APPLICATION NOTE **Genomics & Informatics** Vol. 7(4) 208-211, December 2009

IdMapper: A Java Application for ID Mapping across Multiple Cross-referencing Providers

Hookeun Lee, Hyeonjin Kim and Ungsik Yu*

Lee Gil Ya Cancer and Diabetes Institute, Gachon University of Medicine and Science, Incheon 406-840, Korea

Abstract

We developed an identifier mapping application for bioinformatics research in Java programming language. It is easy to use and provides many usability functionalities that are expected as essentials for a professional application. It supports three widely used mapping services and can convert many ids from one source database into many target databases at once. Id mapping across service providers is possible by remapping the resultant ids. Because it adheres to the NetBeans platform architecture, it can be incorporated into other NetBeans platform applications as an id mapping provider without adaption or modification.

Availability: You can access the application at http://neon.gachon.ac.kr/idmapper.html.

Keywords: ID mapping, java, netbeans, rich client, web services

Introduction

"-omics" style experiments, such as sequencing, microarray, and proteomics, generate enormous amounts of data (Kim *et al*., 2006, Kim *et al*., 2008). To integrate and analyze the data in various respects, multiple tools are used. Each tool supports its own set of identifiers for genes or gene products. Even though there are widely used identifiers, such as Ensembl (Flicek *et al*., 2008), NCBI RefSeq (Pruitt *et al*., 2007), UniProt (The Uniprot Consortium, 2008) and HGNC (Povey *et al*., 2001), there is no universal identifier. When multiple tools and services are used, identifier conversion problems across multiple databases arise frequently. A small research group can not afford its own mapping service and relies on external services. There exist efforts to reconcile identifiers across multiple source databases, such as the David gene ID conversion tool (Huang *et al*., 2008), MatchMiner (Bussey *et al*., 2003), IDconverter (Alibes *et al*., 2007), Onto- translate (Khatri *et al*., 2006), PICR (Cote *et al*., 2007), Synergizer (Brriz *et al*., 2008), and Biomart (Smedley *et al*., 2009).

Services that are provided in the form of a web page can not be incorporated into other applications easily. Only a few services provide API access for batch mapping or incorporation into other applications.

The Protein Identifier Cross-Reference (PICR) service is a web application that provides interactive and programmatic (SOAP and REST) access to a mapping algorithm that uses the UniProt Archive (UniParc) as a data warehouse to offer protein cross-references. The Synergizer is a service for translating between sets of biological identifiers. It can, for example, translate Ensembl Gene IDs to Entrez Gene IDs, or IPI IDs to HGNC gene symbols, and much more. The Synergizer works via a web interface (for users who are not programmers) or through a web service (for programmatic access). BioMart is a query-oriented data management system, developed jointly by the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI). It can be accessed via web services and can be adapted for identifier mapping.

BridgeDb, though not released publically yet, is another attempt at an id mapping framework for bioinformatics applications. BridgeDb lets one add the following capabilities quickly and easily: translate identifiers from one system to another, search references by id or symbol, and link out to online information for an identifier. Applications, such as the PathVisio pathway analysis tool, WikiPathways, CyThesaurus Cytoscape plug-in, and the NetworkMerge Cytoscape plug-in, are utilizing functionalities provided by BridgeDB.

Methods

Instead of developing the system from scratch, we assembled the application by re-using available services, frameworks, and libraries as much as possible.

For ID mapping itself, we used ID mapping services provided by other research groups that allowes programmatic access. We picked the service provided by PICR, Synergizer, and Biomart, as they are recently developed and/or well maintained. Currently, the BridgeDb library implements an ID mapping service for them, and we used BridgeDb's API for the access of id mapping

*Corresponding author: E-mail ungsik@gachon.ac.kr
Tel +82-32-899-6058, Fax +82-32-899-6039
Accepted 12 November 2009

