CLC **Bioinformatics** Database

User manual

Manual for

CLC Bioinformatics Database 3.7

Windows, Mac OS X and Linux

April 12, 2013

This software is for research purposes only.

CLC bio
Finlandsgade 10-12
DK-8200 Aarhus N
Denmark

Contents

Chapter 1

Introduction

The technical set-up of a *CLC Bioinformatics Database* involves both a database back-end and a server middle layer that the Workbench connects to. This server layer also has a web interface which has very basic capabilities for uploading, downloading, browsing and searching for data.

The functionality of the *CLC Bioinformatics Database* can be extended by installation of Server plug-ins. The available plug-ins can be found at http://www.clcbio.com/server_plugins.

1.1 System requirements

The system requirements of *CLC Bioinformatics Database* are:

- Windows XP, Windows Vista, or Windows 7, Windows Server 2003 or Windows Server 2008
- Mac OS X 10.6 or later. However, Mac OS X 10.5.8 is supported on 64-bit Intel systems.
- Linux: Red Hat or SUSE
- 32 or 64-bit
- 256 MB RAM required

Minimum 2GB of RAM recommended, depending on the type of analysis to be performed.

1.2 Licensing

There are three kinds of licenses needed for running the *CLC Bioinformatics Database*:

- A license for the server. This is needed for running the server. The license will allow a certain number of sessions (i.e. number of log-ins from e.g. the web interface or the workbench). The number of sessions is part of the agreement with CLC bio when you purchase a license.
- A special license if you are running the server on a grid system (this is explained in detail in section **??**)

- A license for the workbench. The Workbench is needed to start analyses on the server and view the results. Find the user manuals and deployment manual for the Workbenches at http://www.clcbio.com/usermanuals

The following chapter on installation will give you more information about how to obtain and deploy the license for the server.

Chapter 2

Installation

2.1 Quick installation guide

The following describes briefly the steps needed to set up a *CLC Bioinformatics Database 3.7* with pointers to more detailed explanation of each step.

1. Download and run the server installer. As part of the installation, choose to start the server (section 2.3).
2. Run the license download script that is part of the installation (section 2.7).
3. The script will automatically download a license file and place it in the server installation directory under `licenses`.
4. Restart the server (section 2.8).
5. Log into the server using a web browser (note that the default port is 7777) with username **root** and password **default** (section 3).
6. Change the root password (section 4.1).
7. Configure authentication mechanism and optionally set up users and groups (section 4.2).
8. Add data locations (section 3.2).
9. Download and install plug-ins *in the Workbench* needed for the Workbench to contact the server (section 2.9).
10. Check your server set-up using the **Check set-up** link in the upper right corner as described in section 6.1.1.
11. Your server is now ready for use.

2.2 Installing the database

2.2.1 Download and install a Database Management System

If you do not already have an existing installation of a Database Management System (*DBMS*) you will have to download and install one. The CLC Bioinformatics Database can be used with a

number of different DMBS implementation. Choosing the right one for you and your organization depends on many factors such as price, performance, scalability, security, platform-support, etc.

Information about the supported solutions are available on the links below.

- MySQL: http://dev.mysql.com/downloads/
- PostgreSQL: http://www.postgresql.org/
- Microsoft SQL Server: http://www.microsoft.com/SQL/
- Oracle: http://www.oracle.com/

If you have further questions about the installation of the DBMS, please contact support@clcbio.com. See section 6.2 for guidance on special configurations for the DBMS.

2.2.2 Create a new database and user/role

Once your DBMS is installed and running you will need to create a database for containing your CLC data. We also recommend that you create a special database-user (sometimes called a database-role) for accessing this database.

Consult the documentation of your DBMS for information about creating databases and managing users/roles.

2.2.3 Initialize the database

Before you can connect to your database from a CLC Workbench or Server it must be initialized. The initialization creates the required tables for holding objects, and prepares an index used for searching. Initialization is performed with the CLC Bioinformatics Database Tool (figure 2.1).

- Install the CLC Bioinformatics Database Tool on a client machine, and start the program.
- Fill in the fields with the required information.
  - Hostname: The fully-qualified hostname of the server running the database.
    NOTE: The same hostname must be used every time you connect to the database
  - Port: The TCP/IP listening port on the database server
  - Database name: The name of the database you created in the previous section
  - Username: the name of the user/role you created in the previous section
  - Password: the password for the user/role.
- To re-initializing an existing CLC database you must check the "Delete Existing..." checkbox.
  NOTE: ANY DATA ALREADY STORED IN THE CLC DATABASE WILL BE DELETED.
- Click the Initialize Database button to start the process.

While the program is working the progress-bar will show the status and the transcript will show a log of actions, events and problems. If anything goes wrong, please consult the transcript for

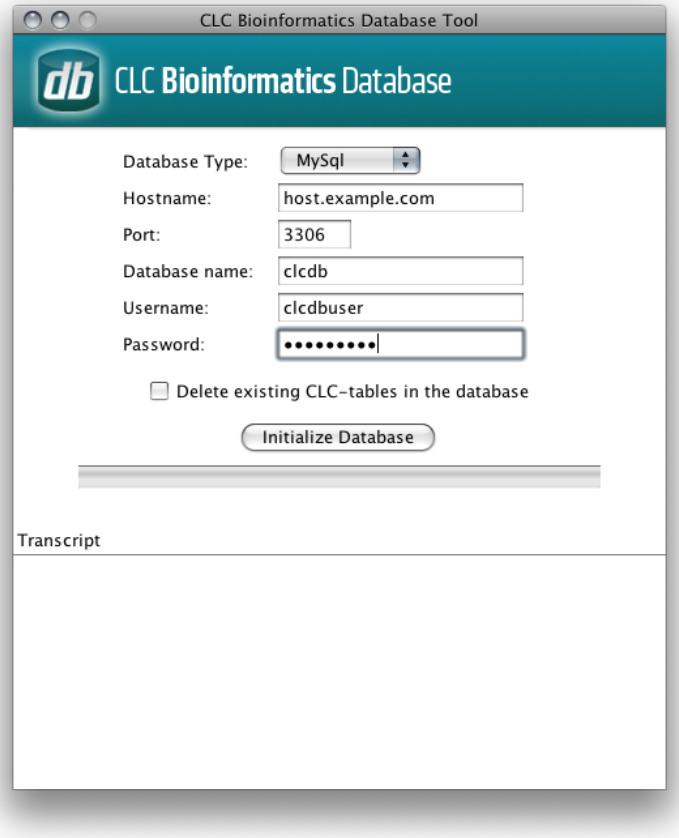

Figure 2.1: *The CLC Bioinformatics Database tool*

more information. If you need assistance, please contact support@clcbio.com, and include the contents of transcript.

If the initialization is successful, the status bar will display this message: *Database successfully initialized*. You can now close the CLC Bioinformatics Database Tool.

2.3 Installing and running the Server

Getting the *CLC Bioinformatics Database* software installed and running involves, at minimum, these steps:

1. Install the software.
2. Ensure the necessary port in the firewall is open.
3. Download a license.
4. Start the Server and/or configure it as a service.

All these steps are covered in this section of the manual. Further configuration information, including for job nodes, grid nodes, and External Applications, are provided in later chapters.

Installing and running the *CLC Bioinformatics Database* is straightforward. However, if you do run into troubles, please refer to the troubleshooting section in Appendix 6.1, which provides tips on how to troubleshoot problems yourself, as well as how to get help.

2.3.1 Installing the Server software

The installation can only be performed by a user with administrative privileges. On some operating systems, you can double click on the installer file icon to begin installation. Depending on your operating system you may be prompted for your password (as shown in figure 2.2) or asked to allow the installation to be performed.

- On Windows 7 or Vista, you will need to right click on the installer file icon, and choose to **Run as administrator**.
- For the Linux-based installation script, you would normally wish to install to a central location, which will involve running the installation script as an administrative user - either by logging in as one, or by prefacing the command with sudo. Please check that the installation script has executable permissions before trying to execute it.

Figure 2.2: *Enter your password.*

Next you will be asked where to install the server 2.3. If you do not have a particular reason to change this, simply leave it at the default setting. The chosen directory will be referred to as the *server installation directory* throughout the rest of this manual.

