Published online 22 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W33–W39***
doi:10.1093/nar/gkp428

WebLab: a data-centric, knowledge-sharing bioinformatic platform

Xiaoqiao Liu, Jianmin Wu, Jun Wang, Xiaochuan Liu, Shuqi Zhao, Zhe Li, Lei Kong, Xiaocheng Gu, Jingchu Luo* and Ge Gao*

Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, China

Received January 30, 2009; Revised April 7, 2009; Accepted May 11, 2009

ABSTRACT

With the rapid progress of biological research, great demands are proposed for integrative knowledge-sharing systems to efficiently support collaboration of biological researchers from various fields. To fulfill such requirements, we have developed a data-centric knowledge-sharing platform WebLab for biologists to fetch, analyze, manipulate and share data under an intuitive web interface. Dedicated space is provided for users to store their input data and analysis results. Users can upload local data or fetch public data from remote databases, and then perform analysis using more than 260 integrated bioinformatic tools. These tools can be further organized as customized analysis workflows to accomplish complex tasks automatically. In addition to conventional biological data, WebLab also provides rich supports for scientific literatures, such as searching against full text of uploaded literatures and exporting citations into various well-known citation managers such as EndNote and BibTex. To facilitate team work among colleagues, WebLab provides a powerful and flexible sharing mechanism, which allows users to share input data, analysis results, scientific literatures and customized workflows to specified users or groups with sophisticated privilege settings. WebLab is publicly available at http://weblab.cbi.pku.edu.cn, with all source code released as Free Software.

INTRODUCTION

To explore mechanisms underlying complex biological processes, high-throughput analysis techniques and multi-disciplinary approaches are becoming main aspects of current biological research. Rapid growth of biological research places great demands on an integrative bioinformatic workbench to help biological researchers to mine knowledge from complex heterogeneous data.

Several bioinformatic analysis systems with intuitive user interface have been implemented in recent years (1–13). While some of them are designed as wrappers for a few specified software packages, a number of systems provide further support to popular bioinformatic analysis tools. Several systems including Taverna (3–5), BioManager (6), Galaxy (7,8), PISE (9), MOWServ (11) and HNB (13) support workflow-based analysis to make complex analysis much easier for non-experts. Moreover, Taverna (3–5), BioManager (6), Galaxy (7,8), PISE (9) and WildFire (10) also allow users to create workflows, increasing the flexibility and customizability.

On the other hand, while the importance of team work for research success is being widely recognized (3,14–16), few existing systems provide enough support for collaborative team work. Some systems allow users to store their input data and analysis results online (1,2,6–8, 11–13), and BioManager (6), Galaxy (7,8) also support users to share their stored data and workflows. Moreover, with the help of some 'Web 2.0' websites, researchers can upload and share their annotation information and workflow online (14,16–18). However, to our best knowledge, no bioinformatic analysis platform with comprehensive supports for data managing,

*To whom correspondence should be addressed. Tel: +86 10 6275 5206; Fax: +86 10 6275 1861; Email: luojc@pku.edu.cn
Correspondence may also be addressed to Ge Gao. Tel: +86 10 6275 1861; Fax: +86 10 6275 1861; Email: gaog@mail.cbi.pku.edu.cn
Present address:
Jianmin Wu, Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Haartmaninkatu 8, Helsinki 00014, Finland

