search of all the relevant databases and tends to be laborious and time consuming. Four different groups (Table 1) have been monitored while performing result annotation.

Our primary concern is to accelerate molecular biology research by *automating* the information retrieval process and by presenting the results in an information digest that can be easily understood.

Methods

Biological databases (Table 2) were mirrored locally in XML format. If not already available, XML dumps were created from the underlying relational or textual representation. The files were processed by a SAX XML parser and each leaf data value, along with its tree path, loaded in an Oracle RDBMS. SQL queries are applied to find candidate mappings and redundant elements across databases.

Barry A. Gusterson

Ela Hunt

Evangelos Pafilis

Torsten Stein

John N. Wilson

Table 1: Groups performing large scale analysis experiments

Prof Anna Dominiczak University of Glasgow	Microarray experiments to study gene expression in rat animal models of hypertension in the context of the Cardiovascular Functional Genomics project
Prof Barry Gusterson University of Glasgow	Microarray experiments to study gene expression in mouse mammary gland development as a model of breast cancer proliferation
Dr Catherine Winchester Yoshitomi Research Institute of Neuroscience, Glasgow	Microarray experiments to study gene expression in a chemically induced rat model of schizophrenia
Dr Mike Turner University of Glasgow	Proteome map of *Trypanosoma brucei* in the context of genetics and genomics

Table 2: Databases containing information used in the analysis of Functional Genomics experimental results

Database	XML file details
NetAffx (Affymetrix)	For each array type there is an XML file available (MAGE-ML format) The sizes range from 16 to 550 MB
Ensembl	A file has been created for human, mouse, rat. The sizes range from 35 to 145 MB
Gene Ontology	XML dump available. Size: 344 MB
PubMed	40 GB compressed size
Mouse Genome Database (MGD)	A 191 MB file has been created containing selected pieces of information
OMIM	A 368 MB xml file has been created by parsing the OMIM flat file
SwissProt	A 852 MB XML file has been downloaded from the website

Results

OMIM, MGD and SwissProt (~1,4 GB of XML files, 32,000,000 leaves in total) have been indexed. We are investigating data mining techniques to find matching paths.

Queries for a specific gene or protein can be performed and reveal their location across databases. Table 3a lists the paths in OMIM, MGD and SwissProt of leaves whose value is the gene symbol 'SOCS3'. Table 3b lists the paths of leaves containing the gene symbol 'SOCS3'.

Besides querying for a specific gene or protein our platform supports pairwise comparison for exact matching of all the leaves in one database against all the leaves of the rest. The comparison has been performed between OMIM and MGD.

On going work

We are currently processing the results of OMIM and MGD comparisons by applying data mining techniques that will return valid database mappings.

References

1. Ela Hunt, Evangelos Pafilis, Inga Tulloch and John Wilson, (2004). Index-Driven XML Data Integration to Support Functional Genomics. Proceeding of the International Workshop on Data Integration in Life Sciences, DILS'04, Lecture Notes in Computer Science 2994. pp95-109

2. http://www.w3.org/TR/REC-xml/

Table 3a: Paths returned after querying the indexed XML leaves for an exact match of the gene symbol 'SOCS3'

```
Mim/Mim-entry/Mim-alternative_titles_and_symbols/Mim-alternative_titles_and_symbols-alt/SOCS3
sptr/entry/dbReference/property/value/SOCS3
sptr/entry/gene/name/SOCS3
```

Table 3b: Paths returned after querying the indexed XML leaves for containing the gene symbol 'SOCS3'

```
Mim/Mim-entry/Mim-description/*SOCS3*
Mim/Mim-entry/Mim-references/Mim-references-item/Mim-reference-title/*SOCS3*
Mim/Mim-entry/Mim-alternative_titles_and_symbols/Mim-alternative_titles_and_symbols-alt/SOCS3
MGI/MGI-entry/References/Reference/title/*SOCS3*
sptr/entry/Reference/citation/title/*SOCS3*
sptr/entry/comment/text/*SOCS3*
sptr/entry/dbReference/property/value/SOCS3
sptr/entry/gene/name/SOCS3
```

CARNEGIE TRUST FOR THE UNIVERSITIES OF SCOTLAND