BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 18 2005, pages 3681–3682
doi:10.1093/bioinformatics/bti587

Gene expression

GeneCruiser: a web service for the annotation of microarray data

Ted Liefeld*, Michael Reich, Joshua Gould, Peili Zhang, Pablo Tamayo and Jill P. Mesirov

Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA

Received on March 23, 2005; revised on May 16, 2005; accepted on July 14, 2005
Advance Access publication July 19, 2005

ABSTRACT

Summary: GeneCruiser is a web service allowing users to annotate their genomic data by mapping microarray feature identifiers to gene identifiers from databases, such as UniGene, while providing links to web resources, such as the UCSC Genome Browser. It relies on a regularly updated database that retrieves and indexes the mappings between microarray probes and genomic databases. Genes are identified using the Life Sciences Identifier standard.

Availability: GeneCruiser is freely available in the following forms: Web service and Web application, http://www.genecruiser.org; GenePattern, GeneCruiser access has been integrated into our microarray analysis platform, GenePattern. http://www.genepattern.org

Contact: liefeld@broad.mit.edu

1 INTRODUCTION

One difficulty facing researchers who use microarray technologies is that to determine information about a microarray feature, such as the gene it represents, its chromosomal location or its molecular function, a researcher must amalgamate this information from a number of publicly available databases. Although vendors have included more of these annotations with their products, this information quickly becomes obsolete.

In addition, researchers often have a list of genes or a genomic category, e.g. tyrosine kinases and would like to find which microarray identifiers correspond to them.

Although numerous standalone applications exist for annotating microarray identifiers, such as DRAGON and Resourcerer, they are not easily integrated with other tools and applications. GeneCruiser was designed to address both the need for identifier annotation and the desire for easy integration by incorporating a public Web Service Description Language (WSDL) defined SOAP web service interface.

2 THE GENECRUISER APPLICATION

GeneCruiser is a web service and web application designed to annotate genomic data in several ways. GeneCruiser allows users to map gene identifiers from genomic databases to Affymetrix probes, find information about Affymetrix probes in genomic databases by keyword searches and locate Affymetrix probes in the human genome using web resources, such as the UCSC Genome Browser.

The GeneCruiser web application facilitates the annotation queries via a web browser-based interface. Desired annotations and identifiers are selectable via HTML forms. IDs may be entered directly into the form or through uploading text files.

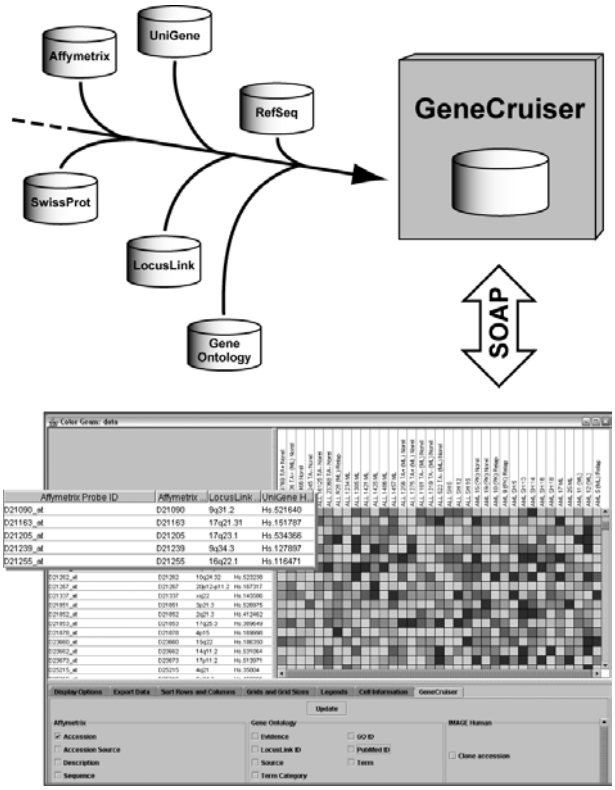

Fig. 1. Overview of GeneCruiser system.

2.1 Databases and annotations

GeneCruiser currently links Affymetrix probe sets to Gene Ontology terms, IMAGE clone IDs and data available in UniGene, LocusLink, RefSeq, SwissProt, and the TIGR human and mouse gene indices. GeneCruiser also provides links from an identifier to its corresponding information in the UCSC Genome Browser, PubMed, GenBank, GeneCards and the Gene Expression Omnibus (Fig. 1). For optimum performance, these databases are automatically downloaded and indexed on a machine local to the GeneCruiser application on a periodic basis.

