Setting up an official mirror of PubMLST.org

In order to mirror the PubMLST.org site, you need a Linux system running PostgreSQL, Perl and the Apache web server. The following instructions were written for Bio-Linux 6, which is based on Ubuntu 10.04 LTS, but the procedure will work on any Linux system, although the paths of various commands and installation locations may vary. Certain Perl modules are required to be installed. Many of these are likely to be already installed on a Bio-Linux system, but if not they can be installed using the package management system (all required Perl modules are available as .deb packages for Debian and Ubuntu Linux).

Required Perl modules

- DBI
- XML::Parser::PerlSAX
- Log::Log4perl
- Log::Dispatch::File
- Error
- Config::Tiny
- BioPerl
- IO::String
- Data::UUID
- List::MoreUtils
- Time::Duration

If running Debian/Ubuntu these can be installed with the following commands:

```
sudo apt-get update

sudo apt-get install libdbi-perl libxml-perl liblog-log4perl-perl liblog-dispatch-perl
liberror-perl libconfig-tiny-perl bioperl libio-string-perl libossp-uuid-perl liblist-
moreutils-perl libtime-duration-perl
```

Required helper applications

- EMBOSS
- BLAST+ (use statically linked version)
- Muscle
- IPCRESS (part of exonerate)
- ImageMagick

These are available as Debian/Ubuntu packages and can be installed with the following:

```
sudo apt-get install emboss muscle exonerate imagemagick
```

A statically-linked version of BLAST+ is available as part of Bio-Linux (http://nebc.nerc.ac.uk/tools/bio-linux/bio-linux-6.0) and can be installed from their repository with the following:

```
sudo apt-get install ncbi-blast+-static
```

Register IP address

Register your server IP address with Keith Jolley (keith.jolley@zoo.ox.ac.uk) so that it will be allowed to make an rsync connection to the primary server. You will be assigned a web address for the mirror dependent on your country, e.g. ukmirror2.pubmlst.org for the second UK mirror site. This will be added to the pubmlst.org DNS so that connections to this address are directed to the IP address of your server.

Configuration

1. Configure PostgreSQL to allow connections from the apache user. The exact configuration changes needed will vary depending on your security requirements, but the following should work for a stand-alone system. The paths to the configuration files will depend on the PostgreSQL version – the following are for version 8.4:

   Edit /etc/postgresql/8.4/main/pg_hba.conf to contain the following:

```
# TYPE  DATABASE        USER            CIDR-ADDRESS    METHOD
local   all             postgres                        ident map=local
local   all             all                             ident map=local
host    all             all             127.0.0.1/32    md5
host    all             all             ::1/128         md5
```

   Edit /etc/postgresql/8.4/main/pg_ident.conf to contain the following:

```
# MAPNAME       SYSTEM-USERNAME         PG-USERNAME
local           postgres                postgres
local           webmaster               postgres
local           www-data                apache
local           www-data                remote
local           bigsdb                  bigsdb
local           bigsdb                  apache
```

Edit /etc/postgresql/8.4/main/postgresql.conf to contain the following:

```
listen_addresses = '*'
stats_temp_directory = '/dev/shm'
```

The latter setting prevents a lot of disk activity by using a ramdisk for temporary stats files (ensure that /dev/shm is available on your system before doing this). You may also need to increase the max_connections, shared_buffers, work_mem, and effective_cache_size parameters within postgresql.conf (see http://www.postgresql.org/docs/8.4/static/runtime-config-resource.html for more details).

2. Create database users 'apache', 'webmaster' and 'remote' using the createuser command. You will need to log in as the postgres user to do this (sudo su postgres).

```
createuser apache
createuser remote
createuser webmaster
```

These database users do not need any special privileges (e.g. permission to create new databases or users).

