Tan et al. BMC Genomics 2013, 14(Suppl 5):S13
http://www.biomedcentral.com/1471-2164/14/S5/S13

BMC Genomics

RESEARCH Open Access

Simple re-instantiation of small databases using cloud computing

Tin Wee Tan[1*], Chao Xie[2], Mark De Silva[2], Kuan Siong Lim[2], C Pawan K Patro[1], Shen Jean Lim[1], Kunde Ramamoorthy Govindarajan[1], Joo Chuan Tong[1,3], Khar Heng Choo[1], Shoba Ranganathan[4], Asif M Khan[5,6*]

From Asia Pacific Bioinformatics Network (APBioNet) Twelfth International Conference on Bioinformatics (InCoB2013)
Taicang, China. 20-22 September 2013

Abstract

Background: Small bioinformatics databases, unlike institutionally funded large databases, are vulnerable to discontinuation and many reported in publications are no longer accessible. This leads to irreproducible scientific work and redundant effort, impeding the pace of scientific progress.

Results: We describe a Web-accessible system, available online at http://biodb100.apbionet.org, for archival and future on demand re-instantiation of small databases within minutes. Depositors can rebuild their databases by downloading a Linux live operating system (http://www.bioslax.com), preinstalled with bioinformatics and UNIX tools. The database and its dependencies can be compressed into an ".lzm" file for deposition. End-users can search for archived databases and activate them on dynamically re-instantiated BioSlax instances, run as virtual machines over the two popular full virtualization standard cloud-computing platforms, Xen Hypervisor or vSphere. The system is adaptable to increasing demand for disk storage or computational load and allows database developers to use the re-instantiated databases for integration and development of new databases.

Conclusions: Herein, we demonstrate that a relatively inexpensive solution can be implemented for archival of bioinformatics databases and their rapid re-instantiation should the live databases disappear.

Database archival Re-instantiation, Cloud computing, BioSLAX, biodb100, MIABi

Background

Other than the big well-funded institutionalized databases, few bioinformatics databases have longevity beyond several years. Small databases are particularly vulnerable. Valuable data and metadata become irretrievably lost leading to irreproducible scientific work and redundant effort [1-3]. This is unnecessary in view of the vast amount of affordable disk space and highly accessible cloud computing power in the market.

Recently, in 2009, the Asia Pacific Bioinformatics Network (APBioNet), Asia's oldest bioinformatics network and pioneer of the annual International Conference on Bioinformatics (InCoB) now in its twelfth year, initiated the effort for Minimum Information about a Bioinformatics Investigation (MIABi), building on earlier ideas [4]. The standards for transparency, provenance and scientific reproducibility amongst the bioinformatics and computational biology community were drafted and published a year later [5]. The MIABi standards were harmonised with the BioDBcore standards of the International Society for Biocuration (ISB) and BioSharing for use of standardized terminologies for infrastructural and informational interoperability [6].

In accordance to the MIABi standards, "when databases are described in a publication, a copy should be frozen, version-labelled and dated for reference", herein, we detail a simple methodology for archival and re-instantiation of small databases. Further, we provide a

* Correspondence: tinwee@bic.nus.edu.sg; asif@perdanauniversity.edu.my
[1]Department of Biochemistry, National University of Singapore (NUS), Singapore
[5]Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, USA
Full list of author information is available at the end of the article

Web interface to demonstrate this functionality, with several exemplar databases to illustrate the utility of the system.

Methods

Briefly, the re-instantiation system consists of a Web portal and a Cloud-based backend. The Web interface allows download of the base live operating system for database developers to build a compressed version of their database, upload their database, boot up a cloud instance, activate database and access the various functionalities (Figure 1). The content of the uploaded databases are vetted for extraneous files or programs that might be malicious before they are allowed for instantiation. Below we describe the different components of the re-instantiation system:

(i) BioSlax. The BioSlax 7.5 (http://www.bioslax.com) live operating system has been developed on a Slackware Linux base distribution called Slax (http://www.slax.org), to which we have added additional modules, including more than 200 bioinformatics software applications. BioSlax is freely downloadable and depositors of databases can choose to use it to develop their databases, or develop on another platform that can be ported to BioSlax. Upon boot-up, BioSlax un-packs, loads and activates all modules, including MySQL, Apache and Webmin servers. Other services can also be loaded and activated by adding additional BioSlax modules. Further, other operating systems will be explored in the future.

