Database Resources Development and Sharing: A Community Approach

Zhi-Liang Hu and James M. Reecy

Department of Animal Science and Center for Integrated Animal Genomics, Iowa State University, 2255 Kildee Hall, Ames, IA 50011.

Abstract

The increasing amount of genomic data generated by research labs calls for powerful bioinformatics solutions to efficiently store, manage and analyze data. As such, the demand for relational databases, both as a resource and research tool, is steadily increasing within the animal research community. Although the development of database resources and tools has been phenomenal in recent years, the gap between the needs of the community and available resources remain large. Here we propose a community approach, under the NRSP-8 infrastructure, to cope with the increasing needs of various labs. In general, the barriers for an average genomics/genetics research lab to develop database tools to serve their own needs include the cost for hardware, software and expertise. The creation of the NRSP-8 bioinformatics site at the Iowa State University (ISU) made it possible to provide a shared hardware platform on a cost-effective basis. However, this does not overcome the challenges associated with the development of the database structure, inputting data and extracting data from the database. Under our scalable infrastructure model, a central database location may provide shared hardware, operational support, and limited consulting services, while individual labs may have direct access to develop, manage and customize their own database tools on a scalable basis in terms of availability of their expertise. Alternatively, labs with smaller database needs could pool funds for shared development and maintenance. Several of our pilot studies seem to demonstrate that a scalable model may be a practical solution for the community.

Introduction

In recent years, the amount of genome research data has been rapidly increasing, which calls for powerful bioinformatics solutions to efficiently store, manage and analyze them. As such, the demand for the use of relational databases, both as a resource and research tool, is steadily increasing within the animal genome research community. Although the development of database resources and tools has been phenomenal in the past few years, the gap between the needs and available resources within the community still remain large. In addition, the needs for database from different laboratories vary greatly, depending on the availability of hardware, software, expertise and size of the lab. Here we propose a community approach, under the NRSP-8 infrastructure, to cope with the increasing database needs of various labs in the animal genome research community.

Materials and Methods

Under the USDA NRSP-8 Bioinformatics Coordination program, we have acquired and set up at the Iowa State University the following computing facilities:

- A Dual processor Red Hat Enterprise Linux server (RHEL) with 1 Terabyte (TB) storage. This computer has been in service since 2004, used mainly to host web sites and databases.
- An 8-node dual processors CentOS Linux cluster computer with 1 Terabyte storage capacity, which has been in service since the fall of 2007. The use of this computer is under active development and will mainly be used for computing jobs such as batch blast, large scale sequence assembly and other CPU intensive jobs.
- On these computers, a number of server engine software are installed, which include: MySQL and Postgres relational database, Apache/PHP web server, Perl/Bioperl programming suite, etc.

Results

World wide web has become a powerful and very useful form for database interface development over internet (Figure 1) for its obvious advantages. Using this scheme as a base structure, we have been experimenting a scalable scheme to meet the challenges of database needs by various labs where the availability of hardware, software and expertise differ. Roughly, this scalable scheme consists three tiers:

1. First Tier: for labs with large operation scales. The lab will have own expertise and personnel to use the NRSP-8 hardware platform to develop and maintain own database applications through remote ssh and web portal (Figure 2a).

2. Second Tier: for labs with medium operation scales. We will help on initial database design and application development, and at the mean time help to train (full or part time) personnel from the lab for using the application, data in/out operations, and future application development (Figure 2b).

3. Third Tier: for labs with small operation scales. We will help to design and implement the database applications following a generic model, and be responsible for future maintenance. The cases in this category are normally small scale laboratory needs and most likely are one-time deal for development (Figure 2c).

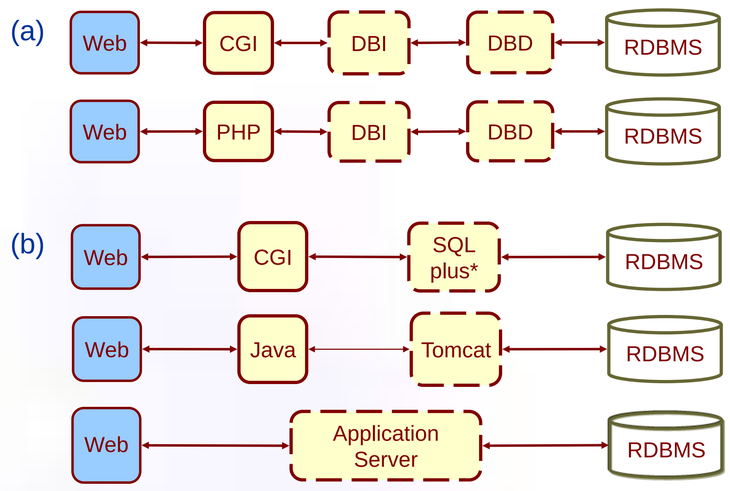
Results (cont'ed)