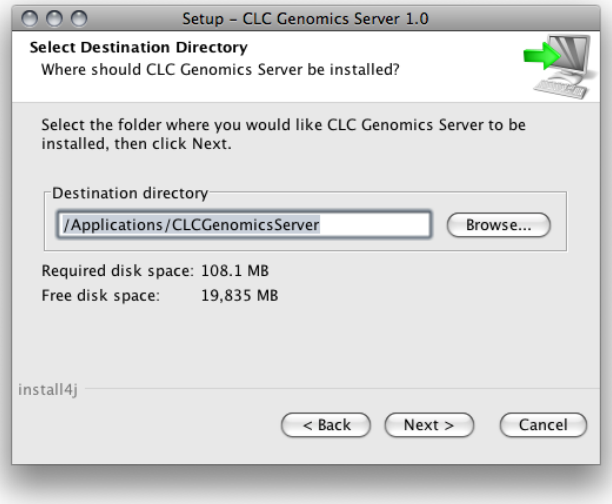

Figure 2.3: *Choose where to install the server.*

The installer allows you to specify the maximum amount of memory the CLC Server will be able to utilize 2.4. The range of choice depends on the amount of memory installed on your system and on the type of machine used. On 32 bit machines you will not be able to utilize more than 2 GB of memory – on 64-bit machines there is no such limit.

If you do not have a reason to change this value you should simply leave it at the default setting.

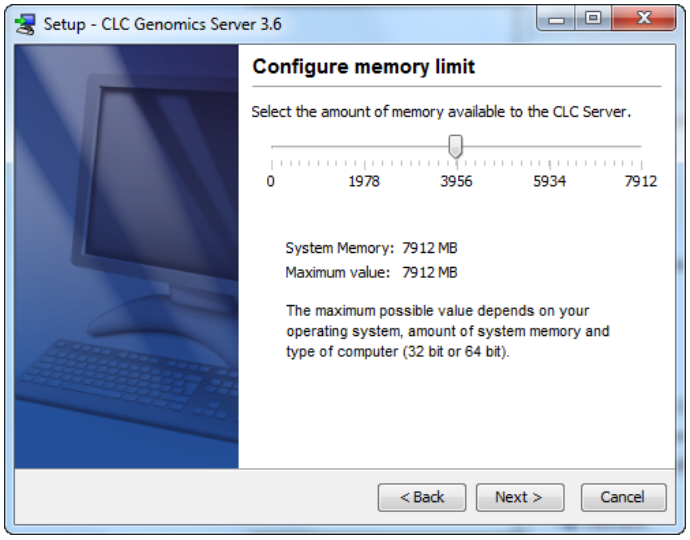

Figure 2.4: *Choose the maximum amount of memory used by the server.*

If you are installing the server on a Windows system you will be able to choose if the service is started manually or automatically by the system.

The installer will now extract the necessary files.

On a Windows system, if you have chosen that the service should be started automatically, the service should also start running at this point. Please note that if you do not already have a license file installed, then the *CLC Bioinformatics Database* process will be running in a limited capacity at this point. Downloading a license is described in section 2.7.

Information on stopping and starting the *CLC Bioinformatics Database* service is provided in section 2.8.

2.4 Silent installation

The installer also has a silent installation mode which is activated by the `-q` parameter when running the installer from a command line, e.g.

```
CLCGenomicsServer_3_0.exe -q
```

On Windows, if you wish to have console output, `-console` can be appended *as the second parameter* (this is only needed when running on Windows where there is no output per default):

```
CLCGenomicsServer_3_0.exe -q -console
```

You can also in silent mode define a different installation directory: `-dir`.

```
CLCGenomicsServer_3_0.exe -q -console -dir "c:\bioinformatics\clc"
```

Note! Both the `-console` and the `-dir` options only work when the installer is run in silent

mode.

The `-q` and the `-console` options work for the Uninstall program as well.

2.5 Upgrading an existing installation

Upgrading an existing installation is very simple. For a single Genomics Server, the steps we recommend are:

- Make sure that nobody is using the server (see section 4.4). A standard procedure would be to give users advance notice that the system will be unavailable for maintenance.
- Install the server in the same installation directory as the one already installed. All settings will be maintained.

If you have a CLC job node setup, you will also need to upgrade the *CLC Bioinformatics Database* software on each job node. Upgrading the software itself on each node is all you need to do. Configurations and plug-ins for job nodes are pushed to them by the master node.

2.5.1 Upgrading major versions

Once you have performed the steps mentioned above, there are a few extra details whenever the release is more than a bug-fix upgrade (e.g. a bug-fix release would be going from version 1.0 to 1.0.1).

First, make sure all client users are aware that they must upgrade their Workbench and server connection plug-in.

For major versions (e.g. going from 1.X to 2.0) a new license needs to be downloaded (see section 2.7), and the server restarted.

2.6 Allowing access through your firewall

By default, the server listens for TCP-connections on port 7777 (See section 3.3 for info about changing this).

If you are running a firewall on your server system you will have to allow incoming TCP-connections on this port before your clients can contact the server from a Workbench or web browser. Consult the documentation of your firewall for information on how to do this.

Besides the public port described above the server also uses an internal port on 7776. There is no need to allow incoming connections from client machines to this port.

2.7 Downloading a license

The *CLC Bioinformatics Database* will look for licenses in the `licenses` folder. This means that all license files should be located in this folder. Check the platform-specific instructions below to see how to download a license file.

2.7.1 Windows license download

License files are downloaded using the `downloadlicense.command` script. To run the script, right-click on the file and choose **Run as administrator**. This will present a window as shown in figure 2.5.

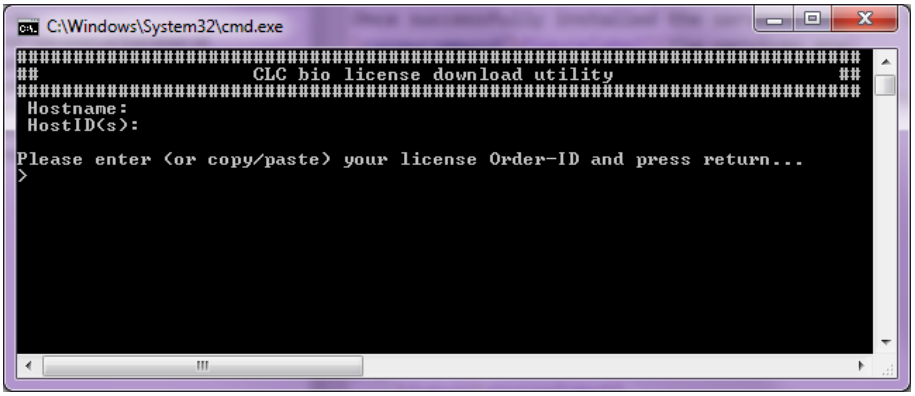

Figure 2.5: *Download a license based on the Order ID.*

Paste the Order ID supplied by CLC bio (right-click to **Paste**) and press Enter. Please contact support@clcbio.com if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense.command` script, it will create a new license file.

Restart the server for the new license to take effect (see how to restart the server in section 2.8.1).

2.7.2 Mac OS license download

License files are downloaded using the `downloadlicense.command` script. To run the script, double-click on the file. This will present a window as shown in figure 2.6.

Paste the Order ID supplied by CLC bio and press Enter. Please contact support@clcbio.com if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense.command` script, it will create a new license file.

Restart the server for the new license to take effect (see how to restart the server in section 2.8.2).

2.7.3 Linux license download

License files are downloaded using the `downloadlicense` script. Run the script and paste the Order ID supplied by CLC bio. Please contact support@clcbio.com if you have not received an Order ID.

Note that if you are *upgrading* an existing license file, this needs to be deleted from the `licenses` folder. When you run the `downloadlicense` script, it will create a new license file.

Restart the server for the new license to take effect. Restarting the server is covered in in

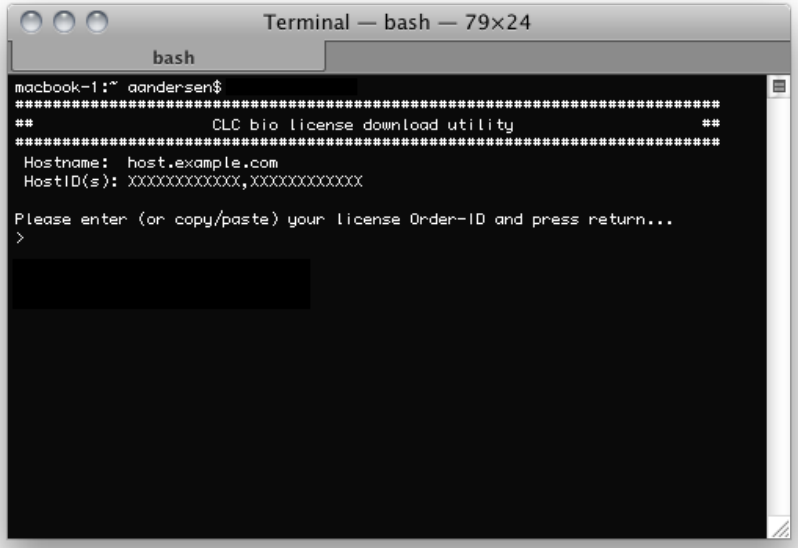

Figure 2.6: *Download a license based on the Order ID.*

section 2.8.3.

