LinkHub: a Semantic Web System for Efficiently Handling Complex Graphs of Proteomics Identifier Relationships that Facilitates Cross-database Queries and Information Retrieval

Andrew K. Smith[1], Kei-Hoi Cheung[1,2,3], Kevin Y. Yip[1], Martin Schultz[1], Mark B. Gerstein[1,4,*]

[1] Department of Computer Science, [2]Center for Medical Informatics, [3]Department of Genetics, [4]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, USA
*Corresponding author

Email: andrew.smith, kei.cheung, yuklap.yip, schultz-martin, mark.gerstein @yale.edu
Phone/Fax: 203-432-6105/360-838-7861 (Andrew K. Smith, Kevin Y. Yip, and Mark B. Gerstein), 203-432-1202/203-432-0593 (Martin Schultz), 203-737-5783/203-737-5708 (Kei-Hoi Cheung)

Abstract. LinkHub is a semantic-web RDF-based system that manages complex graphs of proteomics identifier relationships and allows exploration with web interactive and query interfaces. For efficiency and robustness, we also provide relational-database access and translation between the relational and RDF versions. LinkHub is practically useful in creating small, local hubs on common topics and then connecting these to major portals in a federated architecture; we have used LinkHub to establish such a relationship between UniProt and the North East Structural Genomics Consortium. LinkHub can thus help support loosely coupled, collaborative data integration without requiring explicit coordination or centralization. LinkHub also facilitates queries and access to information spread across multiple databases, acting as "connecting glue" between different identifier spaces. We demonstrate this with example queries discovering “interologs” of yeast protein interactions in the worm and exploring the relationship between gene essentiality and pseudogene content, and also showing how “protein family based” retrieval of documents can be achieved. LinkHub is at hub.gersteinlab.org and hub.nesg.org, with supplements at hub.gersteinlab.org/supplement.

Keywords: semantic web, RDF, information retrieval, interoperation

1 Introduction

Biological research is producing vast amounts of data (e.g. from high throughput experiments such as sequencing projects, and microarray experiments) at a prodigious rate. Most of this data is made freely available to the public, and this has created a large and growing number of internet and world wide web-accessible biological data resources which are characterized by being distributed, heterogeneous, and having a large size variance, i.e. huge, mega-databases such as UniProt [1] down to medium, small or "boutique" databases (e.g., TRIPLES [2]) generated for medium or small scale experiments or particular purposes. Most computational analyses of biological data will require using multiple integrated datasets, and integrated data along with rich, flexible and efficient interfaces to it encourages exploratory data analysis which can lead to serendipitous new discoveries: the sum is greater than its parts. Currently, integration often must be done manually in a labor and time intensive way by finding relevant datasets, understanding them, writing code to combine them, and finally doing the desired analysis. The basic requirements for better, more seamless integrated analysis are uniformity and accessibility; data are ineffectual if scattered among incompatible resources.

Web search engines and hyperlinks are the basic and commonly used ways to find things on the web and navigate web content but they do not enable fine-grained cross-site analysis of data. To improve upon this, one key issue is the need for standardization and its widespread use, and tools supporting and enabling it. Biological data is too vast for brute-force centralization to be the complete solution to data interoperability. We must have standards and systems for people and groups to work independently creating and making data available (although ultimately cooperatively and collaboratively) but still in the end all or most of the pieces of biological knowledge and data are connected together in semantically rich ways. The W3C's (http://www.w3.org) *semantic web* [3, 4] is a promising candidate: it allows web information to be expressed in fine-grain structured ways so applications can more readily and precisely extract and cross-reference key facts and information from it without having to worry about disambiguating meaning from natural language texts. Standard and machine-readable ontologies such as the Gene Ontology [5] are also created and their common use encouraged to further reduce semantic ambiguity, although there is a need to make these ontologies more machine-friendly [6].