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

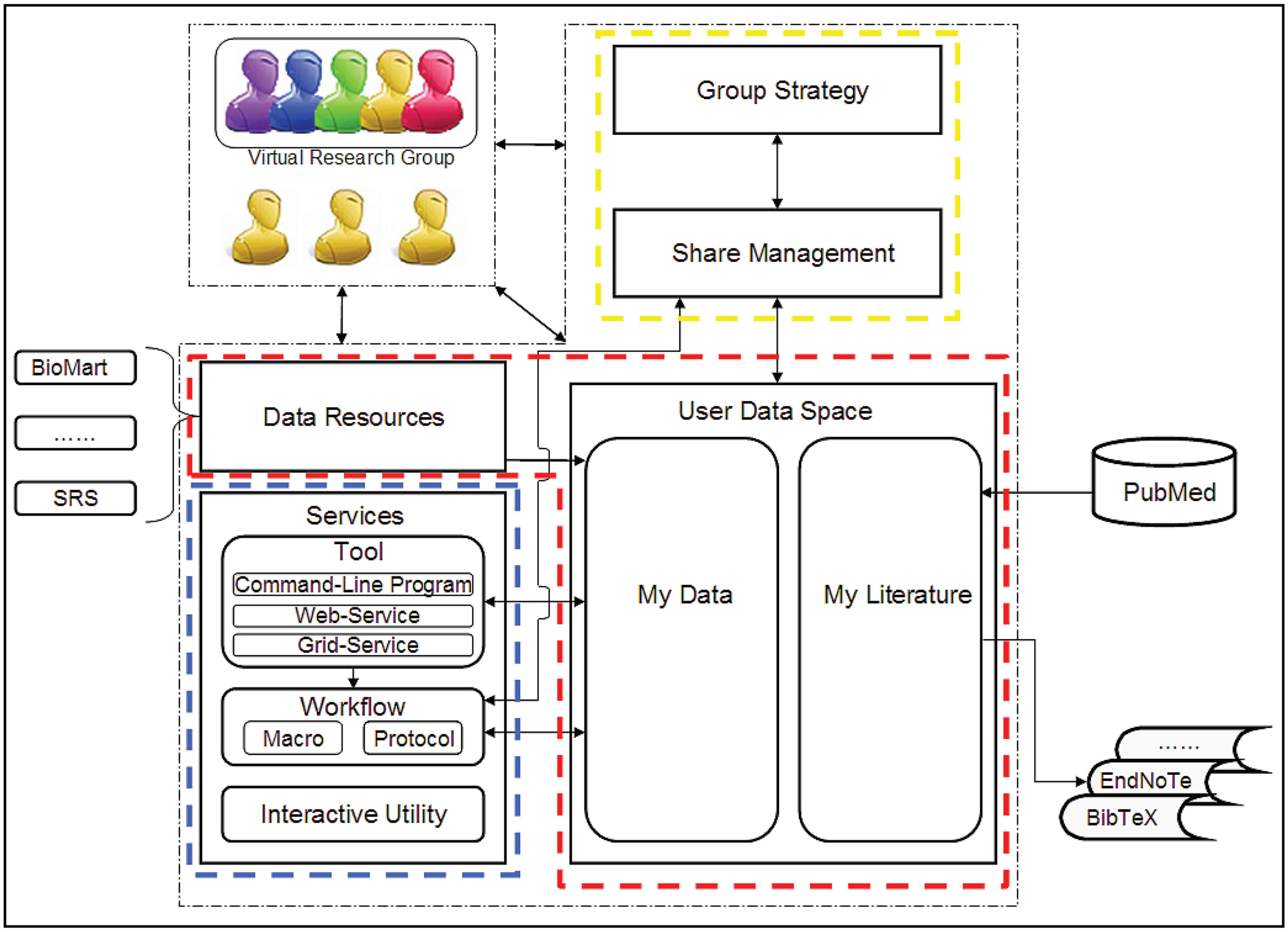


Figure 1. Overview of WebLab architecture. WebLab is comprised of three main functional modules. (i) Data management module (in red frame) maintains the user data space (My Data and My Literature), and also provides supports for accessing remote databases through BioMart and SRS; (ii) Analysis service module (in blue frame) provides a uniform framework to integrate more than 260 popular bioinformatic analysis tools (including command-line program, Web-service and Grid-service), and supports workflow in both Protocol and Macro models; (iii) Team work module (in yellow frame) keeps track of user shared data, literatures and workflows. Virtual research group (VRG) is designed to help collaborators share their works easily.

analyzing and sharing in a web-based integrative environment is publically available to the research community yet.

Here, we have developed a data-centric knowledge-sharing platform WebLab to support biological researchers to efficiently manage, analyze and share their data in an easy-to-use integrative environment. As a data-centric platform, WebLab provides dedicated user space to store and manage input data, analysis results and scientific literatures online. Supports for searching against full text, extracting citation information from PubMed, and exporting citations to EndNote and BibTeX are provided for literature, which is missing in other existing systems. By assembling customized workflows from 260+ integrated bioinformatic tools, complex analysis tasks could be performed automatically. In order to facilitate team work, WebLab provides powerful and flexible sharing mechanism and group strategy. Users can share their data, literatures and customized workflows with specific users or user-groups with sophisticated privilege settings. WebLab is publicly accessible at http://weblab.cbi.pku.edu.cn, with all source code available for downloading freely.

DESIGN AND SYSTEM ARCHITECTURE

To be flexible for further extension and development, WebLab is designed with a modularization approach including three main modules: data management, analysis service and team work (Figure 1).

Data management

As a data-centric platform, WebLab provides a powerful data management system for users to store and manage their data and scientific literatures online.

In their own data space (‘My Data’), users can create a new entry by uploading a file from local disk or retrieving from remote databases through BioMart (19) and