*To whom correspondence should be addressed.

3 GENECRUISER WEB SERVICE

To facilitate integration, GeneCruiser provides a SOAP web service interface allowing other applications to make use of its functionality.

3.1 LSID identifiers

Any service that deals with IDs from multiple disparate sources must deal with the issue of computationally recognizing the context or scope of an identifier. GeneCruiser uses the Life Science Identifier (LSID) (Clark *et al*., 2004; OMG, 2004) as a generic mechanism to allow the transmission of both the identifier and its context.

The LSID is an Object Management Group specification to represent an identifier with its context in the form of a single Universal Resource Name (URN). It uses the following syntax to specify an identifier and its context: urn:lsid:<authority>:<lsid_namespace>:<identifier>:<version> where, urn declares this is a URN, lsid is the URN namespace, <authority> is the issuing authority or source, <lsid_namespace> is the context, <identifier> is a string representation and <version> is an optional version field.

For example, the Affymetrix probe D10537_s_at on the hu6800 chip is represented by the LSID, urn:lsid:affymetrix.com:probeset.hu6800:D10537_s_at.

3.2 Web service interface

The primary interface methods of the GeneCruiser SOAP interface are

(1) annotateProbes(. . .)—retrieve annotations for a list of microarray identifiers.

(2) idsToProbes(. . .)—retrieve microarray IDs for a known accession (e.g. LocusLink, SwissProt).

(3) keywordsToProbes(. . .)—retrieve microarray identifiers based on a keyword search.

The web service also includes several methods to retrieve metadata describing the available resources, databases and fields.

Identifiers for probe sets passed as query parameters to the annotateProbes method may be written as LSIDs or as strings containing the identifier. Strings used in this method are assumed to be Affymetrix probe set identifiers. For the idsToProbes(. . .) method, an ID may be in the form of an LSID in order to provide the identifier's context (i.e. to allow the server to distinguish between IDs from GenBank, SwissProt, etc.), or as a string containing the accession identifier itself, in which case heuristics are used to determine the identifiers' context.

4 INTEGRATION WITH OTHER APPLICATIONS

Other applications can access GeneCruiser as a web service. For example, we have integrated it with the HeatMapViewer in our microarray analysis platform, GenePattern (see Availability section for URL). As users view the heat map image of their microarray data, they may retrieve information about the probes directly from GeneCruiser, allowing them to view annotations from multiple sources simultaneously with their data.

```
// get a GeneCruiserService proxy instance to call the service
GeneCruiserServiceProxy gcProxy = new GeneCruiserServiceProxy(
   "http://www.broad.mit.edu/webservices/genecruiser/services/Annotation");

// get the available query fields retrievable from  the
// server via getDatabaseToFieldsMap()
HashMap availableQueryFields = gcProxy.getDatabaseToFieldsMap();
Set availableDBNames = availableQueryFields.keySet();
Vector availableLocusLinkFields = availableQueryFields.get("LocusLink");

// Available query fields are also retrievable as annotations
String[] queryFields = {
            "LocusLink LocusLink ID",
            "Unigene_Human UniGene Cluster"};

// list probes to retrieve annotations for
String[] probes = new String[2];
// example of using a probe set name written as an LSID
probes[0] = "urn:lsid:affymetrix.com:probeset.hu6800:AF000430_at";
// example of using just the probe set name
probes[1] = "AB002409_at";

// perform the SOAP call to retrieve the annotations
AnnotationResult  queryResult =
            gcProxy.annotateProbes(probes, queryFields);
```

Fig. 2. Java code to retrieve UniGene and LocusLink information for Affymetrix probes using the Java client side library.

Applications can integrate GeneCruiser from any programming language that supports the SOAP protocol, including Perl, Java, C, etc. The interface is available as WSDL (see Availability section for URL) to permit applications to generate client bindings for GeneCruiser using their local web service tool set.

In addition, a client side library is available for integration from the Java programming language. Example code using the GeneCruiser web service via this library is shown in Figure 2.

ACKNOWLEDGEMENTS

The authors wish to thank the following members of the Cancer Program at the Broad Institute: Todd Golub, Ken Ross, Stefano Monti, Justin Lamb, Jim Lerner and Sridhar Ramaswamy.

Conflict of Interest: none declared.

REFERENCES

Clark,T. *et al*. (2004) Globally distributed object identification for biological knowledgebases. *Brief. Bioinformatics*, **5**, 59–70.

OMG (2004) Life Science Identifiers, OMG Adopted Specification, dtc/04-08-02, August 2004.