2.8 Starting and stopping the server

2.8.1 Microsoft Windows

On Windows based systems the *CLC Bioinformatics Database* can be controlled through the *Services* control panel. Choose the service called `CLCBioinformaticsDatabase` and click the start, stop or restart link as shown in figure 2.7.

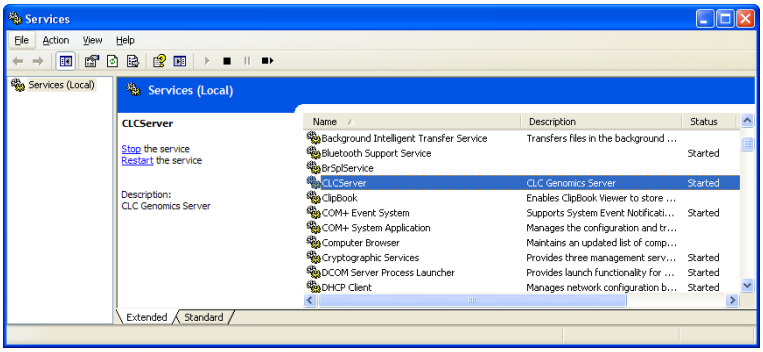

Figure 2.7: *Stopping and restarting the server on Windows by clicking the blue links.*

2.8.2 Mac OS X

On Mac OS X the server can be started and stopped from the command line.

Open a terminal and navigate to the CLC Server installation directory. Once there the server can be controlled with the following commands.

To start the server run the command:

```
sudo ./CLCBioinformaticsDatabase start
```

To stop the server run the command:

```
sudo ./CLCBioinformaticsDatabase stop
```

To view the current status of the server run the command:

```
sudo ./CLCBioinformaticsDatabase status
```

You will need to set this up as a service if you wish it to be run that way. Please refer to your operating system documentation if you are not sure how to do this.

2.8.3 Linux

You can start and stop the Genomics Server from the command line. You can also run it as a service.

Command Line:

Open a terminal and navigate to the CLC Server installation directory. Once there the server can be controlled with the following commands.

To start the server run the command:

```
./CLCBioinformaticsDatabase start
```

To stop the server run the command:

```
./CLCBioinformaticsDatabase stop
```

To view the current status of the server run the command:

```
./CLCBioinformaticsDatabase status
```

You can run these commands as any user, either directly on the command, or by prefacing the commands with sudo. e.g .

```
sudo –u <username> ./CLCBioinformaticsDatabase status
```

Please note: To run the Genomics Server as a non-root user, you should first ensure that the files under your *CLC Bioinformatics Database* installation directory have the appropriate permissions. The easiest way to do this is to recursively change the ownership of all the files in that directory to the user that will run the *CLC Bioinformatics Database* process.

As a service:

On installation a link is created in /etc/init.d to the CLCBioinformaticsDatabase script. At this point, you need to configure it as a service. For example, on Red Hat systems,

- service clcserver-startupscript start

We provide an example wrapper script for running the *CLC Bioinformatics Database* process. You may find this convenient if you plan to run the process as a user other than the root user. The script is called clcserver-startupscript and can be found under the **conf** directory of the *CLC Bioinformatics Database* installation area.

Please refer to your operating system documentation if you require further information about setting up services on your system.

2.9 Installing relevant plug-ins in the Workbench

In order to use the *CLC Bioinformatics Database* from a CLC Workbench, you need to install the CLC Workbench Client Plug-in in the Workbench. This plug-in is needed for logging onto the server and accessing data from the server data locations.

Plug-ins are installed using the Plug-ins and Resources Manager[1], which can be accessed via the menu in the Workbench

Help | **Plug-ins and Resources (**  **)**

or via the **Plug-ins (**  **)** button on the Toolbar.

From within the Plug-ins and Resources Manager, choose the Download Plug-ins tab and click on the CLC Workbench Client Plugin. Then click in the button labelled **Download and Install**.

If you are working on a system not connected to the network, then you can also install the plug-in by downloading the cpa file from the plugins page of our website

http://www.clcbio.com/clc-plugin/

Then start up the Plug-in manager within the Workbench, and click on the button at the bottom of the Plug-in manager labelled **Install from File**.

You need to restart the Workbench before the plug-in is ready for use.

Note that if you want users to be able to use **External applications** (see chapter **??**) on the server, there is a separate plug-in (CLC External Applications Plug-in) that needs to be installed in the Workbench the same way as described above.

[1] In order to install plug-ins on many systems, the Workbench must be run in administrator mode. On Windows Vista and Windows 7, you can do this by right-clicking the program shortcut and choosing "Run as Administrator".

Chapter 3

Configuring and administering the server

3.1 Logging into the administrative interface

Once the server is running, you can log into its administrative interface via a web browser. Most configuration occurs via this interface. Simply type the host name of the server machine you have installed the *CLC Bioinformatics Database* software on, followed by the port it is listening on. Unless you change it, the port number is 7777. An example would be

```
http://clccomputer:7777/
```

The default administive user credentials are:

```
username: root
password: default
```

Use these details the first time you log in.

3.2 Adding locations for saving data

Before you can use the server for doing analyses you will need to add one or more locations for storing your data. The locations can either be simple pointers to a folder on the file system or based on a *CLC Bioinformatics Database*.

To set up a location, open a web browser and navigate to the CLC Server web interface.

Once logged in go to the *Admin* tab and unfold the *Main configuration* section. There are two headings relating to CLC data storage: Database locations and File system locations.

3.2.1 Adding a database location

Before adding a database location, you need to set-up the database. This is described in section 2.2.

Under the **Database locations** heading, click the **Add New Database Location** button to add a new database location (see figure 3.2).

Enter the required information about host, port and type of database. The user name and password refers to the user role on your Database Management System (DBMS), see section

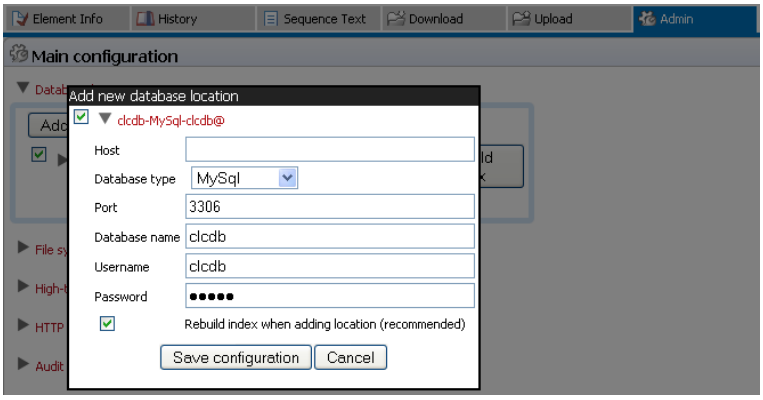

Figure 3.1: *Database location settings.*

2.2. Note that there are two version of Oracle in the list. One is the traditional using SID style (e.g. `jdbc:oracle:thin:@[HOST][:PORT]:SID`) and the other is using thin-style service name (e.g. `jdbc:oracle:thin:@//[HOST][:PORT]/SERVICE`).

Click the *Save Configuration* button to perform the changes. The added database location should now appear in the **Navigation Area** in the left hand side of the window.

3.2.2 Adding a file system location

Under the **File system locations** heading, click the **Add New File Location** button to add a new file system location (see figure 3.2).