(ii) Building databases as BioSlax LZM modules. The BioSlax command "dir2lzm" is used to convert database specific files, other dependencies and other application services into LZM compressed file as BioSlax modules. Alternatively, database depositors can port pre-existing databases built on other platforms onto BioSlax, and save the difference as an LZM Slax module file that includes everything changed from the base LiveOS. The Slax "activate" command is used to unpack and uncompress all files and folders from the LZM file, and restart all relevant processes.

(iii) Web interface. A Web interface has been set up as the portal for users to access the re-instantiation system

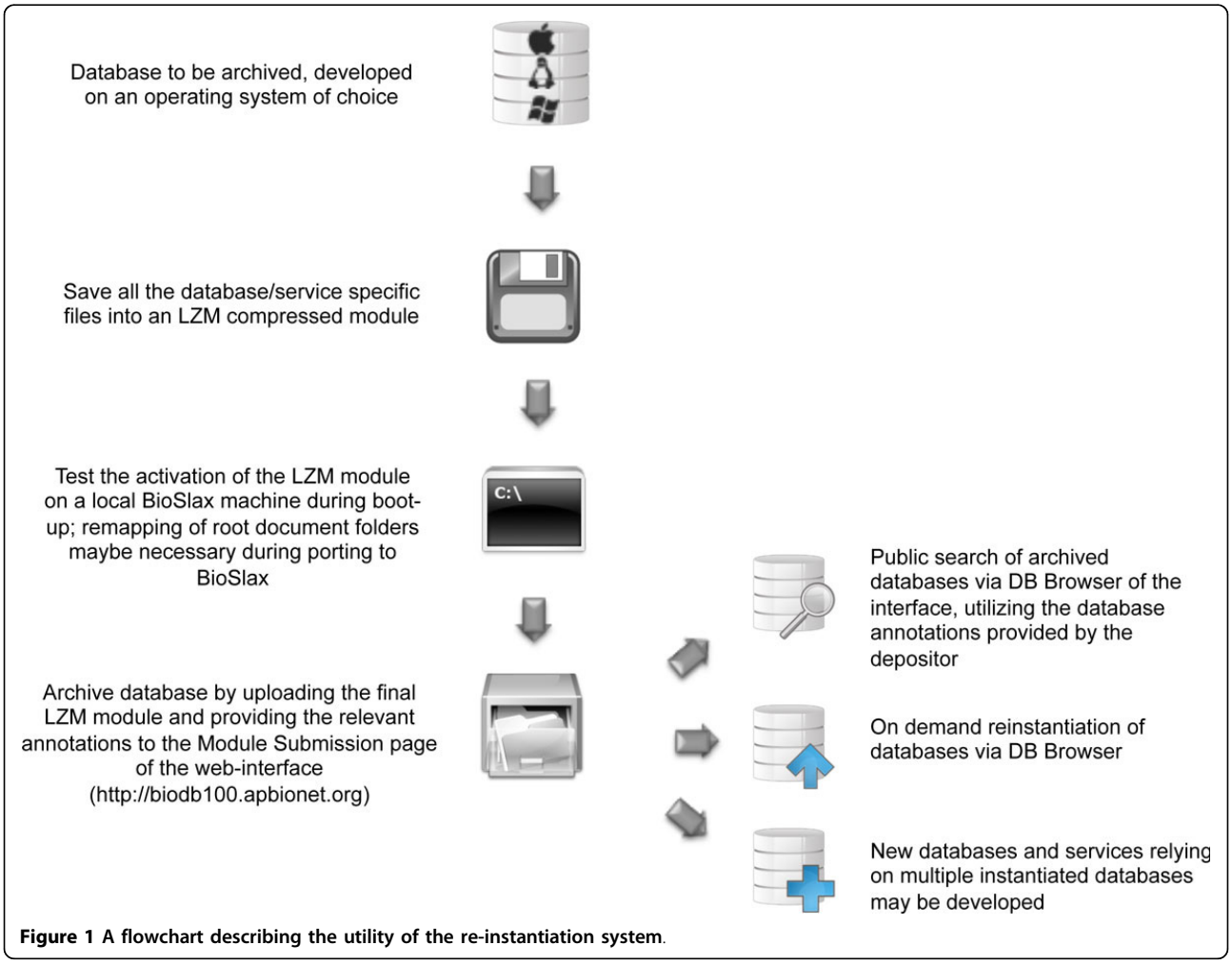

Figure 1 A flowchart describing the utility of the re-instantiation system.