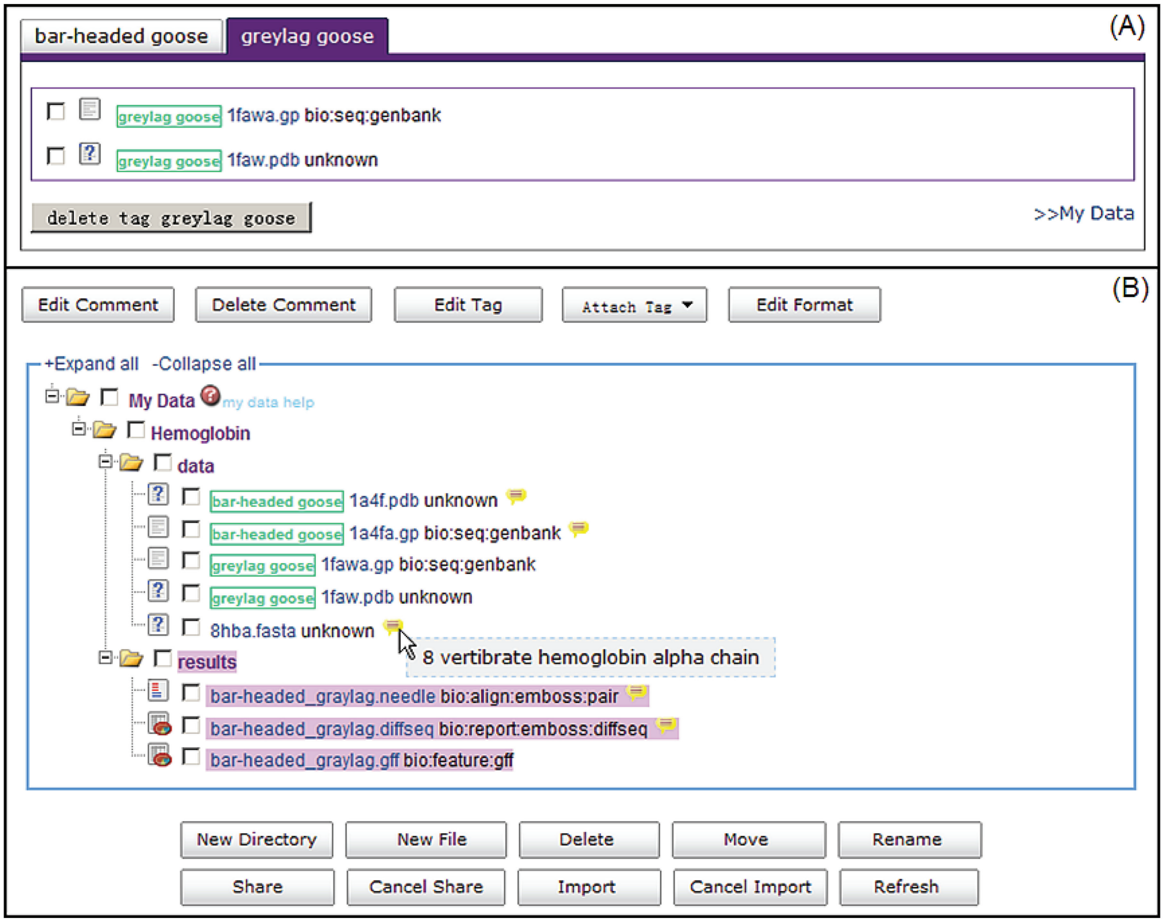

(A)
bar-headed goose
greylag goose
□ greylag goose 1fawa.gp bio:seq:genbank
□ greylag goose 1faw.pdb unknown
delete tag greylag goose
>>My Data
(B)
Edit Comment
Delete Comment
Edit Tag
Attach Tag ▾
Edit Format
+Expand all -Collapse all
□ My Data my data help
□ Hemoglobin
□ data
□ bar-headed goose 1a4f.pdb unknown
□ bar-headed goose 1a4fa.gp bio:seq:genbank
□ greylag goose 1fawa.gp bio:seq:genbank
□ greylag goose 1faw.pdb unknown
□ 8hba.fasta unknown
8 vertibrate hemoglobin alpha chain
□ results
□ bar-headed_graylag.needle bio:align:emboss:pair
□ bar-headed_graylag.diffseq bio:report:emboss:diffseq
□ bar-headed_graylag.gff bio:feature:gff
New Directory
New File
Delete
Move
Rename
Share
Cancel Share
Import
Cancel Import
Refresh

Figure 2. My Data—user-dedicated space to store and manage input data and analysis results. The entries in My Data are presented in the hierarchical tree structure as in local file system. (**A**) WebLab provides an option to add user-defined labels (Tags) to each item in My Data as an annotation. Entries can later be classified and organized according to associated Tags in 'tag view'. (**B**) For each entry, users can create comment, and the detailed content for a comment will be displayed when users put mouse over the yellow icon. The comment will also be shared out along with the data entries.

SRS (20). After data type for the newly created file is specified, WebLab can recognize the format and automatically detect available analysis tools in a context-sensitive approach.

