BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 16 2004, pages 2839–2840
doi:10.1093/bioinformatics/bth277

GDPC: connecting researchers with multiple integrated data sources

Terry M. Casstevens[1,*] and Edward S. Buckler[1,2]

[1]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853-2703, USA, [2]USDA-ARS, Ithaca, NY 14853-2703, USA

Received on October 27, 2003; revised and accepted on April 8, 2004
Advance Access publication April 24, 2004

ABSTRACT

Summary: The goal of this project is to simplify access to genomic diversity and phenotype data, thereby encouraging reuse of this data. The Genomic Diversity and Phenotype Connection (GDPC) accomplishes this by retrieving data from one or more data sources and by allowing researchers to analyze integrated data in a standard format. GDPC is written in JAVA and provides (1) data sources available as web services that transfer XML formatted data via the SOAP protocol; (2) a JAVA API for programmatic access to data sources; and (3) a front-end application that allows users to manage data sources, retrieve data based on filters, sort/group data based on property values and save/open the data as XML files.

Availability: The source code, compiled code, documentation and GDPC Browser are freely available at: www.maizegenetics.net/gdpc/index.html. The current release of GDPC is version 1.0, with updated releases planned for the future. Comments are welcome.

Contact: terry_casstevens@hotmail.com; esb33@cornell.edu

INTRODUCTION

Numerous research projects on genomic diversity and phenotypes have generated valuable data collections, including thousands of QTL mapping studies. These datasets, however, tend to be abandoned after results are published and thus are not easily accessible by subsequent projects or the general public. Ideally, this data would be made publicly available at the conclusion of each project by migrating the collected data to larger, public databases. The Genomic Diversity and Phenotype Connection (GDPC) accelerates the availability of data by providing the infrastructure to create and use connections to multiple data sources. It is possible to use multiple data sources because each connection masks the specifics of its data source. GDPC already has a connection to the maize diversity database, Panzea (Du *et al*. 2003, http://www.panzea.org; Doebley *et al*. 2003), and there is work in progress to create a connection to the comparative cereal database, Gramene (Ware *et al*. 2002; Stein *et al*. 2003, http://www.gramene.org; Fig. 1). Connections to additional data sources are also planned, and other organizations are encouraged to make their data 'GDPC enabled' by developing connections that integrate their data with GDPC. Once these sources are 'GDPC enabled', data from these sources can be analyzed simultaneously with front-end software applications. These applications use the GDPC JAVA API that standardizes access to any 'GDPC enabled' data source. The GDPC Browser is an application that currently uses this API.

IMPLEMENTATION

'GDPC enabled' data sources

Data sources are at the core of GDPC. A data source is considered 'GDPC enabled' when it has been programmed to return data elements in the format understood by GDPC. Making a data source 'GDPC enabled' does not affect any existing access methods to the data source. A 'GDPC enabled' data source is typically designed to be a remote web service that transfers XML formatted data via the SOAP protocol. Data sources can also be accessed using the JAVA JDBC API, and many other ways of accessing data sources could also be designed. The data are always returned in a common format regardless of the access method, which makes it possible to integrate, analyze and view data from multiple sources. This is a significant advantage over other tools that give access to only one data source. The programming that makes a source 'GDPC enabled' is referred to as a 'GDPC connection'. This connection knows the specifics of its data source and masks them from the rest of the system. Researchers can also create connections to their own data. This allows them to integrate and analyze their data with publicly available data. As new GDPC connections to data sources are created, any 'GDPC aware' software application will automatically have access to that data source. Future plans include developing classes that will allow local files to serve as data sources. Also, plans include registering GDPC connections with MOBY Central (Wilkinson and Links, 2002, http://www.biomoby.org) to allow users to look up available data sources via that directory service.

[*]To whom correspondence should be addressed.

