Published online 19 June 2014 *Nucleic Acids Research, 2014, Vol. 42, Web Server issue* ***W473–W477***
doi: 10.1093/nar/gku517

Lynx web services for annotations and systems analysis of multi-gene disorders

Dinanath Sulakhe[1,*], Andrew Taylor[2], Sandhya Balasubramanian[2], Bo Feng[3], Bingqing Xie[3], Daniela Börnigen[2,4], Utpal J. Dave[1], Ian T. Foster[1], T. Conrad Gilliam[1,2] and Natalia Maltsev[1,2]

[1]Computation Institute, University of Chicago/Argonne National Laboratory, Chicago, IL 60637, USA, [2]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA, [3]Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA and [4]Toyota Technological Institute at Chicago, Chicago, IL 60637, USA

Received March 14, 2014; Revised May 25, 2014; Accepted May 26, 2014

ABSTRACT

Lynx is a web-based integrated systems biology platform that supports annotation and analysis of experimental data and generation of weighted hypotheses on molecular mechanisms contributing to human phenotypes and disorders of interest. Lynx has integrated multiple classes of biomedical data (genomic, proteomic, pathways, phenotypic, toxicogenomic, contextual and others) from various public databases as well as manually curated data from our group and collaborators (LynxKB). Lynx provides tools for gene list enrichment analysis using multiple functional annotations and network-based gene prioritization. Lynx provides access to the integrated database and the analytical tools via REST based Web Services (http://lynx.ci.uchicago.edu/webservices.html). This comprises data retrieval services for specific functional annotations, services to search across the complete LynxKB (powered by Lucene), and services to access the analytical tools built within the Lynx platform.

INTRODUCTION

Gaining a greater understanding of molecular mechanisms underlying common multi-gene disorders (e.g. autism, schizophrenia, diabetes) is a major challenge in biomedical research (1). Construction of predictive models of such mechanisms critically depends on the availability of high-throughput genomic data and efficient algorithmic approaches for mining this data with clinical observations and prior knowledge about genotype–phenotype relationships. The unprecedented increase in the production of biological data has led to a number of valuable biological databases. While these databases are very useful in the analysis of data from high-throughput genome-wide associations, expression profiling or next-generation sequencing, accessing these distributed databases can be challenging. Oftentimes, there are a number of databases representing the same classes of information in different non-standard formats with different identifiers and poor cross-references connecting them. The size and frequent changes add to the challenges in using these disparate databases for large high-throughput studies.

To address these challenges, we have developed an integrated Lynx platform that consists of a knowledgebase (LynxKB (2)) which periodically collects various classes of biological data in a structured relational database, analytical tools for the analysis of multi-gene lists and a Lynx web services interface which allows users to query and search LynxKB to retrieve annotations and access the analytical tools. LynxKB integrates many classes of information, including genomic, proteomic, pathways-related, disease-specific, phenotypic, variations, text mining, pharmacogenomics and more from over 35 different public databases (2). It also contains manually curated data collections, including weighted collections of candidate genes extracted from Developmental Brain Disorders Database (DBDB) (3) and LisDB (https://lisdb.ci.uchicago.edu). Currently, Lynx supports only human data (taxonomy: 9606), while mouse and rat related databases are to be integrated in the near future.

Lynx web services are implemented using REST (Representational State Transfer) architecture to provide a simple interface with multiple request types supported (HTTP GET and POST). The results of these RESTful services can be requested in XML and JSON formats for easy consumption.

ARCHITECTURE, DESIGN AND IMPLEMENTATION

LynxKB is integrated and stored in a normalized relational database using MySQL. In order to connect the information between different sources, appropriate cross-reference

*To whom correspondence should be addressed. Tel: +1 630 252 7856; Fax: +1 630 252 5657; Email: sulakhe@mcs.anl.gov

data is also integrated. LynxKB currently has a data volume of more than 800GB stored in the database. In order to provide a comprehensive search capability, we use Apache Lucene (http://lucene.apache.org), whereby appropriate indexes of the data are created for Lucene. An advanced search web service is implemented on top of this Lucene framework.

