SOFTWARE TOOL ARTICLE

REVISED G-Links: a gene-centric link acquisition service [version 2; referees: 2 approved]

Kazuki Oshita, Masaru Tomita, Kazuharu Arakawa

Institute for Advanced Biosciences, Keio University, Fujisawa, 252-0882, Japan

v2 **First published:** 19 Nov 2014, **3**:285 (doi: 10.12688/f1000research.5754.1)
Latest published: 18 Nov 2015, **3**:285 (doi: 10.12688/f1000research.5754.2)

Abstract

With the availability of numerous curated databases, researchers are now able to efficiently use the multitude of biological data by integrating these resources via hyperlinks and cross-references. A large proportion of bioinformatics research tasks, however, may include labor-intensive tasks such as fetching, parsing, and merging datasets and functional annotations from distributed multi-domain databases. This data integration issue is one of the key challenges in bioinformatics. We aim to provide an identifier conversion and data aggregation system as a part of solution to solve this problem with a service named G-Links, 1) by gathering resource URI information from 130 databases and 30 web services in a gene-centric manner so that users can retrieve all available links about a given gene, 2) by providing RESTful API for easy retrieval of links including facet searching based on keywords and/or predicate types, and 3) by producing a variety of outputs as visual HTML page, tab-delimited text, and in Semantic Web formats such as Notation3 and RDF. G-Links as well as other relevant documentation are available at http://link.g-language.org/

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
REVISED version 2 published 18 Nov 2015	✓ report	
version 1 published 19 Nov 2014	? report	✓ report

1 **Mark Ragan**, The University of Queensland Australia, **Alison Anderson**, The University of Queensland Australia **Sriganesh Srihari**, The University of Queensland Australia

2 **Kenji Satou**, Kanazawa University Japan

Discuss this article

Comments (0)

Corresponding author: Kazuharu Arakawa (gaou@sfc.keio.ac.jp)

How to cite this article: Oshita K, Tomita M and Arakawa K. **G-Links: a gene-centric link acquisition service [version 2; referees: 2 approved]** *F1000Research* 2015, **3**:285 (doi: 10.12688/f1000research.5754.2)

Grant information: This work was supported by KAKENHI Grant Number 222681029 from the Japan Society for the Promotion of Science (JSPS), and by funds from the Yamagata Prefectural Government and Tsuruoka City.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 19 Nov 2014, **3**:285 (doi: 10.12688/f1000research.5754.1)

REVISED **Amendments from Version 1**

We addressed the referee comments as follows:

1. the claim that G-Links aims to solve data integration problem is toned down,
2. more recent citations are added for the review of data integration challenges,
3. description of Bio2RDF is added,
4. descriptions of compliance to RESTful model is elaborated,
5. descriptions about multiple IDs pointing to the same resource is added,
6. BioGrid and Cytoscape are listed as similar efforts,
7. limitations of gene-centric approach especially in higher eukaryotes with alternative splicing is added,
8. example programmatic access is added on the website and is referred in the manuscript,
9. update cycle is changed.

See referee reports

Introduction

The use of large-scale data or multi-domain information is becoming a prerequisite in all fields of molecular biology, in light of the advent of high-throughput measurement technologies exemplified by the new generation DNA sequencers, and further driven by the conceptual progress in integrative systems biology approaches. Typical analysis encompasses multiple genes in a pathway or in a regulatory network, uses orthologous gene sets in related organisms, and merges information from multiple-omics layers such as genome, transcriptome, proteome, and metabolome (Arakawa & Tomita, 2013). Bioinformatics researchers therefore need to collect and integrate data from a variety of sources, each with diverse syntax, semantics, protocols, identifiers and naming conventions (Bhagat *et al.*, 2010; Brazas *et al.*, 2012; Katayama *et al.*, 2010). This data integration issue is one of the key challenges in the field of bioinformatics (Aoki-Kinoshita *et al.*, 2015; Katayama *et al.*, 2014; Stein, 2002; Stein, 2008). While the integration of web services under standardized protocols has seen a sound progress over the last few years (Katayama *et al.*, 2011), data integration with efficient cross-domain queries still requires the use of large-scale data warehouses such as BioMart (Smedley *et al.*, 2009) and InterMine (Smith *et al.*, 2012).

