A SOFTWARE ENVIRONMENT FOR SEQUENCE DATA

# Quick Reference Guide for ARB

v6.0.1, July 2016

(based on the ARB release 6.0 of June 2014)

The SILVA databases are recommended

for rRNA-based sequence analysis

---

This document is shared by **The ARB Project**:

***email:*** ***arb@arb-home.de***

***internet: www.arb-home.de***

***A quick reference guide for ARB***

After nearly 10 years of program development, around the millennium change, we finally started to compile some of our knowledge and wrote up a first version of something we called the ARB Reference Guide. Now, another 1.5 decades later, you hold in your hands the 6th edition of this guide. It is still not a comprehensive description of all the functions, buttons, and menus in ARB (and it will never be), but it should help to guide you through the program and explains the central functions of ARB required for your daily work and efficient handling of the software.
When you first open ARB, you will definitely be overwhelmed by the number of options offered. But don't worry, after some hours (or perhaps even days) you will see the light at the end of the tunnel: when you have managed the initial learning curve, you will find out that ARB speeds up sequence alignments, phylogenetic reconstructions, and probe design, tremendously, and offers a cosmos of functions and possibilities.

This version of the guide is based on the official ARB 6.0 release which was provided by the ARB project in June 2014. If you are using an older version, please be aware that the content of some menus has changed. ARB is under continuous development, so tools, menus and programs may change without notice. To get information about recent changes, please refer to the Change log file you can find at www.arb-home.de/downloads.html.
Official releases of ARB can be downloaded for free from www.arb-home.de/downloads.html. Please note that ARB is delivered with several tools and packages from third party authors. Users need to take care concerning the appropriate licence agreements. In general, this should be no problem for academic and non-commercial users. Please see arb_README.txt at www.arb-home.de/downloads.html for further information.

Comprehensive, quality checked and aligned datasets for small and large subunit (SSU & LSU) ribosomal RNA genes specifically designed for ARB are available from the SILVA database project at www.arb-silva.de. The databases include sequences from all three domains of life and are regularly updated based on selected ENA (European Nucleotide Archive) releases.

***General note: Following the reference guide does not guarantee 'publication quality' trees or perfectly working probes/primers. The guide is no substitute for a real understanding of the principles of phylogenetic tree reconstruction and probe design.***

***Acknowledgement***

Authors of the current version of the document are Jörg Peplies and Frank Oliver Glöckner (PI of the SILVA database project). Both have nearly two decades of experience in using ARB for phylogenetic tree reconstruction and probe design, and have organized and taught dozens of international ARB/SILVA workshops.

The authors would like to thank Wolfgang Ludwig (who has born the idea of ARB and acted for most of the software lifetime as its PI) for several years of intensive cooperation and the possibility to participate in the development of ARB, and Ralf Westram (primary ARB developer for more than a decade) for his unremitting patience and expertise with all the “problems” biologists have.

Many thanks also to Falk Warnecke (in memoriam, †2014) who has significantly contributed to the first version of this guide.

Bremen, July 2016

***Technical note***

Version 6.0.1 was compiled in March 2022, after converting the source of this guide to open document text format. There are no intended changes in content compared to the original version 6.0.

# Table of contents

# 1 Installing Linux

For using ARB you need to have a Unix-based operation system like Linux on your computer. Although a lot of people have asked for a Windows version during the years it will never exist. In the past, simply because Windows was not powerful, stable, and flexible enough to handle something like ARB, especially with ten thousands of sequence data entries. And meanwhile, to re-implement something like ARB in a sound way would cost millions of Euros, if possible at all (for Windows). Mac users can be happy, since Mac OSX is a Unix derivate and therefore ARB can be compiled and used on Macs without too much problems. Please remember, that the Mac-version is not officially supported by the developers of ARB, but created and maintained by the ARB-user community. To get more information please refer to the following webpage: http://www.arb-silva.de/documentation/arb-support.

To install Linux on your computer for ARB (productive work), you should have at least 20 GB of free hard disk space (around 6 GB is used by Linux, just 20 MB by ARB and the rest is required for your databases and the PT-server files (see 4.2). When you estimate hard disk space, take into account that the current comprehensive 16S ribosomal RNA reference data set for ARB comprises nearly 600,000 full length sequences (SILVA SSU Ref NR, release 123 from July 2015) needs about 0.6 GB of hard disk space. Database files on disk are always compressed and this results in additional 3.8 GB for every corresponding PT-server on your disk (never compressed). Please also note that ARB loads the database you open to work with plus the PT-servers you start in parallel completely to the main memory (RAM) of your computer (uncompressed)! By a simple calculation you will realize that the bottleneck for usage of ARB is represented by the RAM available in your computer, at least if you intend to work on the complete current rRNA data sets. A table which helps you to estimate your memory requirements can be found here: http://bugs.arb-home.de/wiki/SystemRequirements.

As a Linux platform we recommend to use CentOS or Ubuntu. Both Linux distributions can be downloaded for free from the internet (www.centos.org or www.ubuntu.com). They offer ISO images for download (large single files of about 2-4 GB) which can be used to easily burn e.g. an installation DVD or install it as a virtualized system (see VirtualBox software below). Nowadays, these installation DVDs are so-called “Live DVDs”. If you boot from DVD, a fully operational Linux system will be loaded to the RAM of your machine. It runs virtually only in your main memory, the harddisk will not be touched! With this, you can easily check if everything is working on your computer. Afterwards you just “copy” the whole system to your harddisk. An installation tool will guide you through the quick procedure and just asks some simple questions.

One additional important note: Before you download a Linux system for installation, you have to make another decision which is correlated to your RAM requirements. Operating systems are now available as 32bit and 64bit versions. With simple words, the difference between 32/64bit is the amount of RAM which can be addressed by the system. If you have a 32bit system installed, you can not use more than 4GB of RAM even if there are e.g. 8GB in your machine. A 64bit system has no limitations in this context (only defined by your hardware). With other words, for the large data sets you need a 64bit system. Of course, in such a case you also have to install the 64bit version of ARB on your system,

otherwise ARB can not use more than 4GB of RAM (actually it is just about 3.2 GB), In no case a 64bit ARB will run on a 32bit operating system. However, the need for this decision will dissolve soon because everything is developing into the 64bit direction.

If you already have Windows installed on your computer – no problem – most of the distributions allow you to shrink your Windows partition (for sure this is only possible in case your hard disk is not completely filled up with data) and install Linux on the same hard disk. After the successful installation of Linux the computer will always ask you while booting which operating system you like to start – this is called a dual-boot system. An (easier) alternative is represented by virtual systems which allow you to start e.g. a Linux system within (!) a running windows session. In this context, we recommend the free Oracle VirtualBox (www.virtualbox.org). Limitation here is that your hardware resources (processor cores, RAM) must be split for the two operating systems running "in parallel" on your computer, but for smaller ARB databases (such as the ARB type strain data set of the LTP project, see www.arb-silva.de/projects/living-tree) this is no problem at all, plus, providing powerful hardware is nowadays not a major issue anymore.

# 2 Getting started with Linux

## *2.1 Important commands*

**Login, Logout, …**

After starting your computer you should see a **login screen**: type in your user name ("login name") and password.

Remember that Linux – like all Unix based systems – is **case sensitive**.

How to **logout** depends on the desktop environment you are using. For example Ubuntu is available with Gnome and KDE desktop environments (the latter Ubuntu derivate is named Kubuntu). From e.g. the current Ubuntu (long-term support release 16.04) you can logout (or shutdown) by clicking on the rightmost symbol in the top panel of your desktop.

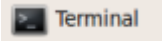
**Shells**

Nowadays, as soon as you get Linux installed, you get a nice graphical interface and rarely if ever need to make use of the so-called terminal mode (aka shell prompt).

However, in Linux the simple, modest terminal is not merely an afterthought, but an extremely powerful tool. While it may be true that you don't need to use it (except for installation and start of ARB), it's not that difficult to learn, and very useful to know. After you got some basics you will find out that several functions sometimes need many mouse clicks, but are only one simple command away.

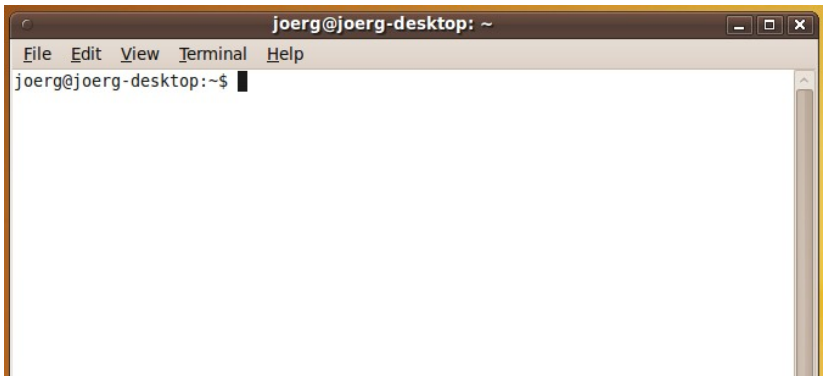
Linux has a variety of shells, which differ mostly in the commands they understand. Nevertheless, the basic commands shown here will be interpreted by all of them (hopefully).

## Talking to a Shell

The way to access the shell (terminal) depends on the Linux system and desktop environment you are using. In e.g. Gnome (Ubuntu 16.04) you should click the 'Search your computer' button on top of the left panel and type 'terminal'. By picking up the icon 

with your mouse you can also move it to the left panel for later easy access.

When you open the shell you then see a cursor (the blinking black box) that expects a command:



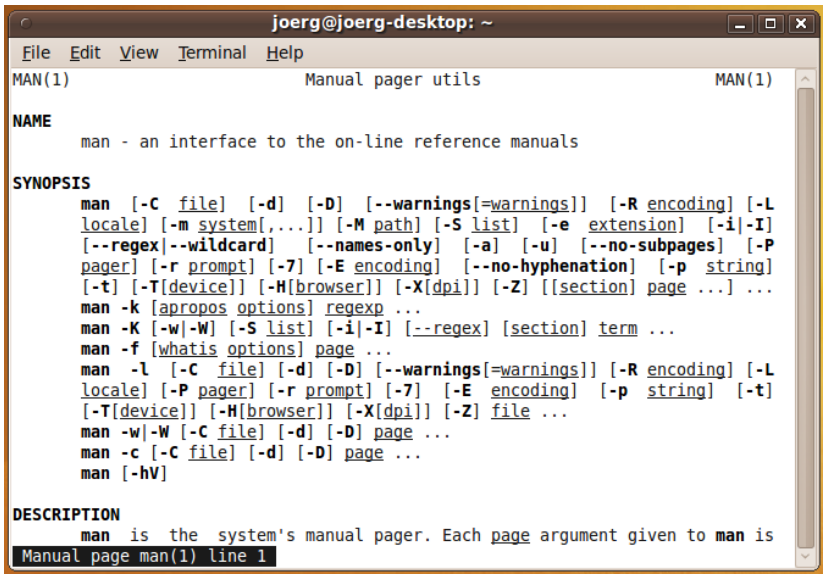

You type in a command line, hit enter, the command is executed and a new prompt signals “waiting for next input”. To logout you should type “exit” and press Enter, or else Ctrl-D

## Obtaining help

### man

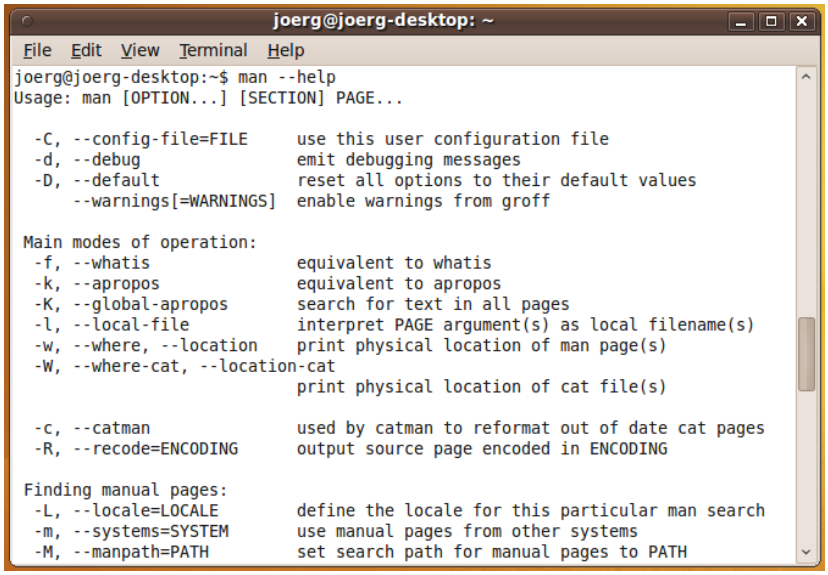
Almost every command in Linux has a help function available on the command line, through the "man" (manual) command.

Try it now to type in `"man man"` (manual of the 'man' command you are just using). The resulting page will describe the command, then describe every option, then give further details about the program, the author, and so on. This information is shown using the `"more"` command (which we'll describe later on). For now, it is sufficient to know that you can use the up and down arrow, PgUp and PgDn keys to move around, and the Q key to quit.




### The "--help" option

Most (but not all) programs have also a --help option which displays a very short description of its main options and parameters. Try typing "man --help" to see what's happening. This will produce more than one screenful of information, so you'll have to use the terminal's scrollbar to see what was displayed.