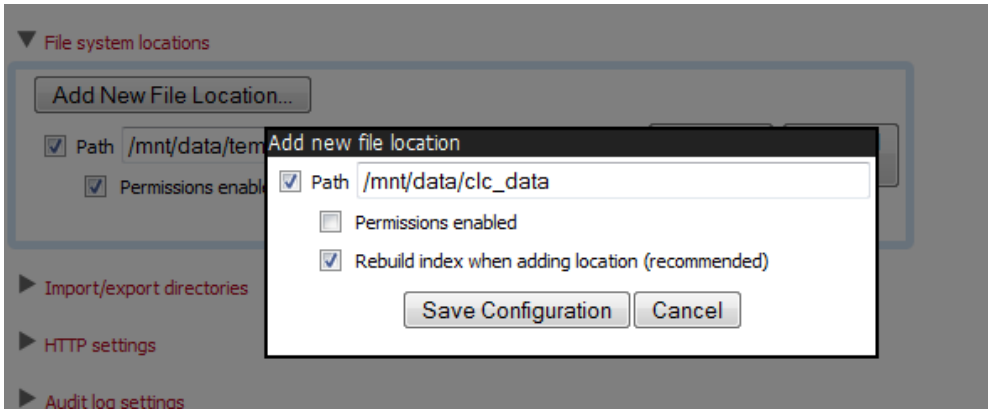

Figure 3.2: *File system location settings.*

In this dialog, enter the path to the folder you want to use for storing the data. The path should point to an *existing* folder on the server machine, and the user *running the server process* needs to have read and write access to the folder. This is usually a dedicated user, or it may be the system's root user if you have not created a dedicated user for this purpose.

The file location(s) configured on the server will be accessible to those working using CLC Workbenches after they log into the server via their Workbench.

Once you have pressed **Save Configuration** (learn more about rebuilding the index in section 3.2.3), this location will be added and it should now appear in the **Navigation Area** in the left hand side of the window. By default it will also appear in the Workbench on next login. You can use the checkbox next to the location to indicate whether it should be visible to your users or not.

You can choose whether access control should be switched on and off. Please see section 5.1 for more information about enabling and setting permissions on *CLC Bioinformatics Database* data folders.

Note that pressing **Remove Location** will only remove the location from this list - it will not delete the folder from your system or affect any data already stored in this folder. The data will be accessible again simply by adding the folder as a new location again.

Important points about the CLC Server data in the file system locations

Any file system locations added here should be folders **dedicated for use** by the *CLC Bioinformatics Database*. Such areas should be directly accessed only by the *CLC Bioinformatics Database*. In other words, files should **not** be moved into these folders, or their subfolders, manually, for example using your standard operating system's command tools, drag and drop, and so on. All the data stored in this areas will be in clc format and will be owned by the user that runs the *CLC Bioinformatics Database* process.

File locations for job node set-ups

When you have a job node set-up, all the job node computers need to have access to the same data location folder. This is because the job nodes will write files directly to the folder rather than passing through the master node (which would be a bottleneck for big jobs). Furthermore, the user running the server must be the same for all the job nodes and it needs to act as the same user when accessing the folder no matter whether it is a job node or a master node.

The data location should be added **after** the job nodes have been configured and attached to the master node. In this way, all the job nodes will inherit the configurations made on the master node.

One relatively common problem faced in this regard is *root squashing* which often needs to be disabled, because it prevents the servers from writing and accessing the files as the same user - read more about this at `http://nfs.sourceforge.net/#faq_b11`.

You can read further about job node setups in section **??**

3.2.3 Rebuilding the index

The server maintains an index of all the elements in the data locations. The index is used when searching for data. For all locations you can choose to **Rebuild Index**. This should be done only when a new location is added or if you experience problems while searching (e.g. something is missing from the search results). This operation can take a long time depending on how much data is stored in this location.

If you move the server from one computer to another, you need to move the index as well. Alternatively, you can re-build the index on the new server (this is the default option when you add a location). If the rebuild index operation takes too long and you would prefer to move the old index, simply copy the folder called `searchindex` from the old server installation folder to the new server.

The status of the index server can be seen in the **User Statistics** pane showing information on where the index server resides and the number of locations currently being serviced.

3.3 Changing the listening port

The default listening port for the CLC Server is 7777. This has been chosen to minimize the risk of collisions with existing web-servers using the more familiar ports 80 and 8080. If you would like to have the server listening on port 80 in order to simplify the URL, this can be done in the following way.

- Navigate to the CLC Server installation directory.
- Locate the file called *server.xml* in the conf directory.
- Open the file in a text editor and locate the following section

```
<Connector port="7777" protocol="HTTP/1.1"
           connectionTimeout="20000"
           redirectPort="8443" />
```

- Change the port value to desired listening port (80 in the example below)

```
<Connector port="80" protocol="HTTP/1.1"
           connectionTimeout="20000"
           redirectPort="8443" />
```

- Restart the service for the change to take effect (see how to restart the server in section 2.8).

3.4 Setting the amount of memory available for the JVM

When running the *CLC Bioinformatics Database*, the Java Virtual Machine (JVM) needs to know how much memory it can use. This depends on the amount of physical memory (RAM) and can thus be different from computer to computer. Therefore, the installer investigates the amount of RAM during installation and sets the amount of memory that the JVM can use.

On **Windows** and **Linux**, this value is stored in a property file called `ServerType.vmoptions` (e.g. `CLCGenomicsServer.vmoptions`) which contains a text like this:

```
-Xmx8000m
```

The number (8000) is the amount of memory the *CLC Bioinformatics Database* is allowed to use. This file is located in the installation folder of the the *CLC Bioinformatics Database* software.

By default, he value is set to 50% of the available RAM on the system you have installed the software on.

You can manually change the number contained in the relevant line of the vmoptions file for your *CLC Bioinformatics Database* if you wish to raise or lower the amount of RAM allocated to the Java Virtual Machine.

3.5 Limiting the number of cpus available for use

A number of the algorithms in the *CLC Bioinformatics Database* will, in the case of large jobs, use all the cores available on your system to make the analysis as fast as possible. If you wish

to restrict this to a predefined number of cores, this can be done with a properties file: Create a text file called *cpu.properties* and save it in the *settings* folder under the *CLC Bioinformatics Database* installation directory.

The cpu.properties file should include one line like this:

maxcores = 1

Restart the *CLC Bioinformatics Database* if you create or change this file for these settings to take effect.

Instead of 1 you write the maximum number of cores that the *CLC Bioinformatics Database* is allowed to use. Please note that this is not a guarantee that the *CLC Bioinformatics Database* will never use more cores than specified, but that will be for very brief and infrequent peaks and should not affect performance of other applications running on your system.

You can download a sample cpu.properties file at http://clcbio.com/files/deployment/cpu.properties.

3.6 Other configurations

3.6.1 HTTP settings

Under the **Admin** () tab, click **Configuration**, and you will be able to specify HTTP settings. Here you can set the time out for the user HTTP session and the maximum upload size (when uploading files through the web interface).

3.6.2 Audit log settings

The audit log records all the actions performed in the web interface and through the Workbenches. Note that data management operations (copying, deleting and adding files) done through the Workbench are not recorded.

3.6.3 Deployment of server information to CLC Workbenches

See the *Deployment manual* at http://www.clcbio.com/usermanuals for information on pre-configuring the server log-in information when Workbench users log in for the first time.

Chapter 4

Managing users and groups

4.1 Logging in the first time - root password

When the server is installed, you will be able to log in via the web interface using the following credentials:

- **User name**: `root`
- **Password**: `default`

Once logged in, you should as a minimum set up user authentication (see section 4.2) and data locations (see section 3.2) before you can start using the server.

For security reasons, you should change the root password (see figure 4.1):

Admin (⚙) | Authentication (🔑) Change root password

Note that if you are going to use job nodes, it makes sense to set these up before changing the authentication mechanism and root password (see section **??**).

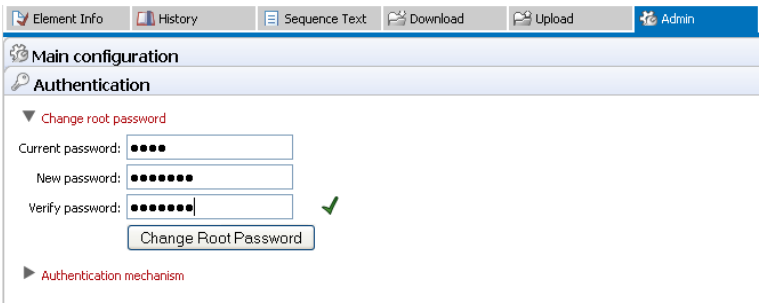

Figure 4.1: *We recommend changing the root password. The verification of the root password is shown with the green checkmark.*

4.2 User authentication using the web interface

When the server is installed, you can log in using the default root password (username=root, password=default). Note that if you are going to use job nodes, it makes sense to set this up first before changing the authentication mechanism and root password (see section **??**).