Lynx is implemented using Service-Oriented Architecture concepts, such that all applications within the Lynx framework are built using web services interfaces. Lynx web services are implemented using the Jersey framework (https://jersey.java.net) (JAX-RS Reference Implementation) (4) and Spring framework (5) to provide RESTful web services. The domain specific datatypes and return types for the web services are modeled and represented as XML schemas (XSDs) using JAXB (6) and are automatically translated into domain specific Java objects that are instantiated with data from the MySQL database. Thus, all the domain specific data types used to hold the queried data and return types for web services are defined as XML schemas. As such, the first step in creating any new web service in Lynx starts with defining all of the data types necessary for that service in a XSD. JAXB also 'marshals' the java objects back into XML or JSON formats as the web service return types. The XML schemas for return data types help users to implement appropriate client scripts, as they know what data structure to expect from the Lynx web services.

WEB SERVICES

Lynx provides a large collection of intuitive RESTful web services to get annotations or to perform analyses for a single gene or a list of genes. These web services can be classified into three broad categories: (i) data retrieval services, (ii) search services and (iii) analytical services. All of these web-services in Lynx are genes-centric, such that users can create a list of genes based on certain criteria using the search services, or retrieve annotations and perform analysis on a list of genes. All of these web services can be accessed via HTTP GET or POST based requests and the results can be requested in XML or JSON format.

Data retrieval services

The Lynx data services allow users to retrieve various classes of annotations (genomic, proteomic, pathways, diseases, phenotypic, toxicogenomic, contextual, interactions, etc.) for a list of genes using Entrez gene IDs or gene symbols. Table 1 below shows a mapping between the specific RESTful resource and sources of the data.

Figure 1 shows the structure of a URL used for a HTTP GET-based data retrieval web service. Lynx currently supports retrieval of seven features for a given set of genes. For example, the URL: http://lynx.ci.uchicago.edu/gediresources/resources/genes/9606/pxn:akt1:cask/pathways retrieves pathway information from various pathway databases for the three genes (PXN, AKT1, CASK). The web services page on the Lynx website provides some example URLs for retrieving the data. By default, these resources provide the data in XML format. All of these data retrieval web services can be accessed using a POST request as well. For example, the following curl command is a HTTP POST request, equivalent of the HTTP GET request above and returns the pathways in JSON format when run from a Linux terminal or a CURL client program:

```
curl -i -H "Content-Type: application/json" -H "Accept: application/json" -X POST -d "{"taxid":"9606", "geneids":["CASK","AKT1", "PXN"]}" http://lynx.ci.uchicago.edu/gediresources/resources/genes/post/pathways
```

Search services

Additionally, Lynx provides a robust genes centric search engine. Users can search against any information about genes, pathways, tissues, diseases and symptoms. The queries can be for a specific functional term or general keyword and can request a fuzzy or strict search. Lynx search service also supports Boolean search operators (AND, OR and NOT) allowing multiple search criteria with refined results. The results returned back are genes centric, thus, allowing users to generate gene lists for a given search criterion and use them with other Lynx analytical services as seed genes or test genes. The results also contain detailed annotations for those genes, including relevant pathways, tissues, diseases and symptoms. Currently, Lynx supports keyword-based searches for direct associations of genes with certain features such as diseases, pathways, tissues and symptoms. We plan to implement the searches against ontology-based hierarchies in the future releases.

The following CURL based POST request is an example of a search to get a list of genes that are associated with autism and seizures:

```
curl -i -H "Content-Type: application/json" -H "Accept: application/json" -X POST -d "{"input": "DISEASES:autism and SYMPTOMS:seizures"}" http://lynx.ci.uchicago.edu/gediresources/resources/search/gene
```

More detailed examples of GET and POST based search services are provided on the Lynx's web services web page.