Since the majority of biological databases are well curated with cross-references, related information can be retrieved *ad hoc* from dispersed databases using hyperlinks. In order to facilitate such processes, web services that collect and provide the cross-reference information from different databases (Diehn *et al.*, 2003; Wu *et al.*, 2013) as well as ID conversion services that assist cross-referencing have been developed (Cote *et al.*, 2007). MyGene.info, for example, provides rapid programmatic access through a RESTful interface for gene-centric queries to retrieve cross-reference information from numerous databases. Gene-centric aggregation, which integrates databases using genes as the primary key, is a highly efficient approach in molecular biology, since the majority of databases have some sort of connection to genes or proteins, due to the success of the "central dogma" of molecular biology. Ideally, a database should be free from predefined schema or primary keys, and should have controlled syntax and semantics. Semantic Web initiatives are therefore collaboratively aiming to provide such framework through HyperText Transfer Protocol (HTTP) with Resource Description Framework (RDF) and Web Ontology Language (OWL) (Katayama *et al.*, 2013). For example, the current release of Bio2RDF resource enables integration and federated queries across 35 datasets (http://download.bio2rdf.org/release/3/release.html). However, the user is required to be familiar with the SPARQL query language, unlike the more intuitive RESTful API approach.

Cross-reference services usually provide database name and identifiers that do not explicitly define the actual location of the data. Moreover, gene-centric data aggregation services usually do not allow querying of gene sets. To this end, here we describe a new RESTful service named G-Links, which gathers Uniform Resource Identifiers (URI) from more than 100 databases in a gene-centric manner, and provide querying interface based on gene sets for hundreds of species. G-Links can be used programmatically as text data, from Semantic Web services, or from graphical HTML pages.

Implementation

G-Links is implemented with Perl programming language and MySQL 5.0, and has a straight-forward RESTful user interface. The server provides a uniform interface based on URL and HTTP in a client-server model, which is stateless and therefore the server does not store any client context information, and the clients and intermediates can cache responses between server update cycles, duration of which is specified by HTML META tag. G-Links internally resolves cross-references in four steps: ID conversion, retrieval of cross-references, filtering and extraction, and formatting of output. G-Links stores all cross-reference information in a gene-centric manner, and for this purpose, it utilizes UniProt IDs as the primary key. Therefore, G-Links first converts the user input to UniProt ID by ID conversion, based on 80 databases supported by UniProt ID Mapping Service (Huang *et al.*, 2011). When a nucleotide or amino acid sequence is given as the query, G-Links searches the corresponding UniProt IDs by sequence similarity search using BLAT (Kent, 2002) against Swiss-Prot database (Bairoch *et al.*, 2004), and when RefSeq ID for bacterial genomes or taxonomy ID is used as the input, G-Links collects all UniProt IDs of genes within the given species based on UniProt taxonomy (http://www.uniprot.org/taxonomy/). In the second step, G-Links collects all text annotations and database cross-references about the gene of interest, gathered from over 130 databases. Here the mapping to Gene Ontology slim (Harris *et al.*, 2004) is pre-computed using map2slim (http://search.cpan.org/~cmungall/go-perl/scripts/map2slim), and resulting URLs for over 30 RESTful bioinformatics analysis web services supported by the G-language Web Services (Arakawa *et al.*, 2010) and Keio Bioinformatics Web Service (KBWS) (Oshita *et al.*, 2011) are generated on-the-fly. KBWS is an European Molecular Biology Open Software Suite (EMBOSS) (Rice *et al.*, 2000) associated software package for accessing popular bioinformatics web services such as BLAST. All cross-references include the URI of the actual location of data, often expressed as Persistent Uniform Resource Locators (PURLs). Retrieved gene set and annotations are optionally filtered in the third step according to user input, and are formatted in the specified output format in the last step.

Results and Discussions

G-Links is available at http://link.g-language.org/ as a RESTful web service, which is suited for resource-centric access and highly accessible via HTTP. Users can rapidly retrieve annotations and cross-references related to a given gene ID, taxonomy ID, or raw sequence data by simply accessing a certain URL. An overview of the URL syntax is presented in Figure 1. For example, the URL to retrieve all annotations and cross references related to the human BRCA1 gene (UniProt ID: BRCA1_HUMAN) is simply http://link.g-language.org/BRCA1_HUMAN (Figure 2). The ID of gene used in this query can be any of the identifiers used in 80 databases supported by G-Links. In this way, multiple URIs can point to the same resource. Programmatic access to this URL can retrieve all 653 annotations and cross-references within 0.2 seconds (tested on dual Xeon X5470 server). G-Links automatically adjusts the output format according to the user context, and outputs the results in human-readable interactive HTML format when accessed from modern HTML browsers, or in Tabular Separated Values (TSV) text format for programmatic access. The HTML format displays a gallery of image resources on the top, such as the pathway maps from KEGG database (Kanehisa *et al.*, 2012), co-expressed gene network from COXPRESdb (Obayashi *et al.*, 2013), and protein 3D structure from Protein Data Bank (Rose *et al.*, 2013), followed by a long table of text annotations and cross-references including database name, ID, and resource URL. Table 1 shows an overview of the categories of databases and web services supported by G-Links output (see http://link.g-language.org/input_list and http://link.g-language.org/output_list for complete listings). In addition to the human-friendly

A. Retrieval of information about Gene or Gene set ID

Syntax: http://link.g-language.org/[GENE ID]/[options…]

1. http://link.g-language.org/BRCA1_HUMAN	Single gene ID
2. http://link.g-language.org/hsa:128,20816955	Multiple gene IDs (can be mixed)
3. http://link.g-language.org/K03553	A set of gene IDs (ex. orthologous groups)
4. http://link.g-language.org/9606	All genes included in an organism

Options

Multiple options can be specified.