## Typing tricks

When you're in the shell prompt, you can use the up- and down-arrow keys to recall previously typed commands (there is a history stored by the system).

If you start typing a filename or directory name, you can press [Tab] and bash will complete the file or directory name for you, assuming that such a file exists and is the only one that starts with the typed-in part. For example, if you type "ls br[Tab]", bash will complete the filename to "brushtopbm", if this file exists and is the only file starting with "br".

## Directory commands

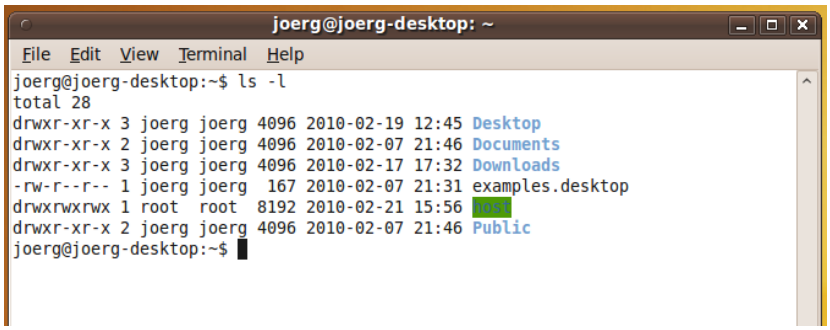
### ls

The "ls" (list) command lists the contents of the current directory. When used from a terminal, it generally uses colours to differentiate between directories, images, executable files etc. As you can see, the prompt reappears at the end.


Like practically all commands in Linux, you can add options to the "ls" command to alter its output or influence its behaviour. An option is preceded by a dash (e.g., "ls -l"). Try out the following variations of the "ls" command, to see different forms of output:

### ls -l

Produces a "long format" directory listing. For each file or directory, it also shows the owner, group, size, date modified and permissions




**`ls -a`**

Lists all the files in the directory, including hidden ones. In Linux, files that start with a period (.) are usually not shown.

**`ls -R`**

Lists the contents of each subdirectory, their subdirectories etc (recursive).

When you want to give more than one option, you can group them together with a single dash. For example, the command `"ls -al"` is the same as `"ls -a -l"`

Some options consist of a word (or words) instead of a letter, and have two dashes instead of one. For example, the command `"ls -l --full-time"` displays the date and time of modification in extended mode.

Finally, some options may also have a value. For example, `"ls -l --sort=size"` sorts the listing by size.

Apart from options (which are preceded by one or two dashes), you can also specify parameters, such as filenames, directory names and so on.

For example with the `"ls"` command, if you don't specify any parameter, it will list the contents of the current directory. However, you could instead give it a parameter specifying what to list. For example if you type in `"ls /Downloads"`, it will list the contents of the `"/Downloads"` directory. Always keep in mind that Linux is case sensitive. The command `"ls /downloads"` would provide no result in this example.

**`mkdir`** new-directory-name

Creates a new directory, "new-directory-name"

**`cd`** directory-name

Goes to the specified directory, making it the "current directory"

**`cd`** change directory

When you don't give a directory name, it goes to your `"home"` directory.

**`rmdir`** directory-name

Removes (deletes) the directory. As a safety measure, the directory must be empty before it can be deleted.

**`rmdir -R`** (recursive) directory-name will also delete nonempty directories

**`pwd`** print working directory

Displays the current directory.

The following sequence of commands (and results) demonstrates the above commands. After displaying the content of joerg's home directory, a new sub-directory called `"test"` is created and the content is listed again. Then we enter the new directory, print the working directory and afterwards display its content (empty), then we go back to the `"home"` directory, and display the current directory content. Finally the `"testing"` directory is removed and the content is listed again.


## File commands

**`cp`** filename1 filename2

**`cp`** filename1 filename2 filename2 (etc) directory

Copies a file, from filename1 to filename2 or (second form) copies one or more files into the specified directory. Warning: if the destination file already exists, it will be overwritten.

**`mv`** filename1 filename2

Renames a file, from filename1 to filename2. Warning: if the second file already exists, it will be overwritten.

**`mv`** filename1 filename2 filename2 (etc) directory

Moves one or more files into the specified directory. Warning: if the directory already contains files with the same names, they will be overwritten.

**`more`** filename

Displays the contents of the specified file onto the screen, allowing you to use the arrow keys, PgUp/PgDown etc to move around (like the `"man"` command).

**`find -name`** filename

Searches for a certain file or directory in the current directory

Another example: After displaying the content of joerg's home directory, the text file `"test_1"` is renamed to `"test_2"`, the "new" content of the folder is listed again and finally the content of the file `"test_2"` is displayed.


### Wildcards

Wherever you can specify a file or directory name in Linux, you can use wildcards. By using one or more special symbols, the shell will find those files which match a pattern, and place them on the command line instead of the pattern itself. The word `"wild card"` refers to the `"Joker"` in a pack of cards, since this card can stand for any other card in many card games. In the same way, the `"wildcard"` character can stand for other letters and characters in a filename.

### Testing Wildcards

To get the hang of wildcards, the best thing to do is to go to a directory which is full of files and try using the `"ls"` command with the wildcards as arguments. As we saw before, the `"ls"` command can take a parameter which tells it what to display. Instead of giving it a directory, we're going to pass it a list of all filenames to display. This list will come from the wildcard patterns which we will see below.

So, before you continue, in the terminal window type the command `"cd /usr/bin".` This will switch to the main directory containing the operating system commands. It's full of files, so it's ideal for our experiments.

The **`*`** wildcard

The first wildcard is the asterisk (`*`). The asterisk stands for zero or more other characters. By placing this wildcard at the beginning, middle or end of a pattern, you can build a pattern which

has the rest of the pattern at one or either end. For example the pattern `"*txt"` means any sequence of letters which ends with `"txt"`.

The **?** wildcard

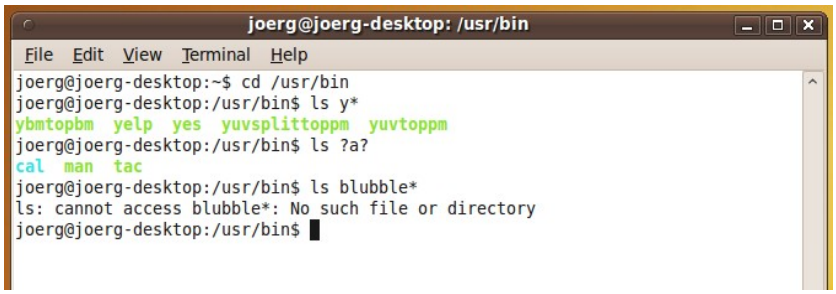
While the * wildcard could stand for zero or more letters or characters, the ? wildcard stands for exactly one. Thus, a pattern of `"???"` stands for filenames which are exactly three characters long. The pattern `"x??"` matches any three-letter filename which starts with `"x"`.

The **[]** wildcard

The square brackets are used to contain a set of characters to match. For example, the pattern `"[ABC]*"` matches any filename which starts in one of the letters A B or C, followed by zero or more characters.

If the first character is an exclamation mark (!) or caret (^), then the pattern matches any character except those given. Thus, the pattern `"[^x]*"` means any filename except those starting with `"x"`.

Instead of individual letters, the set can contain a range. For example, the pattern `"[A-Z]*"` means any filename which starts with an uppercase letter between A and Z inclusive, followed by zero or more other characters, while `"[A-Za-z123]"` means a single character which is an uppercase or lowercase letter, or the digits 1, 2 or 3.




### Wildcards with directories

Wildcards with Linux work on directories too. For example, the pattern `"*/file.txt"` means, all files called `"file.txt"` in any subdirectory.

### Hidden files

Wildcards will not match hidden files unless the wildcard pattern itself starts with a period. Thus, the pattern `".*"` matches all hidden files (hidden files are files which start with a period, such as .profile or .kde2)

### Permissions (Important!)

It is important to protect your files and directories against removal or alteration by yourself or others. Linux keeps track of who owns what file and who can do what to each file. Permissions determine who can use what file or directory. Every file or directory has three types of permissions:

**Read (r):** A user who has read permission for a file may look at its contents or make a copy of it. For a directory, read permission enables a user to find out what files are in that directory.

**Write (w):** A user who has write permission for a file can alter or remove the contents of that file. For a directory, a user can create and delete files in that directory.

**Execute (x):** A user who has execute permission for a file can cause the contents of that file to be executed (provided that the file itself is executable). For a directory, execute permission allows a user to change into that directory.

For each permission, there is a different class of users:

**User (u):** The user who owns the file or directory.

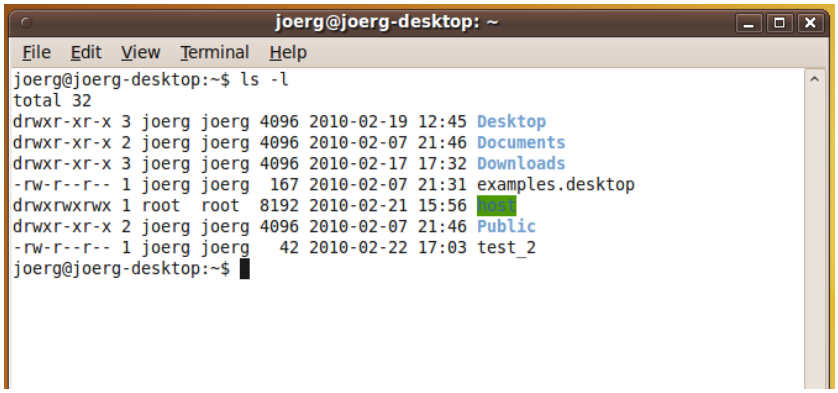
**Group (g):** Several users purposely lumped together so they can share access to each others files.

**Others (o):** The remainder of the authorized users of the system.

**All (a):** combines u, g and o. It sets the given permissions for all three.

As you may recall, the primary command that displays information about files and directories is **`ls -l`**

When doing this, the first column displays the permissions of each file.



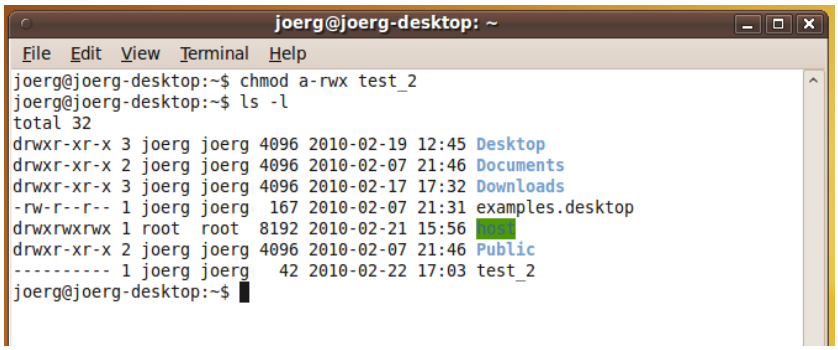

If the first character is a `d`, then the item listed is a directory. If the first character is a `-` then the item is a file. If it is a `l` then it is a link to another file. Characters 2 through 4 refer to the owner's permissions. Characters 5-7 refer to the group permissions. Characters 8-11 refer to the general public's permissions. If you type id at the prompt, you can verify your `userid` and group membership.

**chmod**

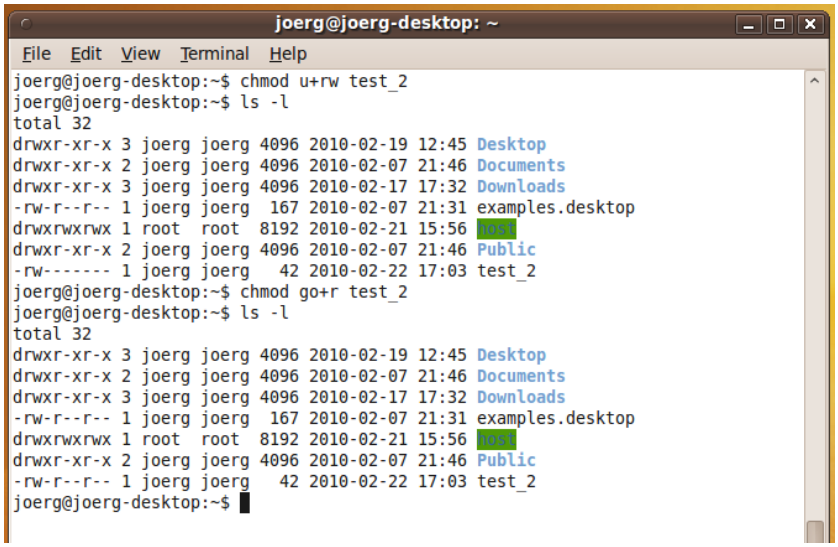
To change permissions on a file, you use the `chmod` (for change mode) command followed by the corresponding arguments. To change permissions, you use the -, + or = signs. These three symbols do the following:

| | |
|---|---|
| + | adds permissions |
| - | removes permissions |
| = | sets the specified permissions, removing the other preset permissions. |

For example, say I wanted to remove all permissions from a file `test_2`. I would type `chmod a-rwx test_2`. Let's break this down. We type `chmod` for change mode. We then type `a` for all classes. We then type `-` to subtract the given permissions. We then type `rwx` to tell it to subtract read, write and execute permissions. Finally, we type the filename.




Notice that the file now has no permissions. Now, say I wanted to allow that all users can read the file `test_2` and I should also be able to change (write) it.