For all three tiers described above, we will offer hardware platform and software to use (see M&M, and other open source, free or low-cost software where it justify). For example, we installed and offer users a "phpMyAdmin" web application for users to manage and access MySQL database remotely (Figure 3). In general, we will be responsible for database maintenance, data storage, backup and recovery (we encourage users periodically download and back up own data as well).

For labs that have personnel who wish to learn, we offer short initial training on database design and interface programming, either on site, or at Iowa State locations. Continued training and support will also be provided online, to answers general questions, solve problems, offer advises on case by case basis.

Under this proposed scheme, we have started, and are working with several labs, each at a different scale, to develop laboratory database/applications during the past year. These include the lab of Dr. John Liu's Aquatic Genomics Lab at the Auburn University; Dr. Diane Spurlock's Cattle Genomics Lab., and Dr. Max Rothschild's Pig Genomics Lab at the Iowa State University, among a number of other labs that are at different stages of development.

Figure 1. Schematic view of different schemes that can be used for programming database-web applications: (a) shows what we currently support; (b) shows the options we may potentially support when the growing needs justify.

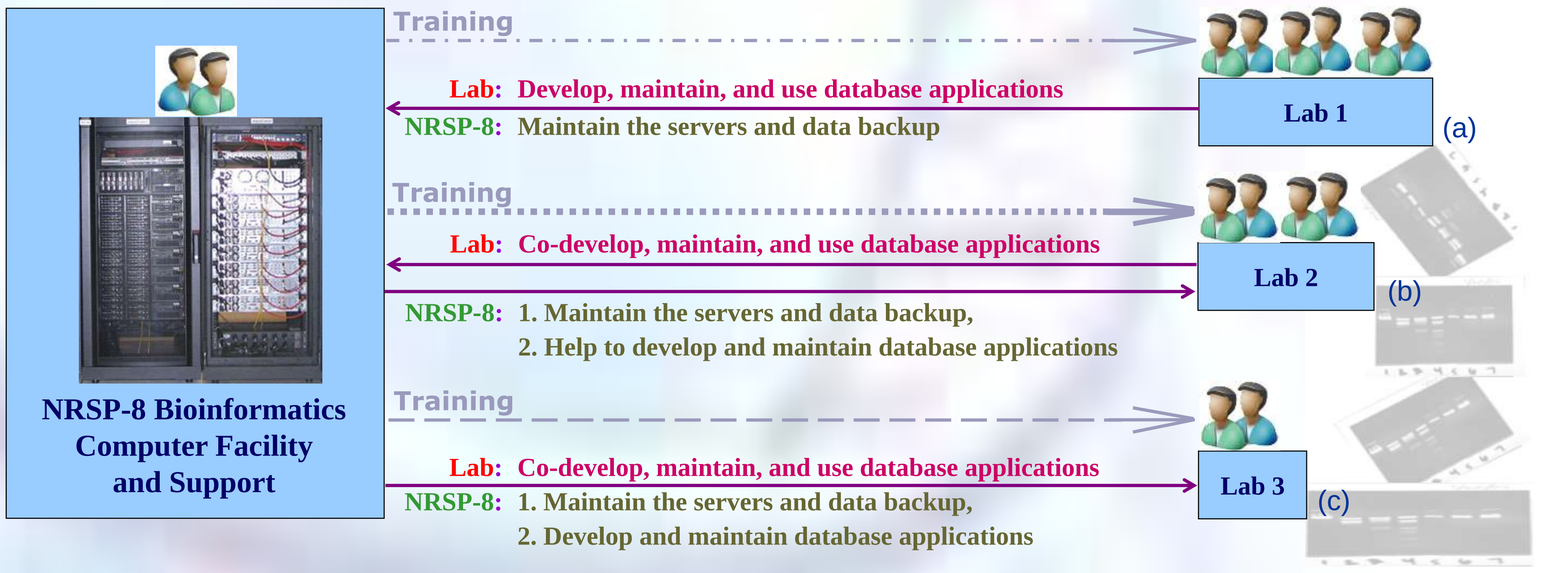


Discussions

In the past a few years, we have received numerous requests in the community for helps on developing and housing databases to serve various research needs, which helped fostering this proposal. While good progress is being made with this approach, we are precocious about possible pitfalls and potential problems down the road, e.g. long term sustainability, etc. On one hand, the database development for genome research is long term and requires continuity, on the other hand, the developmental works is ever changing by nature of research. Towards this end, long term and stable support from NRSP-8 is critical.