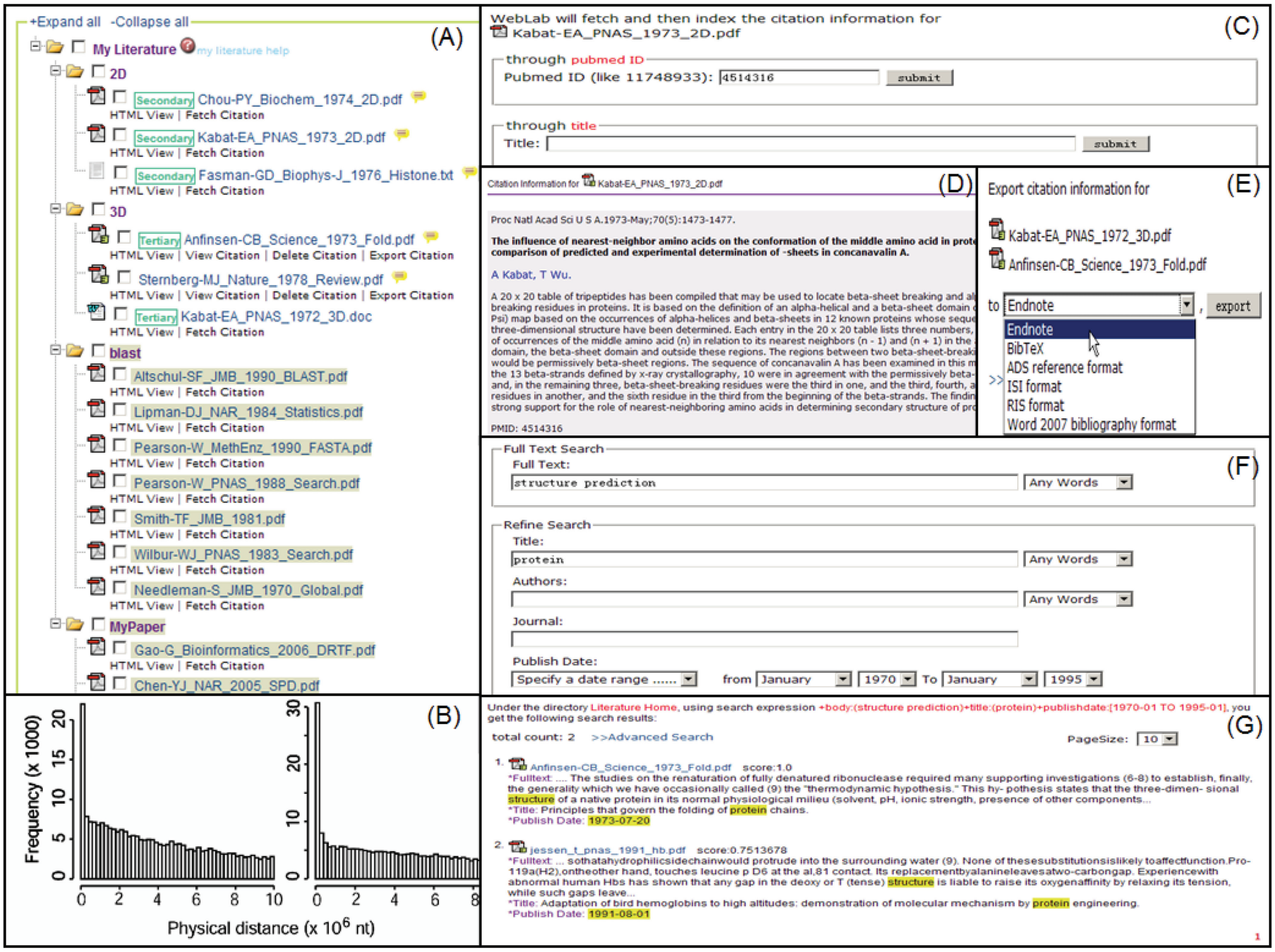
The entries in My Data are presented in hierarchical tree structure as in daily-used local file system. Users can create, rename and delete files and directories (folders) in My Data by simple mouse clicks. Moreover, users could also associate user-defined labels (Tags) or comments to entries in My Data, to classify and organize them in flexible and intuitive ways (Figure 2).

In addition to conventional operations supported in My Data, rich literature-specified functions are provided in 'My Literature'. After uploading literature, WebLab automatically generates HTML preview for a quick check of the paper's content in browser without downloading the whole article. Then, WebLab extracts and indexes full text contents for uploaded articles. When the indexing is done, users can do simple keyword search or complex query search against full text of literatures existing in My Literature. Moreover, citation information could be fetched from NCBI according to PubMed ID or title and to be further associated to respective article in My Literature. All citation information could be easily exported into various well-known citation managers such as EndNote and BibTex (Figure 3).

Analysis service

As an integrative bioinformatic analysis platform, WebLab integrates numerous analysis tools within a uniform framework. In addition to command-line programs, Web-services and Grid-services are also integrated in WebLab with full interoperation (Table S1).

By organizing different tools into a workflow (21), complex analysis tasks are performed in one run. In a workflow, several analysis tools are launched according to previously user-defined rules. Currently, two workflow models with different user interaction abilities are available. In the Protocol model, a workflow is executed stepwise, and the user can tune parameters or options in each step, thus providing maximum flexibility. On the other hand, in the Macro model, after mandatory parameters


+Expand all -Collapse all
(A)
My Literature
my literature help
2D
Secondary Chou-PY_Biochem_1974_2D.pdf
HTML View | Fetch Citation
Secondary Kabat-EA_PNAS_1973_2D.pdf
HTML View | Fetch Citation
Secondary Fasman-GD_Biophys-J_1976_Histone.txt
HTML View | Fetch Citation
3D
Tertiary Anfinsen-CB_Science_1973_Fold.pdf
HTML View | View Citation | Delete Citation | Export Citation
Sternberg-MJ_Nature_1978_Review.pdf
HTML View | View Citation | Delete Citation | Export Citation
Tertiary Kabat-EA_PNAS_1972_3D.doc
HTML View | Fetch Citation
blast
Altschul-SF_JMB_1990_BLAST.pdf
HTML View | Fetch Citation
Lipman-DJ_NAR_1984_Statistics.pdf
HTML View | Fetch Citation
Pearson-W_MethEnz_1990_FASTA.pdf
HTML View | Fetch Citation
Pearson-W_PNAS_1988_Search.pdf
HTML View | Fetch Citation
Smith-TF_JMB_1981.pdf
HTML View | Fetch Citation
Wilbur-WJ_PNAS_1983_Search.pdf
HTML View | Fetch Citation
Needleman-S_JMB_1970_Global.pdf
HTML View | Fetch Citation
MyPaper
Gao-G_Bioinformatics_2006_DRTF.pdf
HTML View | Fetch Citation
Chen-YJ_NAR_2005_SPD.pdf
(B)
Frequency (x 1000)
0
5
10
15
20
0
10
20
30
0
2
4
6
8
10
0
2
4
6
8
Physical distance (x 10^6 nt)
(C)
(D)
(E)
Export citation information for
Kabat-EA_PNAS_1972_3D.pdf
Anfinsen-CB_Science_1973_Fold.pdf
to
export
Endnote
BibTeX
ADS reference format
ISI format
RIS format
Word 2007 bibliography format
(F)
Full Text Search
Refine Search
Any Words
submit
(G)
>>Advanced Search
1