This is the option if you want people to be able to view and use your files without changing them.

### Changing passwords

**`passwd`**

This command allows you to change your login password. You are prompted to enter your current password, and then prompted (twice) to enter your new password. On Linux systems passwords should exceed 6 characters in length, and contain at least one non-alphanumeric character (such as #, %, *, ^, [, or @ etc.).

### Working across networks

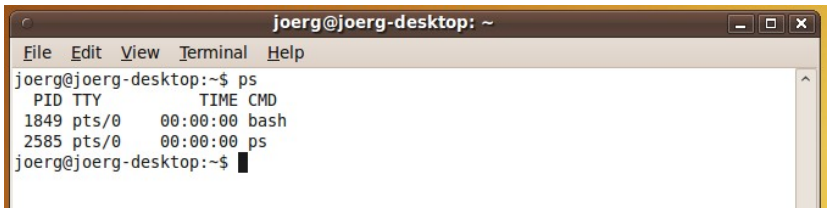
**`ssh`** (secure shell)

Allows you to connect to a remote computer via a secure connection (encrypted). Type in `ssh –X name@host` or `ssh –X –l` name host.

`-X` means that the graphical display will be redirected to your current display (X11 tunnelling).

### Finding processes and killing processes

In case a program “hangs” you can kill it without disturbing other processes on your machine. Type **ps** to see your current processes




To kill one of the processes type in `kill pid – e.g. kill 2476`

This will kill most processes, in case this does not work try `kill -9 pid`

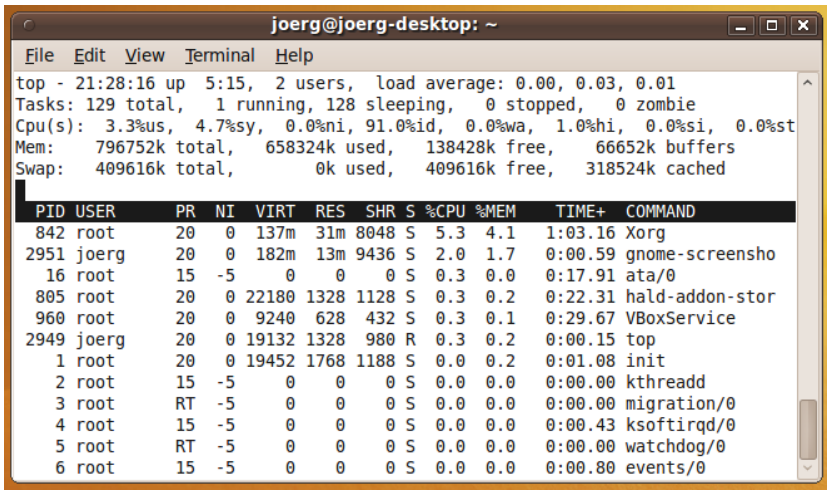
## *2.2 Useful tools*

### System Monitoring

**`top`**

Top is an important system monitoring program that gives you an overview about CPU, memory and swap usage and the currently running processes.

Start Top by typing in `top`:

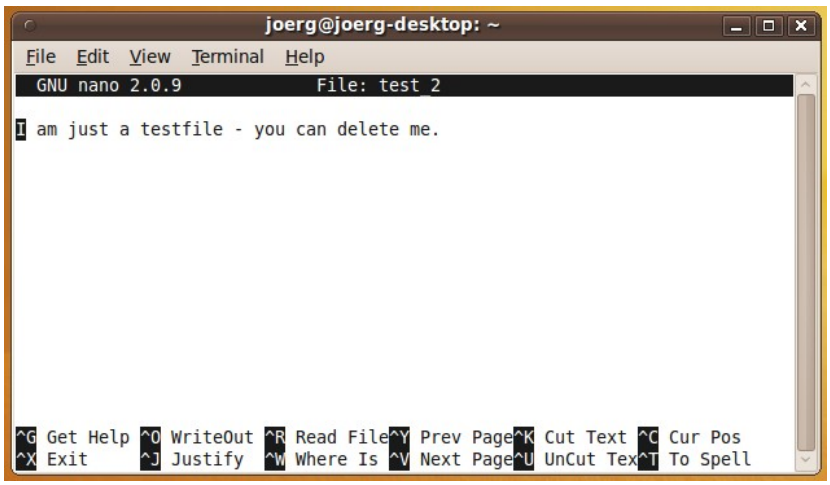



You can use the program for killing processes by typing `k` – the program will ask you to type in the PID to kill and with which signal. If you type in 9 the process will definitely be terminated.

### Text Editing

There are several reasons why you need a command line text editor. Examples are to alter configuration files, write a shell script or to read and edit a sequence file.

A common and simple editor for the Ubuntu shell is **`Nano`**. Type in **`nano`** filename to start it. If the file already exists, `nano` will open it, if not `nano` will create a new file. Now you can type in

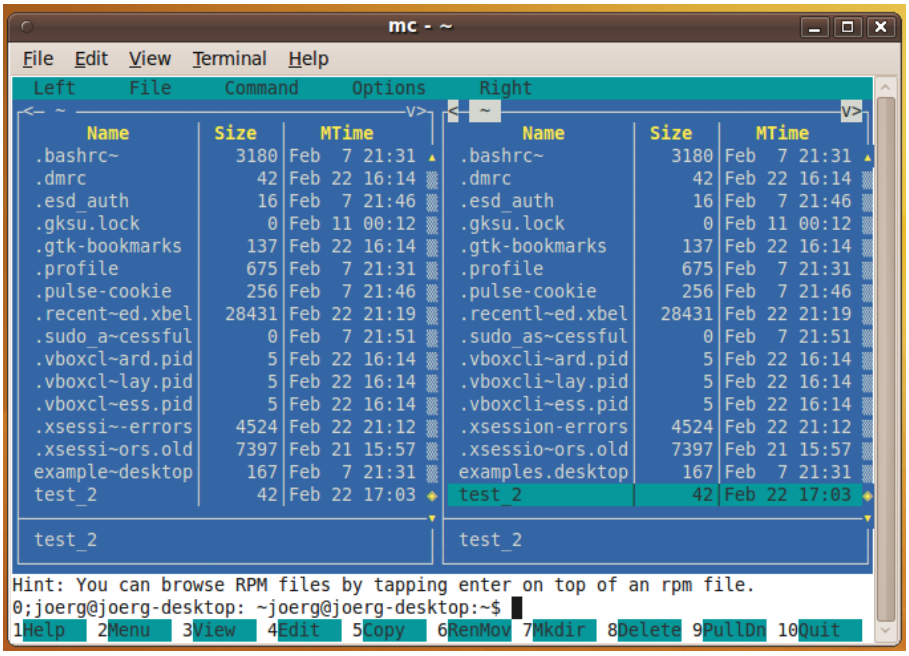



With `Ctrl +G` you will get help

With `Ctrl +X` you will quit without saving (and so on …)

### File Handling

There is a rather old but very powerful and intuitive program called **mc** (Midnight Commander) that can help you with all kind of file operations like copy, rename, edit, permissions etc.. For those who have worked on DOS machines before it is similar to Norton Commander. To start it type `mc`:




After starting you get a view on two directories and by using the cursor and tab buttons you can navigate in the files. By pressing insert you can select single or multiple files for operations with the function keys like copying (F5), moving (F6) etc. Additional options can be found in the pull down menus accessible with the F9 key.

Note: If the Midnight Commander is not installed (e.g. Ubuntu 16.04) type `sudo apt-get install mc` in the Shell for installation (you need an active internet connection). Other missing tools are installed accordingly (of course, you have to provide their short name instead of '`mc`').

If part of the F keys are not working, in e.g. Ubuntu 16.04 go to “Edit” -> “Keyboard Shortcuts …” in the Shell header and change the settings accordingly.

Parts of this manual have been taken from the following original websites:

http://www.sunyocc.edu/ir/linux/linux.htm, http://linux.org.mt/article/terminal

http://www.cs.montana.edu/courses/160/lectures/unix_linux_commands.html

# 3 Installing ARB

## Files needed to install ARB

| File | Comment |
|---|---|
| arb_install.sh | install script |
| arb.xxx.tgz | ARB program archive |

You can find them at www.arb-home.de/downloads.html under "ARB releases". Besides a number of documentation files, there are multiple program archives for the 64bit version of ARB 6, for different Linux distributions (CentOS and Ubuntu) and with and without (NoOPENGL) the rRNA 3D viewer. Choose one of them.

## Install/update ARB

ARB consists of more than 750 files which are installed into a single directory. Creating this directory, copying all data into it, and setting the permissions correctly are done by the installation script `arb_install.sh`.

Copy the downloaded files to a single directory, enter the directory and type `sh arb_install.sh`. If the script will tell you after step 1 (see below) that the ARB directory can not be created, make sure that you are logged in as the superuser/root and repeat the command. In case of an Ubuntu installation, type `sudo` before the command (this provides root privileges for a single command).

Answer all questions asked by the script.

***Steps:***

1. The script will ask you for the path where ARB should be installed

   **default: /usr/arb**

2. The script will ask about the PT_SERVER files location. This is a directory where ARB will store big index files. If possible set the path to a directory, where you have enough space left.

   If you just press enter, the PT_SERVER files will be placed within the ARB directory tree (default)

3. Next question: Who is responsible for the PT_SERVER index files?

   **The best suggestion is to say y**, then all users can build and rebuild PT_SERVERS

4. NameServer installation – trust users?

   **Again, trust your users and say y, if** not they will not be able to import sequences!!

5. Networking

   **In most cases: say s for standalone**

6. **Achieve further installations instructions:** Type in the number corresponding to the shell you are using (bash, csh/tcsh). The installation script will show you the commands you have to add to your shell configuration file. The standard shell in Linux is mainly bash – for details see Note.

**Note:** You can rerun the script many times, it can also be used to change an existing installation.

After the machine tells you:" Have much fun using ARB, ARB Team arb@arb-home.de", you have to make the changes shown as a result of question 6 in the config file for the shell you are using. You can do this either central in `/etc` or individually in the `.cshrc` for tcsh or `.bashrc/.profile` for bash in the home directories of the users.

Just copy the code provided after you have given one of the numbers 1-3 to the end of the corresponding config file and save it. Here, the options are listed again:

### The .cshrc you should add this:

```
setenv ARBHOME /usr/arb
setenv LD_LIBRARY_PATH $ARBHOME/lib
setenv PATH $ARBHOME/bin\:$PATH
```

### The .bashrc or .profile you should add this:

```
ARBHOME=/usr/arb;export ARBHOME
LD_LIBRARY_PATH=${ARBHOME}/lib:${LD_LIBRARY_PATH}
export LD_LIBRARY_PATH
PATH=${ARBHOME}/bin:${PATH}
export PATH
```

**reread the config-files, by opening a new shell or use `source .cshrc/.bashrc`**

Go to a directory where an ARB database is located named xyz.arb,

and **start 'ARB' by typing `arb`**

If you don't have a database you can create your own or download it from www.arb-silva.de.

### There are a number of additional packages required to run (all features) of ARB:

Here is a list of some formerly required packages, some of them are now included in the ARB installation archive. For up-to-date information check the arb_INSTALL.txt file in the download section.

| | |
|---|---|
| xfig | simple drawing program |
| transfig | used to print trees |
| fig2dev | used to print trees (normally part of transfig) |
| gv | previewing trees |
| complete xview | for gde-editor |
| X11 | because ARB is based on X11 |

If you have problems with the normal ARB version using OPENGL (required for the rRNA 3D introduced with the ARB release from 2007), try to install arb_noOPENGL.tgz instead.

### PT_server

When you work with ARB you have to know that some modules use a so called "PT_SERVER" (prefix tree server or positional tree server). For that mysterious thing ARB needs a writeable directory to store the PT_SERVER files (see question 2 in the ARB installation procedure).

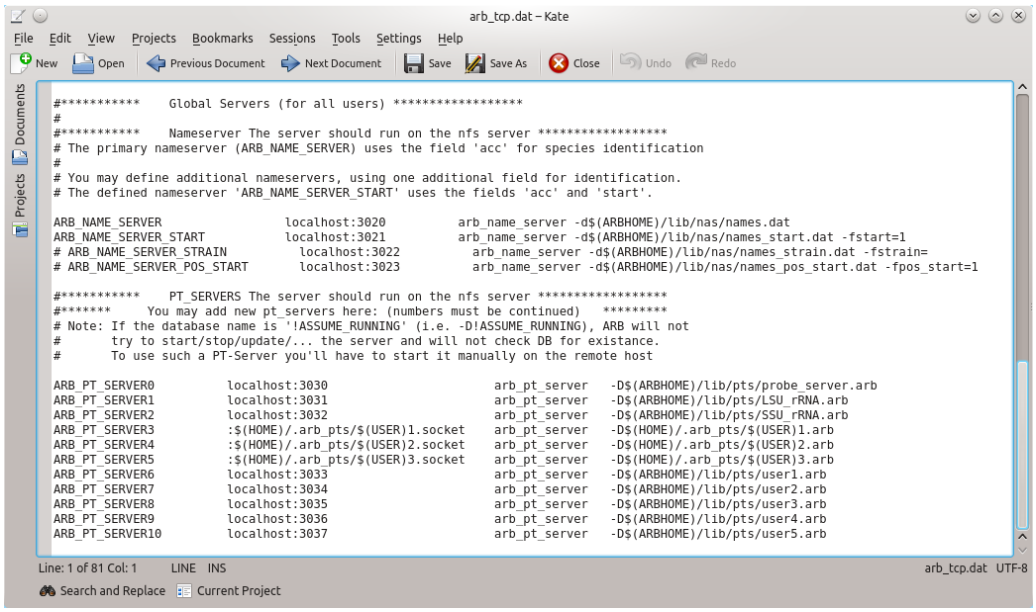
Those files are needed for fast database searches by probe_design, probe_match and the automatic aligner, and can require a lot of disc space, up to several Gigabytes depending on the amount of sequence data in your local ARB database (see chapter 1).