A key abstraction or "scaffold" for representing biological data is the notion of unique identifiers for biological entities and relationships among them. For example, each protein sequence in the UniProt database is given a unique accession, e.g. Q60996, which can be used as a key to access its UniProt sequence record. UniProt sequence records also contain cross-references to related information in other databases, e.g. related Gene Ontology and PFAM identifiers (although the relationship types, e.g. "functional annotation" and "family membership" respectively, are not specified). Two identifiers such as Q60996 and GO:0005634 and the cross-reference between them can be viewed as a single edge between two nodes in a graph, and conceptually then a large, important part of biological knowledge can be viewed as a massive graph whose nodes are biological entities such as proteins, genes, etc. represented by identifiers and the links in the graph are typed and are the specific relationships among the biological entities. Figure 1a is a conceptual illustration of the graph of biological identifier relationships; the problem is that this graph only concretely exists piecemeal or not at all.

A basic problem preventing this graph of relationships from being more fully realized is the problem of nomenclature. Often, there are many synonyms for the same underlying entity caused by independent naming, e.g. structural genomics centers assigning their own protein identifiers in addition to UniProt's. There can also be lexical variants of the same underlying identifier (e.g. GO:0008150 vs. GO0008150 vs. GO-8150). Synonyms are a small part of the overall problem, however, and more generally there are many kinds of relationships including one-to-one and one-to-many relationships. For example, a single Gene Ontology or PFAM identifier can be related with many UniProt identifiers (i.e. they all share the same functional annotation or family membership). PFAM represents an important structuring principle for proteins and the genes they come from, the notion of families (or domains) based on evolution; proteins sharing common PFAM domains are evolutionarily related (called *homologs)* and likely have the same or similar functions. Gene Ontology consists of three widely used structured, controlled vocabularies (ontologies) that describe gene products such as proteins in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. The conceptual graph of identifier relationships is richly connected, and a transitive closure even a few levels deep can lead to indirect relationships with a great number of other entities. Being able to store, manage, and work with this graph of entities and relationships can lead to many opportunities for interesting exploratory analysis and LinkHub is such a system for doing this.

2 Implementation

2.1 LinkHub: a system for loosely coupled, collaborative integration of proteomics identifier relationships

The semantic web is increasingly gaining traction as the key standards-based platform for biological data integration [7, also see http://www.w3.org/2001/sw/hcls/], and since LinkHub is a natural fit to semantic web technologies we use them as the basis of LinkHub. LinkHub is designed based on a semantic graph model, which captures the graph of relationships among biological entities discussed above. To provide a scalable implementation while also exploring semantic web database technologies, we implemented LinkHub in both MySQL (http://www.mysql.com) and Resource Description Framework or RDF (http://www.w3.org/RDF/) databases. LinkHub provides various interfaces to interact with this graph, such as a web frontend for viewing and traversing the graph as a dynamic expandable / collapsible HTML list and a mechanism for viewing particular path types in the graph, as well as with RDF query languages.

Centralized data integration to an extent does make sense, e.g. a lab or organization might want to create a local data warehouse of interconnections among its individual data resources; but it does not want to have to explicitly connect its data resources up to everything in existence, which is impossible. The key idea is that if groups independently maintaining data resources each connect their resources up to some other resource X, then any of them can reach any other through these connections to X, and we can collectively achieve incremental global integration of genomics data in this way. LinkHub is a software architecture and system which aims to help realize this goal by enabling one to create such local minor hubs of data interconnections and connecting them to major hubs such as UniProt in a federated "hub of hubs" framework and this is illustrated in figure 1b.

2.2 Obtaining LinkHub Data

To populate the LinkHub database (described below), Perl web crawlers and other scripts were written to fetch and extract data from different resources. Our running LinkHub instantiation is actively used and populated with data from our labs and related to our research interests, primarily proteomics data but also yeast resources, macromolecular motions, and pseudogenes. UniProt is a major resource used as backbone content. Efficient, exact sequence matching Perl scripts and modules that work by creating custom indexes of UniProt and FASTA-format databases were used to do quick cross-referencing based on exact sequence matches (e.g. to cross-reference structural genomics protein targets to UniProt). We also plan to package the LinkHub code and schemas and release them under an open source license.