WebLab will fetch and then index the citation information for Kabat-EA_PNAS_1973_2D.pdf

through pubmed ID

Pubmed ID (like 11748933): 4514316

through title

Title:

Citation Information for Kabat-EA_PNAS_1973_2D.pdf

Proc Natl Acad Sci U S A.1973-May;70(5):1473-1477.

The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in prote comparison of predicted and experimental determination of -sheets in concanavalin A.

A Kabat, T Wu.

A 20 x 20 table of tripeptides has been compiled that may be used to locate beta-sheet breaking and al breaking residues in proteins. It is based on the definition of an alpha-helical and a beta-sheet domain Psi) map based on the occurrences of alpha-helices and beta-sheets in 12 known proteins whose seque three-dimensional structure have been determined. Each entry in the 20 x 20 table lists three numbers, of occurrences of the middle amino acid (n) in relation to its nearest neighbors (n - 1) and (n + 1) in the domain, the beta-sheet domain and outside these regions. The regions between two beta-sheet-breaki would be permissively beta-sheet regions. The sequence of concanavalin A has been examined in this m the 13 beta-strands defined by x-ray crystallography, 10 were in agreement with the permissively beta- and, in the remaining three, beta-sheet-breaking residues were the third in one, and the third, fourth, a residues in another, and the sixth residue in the third from the beginning of the beta-strands. The findin strong support for the role of nearest-neighboring amino acids in determining secondary structure of pro

PMID: 4514316

Export citation information for: to Endnote

Full Text: structure prediction — Any Words

Title: protein — Any Words

Authors: — Any Words

Journal:

Publish Date: Specify a date range ...... from January 1970 To January 1995

Under the directory Literature Home, using search expression +body:(structure prediction)+title:(protein)+publishdate:[1970-01 TO 1995-01], you get the following search results:

total count: 2 PageSize: 10

1. Anfinsen-CB_Science_1973_Fold.pdf score:1.0
*Fulltext: .... The studies on the renaturation of fully denatured ribonuclease required many supporting investigations (6-8) to establish, finally, the generality which we have occasionally called (9) the "thermodynamic hypothesis." This hy- pothesis states that the three-dimen- sional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components...
*Title: Principles that govern the folding of protein chains.
*Publish Date: 1973-07-20

2. jessen_t_pnas_1991_hb.pdf score:0.7513678
*Fulltext: ... sothatahydrophilicsidechainwould protrude into the surrounding water (9). None of thesesubstitutionsislikely toaffectfunction.Pro-119a(H2),ontheother hand, touches leucine p D6 at the al,81 contact. Its replacementbyalanineleavesatwo-carbongap. Experiencewith abnormal human Hbs has shown that any gap in the deoxy or T (tense) structure is liable to raise its oxygenaffinity by relaxing its tension, while such gaps leave...
*Title: Adaptation of bird hemoglobins to high altitudes: demonstration of molecular mechanism by protein engineering.
*Publish Date: 1991-08-01

Figure 3. WebLab provides rich supports for scientific literatures. (**A**) Literatures in PDF, Microsoft Word and plain text formats could be stored and managed in My Literature. In addition to conventional operations in My Data, rich literature-specified functions are provided in My Literature. (**B**) WebLab provides 'HTML View' for a quick check of the contents in browser without downloading. (**C**) (**D**) Users can fetch citation information from NCBI PubMed repository through PubMed ID or title. (**E**) Citation information can be batch exported in well-known bibliography formats such as Endnote, BibTeX, ADS reference, ISI, RIS and Word 2007 bibliography format. (**F**) (**G**) After the index is built, users can perform search against all literatures in My Literature. The matched literatures will be sorted by their relevant scores and the searched key words are highlighted.

are first inputted by the user, each tool in a workflow will be sequentially executed. Thus, Macro is more suitable for routine analysis. Moreover, an existing Macro could be re-used and treated as a standard analysis tool to define new workflow (recursive definition), which further simplifies users' daily work and increases flexibility. Besides defining their own workflows, users can also use several pre-defined workflows for common analysis tasks such as phylogenetic tree construction and protein function analysis (Figure 4).