A-1. Selection of output data format

Syntax: http://link.g-language.org/[GENE ID]/format=[FORMAT]
tsv (Tabular), html (HTML), json (JSON), rdf (RDF/XML), n3 (Notation3)

1. http://link.g-language.org/BRCA1_HUMAN/format=json	Retrieval of datasets in JSON

A-2. Filtering of gene set

Syntax: http://link.g-language.org/[GENE ID]/filter=[DATABASE:KEYWORD]

1. http://link.g-language.org/NC_000913/filter=eggNOG	Genes with "eggNOG" annotation
2. http://link.g-language.org/K03553/filter=:metabolic	"metabolic" genes
3. http://link.g-language.org/9606/filter=DISEASE:cancer	Genes related to cancer in DISEASE field

A-3. Extraction of necessary information by keyword search

Syntax: http://link.g-language.org/[GENE ID]/extract=[KEYWORD]

1. http://link.g-language.org/9606/extract=dbSNP	Extraction of dbSNP informations

B. Direct input of nucleotide or amino acid sequence

Syntax: http://link.g-language.org/[SEQUENCE]/[options…]

Options

B-1. E-value threshold for BLAT search

Syntax: http://link.g-language.org/[SEQUENCE]/evalue=[E-VALUE] **Default** : 1e-70

B-2. Identity threshold for BLAT search

Syntax: http://link.g-language.org/[SEQUENCE]/identity=[IDENTITY] **Default** : 0.98

B-3. Feeling lucky search against sequence

Syntax: http://link.g-language.org/[SEQUENCE]/direct=[0 or 1] **Default** : 0

If this option is set to "1" (/direct=1), G-Links automatically assigns the ID of top-hit in Swiss-Prot rather than showing the similarity search result.

Figure 1. URL Syntax of G-Links. G-Links is implemented as a RESTful service that can be queried by altering the URL. Full documentation and example queries are available at http://www.g-language.org/wiki/glinks.