2.3 LinkHub Database Models

LinkHub is conceptually based on the semantic web (graph) model and we thus represent it and store it in RDF. However, there could be a potential problem in performance and scalability when using the new RDF database technology, which can be an important impediment to more active and widespread use of the semantic web. In this regard, the creation of high-performance RDF databases should be a research priority of the semantic web community. Thus, while we would ideally use only RDF, to support LinkHub's practical daily use we also model and store its data using relational database technology (MySQL) for efficiency and robustness. A drawback is that relational databases do not naturally model graph structures or provide efficient graph operations for which special procedural codes are necessary (e.g. for the "path type" view described below). The Linkhub relational model includes database tables for identifier types, identifiers, and mappings between identifiers ("mapping type" gives the semantics of the relationship, e.g. "family mapping"). The resource table contains the name, description, and template URL (in which concrete identifiers are interpolated to generate instantiated URLs) of web resources; a separate resource_accepts table lists the identifier types accepted by particular resources, with a link_exceptions table listing exceptions for particular identifiers.

RDF is a popular data model (or ontological language) for the semantic web and is designed to provide a natural representation of a directed graph --- essentially, URIs are used to uniquely name the nodes (which represent objects) and edges (which represent relationships), and literal values may also be used in place of pointed to nodes. In addition, RDF comes with query languages (e.g., RDQL, http://www.w3.org/Submission/RDQL) to allow the user to pose semantic queries against graph data. It is straightforward mapping between the relational and RDF versions of LinkHub and we have written Java code to do this. While there are more advanced ontological languages such as the Web Ontology Language or OWL (http://www.w3.org/TR/owl-features/) that support data reasoning based on Description Logics or DL (http://dl.kr.org/), RDF is a good start and much can be effectively modeled with it. For example, the benefits of representing proteomics data in RDF were discussed [7] and UnitProt data has also recently been made available in

RDF format (http://www.isb-sib.ch/~ejain/rdf/data/).

2.5 LinkHub Web Interfaces

The primary interactive interface to the LinkHub database is a web-based interface (implemented using the so-called AJAX technologies, i.e. DHTML, JavaScript, DOM, CSS, etc.; see http://en.wikipedia.org/wiki/AJAX) which presents subsets or subgraphs of the graph of relationships in a dynamic expandable / collapsible list view. For reasonable interactive computation times this interface allows viewing and exploring of the transitive closure of the relationships stemming from a given identifier interactively one layer at a time: direct edges from the given identifier are initially shown and the user may then selectively expand fringe nodes an additional layer at a time to explore further relationships. Figure 2 is a screenshot of the interface and gives more details of it. The second interface presents results the same as the primary interface but allows viewing of particular path types in the graph. For example, one might want to view all proteins in some database D in the same PFAM family as a given protein; in LinkHub PFAM relationships are stored for UniProt proteins, so one could view the fellow family members of the given protein by viewing all identifier relationship paths (starting from the given protein) matching: Given protein in database D → equivalent UniProt protein → PFAM family → UniProt proteins → other equivalent proteins in database D.

3 Results

3.1 Novel Information Retrieval based on LinkHub Relational Graph Structure

The "path type" interface (and more generally RDF query access) to LinkHub allows one to flexibly retrieve useful subsets of the web documents attached to identifier nodes in the graph based on the graph's relational structure. Normal search engines relying on keyword searches could not provide such access, and LinkHub thus enables novel information retrieval to its known web documents. An important practical use of this "path type" interface is as a secondary, orthogonal interface to other biological databases in order to provide different views of their underlying data. For example, MolMovDB (http://www.molmovdb.org) provides movie clips of likely 3D motions of proteins, and one can access it by PDB (http://www.pdb.org) identifiers. However, an alternative useful interface actually provided by LinkHub is a "family view" where one queries with a PDB identifier and can view all available motion pages for proteins in the same family as the query PDB identifier. We also provide a similar "family view" interface to structural genomics data in http://spine.nesg.org [8]. The system is flexible and other views based on, for example, similar function or other properties are easily imagined.

