Data and text mining

DataBiNS: a BioMoby-based data-mining workflow for biological pathways and non-synonymous SNPs

Young C. Song†, Edward Kawas†, Ben M. Good, Mark D. Wilkinson and Scott J. Tebbutt*

James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research, St. Paul’s Hospital, University of British Columbia, Vancouver, V6Z 1Y6, Canada

Received on September 18, 2006; revised on November 16, 2006; accepted on December 18, 2006

Advance Access publication January 18, 2007

Associate Editor: Alfonso Valcencia

ABSTRACT

Summary: DataBiNS is a custom-designed BioMoby Web Service workflow that integrates non-synonymous coding single nucleotide polymorphisms (nsSNPs) data with structure/function and pathway data for the relevant protein. A KEGG Pathway Identifier representing a specific human biological pathway initializes the DataBiNS workflow. The workflow retrieves a list of publications, gene ontology annotations and nsSNP information for each gene involved in the biological pathway. Manual inspection of output data from several trial runs confirms that all expected information is appropriately retrieved by the workflow services. The use of an automated BioMoby workflow, rather than manual ‘surfing’, to retrieve the necessary data, significantly reduces the effort required for functional interpretation of SNP data, and thus encourages more speculative investigation. Moreover, the modular nature of the individual BioMoby Services enables fine-grained reusing of each service in other workflows, thus reducing the effort required to achieve similar investigations in the future.

Availability: The workflow is freely available as a Taverna SCUFL XML document at the iCAPTURE Centre web site, http://www.mrl.ubc.ca/who/who_bios_scott_tebbutt.shtml.

Contact: stebbutt@mrl.ubc.ca

Supplementary information: Additional information, including test result data, is available from the iCAPTURE Centre web site (see above).

1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) occur when a single base pair in a genome sequence is altered. Non-synonymous SNPs (nsSNPs) are SNPs occurring in the gene-coding regions that result in the alteration of amino acid residues. Such polymorphisms may play a critical role in human disease and drug sensitivity.

Many SNP analysis tools are currently available on the Web; however, each tool provides a disparate and largely disconnected domain of information. The Kyoto Encyclopedia of Genes and Genomes [KEGG; (Kanehisa and Goto, 2000; Kanesiha et al., 2006)] contains resources regarding the interactions between proteins in biological pathways (http://www.genome.jp/kegg); however, it does not contain the information about nsSNPs that map to these proteins. The Entrez dbSNP database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp) contains over 4 million human SNPs, but does not include information about the functional effects of these SNPs in biological pathways. Moreover, using the dbSNP database to separate nsSNPs from synonymous SNPs in a specific pathway requires visual inspection of several web pages in the database, necessitating an immense amount of manual integration. Neither KEGG nor dbSNP contains information about publications, which may indeed provide experimental evidence of the effect of the nsSNPs in the protein pathways.

A research team in the University of California at San Francisco has developed a software pipeline that locates nsSNPs onto protein sequences, functional pathways and comparative protein structure models. Large-scale SNP [LS-SNP; http://alto.compbio.ucsf.edu/LS-SNP (Karchin et al., 2005)] is an annotated database of SNPs which includes information on SNPs found in human genes. One limitation of LS-SNP is that it does not provide users with a direct connection to important resources such as published literature, which may present experimental evidence of the effect of the nsSNPs in the protein pathways. Additional databases have recently become available that also offer tools for investigating the functional implications of SNPs (Reumers et al., 2006; Wang et al., 2006). However, there is no single resource that retrieves all the available information required to investigate the effect of nsSNPs in biological pathways.

We developed the DataBiNS workflow and associated Web Services to integrate information from these disparate resources to better predict the effect of nsSNPs on different biological pathways. Additional versatility and extensibility is provided by using the BioMoby Web Service framework (http://www.biomoby.org)—a protocol developed to enhance Web Service interoperability (Wilkinson and Links, 2002). The BioMoby framework is, in our opinion, more flexible and re-usable than archetypal Web Services such as some of those that we ‘wrap’ in our workflow. Adding semantics into the workflow, in the form of the semantically rich BioMoby services, should facilitate bench scientists or amateur informaticians in extending and/or modifying this workflow to suit...
their specific needs, since the BioMoby framework can be utilized to help guide these modifications.