Once logged in, you can specify how the general user authentication should be done:

Admin | **Authentication** | **Authentication mechanism**

This will reveal the three different modes of authentication as shown in figure 4.2.

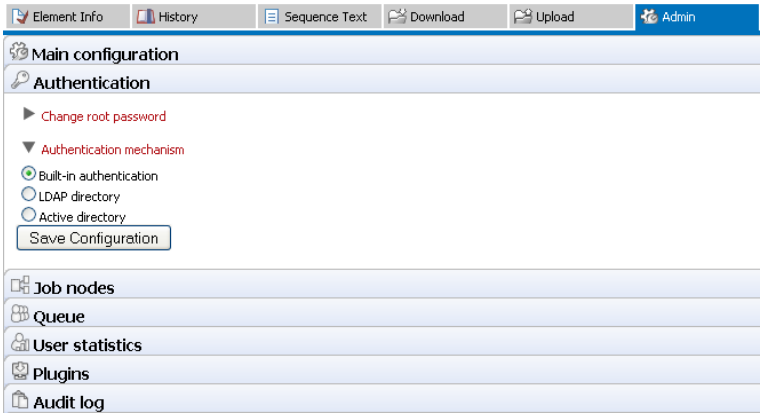

Figure 4.2: *Three modes of user authentication.*

The options are:

- **Built-in authentication**. This option will enable you to set up user authentication using the server's built-in user management system. This means that you create users, set passwords, assign users to groups and manage groups using the web interface (see section 4.2.1) or using the Workbench (see section 4.3.1). All the user information is stored on the server and is not accessible from other systems.
- **LDAP directory**. This option will allow you to use an existing LDAP directory. This means that all information needed during authentication and group memberships is retrieved from the LDAP directory. If needed, the LDAP integration can use Kerberos / GSSAPI.
- **Active directory**. This option will allow you to use an existing Active directory which is Microsoft's LDAP counterpart. This means that all information needed during authentication and group memberships is retrieved from the Active directory.

For the two last options, a settings panel will be revealed when the option is chosen, allowing you to specify the details of the integration.

Note that membership of an administrative group is used to control which users can access the admin part of the web interface. These users will also be able to set permissions on folders (see section 5). For the built-in authentication method, this means adding particular users to the built-in **admin** group. For Active Directory or LDAP, this means designating a group in the box labeled **Admin group name** and adding any users who should be administrators of the CLC Server to this group.

4.2.1 Managing users using the web interface

To create or remove users or change their password:

Admin | **Users and groups** | **Manage user accounts**

This will display the panel shown in figure 4.3.

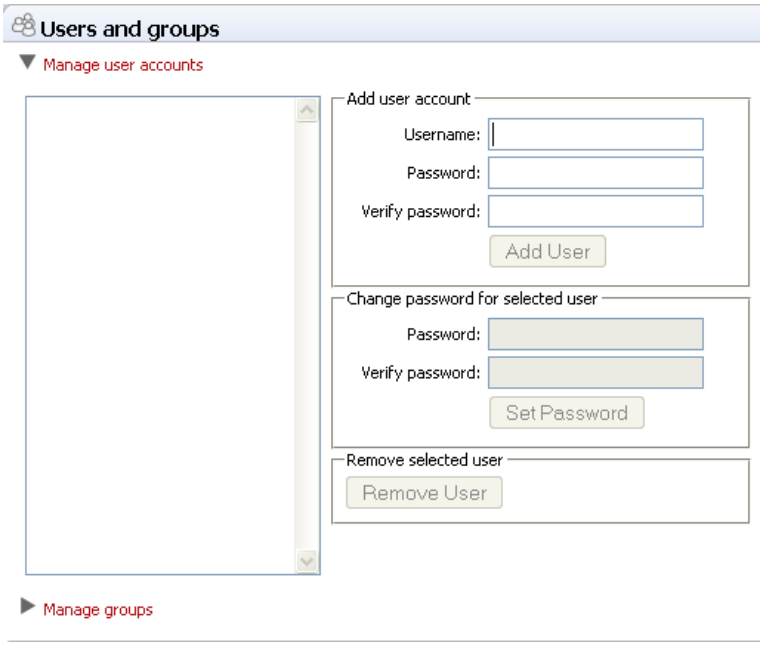

Figure 4.3: *Managing users.*

4.2.2 Managing groups using the web interface

To create or remove groups or change group membership for users:

Admin ( ) | Users and groups ( ) Manage groups

This will display the panel shown in figure 4.4.

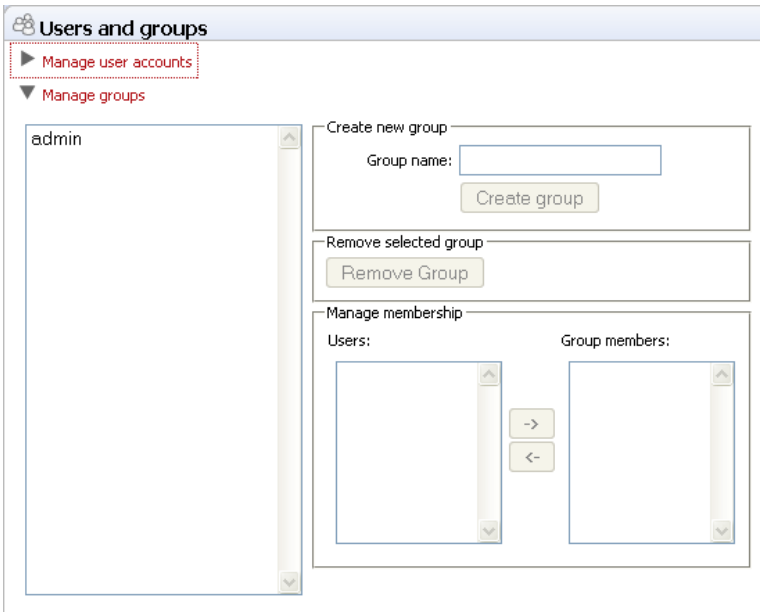

Figure 4.4: *Managing users.*

The same user can be a member of several groups.

Note that membership of the admin group is used for allowing users access to the admin part of the web interface. Users who should have access to the administrative part of the server should

be part of the "admin" group which is the only special group (this group is already created for you).

Note that you will always be able to log in as root with administrative access.

The functionality of this plug-in depends on the user authentication and management system: if the built-in system is used, all the functionality described below is relevant; if an external system is used for managing users and groups, the menus below will be disabled.

4.3 User authentication using the Workbench

Users and groups can also be managed through the Workbench (note that you need to set up the authentication mechanism as described in section 4.2):

File | Manage Users and Groups

This will display the dialog shown in figure 4.5.

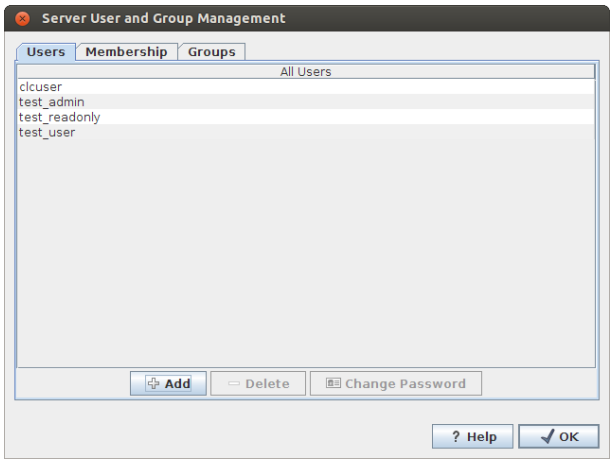

Figure 4.5: *Managing users.*

4.3.1 Managing users through the Workbench

Click the **Add** ( ) button to create a new user. Enter the name of the user and enter a password. You will be asked to re-type the password. If you wish to change the password at a later time, select the user in the list and click **Change password** ( ).

To delete a user, select the user in the list and click **Delete** ( ).

4.3.2 Managing groups through the Workbench

Access rights are granted to groups, not users, so a user has to be a member of one or more groups to get access to the data location. Here you can see how to add and remove groups, and next you will see how to add users to a group.

Adding and removing groups is done in the **Groups** tab (see figure 4.6).