The files are not created within the installation procedure, but later on by going to the **ARB_MAIN window -> Probes -> PT_SERVER Admin -> Build server**. This create/rebuild procedure might take some time, depending on the amount of sequences and the machine (amount of RAM) you are using. You may define a special directory for the PT_SERVER files location, which will prevent loss of servers when installing a new version of ARB.

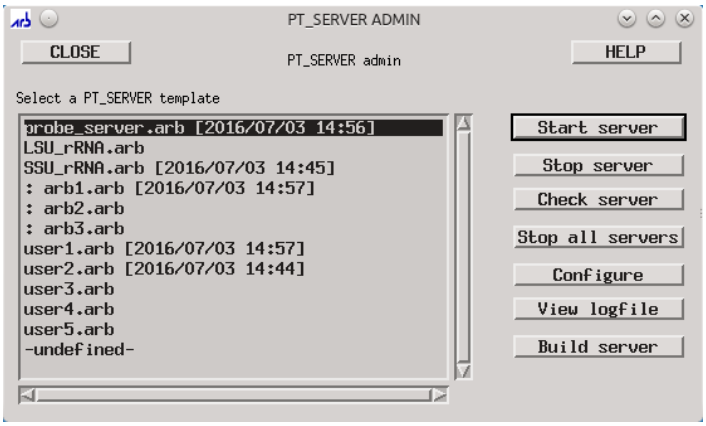
If you are working on a workstation cluster, you can define a central location where all PT_SERVERS are stored and mount it on your local host. All users will than have the same PT_SERVERS on all machines, and the update procedure has to be done only once.

The important configuration file for the PT_SERVER templates is located in the `"ARBHOME"/lib` directory and called **arb_tcp.dat**. It is a simple text file, which can be edited by any kind of texteditor (e.g. nano). More information about the use of the PT-servers can be found in chapter 4.2. Information how to modify the arb_tcp.dat file can be found by opening the arb_tcp.dat file.

**The default arb.tcp.dat file (just a section)**




**Corresponding PT_server admin menu in ARB:**




List of available PT-servers assigned to different functions and users (names can be changed)

# 4 ARB's main functions and windows

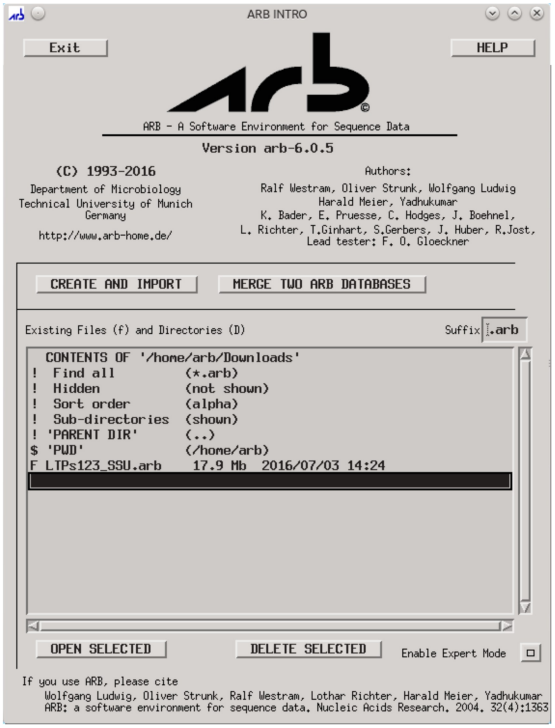
## *4.1 ARB Database*

Computational databases store and organise large amounts of data to make them easily accessible to the user. ARB databases contain sequence data (nucleic acid or protein) and associated information (sequence annotation, source, author etc). The ARB software package is most frequently used to infer phylogenetic trees from ribosomal RNA gene sequence data. However, ARB has been greatly expanded over the past several years, and can now be used for phylogenetic studies of the gene or protein sequences of functional genes as well.

ARB database files are characterised by the suffix .arb (e.g. `database.arb`). One database file can contain thousands of sequences, several alignments, and a number of different trees.

### *4.1.1 Starting ARB*

Start ARB by typing in `arb` in the Shell window (this will work if you have done the installation according to the instructions given in chapter 3. In case this does not work, ask your system administrator for help). You will get the following ARB `Intro` window:

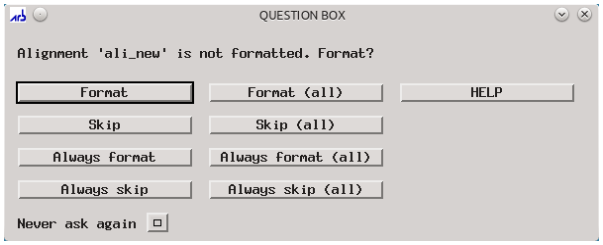



The "ARB_Intro" window

Select the database you want to work with. In case you do not have a database in your home directory you can create your own by clicking on the `Create and Import` button (the sequence import window of ARB will pop up – see 5.1) or download a database from www.arb-silva.de.

After you have selected the database you want to work with click on `OPEN SELECTED` – the main window of ARB (`ARB_MAIN`) will come up.

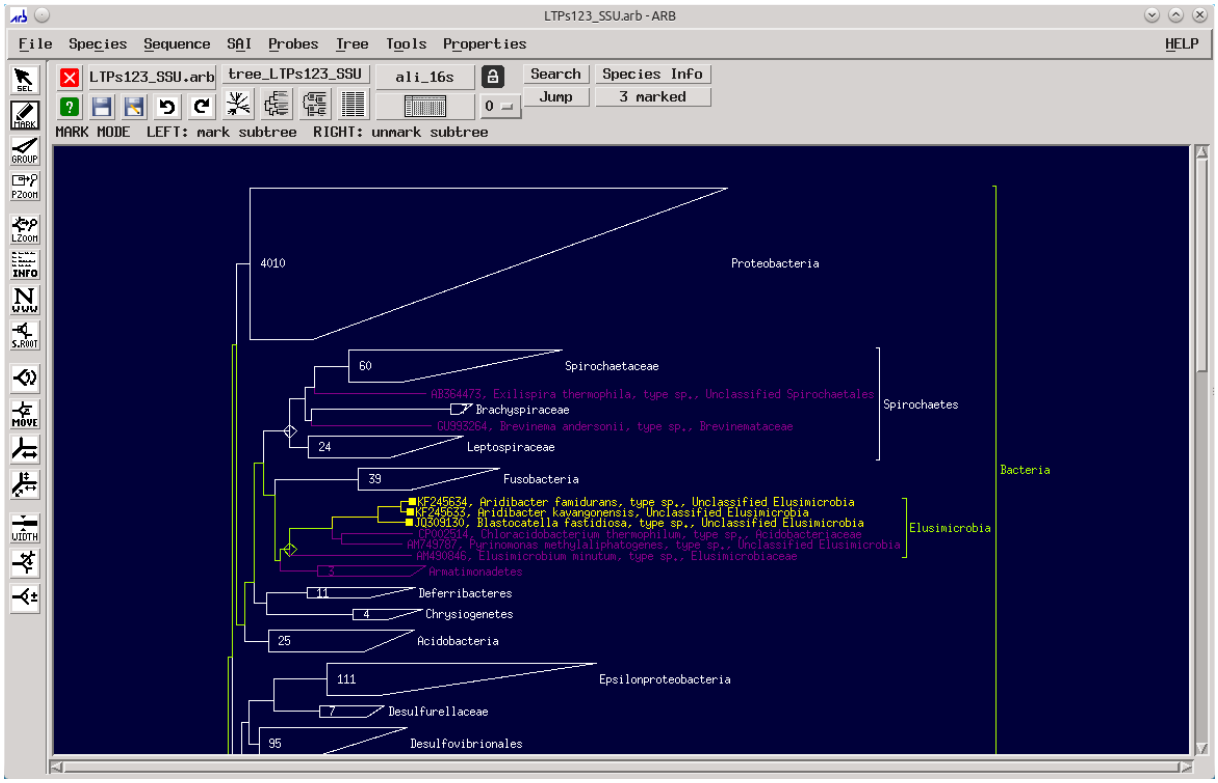
If you get a window asking you "`Alignment 'ali_xxx' is not formatted. Format?`" click on "`Always format (all)`"




The "Format" window

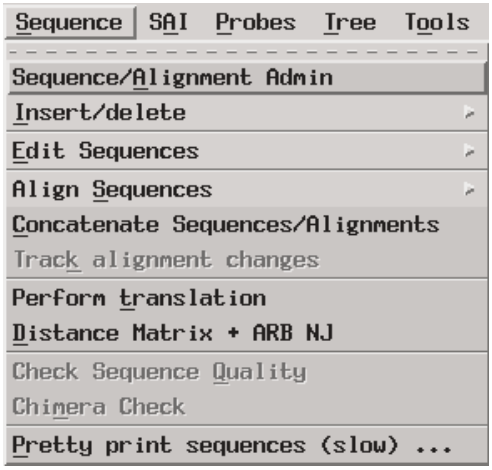
### *4.1.2 Main window ARB_MAIN*

The main ARB window is the graphical interface to the database, and displays one of the trees in your ARB database file. The fields and buttons in the horizontal and vertical menu bars allow you to obtain information from your database, change the appearance of the current tree, and perform other functions which are mostly also available through the pull-down menus.




ARB_MAIN window with standard tree view

***Note: With the ARB release 5.0 some of the common options in the pull-down menus of the ARB main and other windows as well as various selection boxes turned grey and inaccessible (see example on the right). This is a feature, indicating that these options are either not implemented, supported or their proper function can not be guaranteed. In few cases, they also represent fully functional but "dangerous" options for not highly experienced ARB users. If you however would like to access these options the ARB mode can be changed as follows:*** ARB_MAIN window → `Properties` → `Toggle expert mode`



**Horizontal menu bar**

First row:

| | |
|---|---|
| | Close ARB |
| LTPs123_SSU.arb | Current database |
| tree_LTPs123_SSU | Current tree displayed |
| ali_16s | Current alignment |
| Search | Starts the central database management tool: `Search` and `Query` (see 4.1.4) |
| Species Info | Selected species (here: no species selected) |

Second and third row:

| | |
|---|---|
| | Help function |
| | Save database/Save database as |
| | Undo/redo last action |

The next four buttons allow you to change the appearance of the current tree:

| | |
|---|---|
| | Radial tree |
| | Dendrogram style |
| | Hierarchical tree view; helps you to navigate in the tree by showing full taxonomic (group) information of current region always on top |
| | A single click on the button shows you a list of all species in the tree; by clicking again you will get a list of the marked species |
| | Starts the ARB Editor `ARB_EDIT4` (see 6) |
| 0 | Displays the current protection level for the tree |
| Jump | Jumps to the selected species in the tree |
| 3 marked | Displays the number of currently marked species in your ARB database |

**Vertical menu bar**

***Note: The functions of these buttons are explained briefly in the `ARB_MAIN` window underneath the horizontal menu bar (see screenshots of `ARB_MAIN` window above).***


Use this button to select one sequence in the tree → a small square will appear in front of this sequence and the name will be displayed in `Species Info` in the horizontal menu bar


Use this button to mark/unmark one or more sequences in the tree → marked sequences will change colour


Use this button to fold/unfold groups and to create/rename/destroy groups in the tree


Physical Zoom: Use this button to zoom in/out of the tree: press left mouse button and drag to zoom in; use right mouse button to zoom out stepwise


Logical zoom: Click on a node in a tree to hide all sequences or branches which are currently not of interest for you. No information gets lost!! To reset go to `ARB_MAIN` → `Tree` → `Reset zoom` → `Logical zoom`


Opens the `Species information` window of the selected species


Opens a browser window (like Mozilla Firefox), connects to a public database (EMBL/ GenBank) and shows the original entry for the selected species. Settings can be found at `ARB_MAIN` → `Properties` → `Search World Wide Web (WWW)`

***Note: The following tools can be used to change the appearance of the tree, but do not influence the topology and phylogenetic information behind it (even if you think so!):***


"Set root" button: defines a new virtual root for the current tree

Swaps branches


Changes branch widths

Rotates branches

Increases/decreases the angles in a radial tree

***Note: The following tools change the topology of the tree, and therefore the phylogenetic information it contains. They are only recommended for consensus tree reconstructions and experienced users:***