3.2 RDF Queries

We load the RDF-formatted LinkHub dataset into our YeastHub system [9] which uses Sesame (http:// www.openrdf.org) as the native RDF repository. We give two demonstration queries (described intuitively below) written in SeRQL (Sesame implementation of RQL) and executed in YeastHub to show one can effectively do the kinds of interesting preliminary scientific investigation and exploratory analysis commonly done at the beginning of research initiatives (e.g. to see whether they are worth pursuing further). The RDF LinkHub integrated into YeastHub is necessary to support the queries and is used as 'glue' to provide connections (both direct and indirect) between different biological identifiers. It is noteworthy that these queries can be formulated and run in a few hours at most and they roughly duplicate some results from published papers which required extensive effort to combine the necessary data to achieve their results.

Query 1: Finding Worm 'Interologs' of Yeast Protein Interactions

Proteins rarely act in isolation and often interact with one another and other molecules to perform necessary cellular actions. Experimental determinations of protein interactions are expensive and computational methods can leverage them for further interaction predictions. With this query we consider protein interactions in yeast (S. cervisiae) and see how many and which of them are possibly present in their homologs in worm (C. elegans), i.e. as interologs [10] in worm. We start with a dataset containing known and predicted yeast protein interactions (specified between yeast gene names) which is already loaded into YeastHub [11]. For each yeast gene name in this interaction set, we determine its evolutionarily related homologs (via PFAM in LinkHub) in worm by traversing paths in the LinkHub relationship graph of type: yeast gene name → UniProt accession → PFAM accession → UniProt accession → WormBase ID. Finally, we print WormBase (http://www.wormbase.org) ID pairs reachable in this way from an interacting pair of yeast gene names as possible protein interactions. The SeRQL statement together with a portion of its corresponding output can be seen in the paper's web supplement.

Query 2: Exploring Pseudogene Content versus Gene Essentiality in Yeast and Humans

Pseudogenes are genomic DNA sequences similar to normal genes (and usually derived from them) but are not expressed into functional proteins; they are regarded as defunct relatives of functional genes [12,13]. In the queries here we explore the relationship between gene essentiality (a measure of how important a gene is to survival of an organism) and the number of pseudogenes in an organism. We might hypothesize that more essential genes might have larger numbers of pseudogenes, and we explore this idea with queries of the joined YeastHub and LinkHub data. First, YeastHub has the MIPS (http://mips.gsf.de/genre/proj/yeast/) Essential Genes dataset, and we use this as our data on gene essentiality; LinkHub also contains a small dataset of yeast pseudogenes [14].

For each MIPS essential yeast gene name, we determine and count its pseudogenes by traversing paths in the relationship graph of type: yeast gene name → UniProt accession → yeast pseudogene. From the results we can then see if there is a relationship between essentiality and number of pseudogenes. Humans have a large number of known pseudogenes [15] but gene essentiality is difficult to characterize in humans (with many tissue types and developmental states complicating the issue).

Since essentiality is well studied in yeast, one thing we can do is determine the evolutionarily related homologs of yeast essential genes in human, which would perhaps likely be "more important" in a survival sense, and examine them for patterns associated with essentiality. For each MIPS essential yeast gene name, we can find the homologous pseudogenes in human by traversing paths in the LinkHub relationship graph of type: yeast gene name → UniProt accession → PFAM accession → human UniProt Id → UniProt accession → Pseudogene LSID. These queries and results can be seen in the paper supplement, but they show that few yeast essential genes are associated with pseudogenes whereas this is not the case with human. This may reflect the difference in processes of creation of the predominate numbers of yeast and human pseudogenes (duplication vs retrotransposition, see [12, 13]).