Create passwords for the apache and remote database users:

```
psql -c "ALTER USER apache WITH PASSWORD 'remote'"
psql -c "ALTER USER remote WITH PASSWORD 'remote'"
```

Restart the postgresql daemon

```
sudo /etc/init.d/postgresql-8.4 restart
```

3. Create a user account called 'webmaster' (either log in as root or use the sudo command).

```
sudo /usr/sbin/useradd –m –g users –s /bin/zsh webmaster
```

The Z-shell (/bin/zsh) is the default on Bio-Linux but any shell will do, e.g. /bin/bash or /bin/csh.

Create a password for the webmaster account:

```
sudo /usr/bin/passwd webmaster
```

4. The mirror update script expects the web cgi-bin directory to be located at /var/www/cgi-bin. If this doesn't exist, create a symlink here to the real cgi-bin directory (this may be at /usr/lib/cgi-bin).

```
sudo ln -s /usr/lib/cgi-bin /var/www/cgi-bin
```

Create 'mlstdbnet' and 'bigsdb' subdirectories and ensure these are writable by the webmaster account:

```
sudo mkdir /var/www/cgi-bin/mlstdbnet
sudo chown webmaster /var/www/cgi-bin/mlstdbnet
sudo mkdir /var/www/cgi-bin/bigsdb
sudo chown webmaster /var/www/cgi-bin/bigsdb
```

Create other directories required for BIGSdb installation:

```
sudo mkdir -p /usr/local/lib/BIGSdb
sudo chown webmaster /usr/local/lib/BIGSdb

sudo mkdir -p /etc/bigsdb
sudo chown webmaster /etc/bigsdb

sudo mkdir -p /etc/bigsdb/dbases
sudo chown webmaster /etc/bigsdb/dbases
```

5. Create the /home/httpd directory and make it owned by webmaster:

```
sudo mkdir /home/httpd
sudo chown webmaster /home/httpd
```

6. Log in as webmaster (sudo su webmaster) and copy the site as follows:

```
rsync –av --exclude="/tmp" pubmlst.org::pubmlst /home/httpd/pubmlst.org
```

(the above is all one line with a space after pubmlst.org::pubmlst).
7. Create a directory for temporary files

```
mkdir /home/httpd/pubmlst.org/tmp
chmod 777 /home/httpd/pubmlst.org/tmp
```

8. Run the update script to copy the script directory and create the databases:

```
cd  (ensure you're in the webmaster home directory)
/home/httpd/pubmlst.org/mirror/scripts/updatemirror
```

The databases are all prefixed 'pubmlst_' to avoid clashing with other databases you may have on your system.

9. Create an empty log file for use by BIGSdb and ensure it is writable by the web server daemon (www-data on Debian/Ubuntu)

```
sudo touch /var/log/bigsdb.log
sudo chown www-data /var/log/bigsdb.log
```

10. As root (or using the sudo command) create a directory called /usr/local/mlstdbnet and copy the newly downloaded mlstdbnet.conf file to it:

```
sudo mkdir /usr/local/mlstdbnet

sudo cp /home/httpd/pubmlst.org/mirror/conf/mlstdbnet.conf /usr/local/mlstdbnet  (this is one line)
```

The configuration file is suitable for use with Bio-Linux. Check through the configuration and change any settings which are not appropriate. You may need to change the paths of programs such as BLAST to whatever is used on your system.

11. Configure apache and create a virtual host for the web address provided (e.g. ukmirror2.pubmlst.org):

Look in /etc/apache2/mods-enabled and see if include.load and rewrite.load are present. If either are not, enable the appropriate modules with:

```
sudo a2enmod include
sudo a2enmod rewrite
```