Moves branches

Changes branch lengths

Introduction of multifurcations to indicate regions of uncertain topology

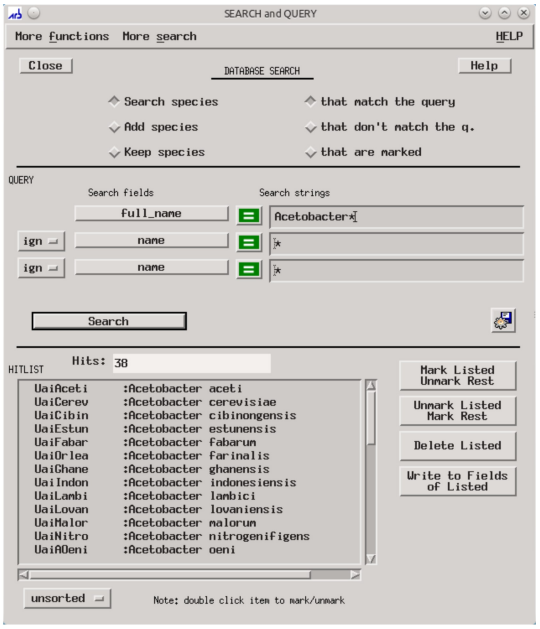
### *4.1.3 ARB database fields*

ARB database fields, a few original examples (ARB databases provided by SILVA have own fields indicated by '_slv' at the end of the field name):

| Database field | Content |
|---|---|
| name[1] | unique identifier of the ARB database entry!! (created by ARB) |
| full_name[2] | name of the sequence/species; this you can edit manually |
| acc[2] | public or ARB internal (own sequences) accession number |
| ali_16S/data | (aligned) sequence data |
| aligned | suggestion: for own seq. fill in your name and the date when you have done the alignment (SILVA uses fields `align_xxx_slv`) |
| ambig | ambiguities calculated by ARB using `count ambiguities` (SILVA uses field `ambig_slv`) |
| ARB_color | stores the information about sequence colors |
| nuc | number of nucleotides; calculated by ARB using `count nucleotides` |
| nuc_term | number of nucleotides coding for the respective rRNA gene;<br>calculated by count nucleotide gene (SILVA uses field nuc_gene_slv) |
| remark | field for your personal remarks |
| tmp | used by various ARB modules |

***[1]Note: Do not edit or change this field manually! Use the ARB function 'Synchronize IDs' instead (see chapter 9.1). Consistency is automatically maintained if you choose 'Generate unique species IDs' in the import procedure, or use the 'Synchronize IDs' function later on (see 5.1 for more information).***
***[2]Note: These fields will automatically be filled/overwritten with information when importing data from e.g. EMBL/GenBank.***
***More information about database fields in SILVA can be found at [https://www.arb-silva.de/documentation/faqs/](https://www.arb-silva.de/documentation/faqs/)***

### *4.1.4 Search and Query*

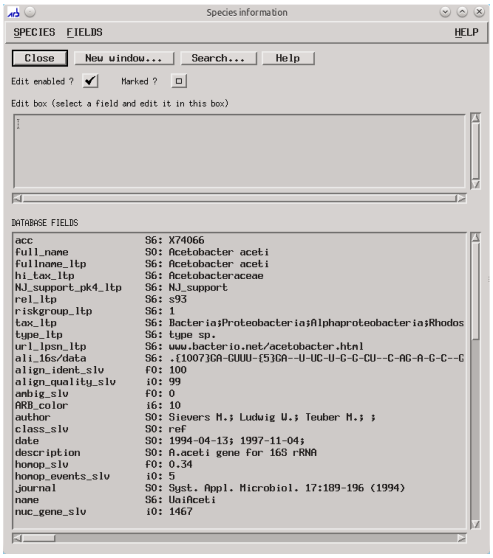



`SEARCH and QUERY` is the central tool for ARB database management.

- `ARB_MAIN` window → `Species` → `Search and Query` or 

- To search for sequences or associated information, type your query in the field `Search strings`. You can add wildcards (*) to either side of your query. ARB will not find entries with text before or after the query unless you include wildcards or you enter the exact match.

- By adjusting the buttons in the field above you can `Add` to or `Keep` `species` in your list, or search for everything that does not match your query.
- By clicking on the  sign, you can search for all entries in a certain field which are unequal to your search string.
- With the buttons on the right side of the Hitlist you can mark or unmark all listed sequences, delete them, or write/add information to a certain field of all of them.




***Note: You can assign different states to the database entries/sequences:***

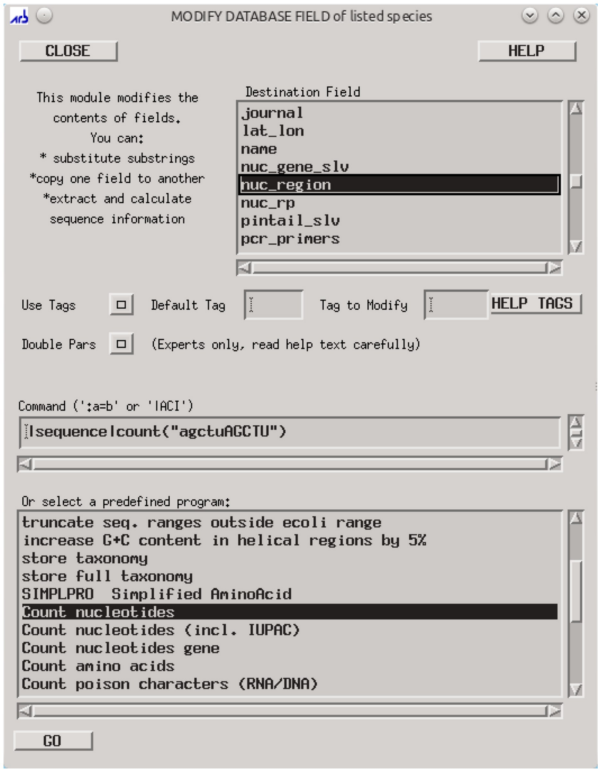
| | |
|---|---|
| Marked | = with asterisk in `Search` and `Query` list |
| Selected | = highlighted black in `Search` and `Query` list |
| Listed | = all entries in the `Search` and `Query` list |

- If you click on one of the listed species, a window will pop up with all corresponding species information available in the ARB database (see picture on the right).

### *4.1.5 Modifying fields of listed species*

Within the `SEARCH` and `QUERY` tool, ARB offers a set of possibilities for quick and easy data modification in batch mode.

- Open the MODIFY DATABASE FIELD of the listed species window: `ARB_MAIN` window → `Species` → `Search and Query` → `More functions` → `Modify Fields of Listed Species`
- Within this tool you can use either predefined programs (e.g. count the number of nucleotides in your sequence, or copy information from one field to another) or write your own application using the ACI (ARB command interpreter) or SRT (Search and replace tool) language. A detailed description can be found in the ARB online help (press HELP). An example for counting all nucleotides of a sequence in the range of the 16S rRNA gene is shown in the screenshot.

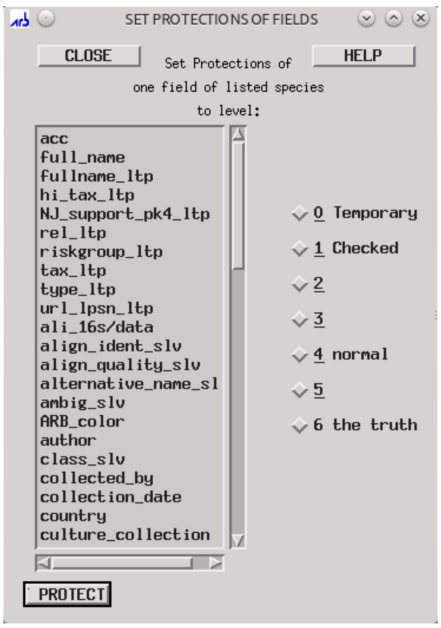


After clicking on the `GO` button, ARB will count the nucleotides of all listed species and write the numbers to the `nuc_region` field.

### *4.1.6 Protection of database fields*

ARB allows you to assign a distinct protection level to each database field. This helps you in database management and to keep the database consistent. Although you can assign a different protection level to every field, practically, you **will in most cases only want to protect the field where the sequence data is stored**. This field is called **ali_xxx/data**, in case you have a 16S alignment: ali_16s/data. The sequences in the Reference databases delivered by the ARB/SILVA project (www.arb-silva.de) have normally protection level 3, which means, that the alignment has been checked by the SILVA quality management system. If you import your own sequences either from a sequencer or from a public database like EBI or Genbank you should import them with level 0 (see 5.1). Since your default working protection level in the ARB_MAIN main window or ARB_EDIT4 is always 0 you can change or modify the sequences without raising your working protection level (to e.g. edit the SILVA alignment, you have to set it to 3 or higher). This makes life easier when you move along the workflow from import to alignment, add species to tree and the refinement of the alignment. As soon as you have worked on the sequences and alignments and think that your result is worthwhile for protection go to the `SEARCH and QUERY` tool and select the sequences you would like to protect and:

- Open the Set Protection of Fields window: `ARB_MAIN` window → `Species` → `Search and Query` → `More functions` → `Set Protection of Fields of Listed Species`
- On the left side you can choose the database field to protect and on the right side the protection level you would like to assign.

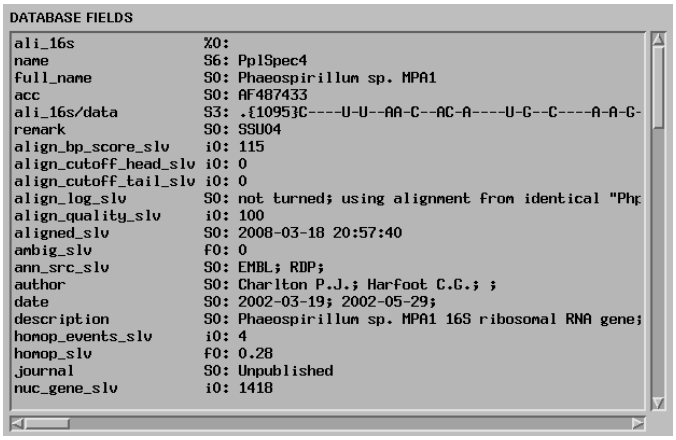



Its up to you how much protection levels you would like to use. Normally 0, 1 (for e.g. a first manually refined, good alignment) and 5 for the final "perfect" alignment are a good choice.

If you have assigned a protection level to your sequences it will be taken into account by all programs in ARB that have the ability to alter or delete sequence information. An example is the `SEARCH and QUERY` tool itself when you e.g. try to use the button `Delete Listed` (see 4.1.4) on sequences with a higher protection level as it is set by default in the ARB_MAIN window  you will get an error message and ARB will not delete them. Only by raising the protection level in the ARB_MAIN window to a value equal or higher than the level of the sequences will allow this operation.

The same accounts for ARB_EDIT4. There you are even able to have different protection levels for the EDIT and ALIGN modus. This is useful in case you want to allow yourself to realign sequences or parts of them which have already been curated (level 5) but on the other hand you want to prevent that bases are changed. The current protection level of the sequences is always shown in the Editor in the field **Xdata** like **5data** for protection level 5 see 6.1 for a screenshot of ARB_EDIT4. Protection levels are also taken into account by the `Fast_Aligner` (see 6.2). By default only sequences with the protection level 0 will be aligned. Sequences with a protection level >0 will not be touched.


As already mentioned, ARB allows you to assign different protection levels to all database fields. This can be used to prevent you from accidentally changing information when working with the database. The current protection levels can bee seen in the `Species Information` field. The SX shows always the respective protection level for the different fields. In the Reference databases released by the SILVA project, by default the field name has always a protection level of S6, the sequence data (ali_16s/data) has S3 and the rest S0 (compare picture on the right which shows part of the Species information window).



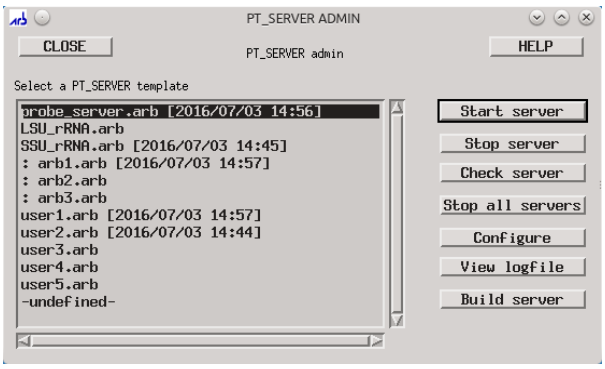

Taken together the protection of database fields in combination with protection levels is a powerful system to prevent you from accidentally altering sequences or alignments when you are working with sequences which have different levels of curation.

### *4.2 PT_Server (Positional Tree Server = Suffix Tree Server)*

Note: The `PT_Server` represents a different (indexed) format of your database, which is necessary for faster performance of sequence search functions within ARB. It is used by the `Fast aligner`, `Probe Design` and `Probe Match` tools, and to search for the closest relatives of a sequence in `Search and Query`.

***Note: the PT_Server has to be build/rebuild independently of your database; saving your ARB database does not affect your PT_Server!!***

***PT_Server build/rebuild***




`PT_Servers` are not delivered together with ARB. Before you can use them for the first time you have to do a `PT_Server` "build".