2 WORKFLOW OVERVIEW

The data-mining tool for biological pathways and non-synonymous SNPs (DataBiNS) workflow begins with an identifier from the Kyoto Encyclopedia of Genes and Genomes (KEGG)—a database containing biological pathway information. The workflow then utilizes LS-SNP to automatically detect missense SNPs that participate in that pathway based on the KEGG identifier (via a SwissProt_ID), and subsequently retrieves the Gene Ontology (GO—the Gene Ontology Consortium—http://www.geneontology.org) terms, PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed), and HapMap SNP information for those loci (via an EntrezGene_ID). The workflow can be easily re-executed in a Web Services workflow application such as Taverna (http://taverna.sourceforge.net) (Oinn et al., 2004; Stevens et al., 2003) at any time, this ensuring facile access to the most up-to-date information.

3 THE CORE BIOMOBY WEB SERVICES

Eight Java-based BioMoby Web Services were created to achieve the DataBiNS workflow, and these are accessible through queries to the public Moby Central Web Service registry at http://mobycentral.icapture.ubc.ca.

getKeggGenesByKeggPathway provides users with the list of genes in the selected KEGG pathway using the publicly available KEGG API.

cconvertKeggGeneId2SwissProtId converts a KEGG gene identifier to a SwissProt identifier using the KEGG API.

getSnpsBySwissProtId retrieves SNPs from their respective genes, and uses the LS-SNP annotated database (http://alto.compbio.ucsf.edu/LS-SNP/) to retrieve the nsSNPs. convertSnp2EntrezGeneId uses NCBI’s eFetch Entrez utility to fetch SNP ‘rs’ IDs.

gGetGenInformationByEntrezGeneId uses NCBI’s eFetch Entrez utility to obtain gene information from the EntrezGene database, including the gene name and synonyms, detailed descriptions and summary.

GeneOntology retrieves GO terms associated with specific genes. The information is retrieved via NCBI’s eFetch Entrez utility and GenNav (http://mor.nlm.nih.gov/perl/gennav.pl).

Gene2PubMed retrieves publication information from PubMed, using NCBI’s eFetch Entrez utility, such as authors, title of the article, the journal the article was published in, and the abstracts for a given gene.

Snp2Frequency retrieves a list of alleles and genotype frequencies for all the nsSNPs in the pathways. The information comes from the International HapMap Project (http://www.hapmap.org) (2003; Thorisson et al., 2005).

4 THE DATABINS WORKFLOW

For our purposes, these eight BioMoby services were pipelined together using Taverna (Oinn et al., 2004)—a Java-based workflow management tool developed by the myGrid project (Stevens et al., 2003). Taverna manages the flow of information from service to service without any human intervention. Taverna workflows can be saved to disk as a SCUFL XML file. Figure 1 (Supplementary Material) shows a diagram of the DataBiNS workflow made by linking the eight different implemented services.

The final result of the workflow is an XML-based set of PubMed, GO terms, EntrezGene, nsSNP and SNP frequency information for genes that are related to one specific biological pathway. Results can also be saved in text file format and Microsoft Excel format.

5 CONCLUSION

The workflow was tested for its efficiency multiple times, across seven KEGG pathways (see Supplementary Material): for example, we tested the KEGG pathway identifications, hsa00051 and hsa04620, which represent a fructose/mannose metabolism pathway and a Toll-like receptor signaling pathway, respectively. Visual inspection of the output data confirmed that all expected information from LS-SNP and Entrez databases were extracted using this model; as such, the workflow performs as efficiently as a human operator, while being capable of operating at a scale beyond that which would be reasonable by hand. The running time varies depending on the pathway chosen—ranging from seconds to minutes.

It is important to note that it was necessary to ‘wrap’ the NCBI e-Utilities Web Service interfaces in order to utilize them effectively. While powerful, the e-Utilities output often takes the form of large XML documents for which specific software must be written to extract the individual data points. By wrapping these large documents in a set of simple, modular and semantically opaque BioMoby Web Services, the output from e-Utilities can be utilized by generic workflow tools such as Taverna. So, while this is not an efficient programming practice in general, we believe that modularization of Web Services, as is common practice in the BioMoby community, significantly increases both their utility and their re-use in ad hoc workflows, as is demonstrated by DataBiNS.

A current limitation in the system is its lack of graphical visualization for the output data. We are in the process of constructing new BioMoby services that output 2D and/or 3D image data in order to enhance the interpretability of these large output datasets.

ACKNOWLEDGEMENTS

We would like to thank Ben Tripp and the anonymous manuscript reviewers for their helpful comments. This research was supported by the National Sanitarium Association (Canada), AllerGen NCE and the Michael Smith Foundation for Health Research. E.K. is supported by an award to MDW from Genome Alberta, in part through Genome Canada. B.M.G. is supported by the Better Biomarkers in Transplantation Project award from Genome British Columbia in part through Genome Canada.

Conflict of Interest: none declared.
REFERENCES