A few popular client-side utilities including Sequence Manipulation Suite (SMS) (22), WebMol (23), Dotlet (24) and JalView (25) are also integrated into WebLab for users to perform interactive work such as editing multiple sequence alignment or visualizing structure. While those utilities could not be incorporated into the workflow like other standard analysis tools due to their interactive nature, they are proved to be useful in daily work. Moreover, users also can keep their favorite analysis tools in their 'My Toolbox' for quick access.

Team work

Collaborations among several researchers in various fields and different locations are recently becoming more and more common, and also crucial for research success. To facilitate collaborative team work, WebLab provides flexible sharing mechanism and group strategy for users to share their data and knowledge.

In WebLab, a user can share almost everything he owns with other users. For entries in My Data and My Literature, both 'read only' and 'read and write' sharing privileges are provided. By employing the reference-count based sharing model, changes in these shared contents

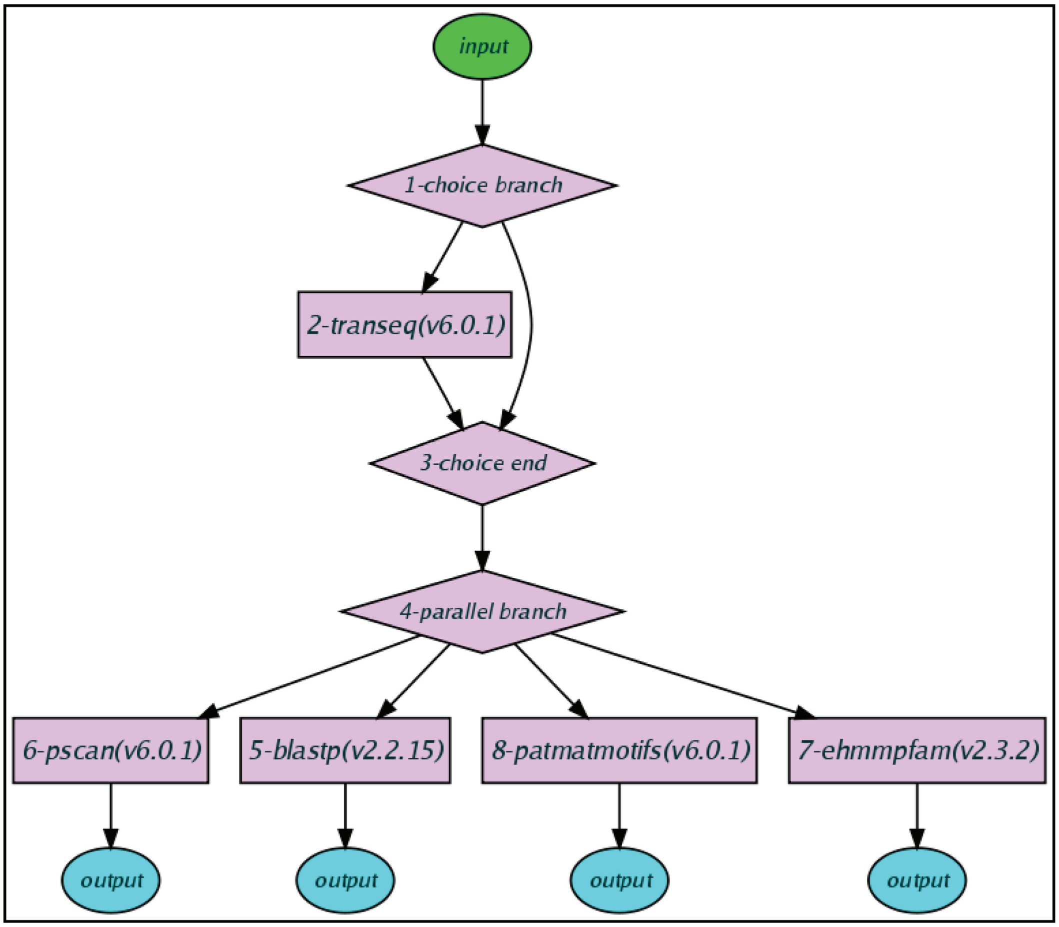


Figure 4. A workflow for protein function analysis. The workflow is represented as Directed Acyclic Graph (DAG) in WebLab. The rectangular node of the DAG means a single tool or a defined Macro; the diamond node stands for operators which control the execution of workflows. WebLab supports two types of operators: condition operator (for example, nodes 1 and 3 in figure) determines following analysis depending on a given condition (for example, the workflow in figure will try to translate the input sequences by *transeq* for DNA sequences); parallel operator (for example, node 4 in figure) is used to share incoming data among several parallel analysis tools or combine several incoming data into one sink.

will be seen by all collaborators simultaneously to assure efficient cooperation among all partners. On the other hand, once a user-defined workflow is shared out, a copy will be made which can be modified without altering the original one, to prevent possible flaw caused by recursive definition of workflow (Figure 5).