Analytical services

Gene lists coming out of high-throughput genomic data analyses (e.g. Next Generation Sequence (NGS) data analysis, gene expression results, Copy-Number Variation (CNV) analysis, expert ranked candidate genes) require subsequent downstream analysis to gain a better understanding of the underlying disease or condition of interest. Here, Lynx's statistical enrichment analysis (20) helps identify the functional categories (e.g. Gene Ontology (GO) terms, diseases, tissues, phenotype, pathways, transcription factor binding sites (21) and enhancers (22)) that are over-represented in the submitted gene set. This enrichment analysis can be performed against all of the human genes or against a specific context (e.g. against genes expressed in a particular tissue or on a particular developmental stage). Lynx's REST based web service can be used to programmatically perform the enrichment analysis by providing the training gene set and the test gene set, selecting the training parameters, *P*-value cutoffs and correction (Bonferroni or False Discovery Rate (FDR)). The example HTTP POST requests shown at http://lynx.ci.uchicago.edu/webservices.html#enrich-genes provide more details on how to access this analytical tool via the REST web service interface. Lynx also provides a tool for network-based gene prioritization for the prediction of

Table 1. Lynx data retrieval Web Services

REST resource	Data sources
Geneinfo	NCBI (7), Ensembl (8), UniGene (7), TRANSFAC (9), RefSeq (10)
Pathways	KEGG (11), Reactome (12), NCI (13), BioCarta, Pathway Commons (14)
Diseases	OMIM, AutDB (15), Schizophrenia Gene Resource (SZGR) (16), Diseases (University of Copenhagen), Cancer gene index, DBDB
Interactions	NCBI (7), MINT (17), KEGG , Reactome (12), NCI (13), BioCarta, GeneWays (18)
Tissues	NCBI UniGene (7)
Symptoms	Human Phenotype Ontology (19)

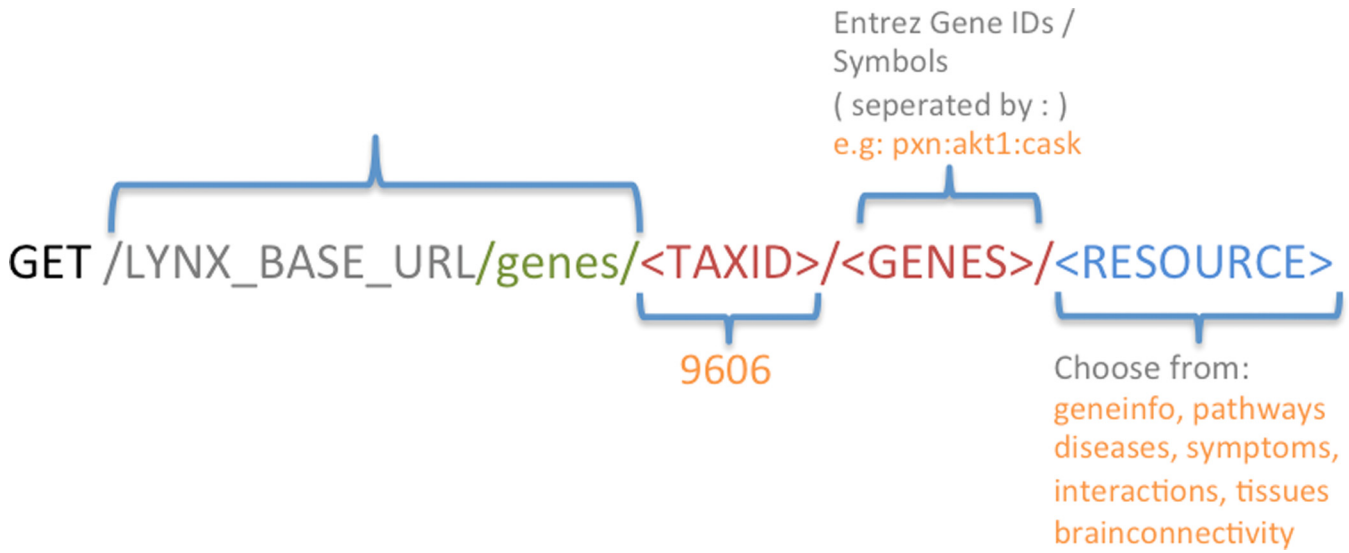

Figure 1. URL Structure to retrieve annotations for a list of genes using HTTP GET.

high-confidence candidate genes from a large set of genes or even from the entire genome for a disease or phenotype of interest. It is based on PINTA (23,24) and provides five different network propagation algorithms (heat kernel diffusion (25), Page Rank with Priors (26), HITS with prior (27), simple random walk K-step markov (27)) while using STRING version 9.0 (28) as the underlying protein interaction network. Detailed examples on how to use this tool via the REST web service interface are provided at http://lynx.ci.uchicago.edu/webservices.html#nw-prioritization.