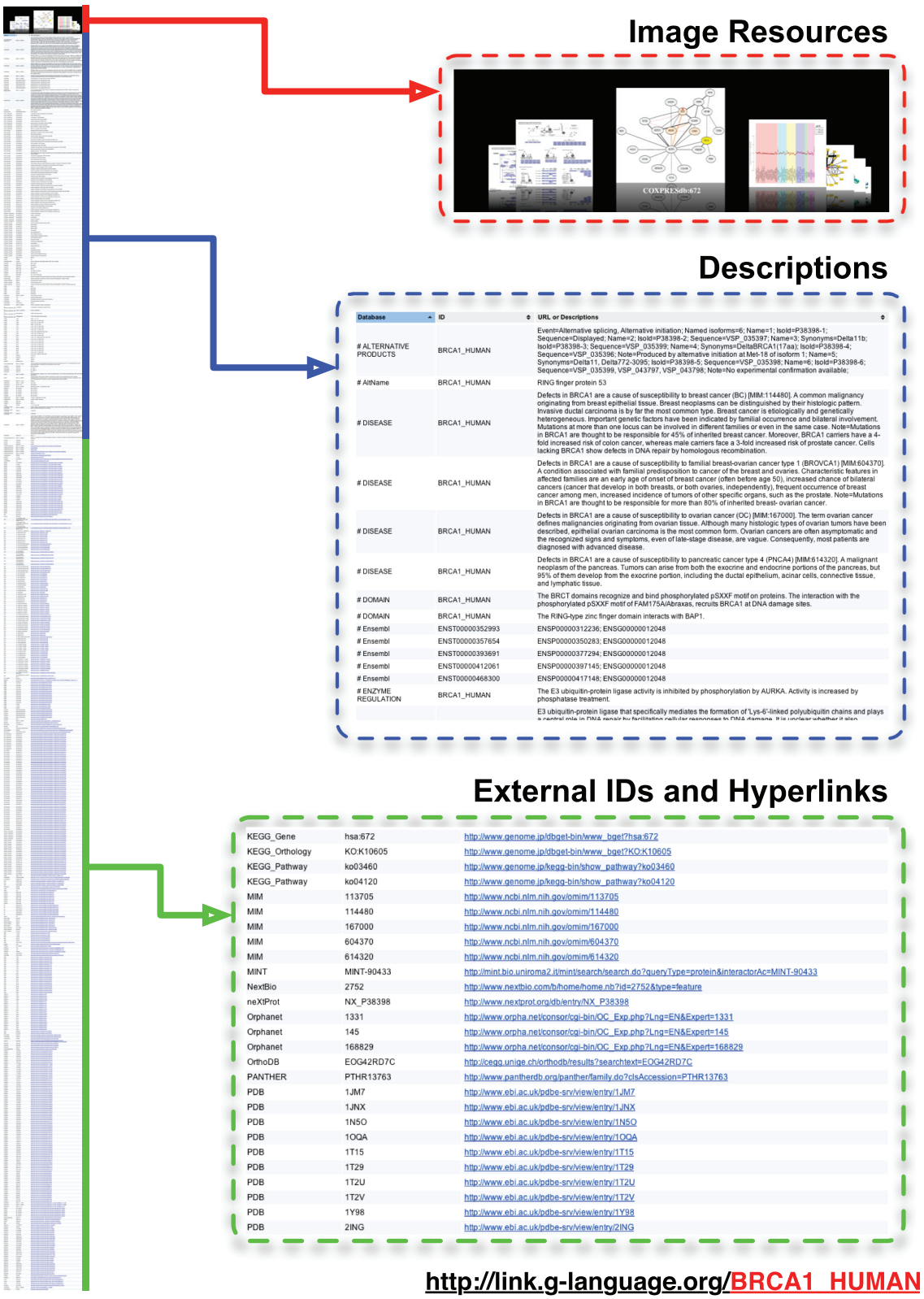

Figure 2. HTML output example of BRCA1_HUMAN (UniProt ID of BRCA1 gene in humans). By default, access to G-Links with web browsers displays the results in interactive HTML, with related image gallery implemented with CoverFlow (http://imageflow.finnrudolph.de/) on the top, followed by a large table of annotations and cross-references.

Table 1. Overview of supported databases and web services in G-Links. Detailed list of Input/Output databases are available at http://link.g-language.org/input_list and http://link.g-language.org/output_list.

Databases (132)				
Genome(11)	Phosphorylation(3)			
Gene(6)	Ortholog(7)	Cluster(1)	Expression(4)	
		SNP(2)	Phylogenesis(2)	
Protein(4)	Structure(5)	Classification(1)	Cluster(4)	
	Family/Domain/Motif(9)	PPI(4)	Enzyme(3)	
Molecular Interaction(2)	Pathway/Reaction(5)	DISEASE/Pathogen/Drug(6)		
Others(15)	Paper(3)	Organisms specific(31)		
Web Services (33)				
	Alignment Local(1)	Data Retrieval Chemistry Data(1)		
	Nucleic Composition(5)	Nucleic CpG Islands(1)	Nucleic Translation(1)	Nucleic Repeats(3)
	Protein Properties(5)	Protein 2D Structure(3)	Protein Composition(3)	Protein Motif(3)
	Protein Localization(4)	Protein Domains(2)	Protein Functional Site(1)	

HTML format and computer readable TSV as well as JavaScript Object Notation (JSON) output, G-Links supports RDF/XML and Notation3 (http://www.w3.org/TeamSubmission/n3/) formats, so that the query results can be readily integrated with Semantic Web technologies. For RDF and Notation3 predicate information is given by EMBRACE Data and Methods (EDAM) ontology.

Likewise other bioinformatics tools such as BioGrid (Chatr-Aryamontri *et al.*, 2015) and Cytoscape (Demchak *et al.*, 2014), G-Links can retrieve information related to gene sets or all genes of organisms, and to filter out non-related genes by keyword search (*filter* option) or to extract necessary fields (*extract* option). Using the filtering option, users can retrieve only the subset of genes related to the given keyword. For example, retrieval of all human (taxonomy ID: 9606) genes having GO slim function including the word "transport" is as simple as http://link.g-language.org/9606/filter=GOslim_function:transport/format=tsv/. Similarly, extraction of only the "DISEASE" annotation of BRCA1 gene is simply http://link.g-language.org/BRCA1_HUMAN/extract=DISEASE. Multiple filtering and extraction conditions can be specified using "|" (vertical bar) as the separator, in order to formulate complex queries. For example, retrieval of SNP information from dbSNP and SNPedia for human genes with known polymorphisms related to breast and ovarian cancer in tabular format is queried as http://link.g-language.org/9606/format=tsv/filter=DISEASE:cancer/filter=:breast|:ovarian|:snps|:polymorphisms/extract=dbSNP|SNPedia.