To create a new group, click the **Add** ( ) button and enter the name of the group. To delete a group, select the group in the list and click the **Delete** ( ) button.

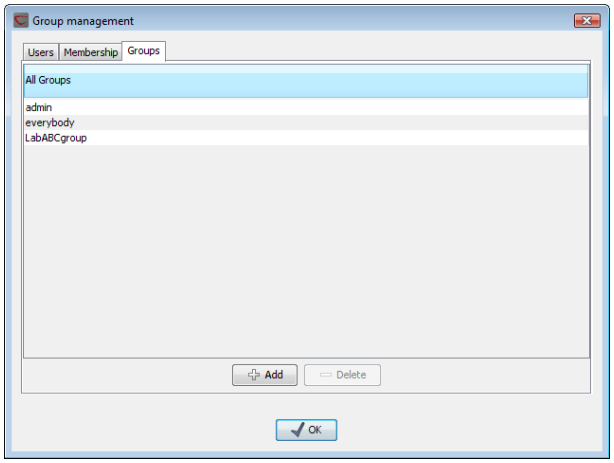

Figure 4.6: *Managing groups.*

4.3.3 Adding users to a group

When a new group is created, it is empty. To assign users to a group, click the **Membership** tab. In the **Selected group** box, you can choose among all the groups that have been created. When you select a group, you will see its members in the list below (see figure 4.7). To the left you see a list of all users.

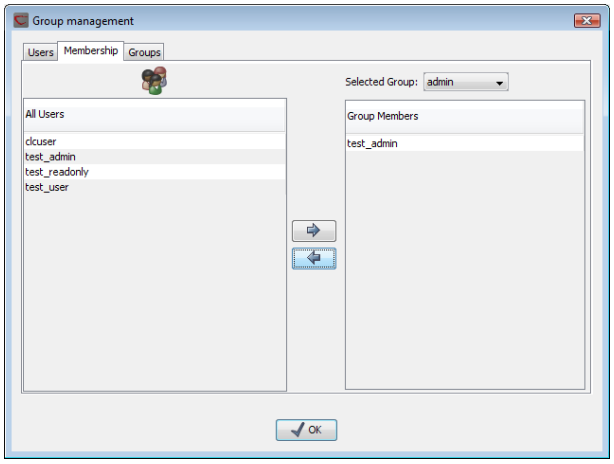

Figure 4.7: *Listing members of a group.*

To add or remove users from a group, click the **Add** ( ) or **Remove** ( ) buttons. To create new users, see section 4.3.1.

The same user can be a member of several groups.

4.4 User statistics

Clicking the **User statistics** panel will show a summary of the current usage of the server. An example is shown in figure 4.8.

You can see the number of users currently logged in, and you can see the number of sessions for each user. The two green dots indicate that this user is logged in twice (e.g. through the Workbench and through the web interface). The other two users have been logged in previously.

Figure 4.8: *The user statistics (user names have been blurred).*

You can also log users off by expanding the user sessions on the + sign and the click **Invalidate Session...**. This will open a confirmation dialog where you can also write a message to the user that will be displayed either in the Workbench or the browser.

Chapter 5

Access privileges and permissions

The *CLC Bioinformatics Database* allows server administrators to control access to the server on several levels:

- Access to the data in the server's **data locations**. This is typically to allow groups of users to store data that cannot be accessed by other users, or to establish reference data sources that are "read-only" for most users.
- Access to **running jobs** on the server. Particular groups of users can be restricted to running only particular types of jobs on the server.
- Access to the **import/export directories**. The server administrator can give users access to browsing designated directories on the server file system. Thus it can be desirable to restrict this access to certain groups of users for security reasons.

The following sections describe each of these levels.

5.1 Controlling access to data

The *CLC Bioinformatics Database* uses folders as the basic unit for controlling access to data, and access is granted (or denied) to groups of users.

On any folder within a location, you can grant two kinds of access to a group:

Read access This will make it possible for the users of the group to see the elements in this folder, to open them and to copy them. Access can be through any route, for example, browsing in the **Navigation Area**, searching or clicking "originates from" in the **History** () of e.g. an alignment.

Write access Changes to an element can be saved **Save** (), and new elements and subfolders can be created.

For a user to be able to access a folder, there has to be at least read access to all the top folders. In the example shown in figure 5.1, to access the *Sequences* folder, the user must have at least read access to both the *Example Data* and *Protein* folders.

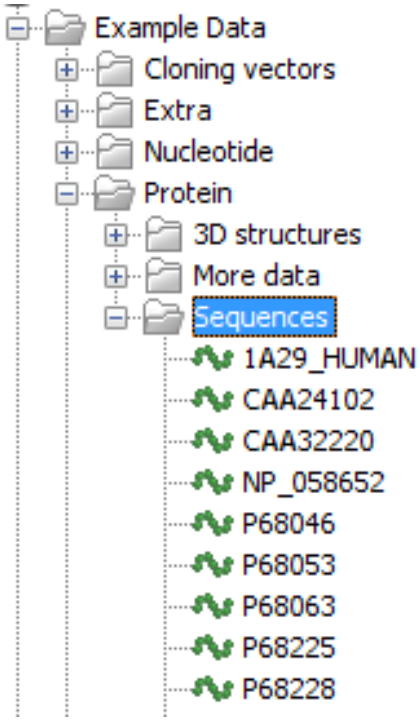

Figure 5.1: *A folder hierarchy on the server.*

However, you can grant read access to the *Example Data* and *Protein* folders and only grant write access to the *Sequences* folder.

Permissions on file system locations must be **explicitly enabled** if they are desired (see section 3.2.2). Please see 5.1.2 for further details about the system behaviour if permission are not enabled and configured.

If permissions are enabled on a file system location, then by default, no groups have read or write access to any area under this location until permissions are configured. Only the *CLC Bioinformatics Database* root user will have access to the data held in the server at this point. In other words, you must set permissions on folders in this file system location before any users will be able to read from or write to it.

5.1.1 Setting permissions on a folder

This step is done from within a CLC Workbench. Start up a copy of a Workbench that has the CLC Workbench Client plugin installed. From within the Workbench, go to the File menu and choose the item **CLC Server Login**. Log into the CLC Server as an administrative user.

You can then set permissions on folders in your database, if you have one, or on folders within file system locations that have had permissions enabled.

right-click the folder ( ) | Permissions

This will open the dialog shown in figure 5.2.

Set the relevant permissions for each of the groups and click **OK**.

If you wish to apply the permissions recursively, that is to all subfolders, check **Apply to all subfolders** in the dialog shown in figure 5.2. **Note** that this operation is only relevant if you wish

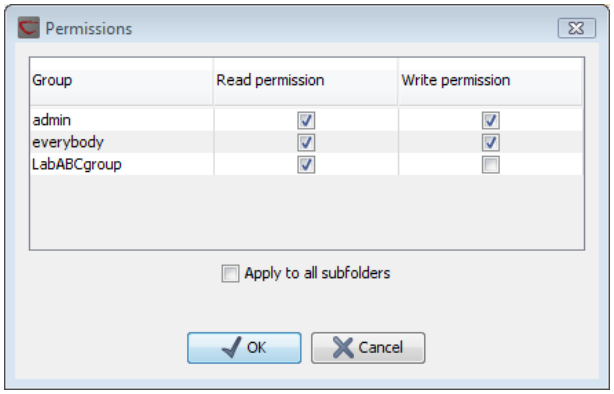

Figure 5.2: *Setting permissions on a folder.*

to clean-up the permission structure of the subfolders. **It should be applied with caution**, since it can potentially destroy valuable permission settings in the subfolder structure.

Permissions on the recycle bin

The recycle bin is conceptually a folder like any else. It is special in the sense that all users must have write access (otherwise they will not be able to delete anything). Because there is one recycle bin for the data from all users, you should be careful granting everybody read access to the recycle bin, since they will then be able to see the data deleted by other users. We recommend only granting read access to administrators.

Only administrators are allowed to empty the recycle bin.

5.1.2 Technical notes about permissions and security

All data stored in *CLC Bioinformatics Database* file system locations are owned by the user that runs the *CLC Bioinformatics Database* process. Changing the ownership of the files using standard system tools is not recommended and will usually lead to serious problems with data indexing and hamper your work on the *CLC Bioinformatics Database*.

One implication of the above ownership setup is that by default, (i.e. without permissions enabled), all users logging into the *CLC Bioinformatics Database* are able to access all data within that file system location, and write data to that file system locations. All files created within such a file system location are then also accessible to all users of the *CLC Bioinformatics Database*.