4 Related Work

The basic conceptual underpinnings of LinkHub, i.e., the importance of biological identifiers and linking them, was given by Karp [16]. LinkHub uses a semantic web approach to build a practical system based on and extending Karp's ideas on database links. The semantic web approach can also be used to implement database integration solutions based on the general approaches of *data warehousing* [17, 18] and *federation* [19, 20, 21]. Essentially, data warehousing focuses on data translation, i.e. translating and combining multiple datasets into a single database, whereas federation focuses on query translation, i.e. translating and distributing the parts of a query across multiple distinct databases and collating their results into one. A methodological overview and comparison of these database integration approaches was discussed in the biomedical context [22]. LinkHub's architecture is a hybrid of these two approaches: individual LinkHub instantiations are a kind of mini, local data warehouse of commonly grouped data and these are connected to large major hubs such as UniProt in a federated fashion; efficiency is gained by obviating the need for all source datasets to be individually connected to the major hubs.

LinkHub differentiates itself by not integrating all aspects of biological data but rather focusing on an important and more manageable high-level structuring principal, namely biological identifiers and the relationships (and relationship types) among them; hyperlinks to identifier-specific pages give access to additional attributes and data in the "Links" section of the LinkHub web interface. In fact, our YeastHub system addressed integration more generally by transforming many datasets to common RDF format and storing and giving RDF query access to them in an RDF database. The problem with YeastHub was that the integration was thin, with rich connections among the integrated datasets being limited. LinkHub is thus useful and complementary to YeastHub in this respect as a "connecting glue" among the datasets in that it makes and stores these cross-references and enables better integrated access to the YeastHub data; the example queries above demonstrated this

LinkHub's primary web interface can be viewed as a kind of "semantic web browser". Other work has also attempted to build browsers for semantic web data, including HayStack [23] and Sealife [24]. LinkHub is a more lightweight browser than HayStack (with a focus on biological relationship browsing) and differs from Sealife by being data-centric (establishing semantic links between data identifiers while treating web documents as metadata associated with the identifiers) as opposed to document-centric (establishing semantic links between terms/phrases appeared in different web documents). Finally, there have been a number of graph database systems and query languages developed through the years but they suffer from being proprietary; none have developed into widely used standard systems. However, it should be pointed out that some of these systems support advanced graph datamining and analysis operations not supported by RDF query languages and these features might be necessary for effective analysis of biological data represented in RDF [25].

5 Conclusion and Future Directions

Our paper demonstrates the natural use of semantic web RDF to inter-connect identifiers of data entries residing in separate web-accessible biological databases. Based on such a semantic RDF graph of biological identifiers and their relationships, powerful cross-database queries, inferences, and semantic data navigation can be performed. In addition, these semantic relationships enable flexible and novel information retrieval access to web documents attached to identifier nodes. LinkHub can be evaluated by considering its current active and practical use in a number of settings. We have already established the "hub of hubs" relationship between UniProt and LinkHub (i.e. UniProt cross-references to our LinkHub). In addition, LinkHub cross-references the targets of the structural genomics initiative (http://targetdb.pdb.org/) to UniProt and serves as a "related links" and "family viewer" gateway for the Northeast Structural Genomics Consortium (http://www.nesg.org) with which we are affiliated; LinkHub also serves as the "family viewer" for MolMobDB. LinkHub is a step towards answering the question "a life science semantic web: are we there yet?" [26]

Currently, LinkHub has limited web document hyperlinks attached to its nodes, and if this could be increased the utility of the novel information retrieval based on the "path type" interface would be enhanced. We are working to leverage the rich information in the LinkHub relational graph for enhanced automated information retrieval to web or scientific literature (MedLine) documents relevant to identifier nodes, e.g. proteomics identifiers, in the graph. A simple search for the identifier itself would likely not give optimal results due to conflated senses of the identifier text and identifier synonyms. In general, we need to consider and query for the key related concepts of an identifier, and these are present in the LinkHub subgraph surrounding the identifier. We consider the web pages attached to the identifiers in the subgraph as a "gold standard" for what additional relevant documents should be like, and we plan to use them as training sets to construct classifiers used to score and rank additional documents for relevance .