DISCUSSION

In Lynx, we have implemented a large integrated biomedical knowledgebase (LynxKB) and a collection of RESTful web services to access this information in a structured format. In studies involving multi-gene disorders or downstream analysis of next-generation sequence (NGS) data that results in large lists of genes, it tends to be extremely difficult for researchers to accumulate all of the annotation information from multiple sources of databases. Lynx's RESTful web services present a useful one-stop service not only for annotations but also for analysis of unknown lists of genes.

Lynx web services can be used in multiple different scenarios by individual researchers or developers building large systems. For example, individual researchers can use CURL on a command line or write a simple Perl script and collect all of the annotations for their genes of interest. The Lynx web application available at http://lynx.ci.uchicago.edu is a perfect example of a large application that is entirely built on top of Lynx's RESTful services. Similarly, developers can consume these web services within a Next-Generation Sequencing analysis platform such as Galaxy [] by writing tools to annotate the Variant Call Format (VCF) files using Lynx data retrieval services; also, they can analyze the genes in these VCF files using Lynx analytical services to find over-represented pathways, diseases, phenotype and other functional categories.

There are other systems such as David that provide similar web services for annotations and enrichment analysis. In comparison, Lynx's REST based web services provide a comprehensive Lucene based Search service that allows developers to fetch any information at a very granular level. Lynx also allows performing Network based gene prioritization (using five different algorithms) that is a unique capability. Lynx's RESTful architecture makes it easy and flexible in writing client applications in any language of interest and also access the web services from a command-line (using CURL) or in a browser (using GET resources).

In the near future, we will be adding more systems biology related functionalities with emphasis on network analysis and addition of contextual data (e.g. expression data) through out the system. We will also provide ontological support for the user queries. We will include, besides the data retrieval and analysis using already integrated ontologies (e.g. GO) additional ontologies, such as anatomical and

developmental ontologies (e.g. developed by Allen Brain Atlas (29), Bgee (30)) as well as additional phenotype and disease-related ontologies and controlled dictionaries (e.g. disease ontology (31), behavioral ontology (https://code.google.com/p/behavior-ontology/)).

We provide detailed documentation with examples (GET and POST request examples) of all of our web services as well as information on any new updates at http://lynx.ci.uchicago.edu/webservices.html. We can be reached via the email address listed on the Lynx website for any specific web service related or general Lynx based support.

ACKNOWLEDGMENTS

The authors would like to acknowledge support from the Autism Genetic Resource Exchange (AGRE) and Autism Speaks. The authors gratefully acknowledge the resources provided by the AGRE consortium and the participating AGRE families. AGRE is a program of Autism Speaks. The authors also acknowledge the computational and storage resources and support provided by the Computation Institute at the University of Chicago.

FUNDING

Mr and Mrs Lawrence Hilibrand, Boler Family Foundation and National Institutes of Health/National Institute of Neurological Disorders and Stroke [NS050375]; Genetic Basis of Mid-Hindbrain Malformations; National Institute of Mental Health [1U24MH081810 to Clara M. Lajonchere (PI)]. Funding for open access charge: National Institutes of Health/National Institute of Neurological Disorders and Stroke [NS050375].
Conflict of interest statement. None declared.