Group permissions on file system locations is an additional layer within the *CLC Bioinformatics Database*, and is not part of your operating system's permission system. This means that enabling permissions, and setting access restrictions on CLC file system locations only affects users accessing data through CLC tools (e.g.using a Workbench, the CLC Command Line Tools, the *CLC Bioinformatics Database* web interface or the Server API). If users have direct access to the data, using for example general system tools, the permissions set on the data in *CLC Bioinformatics Database* has no effect.

5.2 Global permissions

In the server web interface, in the **Admin tab** you can set global permissions for the *CLC Bioinformatics Database* as shown in figure 5.3.

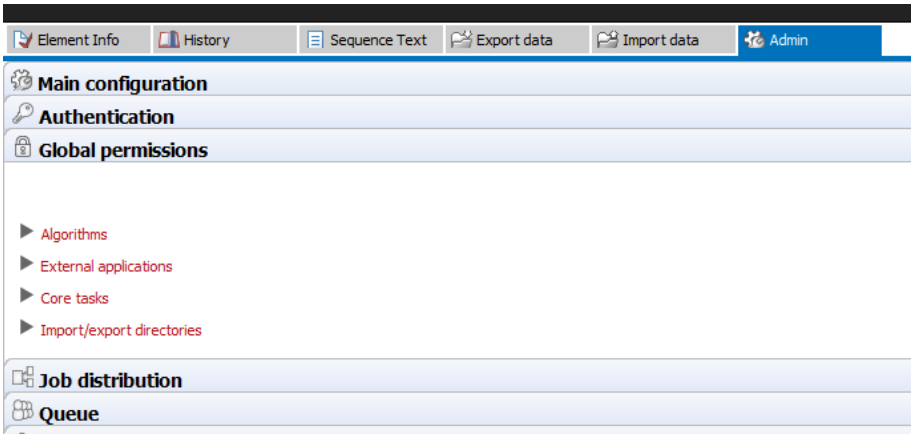

Figure 5.3: *Global permissions.*

Permissions can be set for:

- **Algorithms** All the algorithms running on the server are listed here.
- **External** applications. All the external applications configurations are listed here.
- **Core tasks** These are general import, export and maintenance tasks.
- **Import/export directories** Permissions can be set for each of the directories defined.

You can specify which groups should have access by clicking the **Edit Permissions** button. A dialog will appear like that in figure 5.4. If you choose **Only authorized users from selected groups**, you will be offered a list of groups that you can select (or de-select) to give (or take away) access to that functionality.

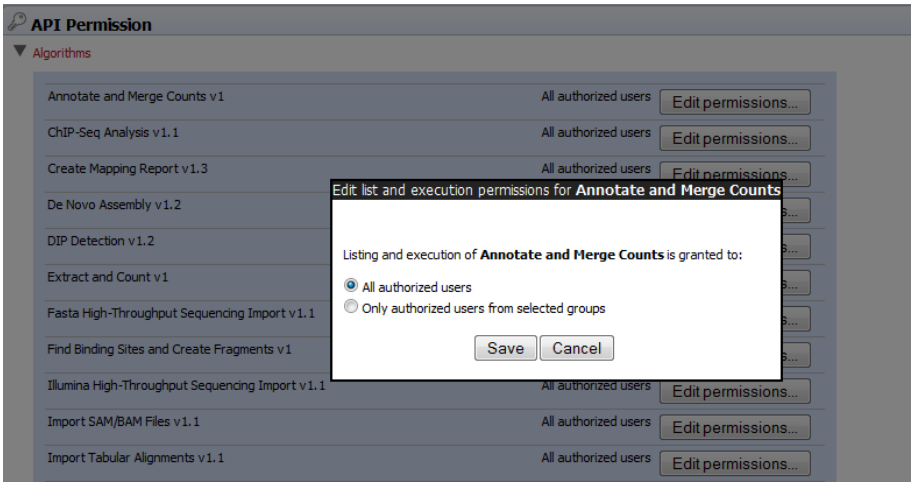

Figure 5.4: *Setting permissions for an alorithm.*

The default configuration is that all users have access to everything.

Chapter 6

Appendix

6.1 Troubleshooting

If there are problems regarding the installation and configuration of the server, please contact support@clcbio.com.

6.1.1 Check set-up

In order to check that your server has been set up correctly, you can run the **Check set-up** tool. Log in on the web interface of the server as an administrator and click the **Check Set-up** link at the upper right corner. This will show a dialog where you click **Generate Diagnostics Report**.

This will show a list of test that are performed on the system as shown in figure 6.1.

If any of the tests fail, it will be shown in the list. You can expand each of the tests to display more information about what the test is checking and information about the error if it fails.

6.1.2 Bug reporting

When contacting support@clcbio.com regarding problems on the server, you will often be asked for additional information about the server set-up etc. In this case, you can easily send the necessary information by submitting a bug report:

> **Log in to the web interface of the server as administrator | report a bug (at the top right corner) | Enter relevant information with as much detail as possible | Submit Bug Report to CLC bio**

You can see the bug report dialog in 6.2.

The bug report includes the following information:

- Log files
- A subset of the audit log showing the last events that happened on the server
- Configuration files of the server configuration

In a job node set-up you can include all this information from the job nodes as well by checking the **Include comprehensive job node info** checkbox in the **Advanced** part of the dialog.

Figure 6.1: *Check system. Failed elements will be marked with a red X. If you have not configured your Server to submit jobs to a local Grid system, or if you have and your setup is configured correctly, you will see a green checkmark beside the Grid setup status item in the diagnostic report.*

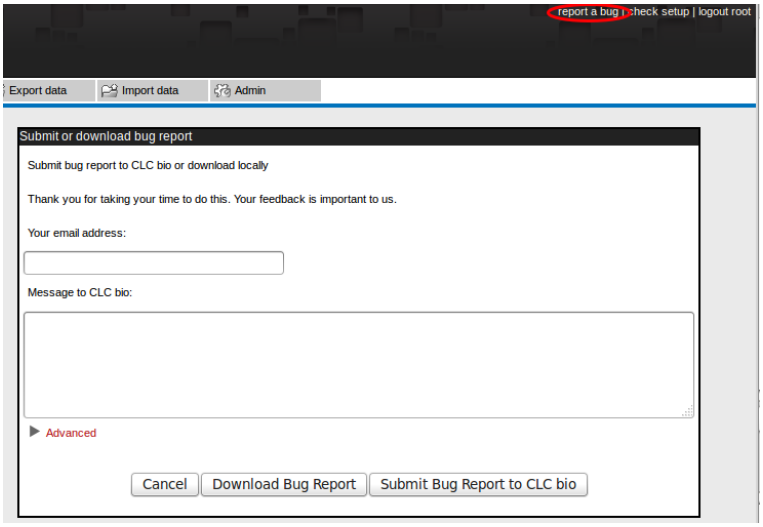

Figure 6.2: *Submitting a bug report to CLC bio.*

If the server does not have access to the internet, you can **Download bug report**. This will create a zip file containing all the information and you can pass that on to CLC bio support. If the server has access to the internet, you can **Submit Bug Report to CLC bio**.

Note that the process of gathering the information for the bug report can take a while, especially for job node set-ups. If a Workbench user experiences a server-related error, it is also possible to submit a bug report from the Workbench error dialog. This report will include the same archive as when submitting a bug report from the web interface. All data sent to support@clcbio.com is treated confidentially.

No password information is included in the bug report.

6.2 Database configurations

6.2.1 Configurations for MySQL

For MySQL we recommend basing your configuration on the example configuration file `my-large.cnf` which is included in the MySQL distribution.

In addition the following changes should be made:

The `max_allowed_packet` should be increased to allow transferring large binary objects to an from the database. This is done by setting the option: `max_allowed_packet = 64M`

InnoDB must be available and configured for the MySQL instance to work properly as the CLC Database. You should enable the options in the InnoDB section of your configuration as suggested below:

```
# You can set ..._buffer_pool_size up to 50 - 80 %
# of RAM but beware of setting memory usage too high
innodb_buffer_pool_size = 256M
innodb_additional_mem_pool_size = 20M
# Set ..._log_file_size to 25 % of buffer pool size
innodb_log_file_size = 64M
innodb_log_buffer_size = 8M
innodb_flush_log_at_trx_commit = 1
innodb_lock_wait_timeout = 50
```

There appears to be a bug in certain versions of MySQL which can cause the cleanup of the query cache to take a very long time (some time many hours). If you experience this you should disable the query log by setting the following option: `query_cache_size= 0`

6.3 SSL and encryption

The *CLC Bioinformatics Database* supports SSL communication between the server and its clients (e.g. Workbenches or the *CLC Server Command Line Tools*). This is particularly relevant if the server is accessible over the internet as well as on a local network.