We also hope to explore how other relevant semantic web-related technologies could be effectively used in LinkHub, in particular named graph [27] and Life Science IDentifier or LSID (http://lsid.sourceforge.net/). Named graph allows RDF graphs to be named by URI, allowing them to be described by RDF statements; named graph could be used to provide additional information (metadata) about identifier mappings, such as source, version, and quality information. LSID is a standard object naming and distributed lookup mechanism being promoted for use on the semantic web, with emphasis on life sciences applications. An LSID names and refers to one unchanging data object, and allows versioning to handle updates. The

LSID lookup system is in essence like what Domain Name Service (DNS) does for converting named internet locations to IP numbers. We could possibly use LSID for naming objects in LinkHub and incorporate LSID lookup functionality. Finally, like software such as Napster and Gnutella did for online file sharing, we plan to explore enhancing LinkHub to enable multiple distributed LinkHub instantiations to interact in peer-to-peer networks for dynamic biological data sharing, possibly using web services technologies such as Web Services Description Language or WSDL (http://www.w3.org/TR/wsdl) and Universal Description, Discovery and Integration (http://www.uddi.org/) for dynamic service discovery, and available peer-to-peer toolkits.

References

1. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al.: **UniProt: the Universal Protein knowledgebase**. *Nucl. Acids Res.* 2004, **32**(90001): D115-119.
2. Kumar A, Cheung KH, Tosches N, Masiar P, Liu Y, Miller P and Snyder M: **The TRIPLES database: a community resource for yeast molecular biology**. *Nucl. Acids. Res.* 2002, **30**(1): 73-75
3. Berners-Lee, T, Hendler J and Lassila O: **The semantic web**. *Scientific American* 2001, **284**(5): 34-43.
4. Antoniou G, Van Harmelen F: *A Semantic Web Primer*. MIT Press; 2004.
5. Ashburner M., Ball C, Blake J, Botstein D, Butler H, Cherry M, Davis A, Dolinski K, Dwight S, Eppig J, et al.: **Gene ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**: 25-29
6. Soldatova LN and King RD **Are the current ontologies in biology good ontologies**. *Nat Biotechnol* 2005, **23**(9): 1095-8.
7. Wang X, Gorlitsky R and Almeida JS: **From XML to RDF: how semantic web technologies will change the design of 'omic' standards**. *Nat Biotechnol* 2005, **23**(9): 1099-103.
8. Goh CS, Lan N, Echols N, Douglas SM, Milburn D, Bertone P, Xiao R, Ma LC, Zheng D, Wunderlich Z, Acton T, Montelione GT, Gerstein M: **SPINE 2: a system for collaborative structural proteomics within a federated database framework.** *Nucleic Acids Res.* 2003, Jun 1;31(11):2833-8.
9. Cheung KH, Yip KY, Smith A, deKnikker R, Masiar A and Gerstein M: **YeastHub: a semantic web use case for integrating data in the life sciences domain.** *Bioinformatics* 2005, **21**(suppl_1): i85-96.
10. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs.** *Genome Res.* 2004, Jun;14(6):1107-18.
11. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, Oct 17;302(5644):449-53.
12. Zhang Z, Gerstein M: **Large-scale analysis of pseudogenes in the human genome**. *Curr Opin Genet Dev.* 2004, Aug;14(4):328-35. Review.
13. Harrison PM, Gerstein M: **Studying genomes through the aeons: protein families, pseudogenes and proteome evolution**. *J Mol Biol.* 2002, May 17;318(5):1155-74. Review.
14. Harrison P, Kumar A, Lan N, Echols N, Snyder M, Gerstein M: **A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution.** *J Mol Biol.* 2002, Feb 22;316(3):409-19.