(http://biodb100.apbionet.org). Remote calls to the cloud server are made by public-private encryption key ssh secure login and remote execution of commands, called from CGI scripts written in bash and perl, with SSH, HTTP perl modules.

(iv) A database of meta-information on uploaded databases and the compressed LZM files. During the upload of the LZMs, meta-information about the submitters are obtained, such as their research group, author identifier (via http://aid.apbionet.org), and document/publication submission (via http://docid.apbionet.org). This meta-information is searchable and end-users can re-instantiate databases identified from the search, where the relevant LZM will be copied into a booted instance of a BioSlax OS and activated dynamically via the Web CGI call to the cloud server.

(v) Xen Hypervisor or vSphere. The cloud computing platform used is the popular open source full hardware virtualization software Citrix XenServer, based on Xen Hypervisor or VMware, based on vSphere, based on vSphere that provide the ability to create, deploy and manage the virtual machines on the cloud. For demonstration, we have set-up five instances of BioSlax virtual machines for remote booting on a private cloud that consists of 3 SuperMicro servers, each with dual Intel Xeon X5690 3.47 GHz 6-core processors (24 virtual processors through Intel Hyper-Threading technology), 145 GB of RAM and 1.8 TB of local storage. Xen's built-in "xe" commands are utilised in the Xenserver to poll halted BioSlax instances, activate on demand or shut down when idle. A Web interface for administrators to remotely start, stop or control virtual machine instances is provided (http://vmc.apbionet.org). We currently assign a public Internet Protocol (IP) address to each instance to enable external access to the databases.

Results

We have successfully used our re-instantiable archival system for half a dozen extinct and extant databases (Table 1). Some databases were previously developed on non-BioSlax platforms, but were ported to BioSlax without difficulty. Hardcoded hyperlinks and directory paths were made relative in order for the unpackaged files to work properly. MySQL databases were compatible while other SQL databases required a SQL dump and recreation in MySQL format. Notably, none of the special SQL function calls not supported by MySQL were detected; otherwise, appropriate SQL rdbms would have to be installed into BioSlax as a separate LZM module when needed. Many databases are accompanied by search functions such as BLAST or other computational features, which were supported by BioSlax for the databases that we tested. Where this is not the case, the relevant programs can be compiled and added to the database LZM modules and activated accordingly. The database re-instantiation time was rapid, within 2 to 4 minutes. All database functionalities tested worked as per the original.

Table 1 Exemplar archived databases.

Exemplar Archived Databases	
Type	**Description**
Extant (available at published URL)	Allergen Atlas [8]
	Customary Medicinal Knowledgebase (CMKb) [9]
	MHC-Peptide Interaction Database-TR version 2 (MPID-T2) [10]
	Signal Peptide Database (SPDB) [11]
	STATdb: A specialised resource for STAT proteins (http://statdb.bic.nus.edu.sg/)
	Sub-Domain (http://chaos.bic.nus.edu.sg/domain/)
	Type III Secretion System Effector Database [12]
	NFκB Base (http://bioslax01.bic.nus.edu.sg/nfkb/)
Extinct (not available at published URL, but copy maintained elsewhere)	MHC-Peptide Interaction Database version T (MPID-T) [13]

List of extant and extinct databases successfully archived for rapid re-instantiation by use of our cloud computing prototype platform (http://biodb100.apbionet.org)

Conclusions

Database longevity is a chronic problem within the bioinformatics community. In this report, we demonstrate that a relatively inexpensive solution can be implemented for archival of bioinformatics databases and their rapid re-instantiation should the live databases disappear. Organisations, such as the APBioNet, can maintain such databases on a low-cost system using cloud computing for a long period. APBioNet will use this re-instantiation platform for their future InCoB conference submissions, as part of its MIABi compliance, as well as part of its larger effort in the BioDB100 project [7], to build 100 MIABi-compliant interoperable databases.

List of abbreviations used

(APBioNet): Asia Pacific Bioinformatics Network; (InCoB): International Conference on Bioinformatics; (MIABi): Minimum Information about a Bioinformatics Investigation; (ISB): International Society for Biocuration.

Competing interests

TWT and AMK are honorary directors of Asia Pacific Bioinformatics Network Limited. The authors declare that they have no competing interests.