The default configuration of the server does not use SSL.

6.3.1 Enabling SSL on the server

A **server certificate** is required before SSL can be enabled on the *CLC Bioinformatics Database*. This is usually obtained from a *Certificate Authority* (CA) like Thawte or Verisign (see http://en.wikipedia.org/wiki/Certificate_authorities).

A **signed certificate** in a `pkcs12` keystore file is also needed. The keystore file is either provided by the CA or it can be generated from the private key used to request the certificate and the signed-certificate file from the CA (see section 6.3.1).

Copy the keystore file to the conf subdirectory of the *CLC Bioinformatics Database* installation folder.

Next, the `server.xml` file in the `conf` subdirectory of the *CLC Bioinformatics Database* installation folder has to be edited to enable SSL-connections. Add text like the following text to the `server.xml` file:

```
<Connector port="8443" protocol="HTTP/1.1" SSLEnabled="true"
               maxThreads="150" scheme="https" secure="true"
               clientAuth="false" sslProtocol="TLS"
               keystoreFile="conf/keystore.pkcs12" keystorePass="tomcat"
               keystoreType="PKCS12"
/>
```

Replace `keystore.pkcs12` with the name of your keystore file, and replace `tomcat` with the password for your keystore.

The above settings make SSL available on port 8443. The standard (non-SSL) port would still be 7777, or whatever you may have configured it to. If only SSL connection should be allowed, the connector entry for port 7777 can safely be removed from the `server.xml` file.

Self-signed certificates can be generated if only connection encryption is needed. See `http://www.akadia.com/services/ssh_test_certificate.html` for further details.

Creating a PKCS12 keystore file

If the certificate is not supplied in a pkcs12 keystore file, it can be put into one by combining the private key and the signed certificate obtained from the CA by using *openssl*:

```
openssl pkcs12 -export -out keystore.pkcs12 -inkey private.key -in certificate.crt -name "tomcat"
```

This will take the private key from the file `private.key` and the signed certificate from `certificate.crt` and generate a pkcs12-store in the `keystore.pkcs12` file.

6.3.2 Logging in using SSL from the Workbench

When the Workbench connects to the *CLC Bioinformatics Database* it automatically detects if Secure Socket Layer (SSL) should be used or not.

If SSL is detected, the server's certificate will be verified and a warning is displayed if the certificate is not signed by a recognized Certificate Authority (CA) as shown in figure 6.3.

When such an "unknown" certificate has been accepted once, the warning will not appear again. It is necessary to log in again once the certificate has been accepted.

When logged into a server, information about the connection can be viewed by hovering the connection icon on the status-panel as shown in figure 6.4.

The icon is gray when the user is not logged in, and a pad lock is overlayed when the connection is encrypted via SSL.

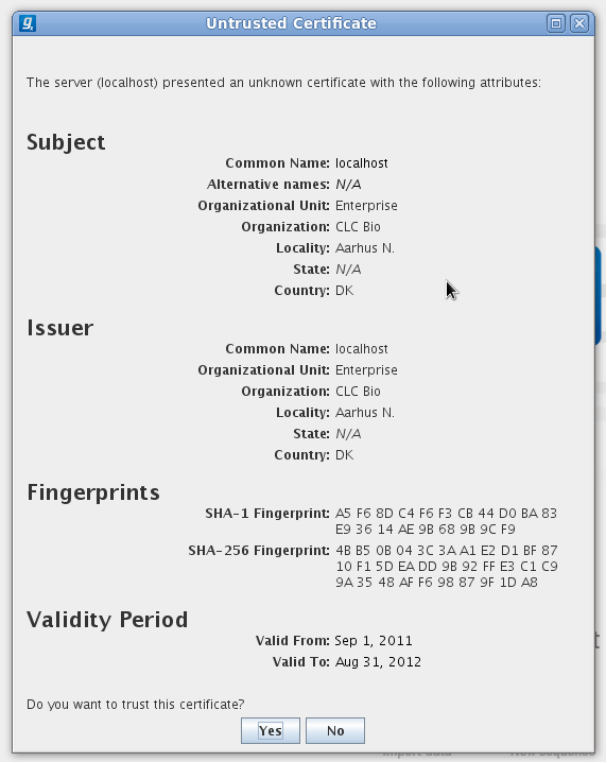

Figure 6.3: *A warning is shown when the certificate is not signed by a recognized CA.*

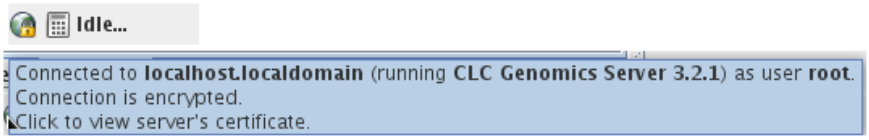

Figure 6.4: *Showing details on the server connection by placing the mouse on the globe.*

6.3.3 Enabling redirection from non-ssl port to ssl-enabled port

If SSL is to be mandatory, it is possible to get the non-ssl port (7777) to forward to the ssl port (8443), rather than closing down the non-ssl port. This is done by adding the following section to the `conf/web.xml` file:

```
<security-constraint>
<display-name>Security Constraint</display-name>
<web-resource-collection>
      <web-resource-name>SSL-restricted Area</web-resource-name>
      <url-pattern>/*</url-pattern>
 </web-resource-collection>
<user-data-constraint>
      <transport-guarantee>CONFIDENTIAL</transport-guarantee>
</user-data-constraint>
</security-constraint>
```

The above section is commented out in the supplied `conf/web.xml` file, and just has to be commented in in order for redirection to work. Please note that this redirection only works for the browser accessing the web interface. When connecting from the Workbench, the correct port has

to be specified.

6.3.4 Logging in using SSL from the *CLC Server Command Line Tools*

The *CLC Server Command Line Tools* will also automatically detect and use SSL if present on the port it connects to. If the certificate is untrusted the `clcserver` program will refuse to login:

```
./clcserver -S localhost -U root -W defaullt -P 7778
Message: Trying to log into server
Error: SSL Handshake failed. Check certificate.
Option                                  Description
------                                  -----------
-A <Command>                            Command to run. If not specified the list of commands on the server will be returned.
-C <Integer>                            Specify column width of help output.
-D <Boolean>                            Enable debug mode (default: false)
-G <Grid Preset value>                  Specify to execute on grid.
-H                                      Display general help.
-I <Algorithm Command>                  Get information about an algorithm
-O <File>                               Output file.
-P <Integer>                            Server port number. (default: 7777)
-Q <Boolean>                            Quiet mode. No progress output. (default: false)
-S <String>                             Server hostname or IP-address of the CLC Server.
-U <String>                             Valid username for logging on to the CLC Server
-V                                      Display version.
-W <String>                             Clear text password or domain specific password token.
```

In order to trust the certificate the `sslStore` tool must be used:

```
./sslStore -S localhost -U root -W defaullt -P 7778
The server (localhost) presented an untrusted certificate with the following attributes:
SUBJECT
=======
Common Name        : localhost
Alternative Names  : N/A
Organizational Unit: Enterprise
Organization       : CLC Bio
Locality           : Aarhus N.
State              : N/A
Country            : DK

ISSUER
=======
Common Name        : localhost
Organizational Unit: Enterprise
Organization       : CLC Bio
Locality           : Aarhus N.
State              : N/A
Country            : DK

FINGERPRINTS
============
SHA-1              : A5 F6 8D C4 F6 F3 CB 44 D0 BA 83 E9 36 14 AE 9B 68 9B 9C F9
SHA-256            : 4B B5 0B 04 3C 3A A1 E2 D1 BF 87 10 F1 5D EA DD 9B 92 FF E3 C1 C9 9A 35 48 AF F6 98 87 9F 1D A8

VALIDITY PERIOD
===============
Valid From         : Sep 1, 2011
Valid To           : Aug 31, 2012
Trust this certificate? [yn]
```

Once the certificate has been accepted, the `clcserver` program is allowed to connect to the server.

Bibliography

Index