15. Zhang Z, Harrison PM, Liu Y, Gerstein M: **Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome.** *Genome Res.* 2003, Dec;13(12):2541-58.
16. Karp PD: **Database links are a foundation for interoperability**. *Trends Biotechnol.* 1996, Aug;14(8):273-9. Review.
17. Agrawal D, El Abbadi A, Singh AK, Yurek T: **Efficient View Maintenance at Data Warehouses.** In *SIGMOD Conference*: 417-427 1997.
18. Zdobnov EM, Lopez R, Apweiler R, Etzold T: **The EBI SRS server--recent developments.** *Bioinformatics* 2002, Feb;18(2):368-73. PMID: 11847095
19. Sheth AP, Larson JA: **Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases.** *ACM Comput. Surv.* 1990, 22(3): 183-236.
20. Haas LM, Schwarz PM, Kodali P, Kotlar E, Rice JE, Swope WC: **DiscoveryLink: A system for integrated access to life sciences data sources.** *IBM Systems Journal* 2001, Volume 40, Number 2.
21. Kolatkar PR, Sakharkar MK, Tse CR, Kiong BK, Wong L, Tan TW, Subbiah S: **Development of software tools at BioInformatics Centre (BIC) at the National University of Singapore (NUS).** *Pac Symp Biocomput.* 1998,;:735-46. PMID: 9697226
22. Sujansky W: **Heterogeneous database integration in biomedicine**. *J Biomed Inform*. 2001, Aug;34(4):285-98.
23. Quan D, Huynh D, Karger DR:. **Haystack: A platform for authoring end user semantic web applications**. In Proc. 2nd International Semantic Web Conference, 2003.
24. Schroeder M, Burger A, Kostkova P, Stevens R, Habermann B, Dieng-Kuntz R:.**Sealife: a semantic grid browser for the life sciences applied to the study of infectious diseases**. Stud Health Technol Inform. 2006;120:167-78.
25. Angles R, Gutiérrez C: **Querying RDF Data from a Graph Database Perspective**. In ESWC: 346-360, 2005.
26. Neumann E: **A life science Semantic Web: are we there yet?** Sci STKE. 2005 May 10;2005(283):pe22.
27. Carroll J., Bizer C, Hayes P and Stickler P: **Named Graphs**. *Web Semantics* 2005, **3**(4): 247-67.

Figures

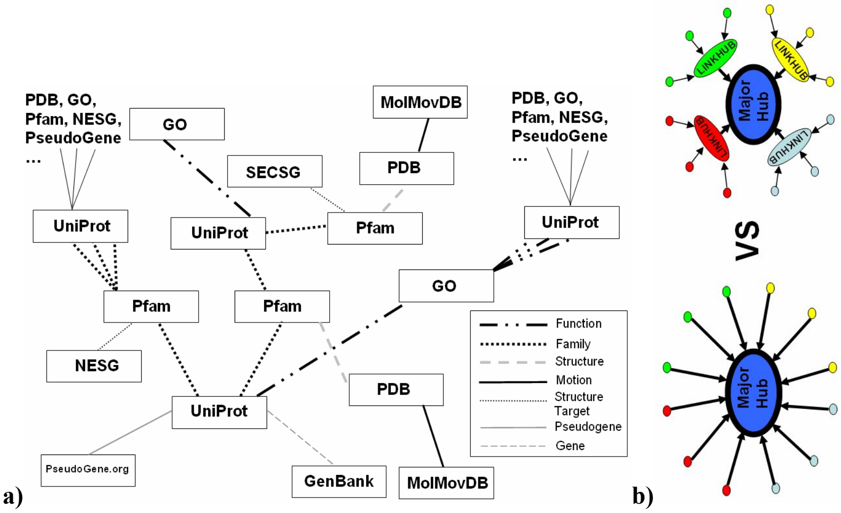
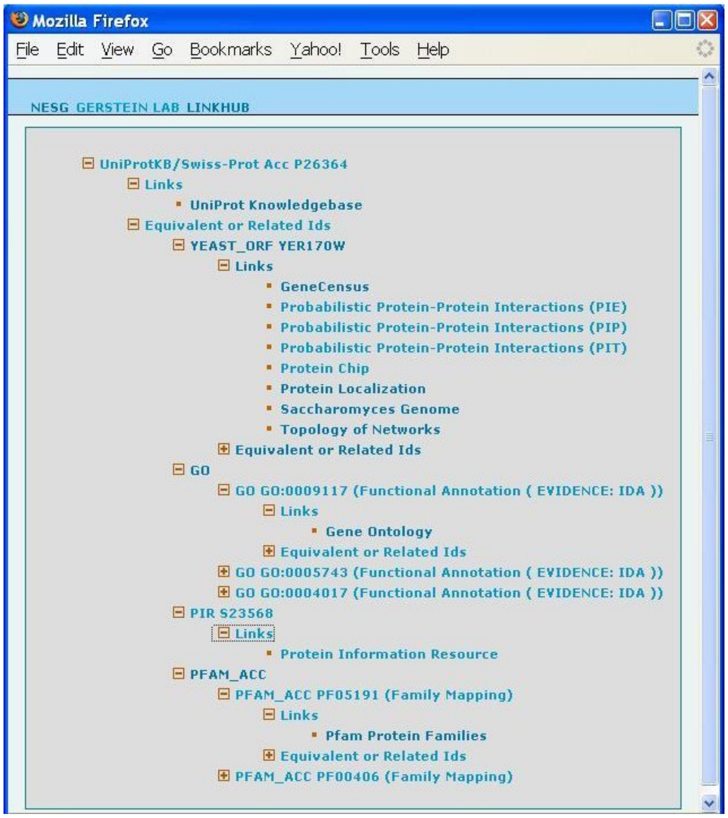