`ARB_MAIN` window → `Probes` → `PT_SERVER admin` … → Select the name (template) of your `PT_Server` → Click on `Build server`

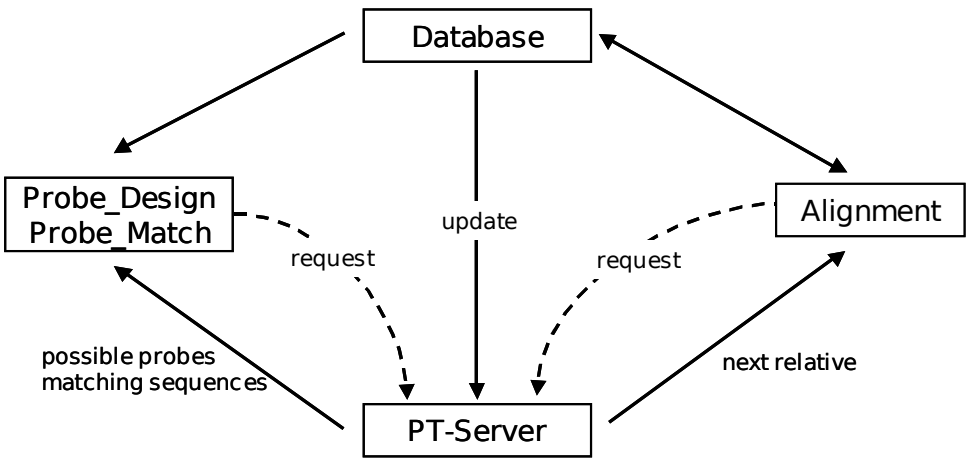
The calculation process can take up to several hours (depends on RAM available and size of the database)! After the process is finished there will be a pop up a message box informing you that your server has been created., plus on the command line where you have started ARB you will see the line "ok, server is running" at the end. The `PT_server` will run for several hours in the background after you have updated or used it. It will then terminate to save resources on your machine, and is automatically restarted when you do your first alignments or `Probe Design`/`Probe Match` operations. The restart of the `PT_server` may take up to several minutes, depending on your machine.

***Note: ARB is delivered with some general templates (names) for `PT_servers`, including `LSU_rRNA.arb` (large subunit = 23/28S), `SSU_rRNA.arb` (small subunit = 16/18S), and `user1.arb` to `user5.arb`. These are just empty templates; you can fill them with any kind of nucleic acid information, rename them, or create new templates according to your needs.***

***Note: An update (build = rebuild) of the PT_Server for use with the `Fast_Aligner` tools should be performed only after you are sure about the <u>correct alignment</u> of any newly added sequences!! Alignment is not required for the `Probe_Design` or `Probe_Match` functions.***

***Please be also aware that in case you generate new names in the database you have to update the PT_server as well. The consistency of the names is crucial for successfully working with ARB.***

***ARB's internal architecture:***



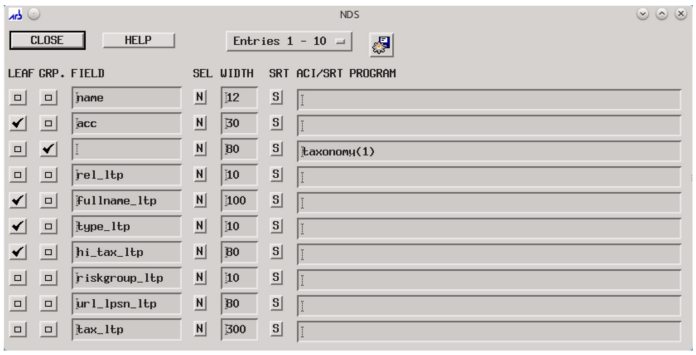
### *4.3 Trees*

As mentioned in the introduction, an ARB database can hold several trees. You can access them via the Select A Tree button in the horizontal menu bar in the ARB main window 


#### *4.3.1 NDS (Node information)*

***Note:*** You can change/adjust the information displayed in the trees. You might want to see for example the full names, accession numbers, sequence lengths, etc.

- `ARB_MAIN` window → `Tree` → `NDS (Node display setup)`
- By activating the `LEAF` and/or `GRP.` button you can select which information is shown in the tree for leafs and groups; with the `SEL` button you can change the `FIELD` to be displayed
- With the `WIDTH` button you can adjust the number of characters shown of each field in the tree




- SRT (search and replace tools) shows a list of predefined operations which can be performed on the information in the fields: e.g. you may choose to show only the first character of the genus name, followed by the complete species name (A. BB)
- ACI (ARB command interpreter)/SRT Program: here you can define your own programs. For example, `:*]=:*[*=*1` will hide the [XXX] tags in the tree when the strain name is shown. Further information about SRT and ACI can be found in the ARB online manual.

#### *4.3.2 Printing trees*

- `ARB_MAIN` window → `Tree` → `Print tree`
- `EXPORT ALL`
- `REMOVE HANDLE`

| Destination | |
|---|---|
| Printer | lpr –h will print on default printer |
| File (Postscript) | redirects the output to a postscript file; currently, you have to give the absolute destination path (e.g. `/home/username/tree_test.ps`) |
| Preview | Gives you a preview of your output; **Note:** Ghostview (package gv) has to be installed on your system |

# 5 Import of sequences and creating a new database

## 5.1 *Import of sequences to an existing database (e.g. 16S rRNA)*

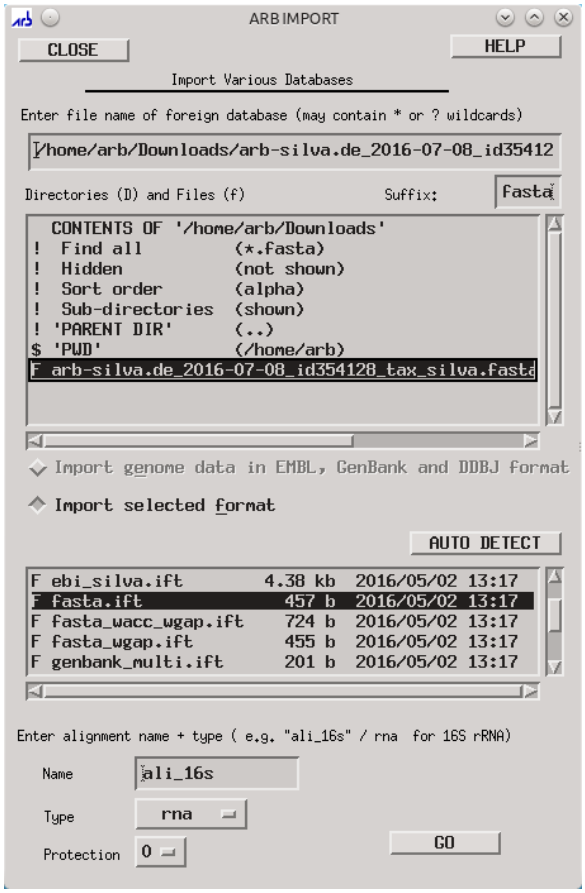
Before you can analyse a new sequence, or amend the ARB database with recently published sequences, you have to import them and align them with the existing sequences/alignment. Note that these sequences need to be in a certain recognizable format before you are able to import them; ARB provides import filters for a selection of different sequence formats.

Alternatively, you can download already aligned rRNA sequences plus detailed meta information from the SILVA database project (www.arb-silva.de) in the .arb format and merge them into your personal database (see 9.3).

### 5.1.1 *Formats of sequences to import*

| Format | Example |
|---|---|
| FASTA | `>NZ1`<br>`AGAGTTTGATCATGGCTCAGGACGAACGCT … GTAACAAGGTA` |
| EMBL | `ID  PMIFAM16S  standard; DNA; PRO; 1503 BP.`<br>`XX`<br>`AC  X62912;`<br>`XX`<br>`SV  X62912.1`<br>`XX`<br>`DT  15-JAN-1992 (Rel. 30, Created)`<br>`DT  23-AUG-1994 (Rel. 40, Last updated, Version 10)`<br>`XX`<br>`DE  P.marina 16S rDNA`<br>`XX`<br>`KW  16S ribosomal DNA.`<br>`XX`<br>`OS  Pirellula marina`<br>`OC  Bacteria; Planctomycetales; Planctomycetaceae; Pirellula.`<br>`XX`<br>`RN  [1]`<br>`RP  1-1503`<br>`RA  Stackebrandt E.;`<br>`RT  ;`<br>`RL  Submitted (31-OCT-1991) to the EMBL/GenBank/DDBJ databases.`<br>`RL  E. Stackebrandt, Dept of Microbiology, University of Queensland, St Lucia`<br>`RL  4067, AUSTRALIA`<br>`XX`<br>`RN  [3]`<br>`RA  Liesack W., Soeller R., Stewart T., Haas H., Giovannoni S.,`<br>`RA  Stackebrandt E.;`<br>`RT  "The influence of tachytelically (rapidly) evolving sequences on the`<br>`RT  topology of phylogenetic trees- intrafamily relationships and phylogenetic`<br>`RT  position of Planctomycetaceae as revealed by comparative analysis of 16S`<br>`RT  ribosomal RNA sequences";`<br>`RL  Syst. Appl. Microbiol. 15:357-362(1992).`<br>`XX`<br>`FH  Key  Location/Qualifiers`<br>`FH`<br>`FT  source  1..1503`<br>`FT  /db_xref="taxon:124"`<br>`FT  /organism="Pirellula marina"`<br>`FT  /strain="IFAM 1313"`<br>`FT  /clone_lib="M13"`<br>`XX`<br>`SQ  Sequence 1503 BP; 366 A; 346 C; 486 G; 305 T; 0 other;`<br>`  caattgaagg gtttgattct ggctcagaat gaacgttggc ggcatggatt aggcatgcaa  60`<br>`//` |

**Procedure for importing raw sequences (e.g., obtained directly from the sequencer/sequencing company):**




- Save sequences in the simple FASTA format on your computer:
  - Simple text format without carriage return
  - Use short names (not more than 8 characters) for file and sequence names (don't use special characters or symbols!)
  - Use a simple text editor to create/handle these files (files created/ handled with Microsoft Word will cause problems!)
- `ARB_MAIN` window → `File` → `Import` → `Import from external format`

  → The `ARB IMPORT` window will appear

  → Move to the folder which contains the sequences to import

  → Click on the file you want to import (to highlight it)

  If you can't find your sequences, you can use the suffix field. By typing e.g. `.fasta` ARB will show you only the files with the suffix `.fasta`

  → Press the `AUTO DETECT` button (ARB should recognize the file format - otherwise you might have a problem). Please note that there are multiple FASTA import filters available, resulting in a message "Several import filters matched …". Simply select the proper one manually. Also in case you import sequence files without converting them to FASTA format before, you have to switch the filter manually to `universal_dna.ift`

  Enter the correct alignment name (default is the standard alignment, normally you don't touch this) and, even more `importantly`, the type of your data – RNA (for 16S, 23S, or 5S data), DNA or Protein for functional genes. **We strongly recommend choosing `Protection` level 0 for newly imported sequences!**

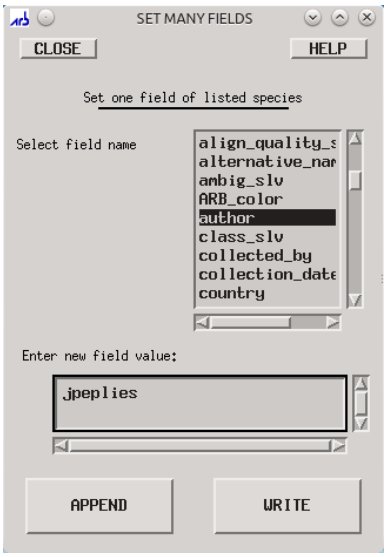
  → Press `GO`
- A question box appears: To retain the consistency of your database, you should allow ARB to automatically generate unique identifiers (see 4.1.3) using the accession number of the sequence (if no public accession number is provided by the import file, ARB will create a temporary ARB internal accession number indicated by the prefix "ARB_"). Click on `Generate unique species IDs`

- If you have more than one alignment type in your database and you provide no alignment name before (see above), a second question box will now pop up: “There are more than one possible alignment targets”. Choose a destination alignment or `ABORT`.
- The `Search and Query` window will pop up with your imported sequence(s) appearing in the list. They are already marked.
  → `Mark Listed, Unmark Rest` (to make sure that only the listed sequences are marked)
- Now it's time to tag your newly imported sequences so that you will be able to trace them later on.
  → `Write to Fields of Listed`
- The `SET MANY FIELDS` window will pop up; select e.g. `author` in the list of fields and type your name into the box below

  **Note:** This should only be done if you import raw sequence data,

  otherwise you will overwrite existing information!

  (Enter new field value) → `WRITE`
- Repeat the last step with e.g. the field `aligned` and type e.g. today's date into that field; do not forget to click on `WRITE`!
  → Close the `SET MANY FIELDS` window by clicking on `CLOSE`




**Procedure for importing sequences from ENA (European Nucleotide Archive) by EMBL-EBI**

- Go to http://www.ebi.ac.uk/ena
- Directly type in keywords describing the sequences you like to retrieve. If you have an accession number use it! Click on `Search` to get your results. Alternatively, you can switch to the second header `Search & Browse` for advanced options.
- In case of a keyword search, you will probably get hits within multiple sub-databases of ENA. On the left select 'Sequence (Release)' and now you can follow single entries via their acc on the right or download all hits together in a single multi-EBI file by clicking on 'Text' also on the right (provides text file in EBI format). Check your downloaded sequences with a text editor (e.g. nano in Ubuntu). They must have a **correct EMBLformat** (see formats) and should not contain any HTML tags or other unusual characters. If this is the case, check your settings in SRS and repeat the procedure.
- The downloaded file can now be renamed (for better organization of your data) and imported into your ARB database (see above) using the import filter `ebi_multi.ift`

***Note:*** Windows/Mac carriage returns are different than Linux carriage returns. If you see **^M** at the end of the lines (only with the `joe` editor) the file has been saved on a Windows machine. You can use the `dos2unix` command (only for DOS/Windows) to convert them.