Authors' contributions

TWT conceived the study, participated in its design and coordination. MDS and KSL implemented the prototype system. CX, CPKP, SJL, KRG, JCT, KHC, SR and AMK contributed one or more databases. AMK, MDS, KSL and TWT drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank ISB and BioSharing for the BioDBcore initiative and for their comments and cooperation; ASEAN SubCommittee on Bioinformatics for BioSlax endorsement; our graduate students in the Life Science Integrated Modules of NUS for testing our archival and re-instantiation system. We also thank Hadia Syahirah Abd Raman of Perdana University for her help in preparation of the manuscript.

Declaration
Publication of this article was funded by the EUAsiaGrid Project (FP7 of the EC) awarded to TWT.
This article has been published as part of *BMC Genomics* Volume 14 Supplement 5, 2013: Twelfth International Conference on Bioinformatics (InCoB2013): Computational biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/14/S5.

Authors' details
[1]Department of Biochemistry, National University of Singapore (NUS), Singapore. [2]Bioinformatics Centre, Life Science Institute, NUS, Singapore. [3]Computing Science Department, Institute of High Performance Computing, Singapore. [4]Department of Chemistry and Biomolecular Sciences, Macquarie University, Australia. [5]Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, USA. [6]Perdana University Graduate School of Medicine, Selangor, Malaysia.

Published: 16 October 2013

References

1. Wren JD: **404 not found: the stability and persistence of URLs published in MEDLINE.** *Bioinformatics* 2004, **20(5)**:668-672.
2. Veretnik S, Fink JL, Bourne PE: **Computational biology resources lack persistence and usability.** *PLoS Comput Biol* 2008, **4(7)**:e1000136.
3. Anderson NR, Tarczy-Hornoch P, Bumgarner RE: **On the persistence of supplementary resources in biomedical publications.** *BMC Bioinformatics* 2006, **7**:260.
4. Ranganathan S, Eisenhaber F, Tong JC, Tan TW: **Extending Asia Pacific bioinformatics into new realms in the "-omics" era.** *BMC Genomics* 2009, **10(Suppl 3)**:S1.
5. Tan TW, Tong JC, Khan AM, de Silva M, Lim KS, Ranganathan S: **Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information About a Bioinformatics investigation (MIABi).** *BMC Genomics* 2010, **11(Suppl 4)**:S27.
6. Gaudet P, Bairoch A, Field D, Sansone SA, Taylor C, Attwood TK, Bateman A, Blake JA, Bult CJ, Cherry JM, *et al*: **Towards BioDBcore: a community-defined information specification for biological databases.** *Nucleic Acids Res* 2011, **39(Database)**:D7-10.
7. Ranganathan S, Schonbach C, Nakai K, Tan TW: **Challenges of the next decade for the Asia Pacific region: 2010 International Conference in Bioinformatics (InCoB 2010).** *BMC Genomics* 2010, **11(Suppl 4)**:S1.
8. Tong JC, Lim SJ, Muh HC, Chew FT, Tammi MT: **Allergen Atlas: a comprehensive knowledge center and analysis resource for allergen information.** *Bioinformatics* 2009, **25(7)**:979-980.
9. Gaikwad J, Khanna V, Vemulpad S, Jamie J, Kohen J, Ranganathan S: **CMKb: a web-based prototype for integrating Australian Aboriginal customary medicinal plant knowledge.** *BMC Bioinformatics* 2008, **9(Suppl 12)**:S25.
10. Khan JM, Cheruku HR, Tong JC, Ranganathan S: **MPID-T2: a database for sequence-structure-function analyses of pMHC and TR/pMHC structures.** *Bioinformatics* 2011, **27(8)**:1192-1193.
11. Choo KH, Tan TW, Ranganathan S: **SPdb–a signal peptide database.** *BMC Bioinformatics* 2005, **6**:249.
12. Tay DM, Govindarajan KR, Khan AM, Ong TY, Samad HM, Soh WW, Tong M, Zhang F, Tan TW: **T3SEdb: data warehousing of virulence effectors secreted by the bacterial Type III Secretion System.** *BMC Bioinformatics* 2010, **11(Suppl 7)**:S4.
13. Tong JC, Kong L, Tan TW, Ranganathan S: **MPID-T: database for sequence-structure-function information on T-cell receptor/peptide/MHC interactions.** *Appl Bioinformatics* 2006, **5(2)**:111-114.

doi:10.1186/1471-2164-14-S5-S13
Cite this article as: Tan *et al.*: Simple re-instantiation of small databases using cloud computing. *BMC Genomics* 2013 **14**(Suppl 5):S13.