Fig. 1. Graph of Relationships and Hub of Hubs Organization. a) A conceptualization of the

semantic graph of interrelationships among biological identifiers, with boxes representing biological identifiers (originating database names given inside) and different edge types representing different kinds of relationships b) LinkHub as an enabler of an efficient "hub of hubs" organization of biological data. The different colors represent different labs, organizations, or logical groupings of data resources.


Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
NESG GERSTEIN LAB LINKHUB

```
⊟ UniProtKB/Swiss-Prot Acc P26364
    ⊟ Links
        ▪ UniProt Knowledgebase
    ⊟ Equivalent or Related Ids
        ⊟ YEAST_ORF YER170W
            ⊟ Links
                ▪ GeneCensus
                ▪ Probabilistic Protein-Protein Interactions (PIE)
                ▪ Probabilistic Protein-Protein Interactions (PIP)
                ▪ Probabilistic Protein-Protein Interactions (PIT)
                ▪ Protein Chip
                ▪ Protein Localization
                ▪ Saccharomyces Genome
                ▪ Topology of Networks
            ⊞ Equivalent or Related Ids
        ⊟ GO
            ⊟ GO GO:0009117 (Functional Annotation ( EVIDENCE: IDA ))
                ⊟ Links
                    ▪ Gene Ontology
                ⊞ Equivalent or Related Ids
            ⊞ GO GO:0005743 (Functional Annotation ( EVIDENCE: IDA ))
            ⊞ GO GO:0004017 (Functional Annotation ( EVIDENCE: IDA ))
        ⊟ PIR S23568
            ⊟ Links
                ▪ Protein Information Resource
        ⊟ PFAM_ACC
            ⊟ PFAM_ACC PF05191 (Family Mapping)
                ⊟ Links
                    ▪ Pfam Protein Families
                ⊞ Equivalent or Related Ids
            ⊞ PFAM_ACC PF00406 (Family Mapping)
```

Fig. 2. The basic DHTML list interface to LinkHub. Here, the data and relationships for UniProt identifier P26364 are presented. P26364 is presented at the root of the list, and lower levels contain information on additional related identifiers. Each identifier has two subsections: Links which gives a list of hyperlinks to web documents directly relevant to the identifier; and Equivalent or Related Ids which contains a list of additional identifiers related to the first identifier (the relationship type if it exists is given in parentheses; a synonym relationship is assumed if no relationship is given). The identifiers in the Equivalent and Related Ids section may themselves be further related to other identifiers which will have their own Links and Equivalent or Related Ids sections, ad nauseum. The initial display shows the transitive closure of the root identifier one level deep, and dynamic callbacks to the server retrieve additional data when the user clicks on identifiers whose subsections have not yet been loaded; in this way, the user can explore the relationship paths he desires without performance penalties (of loading the whole graph) or 'information overload'. The interface is dynamic, and a '+' list icon can be expanded to view the hidden underlying content, and a '-' list icon can be clicked to hide the content.