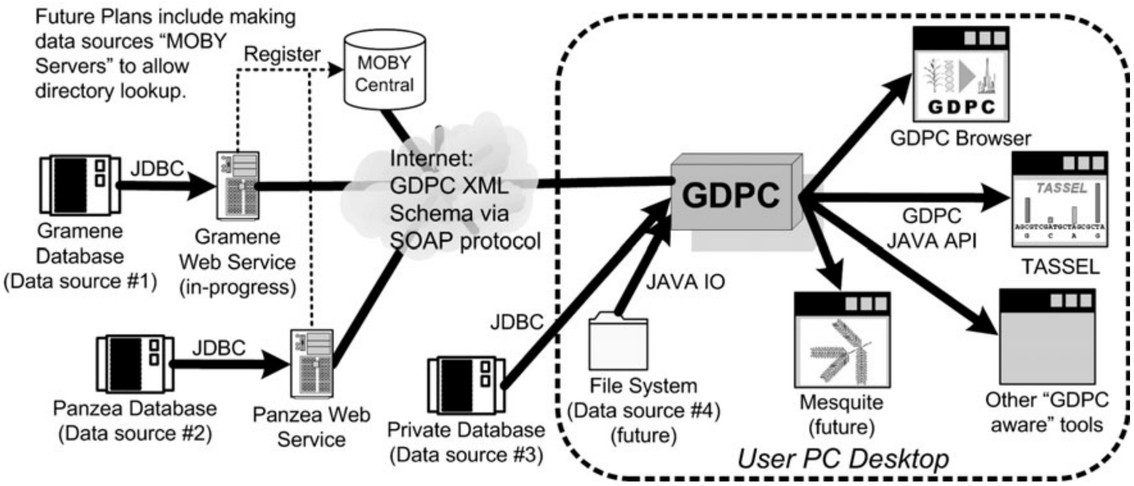
Future Plans include making data sources "MOBY Servers" to allow directory lookup.
Register
MOBY Central
JDBC
Gramene Database (Data source #1)
Gramene Web Service (in-progress)
Internet: GDPC XML Schema via SOAP protocol
JDBC
Panzea Database (Data source #2)
Panzea Web Service
JDBC
Private Database (Data source #3)
GDPC
JAVA IO
File System (Data source #4) (future)
GDPC Browser
GDPC JAVA API
TASSEL
Mesquite (future)
Other "GDPC aware" tools
User PC Desktop

Fig. 1. As shown here, genomic diversity and phenotype data can be retrieved, integrated and analyzed from multiple data sources using GDPC.

GDPC JAVA API

Several research projects make data available via websites, but programmatic ways to retrieve the data do not generally exist. In contrast, GDPC provides a JAVA API that standardizes access to data regardless of the underlying format. Applications that use this API are described as being 'GDPC aware'. Programmers can use this API to develop front-end applications that perform algorithms relevant to their project. Since GDPC masks the specifics of the different data sources, programmers only need to focus on their individual project goals. The GDPC Browser and TASSEL (Buckler, 2003, http://www.maizegenetics.net/bioinformatics/tasselindex.htm) are examples of applications currently using this API. Future plans include development of a Mesquite module (Maddison and Maddison, 2003, http://mesquiteproject.org/mesquite/mesquite.html) that will make any 'GDPC enabled' data source accessible from Mesquite.

GDPC Browser

The GDPC Browser is a front-end application that allows users to retrieve, view and group genomic diversity and phenotypic data based on property values. With this application, users can manage one or more data sources and retrieve data from these sources based on user-defined filters. Once retrieved, the data elements' properties can be viewed by selecting the elements. The different types of data elements are taxa, loci, environment experiments, genotype experiments, localities, genotypes and phenotypes. Working lists of these elements can be created and sorted based on the needs of the researcher. These working lists can then be saved as XML files, and later opened to restore the data elements. Data can also be exported to other formats chosen by the user.

CONCLUSION

Each year much effort and great expense goes into collecting genomic diversity and phenotype data. Rather than sitting idle in remote databases, the data would prove far more valuable if it were maintained in a way that allowed others to continually improve and reuse it. In time, these publicly available data sources would improve in quality and the datasets would grow larger. GDPC provides access to such data collections by retrieving data from multiple data sources and by allowing researchers to analyze integrated data in a standard format.

ACKNOWLEDGEMENTS

This project is supported by a cooperative agreement with the USDA-ARS.

REFERENCES

Buckler,E.S. (2003) TASSEL: Trait Analysis by aSSociation, Evolution, and Linkage.

Doebley,J., Buckler,E., Gaut,B., Goodman,M., Kresovich,S., Muse,S. and Weir,B. (2003) Panzea: maize diversity.

Du,C., Buckler,E., and Muse,S. (2003) Development of a maize molecular evolutionary genomic database. *Comput. Funct. Genom.*, **4**, 246–249.

Maddison,W.P. and Maddison,D.R. (2003) Mesquite: a modular system for evolutionary analysis.

Stein,L., McCouch,S.R. and Cartinhour,S. (2003) Gramene: a resource for comparative grass genomics.

Ware,D., Jaiswal,P., Ni,J., Pan,X., Chang,K., Clark,K., Teytelman,L., Schmidt,S., Zhao,W., Cartinhour,S., McCouch,S. and Stein,L. (2002) Gramene: a resource for comparative grass genomics. *Nucleic Acids Res.*, **30**, 103–105.

Wilkinson,M.D. and Links,M. (2002) BioMoby: an open source biological web services.