Create a file called /etc/apache2/sites-available/pubmlst.org with the following contents (replace the ukmirror2.pubmlst.org hostname with the name of your mirror and add an appropriate Email address for the server admin):

```
<VirtualHost *:80>
ServerName ukmirror2.pubmlst.org
ServerAdmin keith.jolley@zoo.ox.ac.uk
DocumentRoot /home/httpd/pubmlst.org
CustomLog /var/log/apache2/pubmlst.org_access.log combined
Alias /images/ /home/httpd/pubmlst.org/images/
 <Directory /home/httpd/pubmlst.org>
  Options +Includes -Indexes
  DirectoryIndex index.shtml index.html
  AllowOverride All
 </Directory>
RewriteEngine on
RewriteRule ^/cgi-bin/(.+mlstdbnet.pl) /perl/$1 [R,L]
RewriteRule ^/cgi-bin/(.+bigsdb.pl) /perl/$1 [R,L]
RewriteRule ^/cgi-bin/(.+bigscurate.pl) /perl/$1 [R,L]
RewriteRule ^/cgi-bin/(.+agdbnet.pl) /perl/$1 [R,L]
ErrorDocument 404 /errors/404.shtml
ErrorDocument 403 /errors/403.shtml
ErrorDocument 401 /errors/401.shtml
</VirtualHost>
```

Ensure mod_perl is installed:

```
sudo apt-get install libapache2-mod-perl2
```

Add the following to /etc/apache2/httpd.conf:

```
PerlSwitches -T

Alias /perl/ /usr/lib/cgi-bin/
<Location /perl>
    SetHandler perl-script
    PerlResponseHandler ModPerl::Registry
    PerlOptions +ParseHeaders
    Options +ExecCGI
</Location>
```

Enable the new site configuration:

```
sudo a2ensite pubmlst.org
```

Add a file called /etc/apache2/modperl_startup.pl with the following contents:

```
#!/usr/bin/perl
use lib "/var/www/cgi-bin/mlstdbnet/lib";
use lib "/var/www/cgi-bin/mlstdbnet/Plugins";
1;
```

Add the following line to the bottom of /etc/apache2/apache2.conf

```
PerlRequire "/etc/apache2/modperl_startup.pl"
```

Restart the web server daemon:

```
sudo /etc/init.d/apache2 restart
```

12. Enable web traffic (port 80) through the firewall (if running). On Bio-Linux 6, you can do this using the command gufw (use the preconfigured option: Allow In Service HTTP).

Chart Director

BIGSdb will optionally use the ChartDirector graphics library if it is installed - some plugins also require it. This is a commercial library available from http://www.advsofteng.com/ and we can not distribute it. You can download and install this without a license for evaluation, although the charts will display a banner stating that the software is unregistered. Registration currently costs US$99.

To enable ChartDirector, set 'chartdirector=1' in the /etc/bigsdb.conf file.

Scheduling update and cleanup jobs

Updates should be scheduled to run once a day. Use the updatemirror script you used to copy the site and databases over (/home/httpd/pubmlst.org/mirror/scripts/updatemirror). You should also regularly clean up any BIGSdb temporary files in the web temp and secure temp directories (/home/httpd/pubmlst.org/tmp and /var/tmp). The following entries in /etc/crontab will schedule a daily update at 21:00 and clean out the two temporary directories of files older than 7 days (run every day at 06:00 and 06:10).

```
# m h dom mon dow user command
0 21   * * *   webmaster    /home/httpd/pubmlst.org/mirror/scripts/updatemirror
0 6    * * *   root    find /var/tmp/ -name '*BIGSdb_*' -type f -mmin +10080 -exec rm -f {} \;
2>/dev/null
10 6   * * *   root    find /home/httpd/pubmlst.org/tmp/ -type f -mmin +10080 -exec rm -f {}
\; 2>/dev/null
```

Prevent log file from getting too large

Set the log file to auto rotate by adding a file called 'bigsdb' with the following contents to /etc/logrotate.d:

```
/var/log/bigsdb.log {
  weekly
  rotate 4
  compress
  copytruncate
  missingok
  notifempty
  create 640 root adm
}
```

Running the offline job manager

Some plugins require a long time to run their jobs and these are consequently run offline. By default offline jobs are not enabled on mirror sites. If you'd like to run offline jobs, e.g. to enable Genome Comparator, you may do so by following the instructions at http://pubmlst.org/software/database/bigsdb/installation/. You will need to download the latest BIGSdb package and create a bigsdb_jobs database as described.