REFERENCES

1. Sulakhe,D., Balasubramanian,S., Xie,B., Berrocal,E., Feng,B., Taylor,A., Chitturi,B., Dave,U., Agam,G., Xu,J. *et al.* (2014) High-throughput translational medicine: challenges and solutions. *Adv. Exp. Med. Biol.*, **799**, 39–67.
2. Sulakhe,D., Balasubramanian,S., Xie,B., Feng,B., Taylor,A., Wang,S., Berrocal,E., Dave,U., Xu,J., Bornigen,D. *et al.* (2014) Lynx: a database and knowledge extraction engine for integrative medicine. *Nucleic Acids Res.*, **42**, D1007–D1012.
3. Mirzaa,G. M., Millen,K. J., Barkovich,A. J., Dobyns,W. B. and Paciorkowski,A. R. (2014) The Developmental Brain Disorders Database (DBDB): A curated neurogenetics knowledge base with clinical and research applications. *Am. J Med. Genet. Part A*,**164**, 1503–1511.
4. Potociar,M. (2009) JSR 311: JAX-RS: the java API for RESTful web services, *Technical Report*.
5. Johnson,R., Hoeller,J., Arendsen,A., Risberg,T. and Kopylenko,D. (2005). *Professional Java Development with the Spring Framework*. Wrox Press Ltd, Birmingham, UK.
6. Kawaguchi,K., Vajjhala,S. and Fialli,J. (2006) *The Java Architecture for XML Binding (JAXB) 2.1*. Sun Microsystems, Inc.
7. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **42**, D7–D17.
8. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G. and Fairley,S. (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
9. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
10. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
11. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
12. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
13. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37**, D674–D679.
14. Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,O., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
15. Basu,S.N., Kollu,R. and Banerjee-Basu,S. (2009) AutDB: a gene reference resource for autism research. *Nucleic Acids Res.*, **37**, D832–D836.
16. Jia,P., Sun,J., Guo,A.Y. and Zhao,Z. (2010) SZGR: a comprehensive schizophrenia gene resource. *Mol. Psychiatry*, **15**, 453–462.
17. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
18. Rzhetsky,A., Iossifov,I., Koike,T., Krauthammer,M., Kra,P., Morris,M., Yu,H., Duboue,P.A., Weng,W., Wilbur,W.J. *et al.* (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.*, **37**, 43–53.
19. Kohler,S., Doelken,S.C., Rath,A., Ayme,S. and Robinson,P.N. (2012) Ontological phenotype standards for neurogenetics. *Hum. Mutat.*, **33**, 1333–1339.
20. Xie,B., Agam,G., Sulakhe,D., Maltsev,N., Chitturi,B. and Gilliam,T.C. (2012) In: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine. ACM, New York, NY. pp. 564–566.
21. Gotea,V., Visel,A., Westlund,J.M., Nobrega,M.A., Pennacchio,L.A. and Ovcharenko,I. (2010) Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.*, **20**, 565–577.
22. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
23. Nitsch,D., Tranchevent,L.-C., Goncalves,J.P., Vogt,J.K., Madeira,S.C. and Moreau,Y. (2011) PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Res.*, **39**, W334–W338.
24. Nitsch,D., Gonçalves,J.P., Ojeda,F., De Moor,B. and Moreau,Y. (2010) Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, **11**, 460.
25. Saad,Y. (1992) Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, **29**, 209–228.
26. Page,L., Brin,S., Motwani,R. and Winograd,T. (1999) The PageRank citation ranking: bringing order to the web, *Technical Report*, Stanford InfoLab.
27. Chen,J., Aronow,B.J. and Jegga,A.G. (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, **10**, 73.
28. Franceschini,A., Szklarczyk,D., Frankild,S., Kuhn,M., Simonovic,M., Roth,A., Lin,J., Minguez,P., Bork,P. and von Mering,C. (2013) STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
29. Hawrylycz,M.J., Lein,E.S., Guillozet-Bongaarts,A.L., Shen,E.H., Ng,L., Miller,J.A., van de Lagemaat,L.N., Smith,K.A., Ebbert,A., Riley,Z.L. *et al.* (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, **489**, 391–399.
30. Bastian,F., Parmentier,G., Roux,J., Moretti,S., Laudet,V. and Robinson-Rechavi,M. (2008), *Data Integration in the Life Sciences*. Springer, Berlin-Heidelberg, Vol. **5109**, pp. 124–131.

31. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res*, **40**, D940–D946.