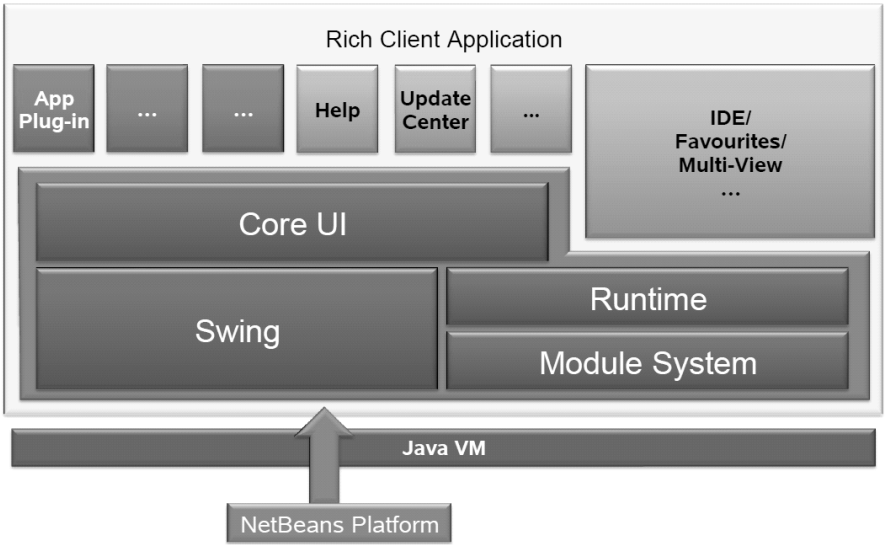

Fig. 1. NetBeans Platform Rich Client Application Architecture. The platform provides various kinds of libraries for GUI application development.

services. Three mapping services were incorporated into one application, and relevant metadata and mapping results are fetched from each service provider on the fly; no data are stored in the application.

Implementation

We developed the application as a NetBeans Platform-based standalone Java Web Start rich client application.

We selected the Java environment because the developed application can be utilized under most operating systems, such as Windows, Linux, and Mac OS, provided that the Java Runtime Environment is installed. Java Web Start technology enables the standalone Java software applications to be deployed with a single click over the network. Java Web Start ensures that the most current version of the application, as well as the correct version of the Java Runtime Environment (JRE), will be deployed.

Most desktop applications have similar features, such as menus, toolbars, status bars, progress visualization, data displays, customization settings, the saving and loading of user-specific data and configurations, splash screens, About boxes, internationalization, help systems, and so on. For these and other typical client application features, a rich client platform, which is an application lifecycle environment and a basis for desktop applications, provides a framework with which the features can quickly and simply be put together. A rich client platform also frees the developer from being concerned with tasks that have little to do with the application's business logic. Currently, the NetBeans Platform and Eclipse rich client platform are in wide use for Java-based application development. Recently, some bioinformatics-related applications, such as BioClipse, Quantitative Biology Tool, SpectraSuite, InstantJChem, and ChipInspector, have been developed using the rich client platform. It is expected that utilization of the rich client platform for applications development will become popular.

We chose the NetBeans platform because there already exist many Swing-based bioinformatics applications. We plan to develop a general purpose bioinformatics application framework, covering various areas by adapting and integrating these existing applications as needed.

We designed the application functionality and layout, developed the core mapping function, and glued those to the NetBeans platform. The NetBeans platform provides quite a lot of libraries for applications development (Fig. 1), and we tried to utilize the functionalities as much as possible and minimize custom developments. After the initial steep learning curve, we were able to develop a high-quality, robust, and extensible application without much housekeeping code development effort.

By using the Swingx JXTable, the resulting mapping table can be customized further by column visibility control, row sorting, and column ordering. The copy/paste/save supports are provided for the resulting mapping table with user-defined column and row selection (Fig. 2).