The gene-centric approach is effective for data aggregation from a variety of databases, especially for prokaryotes, where the genes, transcripts, and proteins are mostly synonymous. On the other hand, this approach can be a limitation for many questions in eukaryote systems biology that require a transcript-centric approach due to the large complexity and diversity of transcriptome regulated by alternative splicing (Nilsen & Graveley, 2010). Currently G-Links lists information of all transcript isoforms, their structures and other annotations, and therefore the gene-centric information can be queried from the identifiers related to the isoforms, but not necessarily the other way around.

Conclusions

By serving as a data hub of linked open biological data, G-Links can be a starting point in retrieval of gene-centric information. Users can quickly obtain related links and annotations of a gene of interest either graphically via HTML or programmatically via REST interface, such as the orthologs, Gene Ontology terms, protein structure, pathways, SNPs, and publications.

Software availability

Software access

G-Links is a RESTful service with base URL http://link.g-language.org/. Detailed documentation is available at http://www.g-language.org/wiki/glinks including service description, syntax, list of all available options, example queries (URLs) and sample scripts for programmatic access in Perl, Ruby, Python, and Java. Examples of programmatic access from the UNIX commandline for Gene Ontology classification of all genes in *E. coli*, as well as for specific set of genes of interest for possible Gene Ontology enrichment analysis, or KEGG BRITE enrichment analysis are also provided. Comprehensive lists of supported input/output databases and web services are available at http://link.g-language.org/input_list and http://link.g-language.org/output_list. Internal database of G-Links is regularly updated every six months, and only the latest version of each resource is accessible, and the source code is freely available from GitHub repository (http://github.com/cory-ko/G-Links).

Latest source code
http://github.com/cory-ko/G-Links

Source code as at the time of publication
https://github.com/F1000Research/G-Links/releases/tag/v1.0

Archived source code as at the time of publication
http://dx.doi.org/10.5072/zenodo.12701 (Oshita & Arakawa, 2014).

License
MIT License

Author contributions
KO and KA conceived and designed the software, and KO implemented the software. MT provided supervision for the study. KO and KA drafted the manuscript, and all authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests
No competing interests were disclosed.

Grant information
This work was supported by KAKENHI Grant Number 22681029 from the Japan Society for the Promotion of Science (JSPS), and by funds from the Yamagata Prefectural Government and Tsuruoka City.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
The development of web services described in this work was significantly facilitated by the DBCLS BioHackathon 2012 hosted by the National Bioscience Database Center/Database Center for Life Science (NBDC/DBCLS). We thank the members of G-language Project at the Institute for Advanced Biosciences, Keio University, and elsewhere, for their extremely valuable feedback and support.

References

Aoki-Kinoshita KF, Kinjo AR, Morita M, *et al.*: **Implementation of linked data in the life sciences at BioHackathon 2011.** *J Biomed Semantics.* 2015; **6**: 3.
PubMed Abstract | Publisher Full Text | Free Full Text

Arakawa K, Kido N, Oshita K, *et al.*: **G-language genome analysis environment with REST and SOAP web service interfaces.** *Nucleic Acids Res.* 2010; **38**(Web Server issue): W700–705.
PubMed Abstract | Publisher Full Text | Free Full Text

Arakawa K, Tomita M: **Merging multiple omics datasets in silico: statistical analyses and data interpretation.** *Methods Mol Biol.* 2013; **985**: 459–470.
PubMed Abstract | Publisher Full Text

Bairoch A, Boeckmann B, Ferro S, *et al.*: **Swiss-Prot: juggling between evolution and stability.** *Brief Bioinform.* 2004; **5**(1): 39–55.
PubMed Abstract | Publisher Full Text

Bhagat J, Tanoh F, Nzuobontane E, *et al.*: **BioCatalogue: a universal catalogue of web services for the life sciences.** *Nucleic Acids Res.* 2010; **38**(Web Server issue): W689–694.
PubMed Abstract | Publisher Full Text | Free Full Text

Brazas MD, Yim D, Yeung W, *et al.*: **A decade of Web Server updates at the Bioinformatics Links Directory: 2003–2012.** *Nucleic Acids Res.* 2012; **40**(Web Server issue): W3–W12.
PubMed Abstract | Publisher Full Text | Free Full Text

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, *et al.*: **The BioGRID interaction database: 2015 update.** *Nucleic Acids Res.* 2015; **43**(Database issue): D470–478.
PubMed Abstract | Publisher Full Text | Free Full Text

Cote RG, Jones P, Martens L, *et al.*: **The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases.** *BMC Bioinformatics.* 2007; **8**: 401.
PubMed Abstract | Publisher Full Text | Free Full Text