So far, both the progress we made and feedback from our collaborators are positive and encouraging. While we are optimistic about the good turn-outs, we will continue to put in efforts for improving the schemes in coping with evolving situation in the community. It is our hope that our efforts will provide a useful model to solve the community research needs for database support on cost-efficient basis.

Figure 2. Different tiers of database support depending upon the scale of the operations and funds available from individual labs.



Acknowledgements

This project has been a team work among many people involved. Our thanks are due, but not limited, to colleagues at NRSP-8 Bioinformatics Coordination program for their various support, and to all collaborators who willingly worked with us on this endeavor.

Colleagues: Co-coordinators: Susan Lamont, Max Rothschild, Chris Tuggle; Reecy Lab students: LaRon Hughes and Eric Fritz; Iowa State University BCB Lab students: Yves Sucaet, Fadi Towfic, and Deepak Reyon. Discussions at times with all species coordinators and others, Ernie Bailey, Noelle Cockett, Jerry Dodgson, John Liu, Max Rothschild, James Womack, and Joan Lunney have been useful too.

Collaborators: John Liu from Auburn University, Diane Spurlock and Max Rothschild from Iowa State University, and personnel from their laboratories: Shao-Lin Wang, Mary Healey, Benny E Mote and Bin Fan.

This research is supported by the USDA NRSP-8 Bioinformatics Coordination Project.

The cartoons below only meant to stress that: (1) Good understanding of the roles of computer in genome research is essential. (2) Good teamwork can never been taken lightly.

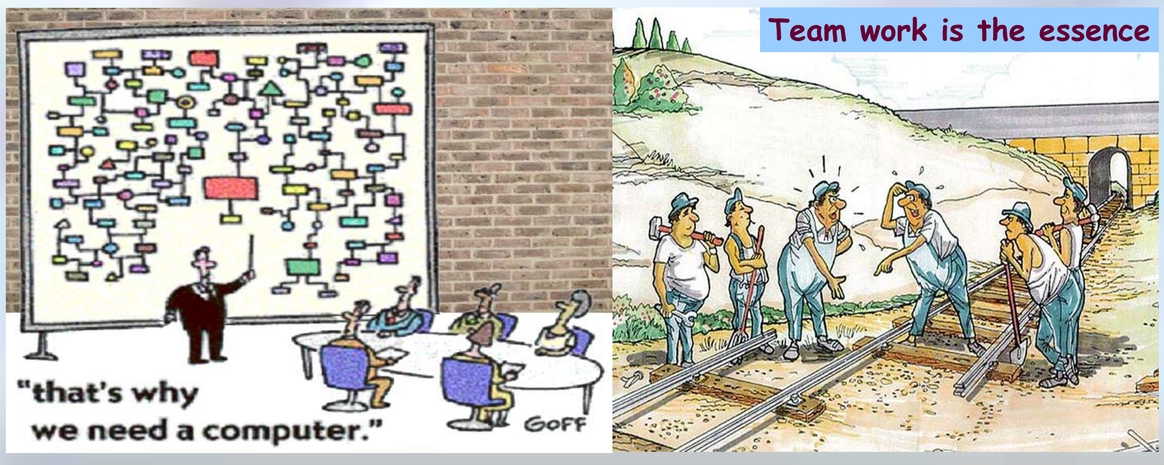


Figure 3. A sample screen shot of a phpMySQL window, a light weighted yet powerful tool. "phpMySQL" is a user-friendly shareware that can be conveniently used for remotely developing and managing MySQL databases.