Results and Discussion

The developed application is simple and easy yet powerful for ID conversion. It supports three widely used mapping services and can convert many ids from one

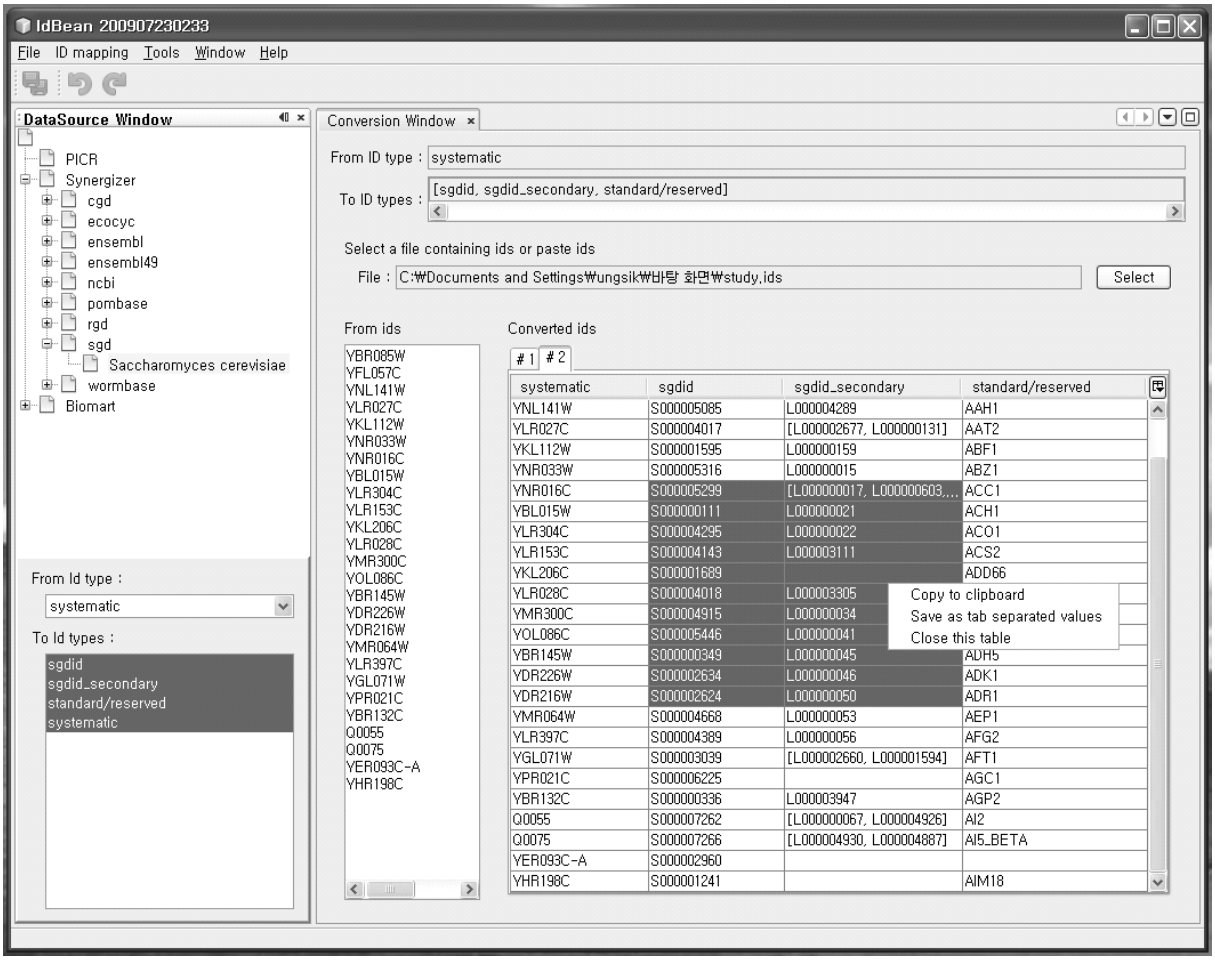

Fig. 2. IdMapper sample mapping result. Start IdMapping by selecting "New ID mapping" action from the "ID mapping" menu bar. Specify the DataSource, source Id type, and target id types in the "DataSource window." Upload or paste the ids for conversion into the "Conversion Window" and select "Perform mapping" from the "ID mapping" menu bar. It is possible to select and copy parts of the mapping table.

source database into many target databases at once. By copying the resulting mapped ids and mapping again against other cross-referencing providers, it is possible to map the ids further.