Avoid using Microsoft Word etc. for editing of your sequences. These tools might introduce hidden characters which will cause problems in the ARB aligner. If you have your sequences stored as a Word file save them as plain text format without carriage return (.txt) first.

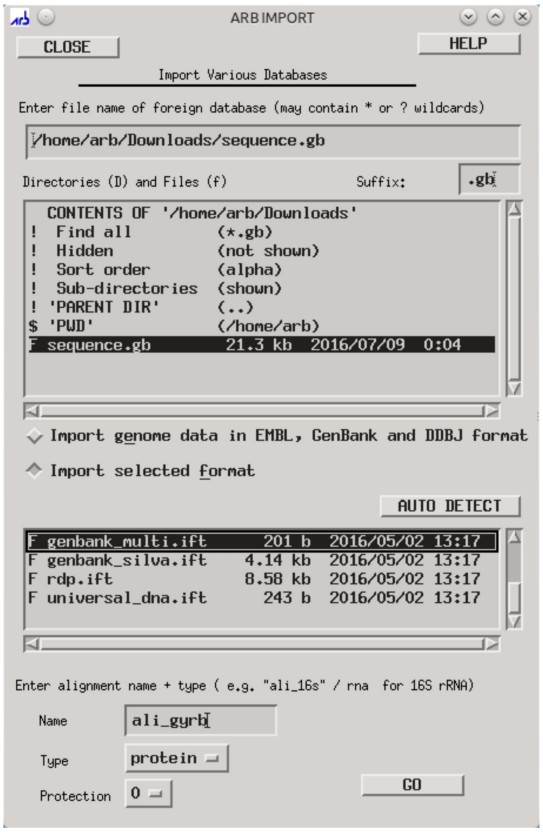
***Note:*** Annotation quality in non-curated databases like ENA/GenBank/DDBJ is in most cases not reliable! You have to play around with your queries to obtain an optimal sensitivity and specificity of your search.

### *5.2 Creating a new database (e.g. for proteins)*

To set up a new database for a specific gene (protein or DNA) a good selection of sequences has to be first downloaded from the public databases like ENA or GenBank. Since keyword searches against the annotation (gene product) will often not return a comprehensive set of related proteins due to annotation errors and inconsistencies, a sequence based search (e.g. BLAST) should be preferred. This can be easily done on the NCBI webpage (blast.ncbi.nlm.nih.gov/).

- Take you sequence of interest as a template and perform an e.g. blastp (Protein BLAST) search against the nr (Non-redundant) protein database of NCBI
- Select the sequences which should be used for phylogenetic reconstruction from the hitlist. After this, click on `Download` (on top of the list) and select `Genbank (complete sequence)` to download sequences in Genbank format
- Save the file on your home computer and rename it for better organization

Start ARB by typing in `arb` in the Shell window (this will work if you have done the installation according to the instructions given in 3. In case this does not work, ask your system administrator for help). You will get the `ARB Intro` screen (see 4.1.1). Click on the `CREATE AND IMPORT` button (a sequence import window similar to 5.1 will pop up).




→ Click on `D '$HOME'` in the upper list, and then scroll down

→ Click on the file you want to import

If you can't find your sequences, you can use the suffix field. By typing e.g. `.gb` ARB will show you only the files with the suffix `.gb`

→ Push the `AUTO DETECT` button (ARB should recognize the file format).

Enter the correct alignment name and, even more importantly, the type of your data – DNA or Protein for functional genes. **We strongly recommend choosing protection level 0 for newly imported sequences!**

→ Press `GO`

- A question box appears: To retain the consistency of your database, you should allow the `Name_server` to generate new short names (the unique identifier, see 4.1.3). Click on `Generate unique species IDs.`
- Another question box will appear asking you if and which additional field to use to create the unique IDs (compare 9.1).
- A window will pop up (in the background) giving you the advice: "`Your database contains no tree`". Click on OK.
- The `ARB_MAIN` window pops up showing the imported sequences in the list mode. All sequences are marked.
- Go to `ARB_MAIN` window → `Species` → `Search and Query` and search for all species in the database (see 4.1.4).
- Tag your newly imported sequences in order to be able to trace them later on (see 5.1.1). If you need an additional field like e.g. `aligned`, open again the `Search and Query` window → click on one of the entries → the `SPECIES INFORMATION` window will pop up. Go to `FIELDS` → `Create fields` … and type in a name for the new field in `FIELD NAME`.

**Important remark:** Fields can have different formats (`FIELD TYPE`). If you e.g. would like to introduce a field called “water_depth” and later on to search for all sequences below a given depth with “>” or “<”, this will not work if you choose for the field format `Ascii Text`. Then, the number is just a “word” - you rather should choose `Numerical` in this case.

You also can change the format of a selected field later. To do so, first toggle the expert mode (`ARB_MAIN` → `Properties`) and then you can access the `Convert fields` … option (`SPECIES INFORMATION` window → `FIELDS`).

- Finally click on Create.


Create a new field in the database

# 6 Aligning sequences

## *6.1 Align DNA/RNA sequences according to a seed alignment using the ARB_Editor (ARB_EDIT4)*

For protein sequences please have a look at the note at the end of chapter 6.4.

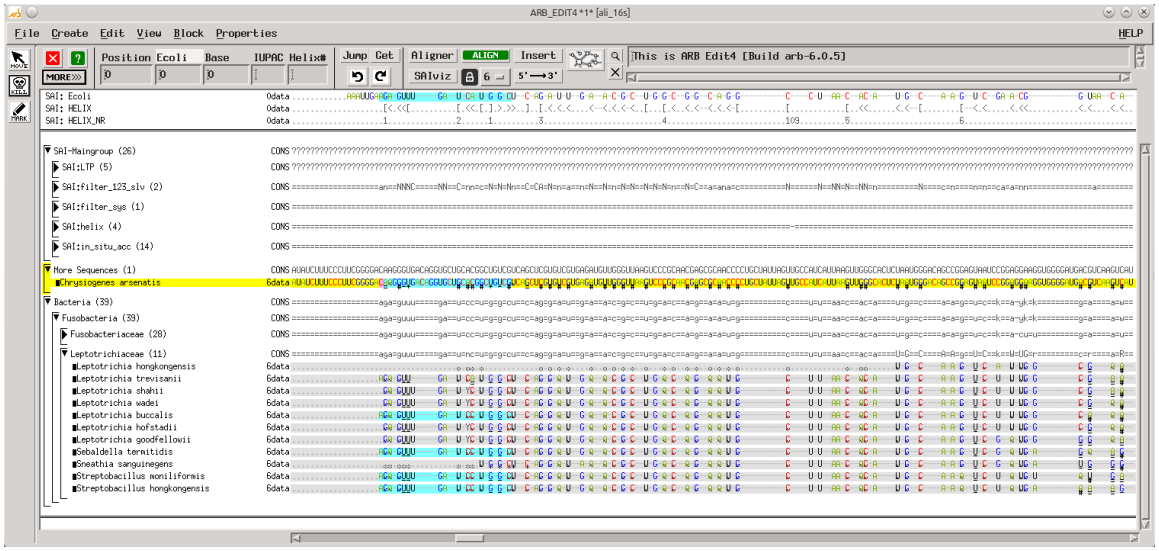
Open the editor by:

`ARB_MAIN` → `Sequence` → `Edit Sequences` → `Using marked species and tree` or

by clicking on the  button in the `ARB_MAIN` window

- All new sequence(s) which are not in a tree so far will be listed under `More Sequences`. If you have also selected some reference sequences, they will be grouped according to their phylogeny provided by the currently selected tree

***Note:*** Click on the little black triangles to open and close the groups




Selection of a specific sequence in `ARB_EDIT4`

***Note:*** Your new sequence(s) are unaligned – that means: all bases are located on the left-hand side of the window (without gaps)

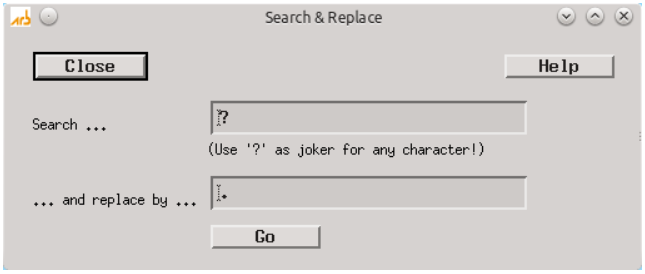
- Select your new sequence(s) by right-hand mouse clicking (their background colour will change to yellow)

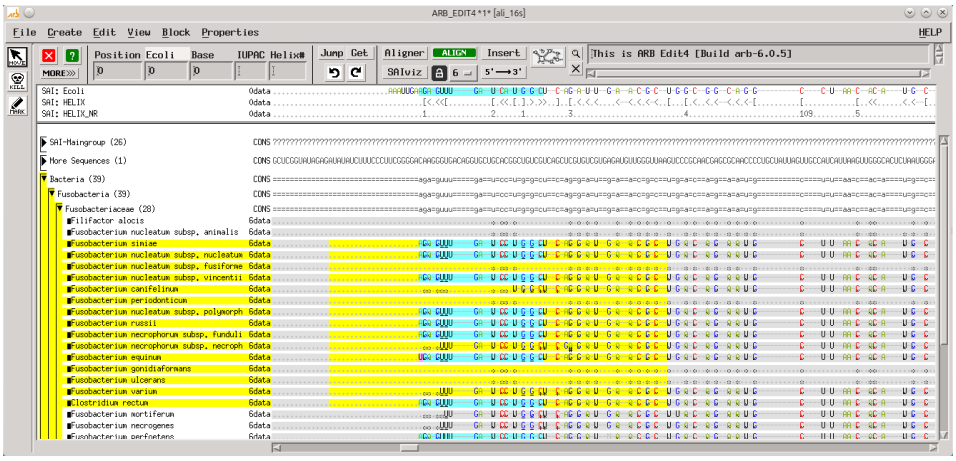
  → all `BLOCK` operations just affect the currently selected sequences/regions
- You can now change lower case letters to capitals by → `Block` → `Change to upper case`
- Replace T's by U's by → `Block` → `Search & Replace`
- The Search & Replace window will pop up: `Search… T … and replace by … U`

  → Press `GO`

  → Close the window

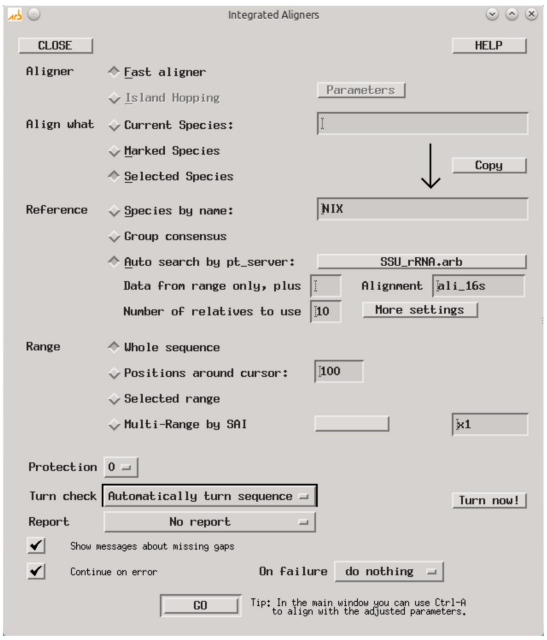
***Note:*** This function can also be used if you want to delete parts of one or more sequences (e.g. to remove the vector sequences or bases with low quality). Select the regions you want to delete by moving the cursor with the right mouse button pressed – this region will change background colour to yellow. Then open the `Search & Replace` window: `Search for ? and replace by .`




Example for the selection of a region within a set of sequences

### *6.2 Automated alignment using the ARB Tool `Fast aligner` (works only for DNA sequences)*




If your new sequence is still selected = yellow

`ARB_EDIT4` → `Edit` → `Integrated Aligners`

The `Integrated Aligners` window will appear:

- → `Align what` → `Selected Species`
- → `Reference` → `Auto search by pt_server:` (→ select appropriate PT_Server from the list)
- → Set the number of relatives to use to about `10`
- → `Range` → `Whole sequence`
- → Press `GO`

***Note:*** If your sequences are reverse complementary, you will be asked if the Aligner should turn them around (`Turn check` → `User acknowledgement`). If you don't want to be asked all the time, you can prevent this question by switching to `Automatically turn sequence`

If you have only unaligned sequences in the Editor and all of them are marked with e.g. the `Search and Query` tool you can select `Align` → `Marked Species`

***Note:*** If you haven't replaced all `Ts` by `Us` already, the `Fast_Aligner` will do this if it has to turn the sequence.

***Note: In case you have build/rebuild your `PT_Server` after you imported your unaligned sequences in ARB, the aligner will most probably fail to align your sequences when you select 'Auto search by pt_server'! Why? The aligner will try to get the next relative for your unaligned sequence by asking the PT-server. The PT_server will report back that your newly imported unaligned sequence is the next relative and than the `Fast aligner` will correctly align your unaligned sequence against itself which will not lead to any alignment!!***

***Note: Due to some sever limitations of the ARB internal aligner, a significantly improved aligner called SINA was developed within the scope of the SILVA database project. An online version is available at [www.arb-silva.de/aligner](www.arb-silva.de/aligner).***

### *6.3 Improving the alignment*

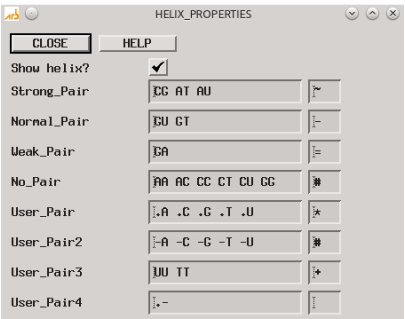
***Note:*** Manual refinement is, unfortunately, necessary (e.g. sequences must begin and end with `...` characters and this is not always the case after the `Fast_Aligner` has done its job).