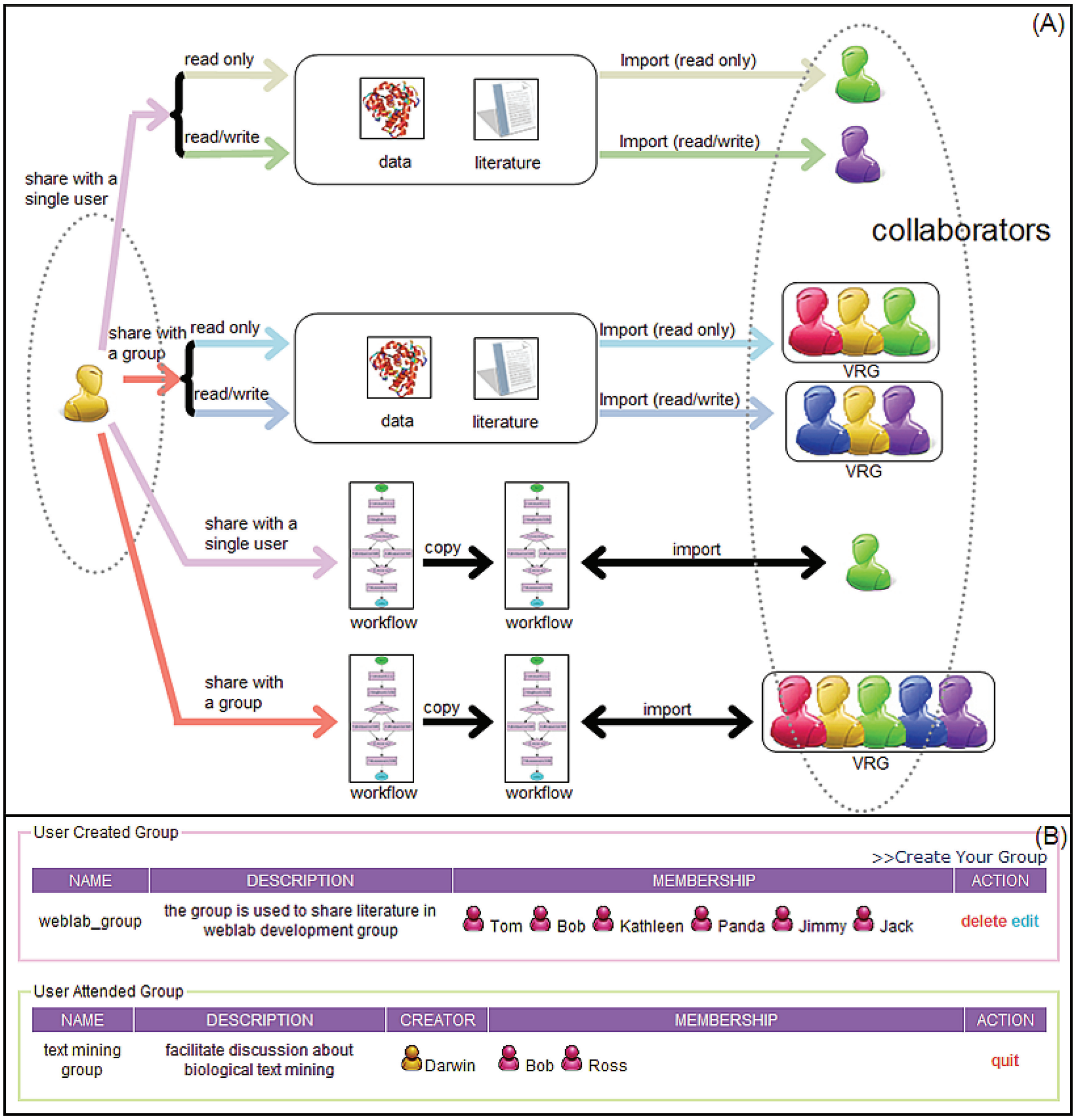
Groups are designed for colleagues who work closely together. Any user can set up a new virtual research group (VRG) and invite other users to join the new VRG. A member of a VRG can also share with other members in the VRG, by employing similar operations used for sharing with the normal user (Figure 5).

IMPLEMENTATION AND AVAILABILITY

Given the heavy computational load, WebLab is implemented as a loosely coupled distributed system. The portal server holds the web interface and acts as a proxy to users' requests. With dispatch daemon running, several backend computing servers execute the required operations following the request from WebLab portal server. The results will be sent back to the portal server after the analysis is finished and saved into database maintained by the portal server. Call-back mechanism is widely used in WebLab system to increase the flexibility. Adding a new tool does not require writing additional codes besides changing an XML format configuration file.

WebLab was developed using Java 1.5, providing it with the platform-independent advantage. WebLab uses Apache Tomcat as container for Java Servlet and JSP, MySQL as backend database system to store user data and other necessary information. WebLab also uses Graphviz (http://www.graphviz.org) to produce figures, and Lucene (http://lucene.apache.org) as information retrieval library to build index and search information.