Demchak B, Hull T, Reich M, *et al.*: **Cytoscape: the network visualization tool for GenomeSpace workflows [version 2; referees: 3 approved].** *F1000Res.* 2014; **3**: 151.
PubMed Abstract | Publisher Full Text | Free Full Text

Diehn M, Sherlock G, Binkley G, *et al.*: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res.* 2003; **31**(1): 219–223.
PubMed Abstract | Publisher Full Text | Free Full Text

Harris MA, Clark J, Ireland A, *et al.*: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res.* 2004; **32**(Database issue): D258–261.
PubMed Abstract | Publisher Full Text | Free Full Text

Huang H, McGarvey PB, Suzek BE, *et al.*: **A comprehensive protein-centric ID mapping service for molecular data integration.** *Bioinformatics.* 2011; **27**(8): 1190–1191.
PubMed Abstract | Publisher Full Text | Free Full Text

Kanehisa M, Goto S, Sato Y, *et al.*: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic Acids Res.* 2012; **40**(Database issue): D109–114.
PubMed Abstract | Publisher Full Text | Free Full Text

Katayama T, Arakawa K, Nakao M, *et al.*: **The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium*.** *J Biomed Semantics.* 2010; **1**(1): 8.
PubMed Abstract | Publisher Full Text | Free Full Text

Katayama T, Wilkinson MD, Aoki-Kinoshita KF, *et al.*: **BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains.** *J Biomed Semantics.* 2014; **5**(1): 5.
PubMed Abstract | Publisher Full Text | Free Full Text

Katayama T, Wilkinson MD, Micklem G, *et al.*: **The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies.** *J Biomed Semantics.* 2013; **4**(1): 6.
PubMed Abstract | Publisher Full Text | Free Full Text

Katayama T, Wilkinson MD, Vos R, *et al.*: **The 2nd DBCLS BioHackathon: interoperable bioinformatics Web services for integrated applications.** *J Biomed Semantics.* 2011; **2**: 4.
PubMed Abstract | Publisher Full Text | Free Full Text

Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res.* 2002; **12**(4): 656–664.
PubMed Abstract | Publisher Full Text | Free Full Text

Nilsen TW, Graveley BR: **Expansion of the eukaryotic proteome by alternative splicing.** *Nature.* 2010; **463**(7280): 457–463.
PubMed Abstract | Publisher Full Text | Free Full Text

Obayashi T, Okamura Y, Ito S, *et al.*: **COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals.** *Nucleic Acids Res.* 2013; **41**(Database issue): D1014–1020.
PubMed Abstract | Publisher Full Text | Free Full Text

Oshita K, Arakawa K, Tomita M: **KBWS: an EMBOSS associated package for accessing bioinformatics web services.** *Source Code Biol Med.* 2011; **6**: 8.
PubMed Abstract | Publisher Full Text | Free Full Text

Oshita K, Arakawa K: **G-Links: F1000Research/G-Links.** *Zenodo.* 2014.
Data Source

Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet.* 2000; **16**(6): 276–277.
PubMed Abstract | Publisher Full Text

Rose PW, Bi C, Bluhm WF, *et al.*: **The RCSB Protein Data Bank: new resources for research and education.** *Nucleic Acids Res.* 2013; **41**(Database issue): D475–482.
PubMed Abstract | Publisher Full Text | Free Full Text

Smedley D, Haider S, Ballester B, *et al.*: **BioMart--biological queries made easy.** *BMC Genomics.* 2009; **10**: 22.
PubMed Abstract | Publisher Full Text | Free Full Text

Smith RN, Aleksic J, Butano D, *et al.*: **InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data.** *Bioinformatics.* 2012; **28**(23): 3163–3165.
PubMed Abstract | Publisher Full Text | Free Full Text

Stein L: **Creating a bioinformatics nation.** *Nature.* 2002; **417**(6885): 119–120.
PubMed Abstract | Publisher Full Text

Stein LD: **Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges.** *Nature Rev Genet.* 2008; **9**(9): 678–688.
PubMed Abstract | Publisher Full Text

Wu C, Macleod I, Su AI: **BioGPS and MyGene.info: organizing online, gene-centric information.** *Nucleic Acids Res.* 2013; **41**(Database issue): D561–565.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status:  

Version 2

Referee Report 20 November 2015

doi:10.5256/f1000research.6534.r11277

Mark Ragan
Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, 4072, Australia

The Authors have responded satisfactorily to our queries and concerns.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 22 January 2015

doi:10.5256/f1000research.6153.r7340

Kenji Satou
Institute of Science and Engineering, Kanazawa University, Kanazama, Japan

As described in this paper, G-Links system provides a sophisticated way of accessing gene-related information scattered in various databases. The revisions recommended by the first reviewer are still helpful. I think this paper is worth indexing after following the recommended revisions as much as possible.