We plan to extend the source database to the other id mapping services that provide programmatic access, such as CRONOS (Waegele *et al*., 2009). In that respect, we will use JAX-WS directly instead of the BridgeDb library, which currently supports only a limited number of mapping services.

The application is now a standalone Web Start application, but it can be converted into a NetBeans platform plug-in with simple re-packaging. The converted plug-in can be used as an ID mapping provider for other bioinformatics applications.

There are many bioinformatics Java GUI applications, but combining each application with another is next to impossible for most cases, because each application implements its own way of "plumbing." Programs that are developed as plug-ins for a rich client platform can be re-used and work together seamlessly with other independently developed plug-ins. This application can be used as an ID mapping provider plug-in for other bioinformatics applications. Applications development in this direction will facilitate the merging and integration of individual modules into a large bioinformatics applications framework that includes many areas.

Acknowledgments

This work was supported by a grant (2007-03983) from the Korea Science and Engineering Foundation funded by the Ministry of Education, Science and Technology of Korea government.

References

Alibes, A., Yankilevich, P., Cañada, A., and Díaz-Uriarte, R. (2007). IDconverter and ICDlight: Convertsion and annotation of gene and protein IDs. *BMC Bioinformatics* 8, 9.

BioClipse. *http://www.bioclipse.net.*

BridgeDb. *http://www.bridgedb.org.*

Berriz, G.F., and Roth, F.P. (2008). The Synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics* 24, 2272-2273.

Bussey, K.J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W.C., Zeeberg, B., Ajay, W., and Weinstein, J.N. (2003). MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biology* 4, R27.

ChipInspector. *http://www.genomatix.de/products/ChipInspector/index.html.*

Cote, R.G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., and Jermjakob, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* 8, 401.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K.L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A.J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Hubbard, T.J.P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., and Searle, S. (2008). Ensembl 2008. *Nucl. Acids Res.* 36, D707-714.

Huang, D.W., Sherman, B.T., Stephens, R., Baseler, M.W., Lane, H.C., and Lempicki, R.A. (2008). DAVID gene ID conversion tool. *Bioinformation* 2, 428-430.

InstantJChem. *http://www.chemaxon.com/instantjchem/.*

JAX-WS. *https://jax-ws.dev.java.net/.*

Khatri, P., Desai, V., Tarca, A.L., Sellamuthu, S., Wildman, D.E., Romero, R., and Draghici, S. (2006). New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate. *Nucl. Acids Res.* 34, W626-W631.

Kim, C., Choi, J., and Yoon, S. (2008). Microarray data analysis of perturbed pathways in breast cancer tissues. *Genomics & Informatics* 6, 210-222.

Kim, K., Chung, H., Jeung, H., Shin, J., Kim, T., and Rha, S. (2006). Significant gene selection using integrated microarray data set with batch effect. *Genomics & Informatics* 4, 103-109.

NetBeans Platform. *http://platform.netbeans.org.*

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush,M., and Wain, H. (2001). The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.* 109, 678-680.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucl. Acids Res.* 35, D61-65.

Quantitative Biology Tool. *http://www.semanticbits.com/what_we_do/software_solutions/qbt.php.*

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). BioMart-biological queries made easy. *BMC Genomics* 10, 22.

SpectraSuite. *http://www.oceanoptics.com/Products/spectrasuite.asp.*

Sun Java. *http://java.sun.com.*

SwingLabs. *http://swinglabs.org.*

The Uniprot Consortium (2008). The universal protein resource (UniProt). *Nucl. Acids Res.* 36, D190-195.

Waegele B., Dunger-Kaltenbach I., Fobo G., Montrone C., Mewes, H.W., and Ruepp, A. (2009). CRONOS: the crossreference navigation server. *Bioinformatics* 25, 141-143.