WebLab is publicly accessible at http://weblab.cbi.pku.edu.cn and is compatible with the most common web browsers such as Mozilla Firefox (version 2 and 3) and Internet Explorer (version 6, 7 and 8). Online HTML and video tutorials are being actively maintained and updated. The source code of WebLab is released as



User Created Group

>>Create Your Group

NAME	DESCRIPTION	MEMBERSHIP	ACTION
weblab_group	the group is used to share literature in weblab development group	Tom Bob Kathleen Panda Jimmy Jack	delete edit

User Attended Group

NAME	DESCRIPTION	CREATOR	MEMBERSHIP	ACTION
text mining group	facilitate discussion about biological text mining	Darwin	Bob Ross	quit

Figure 5. Team work support in WebLab. (**A**) In WebLab, a user can share with his colleagues almost everything he owns. Analysis data and scientific literatures can be shared out with 'read only' or 'read and write' privileges. On the other hand, a different sharing strategy is applied for workflow to prevent possible flaw caused by recursive definition of workflow, i.e. once a user-defined workflow is shared out, the collaborators will get a copy of this workflow, which can be edited without altering the content of the original one. (**B**) Through group strategy provided by WebLab, users can create their own virtual research groups (VRG) in web environment. The group creator is privileged to edit or to delete the whole group. And a group member can also choose to quit from the group freely.

'Free Software' under the GNU General Public License version 3 (GPLv3), and freely available for downloading.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

China National High-tech 863 Programs [2006AA02Z334, 2006AA02Z314, 2006AA02A312]; China high-tech platform program. Funding for open access charge: China National High-tech (863) Program (2006AA02Z334).

Conflict of interest statement. None declared.

REFERENCES

1. Carver,T. and Bleasby,A. (2003) The design of Jemboss: a graphical user interface to EMBOSS. *Bioinformatics*, **19**, 1837–1843.
2. Sarachu,M. and Colet,M. (2005) wEMBOSS: a web interface for EMBOSS. *Bioinformatics*, **21**, 540–541.
3. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*., **34**, W729–W732.
4. Lanzen,A. and Oinn,T. (2008) The Taverna Interaction Service: enabling manual interaction in workflows. *Bioinformatics*, **24**, 1118–1120.

5. Oinn,T., Addis,M., Ferris,J., Marvin,D., Senger,M., Greenwood,M., Carver,T., Glover,K., Pocock,M.R., Wipat,A. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.
6. Cattley,S. and Arthur,J.W. (2007) BioManager: the use of a bioinformatics web application as a teaching tool in undergraduate bioinformatics training. *Brief Bioinform.*, **8**, 457–465.
7. Blankenberg,D., Taylor,J., Schenck,I., He,J., Zhang,Y., Ghent,M., Veeraraghavan,N., Albert,I., Miller,W., Makova,K.D. *et al.* (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res.*, **17**, 960–964.
8. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
9. Garcia Castro,A., Thoraval,S., Garcia,L.J. and Ragan,M.A. (2005) Workflows in bioinformatics: meta-analysis and prototype implementation of a workflow generator. *BMC Bioinformatics*, **6**, 87.
10. Tang,F., Chua,C.L., Ho,L.Y., Lim,Y.P., Issac,P. and Krishnan,A. (2005) Wildfire: distributed, Grid-enabled workflow construction and execution. *BMC Bioinformatics*, **6**, 69.
11. Navas-Delgado,I., Rojano-Munoz Mdel,M., Ramirez,S., Perez,A.J., Andres Leon,E., Aldana-Montes,J.F. and Trelles,O. (2006) Intelligent client for integrating bioinformatics services. *Bioinformatics*, **22**, 106–111.
12. Shah,S.P., He,D.Y., Sawkins,J.N., Druce,J.C., Quon,G., Lett,D., Zheng,G.X., Xu,T. and Ouellette,B.F. (2004) Pegasys: software for executing and integrating analyses of biological sequences. *BMC Bioinformatics*, **5**, 40.
13. Crass,T., Antes,I., Basekow,R., Bork,P., Buning,C., Christensen,M., Claussen,H., Ebeling,C., Ernst,P., Gailus-Durner,V. *et al.* (2004) The Helmholtz Network for Bioinformatics: an integrative web portal for bioinformatics resources. *Bioinformatics*, **20**, 268–270.
14. Waldrop,M. (2008) Big data: Wikiomics. *Nature*, **455**, 22–25.
15. Gilbert,W. (1991) Towards a paradigm shift in biology. *Nature*, **349**, 99.
16. Roure,D.D., Goble,C. and Stevens,R. (2007) Designing the my Experiment virtual research environment for the social sharing of workflows. *Third IEEE International Conference on e-Science and Grid Computing (e-Science 2007)*. Bangalore, India, pp. 603–610.
17. Huss,J.W. 3rd, Orozco,C., Goodale,J., Wu,C., Batalov,S., Vickers,T.J., Valafar,F. and Su,A.I. (2008) A gene wiki for community annotation of gene function. *PLoS Biol.*, **6**, e175.
18. Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
19. Durinck,S.M.Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **15**, 3439–3440.
20. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server—recent developments. *Bioinformatics*, **18**, 368–373.
21. Tiwari,A. and Sekhar,A.K. (2007) Workflow based framework for life science informatics. *Comput. Biol. Chem.*, **31**, 305–319.
22. Stothard,P. (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*, **28**, 1102–1104.
23. Walther,D. (1997) WebMol—a Java-based PDB viewer. *Trends Biochem. Sci.*, **22**, 274–275.
24. Junier,T. and Pagni,M. (2000) Dotlet: diagonal plots in a web browser. *Bioinformatics*, **16**, 178–179.
25. Clamp,M., Cuff,J., Searle,S.M. and Barton,G.J. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.