Minor comment: Isn't the number 22681029?

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 04 Nov 2015

Kazuharu Arakawa, Keio University, Japan

Minor comment: Isn't the number 22681029?

The KAKENHI Grant Number is revised accordingly.

Competing Interests: There is no competing interests.

Referee Report 27 November 2014

doi:10.5256/f1000research.6153.r6750

Mark Ragan, **Sriganesh Srihari**, **Alison Anderson**
Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, 4072, Australia

This paper introduces a simple approach to data integration that can assist bioinformatics researchers. The RESTful API is easy-to-use and allows gene-centric linking of information from a very large numbers of data sources.

We recommend the following revisions:

Scope
The authors rightly present data integration as a key challenge in bioinformatics (Abstract), and present their work as aiming to solve this problem. However, data integration is a much deeper problem than just data aggregation.

Literature review
The authors might mention earlier link-based aggregators, *e.g.* BioMOBY.

Gene centricity
A gene-centric approach is appropriate for some organisms (notably bacteria and archaea) and problems, but many questions in eukaryote systems biology require a transcript-centric approach. The authors might mention this (significant) limitation.

Aliases and catching errors
AOF2 is an alias for KDM1A. A search for KDM1A works fine. A search for AOF2 using the search box on the website or using http://link.g-language.org/AOF2 returns a blank page with no error message. A search using http://link.g-language.org/AOF2_HUMAN gives an inappropriate error message (contact root@localhost).

The article should describe whether and how the system tries to deal with gene aliases, and appropriate feedback should be given to the user when no results are found.

Website
G-Links appears to be an extension of the G-language Project. Having the G-Links homepage within the G-language Project wiki is confusing, as menu-items on the left are relevant to the latter. We also suggest that the 'Quick Star' should be near the top of the page to facilitate ease of use.

Comments on current Semantic Web technologies
The last sentence of the second paragraph of the Introduction states that with current Semantic Web

technologies, cross-domain queries require extensive reasoning or manual curation of ontologies. This is not an accurate description of the current state of Semantic Web technologies. The authors might alternatively make reference to the latest version (July 2014) of the Bio2RDF resource, which enables integration and federated queries across 35 datasets, and suggest that a limitation (in comparison to their RESTful API approach) is that the user is required to be familiar with the SPARQL search language.

Example use cases
The use cases provided in the manuscript are quite simple and don't really demonstrate why the G-Links approach can be more powerful than using (for instance) GeneCards to get an overview of a gene such as BRCA1, or just going to the KEGG website and running a search on 'cancer'. A more comprehensive use case that shows, for example, how easily results from a gene list search in (for instance) TSV format can be programmatically searched to identify common elements for the genes of interest, might better demonstrate the utility of G-Links.

Data integration as a current challenge
The references given to support the claim that data integration is currently a major challenge for the bioinformatics field are 12 and 6 years old. Several features claimed as advantages for G-Links (linking to multiple databases, retrieving information on genes from multiple organisms, filtering by keywords, extracting by fields) are reasonably common, *e.g.* in BioGrid or Cytoscape.

Licensing
If any linked resources have license rights that need to be adhered to, this information should be brought to the user's attention, perhaps using a note on the G-Links website.

Constraints
For best practice, the system should meet the constraints for RESTful – stateless, cacheable and so on; these might be described briefly in the paper.

The authors might mention that multiple URIs can point to the same resource.

A resource can exist in different versions; does the system capture and display the version from which information has been captured?

Finally, we recommend that the authors present a Semantic Web-specific example for their statement that 'G-links can be used programmatically as text data, from Semantic Web services..'.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Author Response 04 Nov 2015

Kazuharu Arakawa, Keio University, Japan

We would like to thank the reviewer for thorough review, and apologize for the extreme delay in our revision. Following are point-by-point comments for our revision.

Scope

The authors rightly present data integration as a key challenge in bioinformatics (Abstract), and present their work as aiming to solve this problem. However, data integration is a much deeper problem than just data aggregation.

We have toned down this claim as follows: “We aim to provide an identifier conversion and data aggregation system as a part of solution to solve this problem”.

Literature review
The authors might mention earlier link-based aggregators, *e.g.* BioMOBY.

BioMOBY is a registry of bioinformatics web services, and is not a link-based aggregators. It allow the suggestion and discovery of web services based on the data type (ex. to discover BLAST services from a FASTA file – discovery of a service for a given input, or NCBI data retrieval services that produces a FASTA file – discovery of a service for a given output.). As an example of earlier link-based aggregators, we have mentioned MyGene.info service.

Gene centricity
A gene-centric approach is appropriate for some organisms (notably bacteria and archaea) and problems, but many questions in eukaryote systems biology require a transcript-centric approach. The authors might mention this (significant) limitation.

We thank the Reviewer for this important comment. We have added the following paragraph in the manuscript to clarify this limitation.
“The gene-centric approach is effective for data aggregation from a variety of databases, especially for prokaryotes, where the genes, transcripts, and proteins are mostly synonymous. On the other hand, this approach can be a limitation for many questions in eukaryote systems biology that require a transcript-centric approach due to the large complexity and diversity of transcriptome regulated by alternative splicing {Nilsen, 2010 #178}. Currently G-Links lists information of all transcript isoforms, their structures and other annotations, and therefore the gene-centric information can be queried from the identifiers related to the isoforms, but not necessarily the other way around.“

Aliases and catching errors
AOF2 is an alias for KDM1A. A search for KDM1A works fine. A search for AOF2 using the search box on the website or using http://link.g-language.org/AOF2 returns a blank page with no error message. A search using http://link.g-language.org/AOF2_HUMAN gives an inappropriate error message (contact root@localhost).

The system is modified to provide 404 and 500 HTML errors with corresponding error messages when ID cannot be resolved. Redirection for aliases are likewise revised, and now http://link.g-language.org/AOF2_HUMAN shows a list of redirections.

The article should describe whether and how the system tries to deal with gene aliases, and appropriate feedback should be given to the user when no results are found.

Error handling is updated as described above. We do not specifically implement support for gene symbols, but UniProt aliases are now correctly handled. For gene symbols, ChiTaRS database contains the information for model eukaryotes, so all information about KDM1A can be retrieved from http://link.g-language.org/KDM1A, although it takes a little while to load all information.

Website
G-Links appears to be an extension of the G-language Project. Having the G-Links homepage within the G-language Project wiki is confusing, as menu-items on the left are relevant to the latter. We also suggest that the 'Quick Star' should be near the top of the page to facilitate ease of use.

The website (http://link.g-language.org/) is now not redirected and is given a dedicated page, and the 'Quick Start' menu is moved to the top as suggested.

Comments on current Semantic Web technologies
The last sentence of the second paragraph of the Introduction states that with current Semantic Web technologies, cross-domain queries require extensive reasoning or manual curation of ontologies. This is not an accurate description of the current state of Semantic Web technologies. The authors might alternatively make reference to the latest version (July 2014) of the Bio2RDF resource, which enables integration and federated queries across 35 datasets, and suggest that a limitation (in comparison to their RESTful API approach) is that the user is required to be familiar with the SPARQL search language.

Revised accordingly.

Example use cases
The use cases provided in the manuscript are quite simple and don't really demonstrate why the G-Links approach can be more powerful than using (for instance) GeneCards to get an overview of a gene such as BRCA1, or just going to the KEGG website and running a search on 'cancer'. A more comprehensive use case that shows, for example, how easily results from a gene list search in (for instance) TSV format can be programmatically searched to identify common elements for the genes of interest, might better demonstrate the utility of G-Links.

We now provide illustrative examples of programmatic access from the UNIX commandline for Gene Ontology classification of all genes in *E.coli*, as well as for specific set of genes of interest for possible Gene Ontology enrichment analysis, or KEGG BRITE enrichment analysis in the website, and it is also mentioned in the text.

Data integration as a current challenge
The references given to support the claim that data integration is currently a major challenge for the bioinformatics field are 12 and 6 years old.

Two more latest reviews are added.

Several features claimed as advantages for G-Links (linking to multiple databases, retrieving information on genes from multiple organisms, filtering by keywords, extracting by fields) are reasonably common, *e.g.* in BioGrid or Cytoscape.

We have removed the mention that these are advantages, and mentioned that these features are

common as in BioGrid and Cytoscape.

Licensing
If any linked resources have license rights that need to be adhered to, this information should be brought to the user's attention, perhaps using a note on the G-Links website.

The information is added to the website.

Constraints
For best practice, the system should meet the constraints for RESTful – stateless, cacheable and so on; these might be described briefly in the paper.

Brief description is added to the manuscript: "The server provides a uniform interface based on URL and HTTP in a client-server model, which is stateless and therefore the server does not store any client context information, and the clients and intermediates can cache responses between server update cycles, duration of which is specified by HTML META tag."

The authors might mention that multiple URIs can point to the same resource.

Revised accordingly.

A resource can exist in different versions; does the system capture and display the version from which information has been captured?

G-Links only provides the latest resource, and the manuscript is revised to include this information.

Finally, we recommend that the authors present a Semantic Web-specific example for their statement that 'G-links can be used programmatically as text data, from Semantic Web services..'.

Since the use of RDF data requires storage of data obtained from G-Links in a triple store with other semantic web resources, we have removed the claim "from Semantic Web services".

Competing Interests: There is no competing interest.