What's the difference?

| | |
|---|---|
| . | means: no (sequence) information available for this position |
| - | means: no base at this position (= a gap) |

Editing the alignment:

| Button combinations | Action |
|---|---|
| Cursor buttons | The cursor moves around in the sequences |
| Ctrl+Cursor buttons | The cursor jumps over blocks of bases or gaps. This makes moving around in the sequences faster. |
| Shift+Cursor buttons | Pushes and pulls blocks of adjacent bases |
| Alt+Cursor buttons | Pushes and pulls bases towards the cursor position or pushes them over gaps |
| Middle mouse button | Moves the alignment window left/right/up/down |
| Right mouse button | Selects or unselects sequences or regions (yellow) |




- Now try to align your new sequence as well as possible with respect to the secondary structure information (ribosomal RNA) shown as helix symbols (see screenshot) under the bases and - if available - reference sequences. The helix symbols can be found and changed (if you like) the ARB_EDIT4 window at `Properties` → `Helix Settings`

- After finishing the manual alignment process, close the ARB_EDIT4 window by clicking on the red Quit button (or: → File → QUIT)
- Save your work: → File → Save whole database as … (in the ARB_MAIN window) or using the File → Quicksave changes option.

***Note:*** For a more detailed description of the ARB_EDIT4 Aligner please refer to the ARB_EDIT4 Manual at http://www.arb-home.de/arb_edit.html

**Procedure for a manual refinement of your alignment:**

**A)** Search for closely related sequences in your **ARB database** as references for the manual refinement of the alignment:

- Select your sequence of interest in the Search and Query window by clicking on it
  → More search → Search Next Relatives of SELECTED Species in PT_Server
- The Search Next Neighbours of Selected window will appear
- Select the appropriate PT_Server from the list
- Results: Select a number of sequences you want to use as reference (5 to 10 is suggested). Usually there is no need to play around with the other default settings.
  → SEARCH (Starting the PT_Server the first time can take several minutes, subsequent use will be much faster)
- The closest relatives in the database will appear in Hits → Move to hitlist
- All found sequences are now transferred to the list in the Search and Query window, and thereby your sequence of interest is removed. Add it back by:
  → Add species → that are marked → Search
- Now the closest relatives plus your new sequence should appear in the Search and Query list! (if not you have to repeat the procedure; use the tags to trace back your sequence of interest!)

**B)** Select closely related sequences in your **tree** as reference for the manual refinement of the alignment:

- If you already know the phylogenetic position of your sequences, you can go to the ARB_MAIN window and mark the corresponding group in the tree or use the Search and Query tool to search for certain sequences or groups.
- If you have done a crude automatic alignment with some manual refinement you can use the "Quick-add sequences to existing tree" tool of ARB (→ Tree → Add Species to Existing Tree → ARB Parsimony (Quick add marked)) to do a preliminary phylogenetic assignment of your sequences.
- After you have done this, you should mark 10 or 20 sequences or the complete group or subgroup related to your sequence of interest and go back to the ARB_EDIT4 Aligner to spend some time by manually optimizing the alignment
- An alternative and easy method in ARB to select some references for the manual refinement of your alignment is → Sequence → Edit Sequences → … plus relatives

- A window pops up where you can insert how many sequences you want to use as references (e.g. 10) → `OK` → `ARB_EDIT4` window will pop up

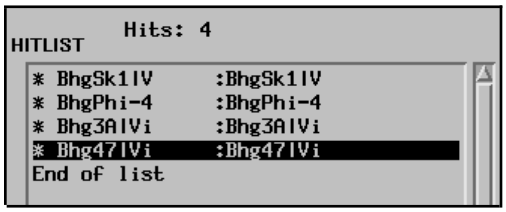
***Note: The best possible alignment of your sequences is a prerequisite for doing good phylogenetic reconstructions. The most sophisticated treeing algorithms will not deliver anything reasonable without a good alignment. Thus, spend more time on alignments than on applying the 20th variation of a distance matrix algorithm with the 10th version of Kimura 10.000 parameter correction.***

### *6.4 De Novo alignments in ARB with ClustalW (v1.83)*

ARB can also do *de novo* multiple alignments by calling the ClustalW program. This can be used for DNA as well as protein sequences.

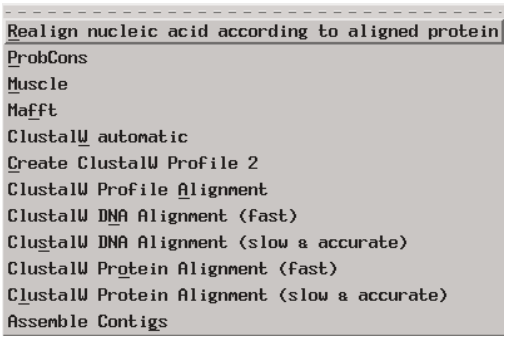
***Important Note: When ARB calls external programs it is necessary that all names are free of any special characters and even dots or dashes! In case names are longer than 8 characters or do contain anything except characters or numbers they will be truncated or corrupted by e.g. ClustalW. This in turn will lead to a duplication of the sequence data (with different names) when reimporting them. The best option in this respect is to never touch the ARB "name" field by hand and use the name server of ARB to create the names (unique identifier).***

- Go to `ARB_MAIN` window → `Species` → `Search and Query` and search for all species in the database (see 4.1.4).
- If you find names containing **-, |, ., %,** etc. proceed with "generate new names" to remove them.




Example for poisonous characters in the `name` that need to be removed

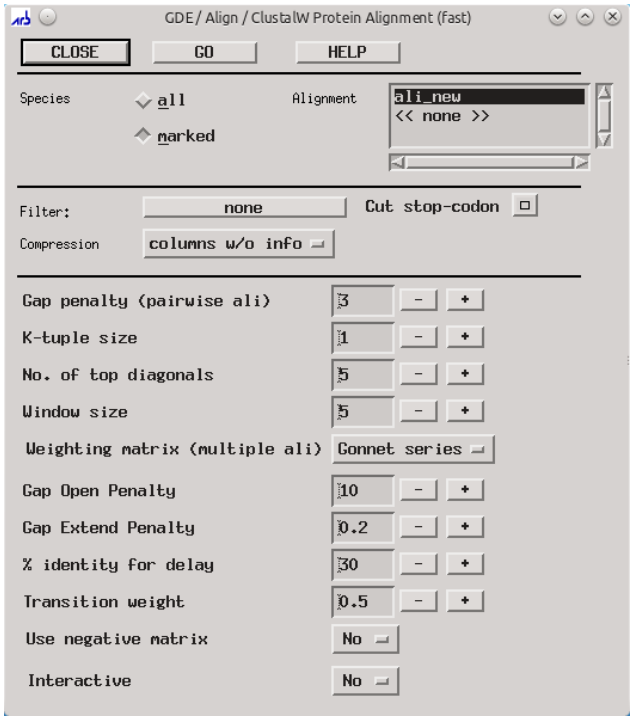
- Go to `ARB_MAIN` window → `Sequence` → `Align Sequences`, the following window will pop up:

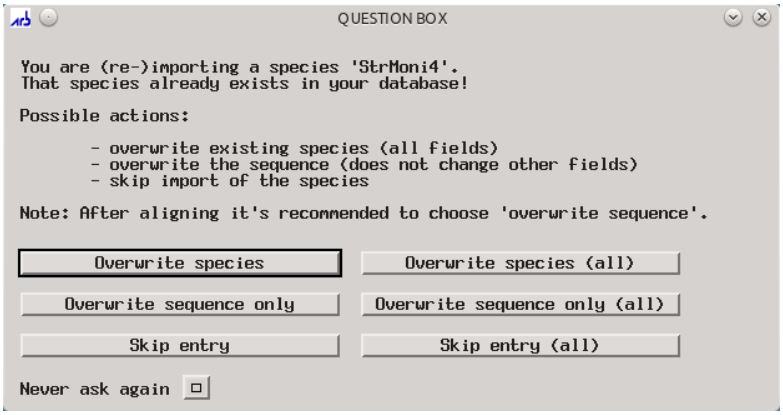



The `Align Sequences` menu

- Select the appropriate tool and parameters
- The corresponding window will pop up (on the right: `ClustlW Protein Alignment (fast)`).

  Use the default settings as a first try, for more information about parameters and settings please refer to the ClustalW manual at http://www.clustal.org/clustal2/#Documentation
- Press `GO` to start the alignment process. A bash xterm window will pop up where you can monitor the progress
- After the program is finished press `Return`
- A window will pop up asking you what to do with the aligned sequences.




The ClustalW reimport sequences window

- As it is recommended, click on `Overwrite sequence only (all)`, which means ARB will reimport all the sequences and overwrite existing sequence information (you overwrite the unaligned sequence information with the aligned one – nothing else will be changed, don't worry)
- Proceed by opening ARB_EDIT4 (see 6.1)

  → `Sequence` → `Edit Sequences` → `using marked species and tree` or by clicking on the respective button  in the `ARB_MAIN` window

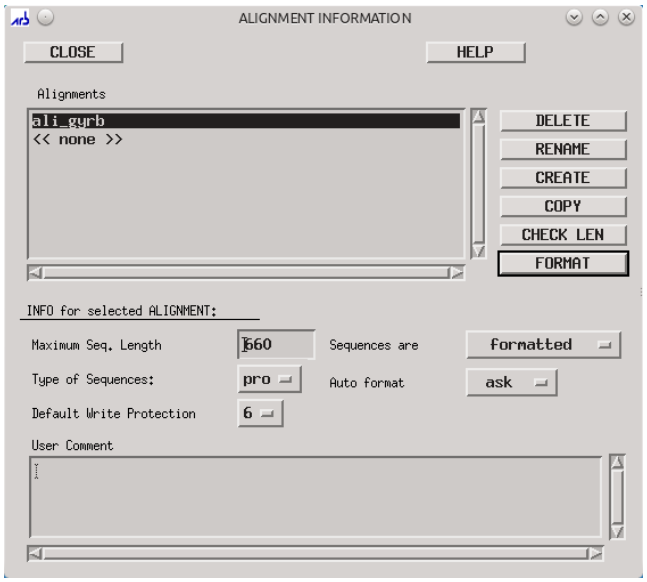
  Your new sequence(s) will be arranged under `More Sequences`. Click on the little black triangles to open and close the groups.

  All sequences should now be aligned. Nevertheless, you have to **format the alignment** and **remove the gaps at the end of the sequences** (if so) before you can reconstruct you first trees!

### *6.4.1 Formatting the aligned sequences*

As you can see, in the editor there are “white spots” at the end of the alignment. This happens if the space for the sequences was not enough and had to be extended by the alignment procedure. These “white spots” are undefined regions and need to be filled up by dots (.).


Go to `ARB_MAIN` window → `Sequence` → `Sequence/Alignment Admin` and open the `ALIGNMENT INFORMATION` window, click on `FORMAT` to format the alignment and remove these undefined regions.




The `ALIGNMENT INFORMATION` window

### *6.4.2 Remove gaps introduced by ClustalW or other programs*

ClustalW fills up missing sequence information with gaps (-). These gaps at the beginning or end of the sequences have to be removed manually since gaps have a different meaning in phylogenetic reconstruction than missing information which is represented in ARB by dots (.).

```
-CONS   ====----------------------------???id?aaahihiaffaddiidah
0data   ------------------------------------------------MYYLNKM
0data   ------------------------------------------------MYYLNKM
0data   ------------------------------------------------MYYMNQM
0data   ------------------------------------------------MYYLNKM
0data   ---------------------------------------MVEMRYFDKYAQLIYTGK
0data   -------------------------------MTIKVLNEPSPKLLTTWYAEQVTQGK
0data   -------------------------------MTIKVLNEPSPKLLTTWYAEQVTQGK
0data   -------------------------------MTIKVLNEPSPKLLTTWYAEQVTQGK
0data   ----------------------------MDLVTIKILNEPSPKLLTTWYAEQVTQGK
0data   -------------------------------MTIKVLNEPSPKLLTTWYAEQVTQGK
0data   -------------------------------MTIKVLNEPSPKLLTTWYAEQVTQGK
0data   ----------------------------MDSVTIKVLNEPSPKLLTTWYAEQVTQGK
```

Gaps at the beginning of the sequences

To change gaps into dots open `ARB_EDIT4` with your sequences and press CTRL+'.' on the keyboard in the `ALIGN` modus to change all concatenated gaps into dots. If you do this in the consensus (`-CONS`) it will be changed for all sequences. Be careful to not shift the alignment!

General Remark: If you have refresh/display problems, fill up the `Editor` window with more sequences or make the widow smaller.

**Now your sequences are prepared for tree reconstruction.**

***Note: You can also use the ARB Tool Fast aligner as described in 6.1 and 6.2 to align additional protein sequences to an existing alignment. The only difference is that the PT-server can not handle protein sequences and therefore the automatic selection of the next relative (Reference) is not possible. You have to select the reference sequence (to which it will be aligned) by hand!***

### *6.5 De Novo alignments using external programs like MUSCLE or MAFFT*

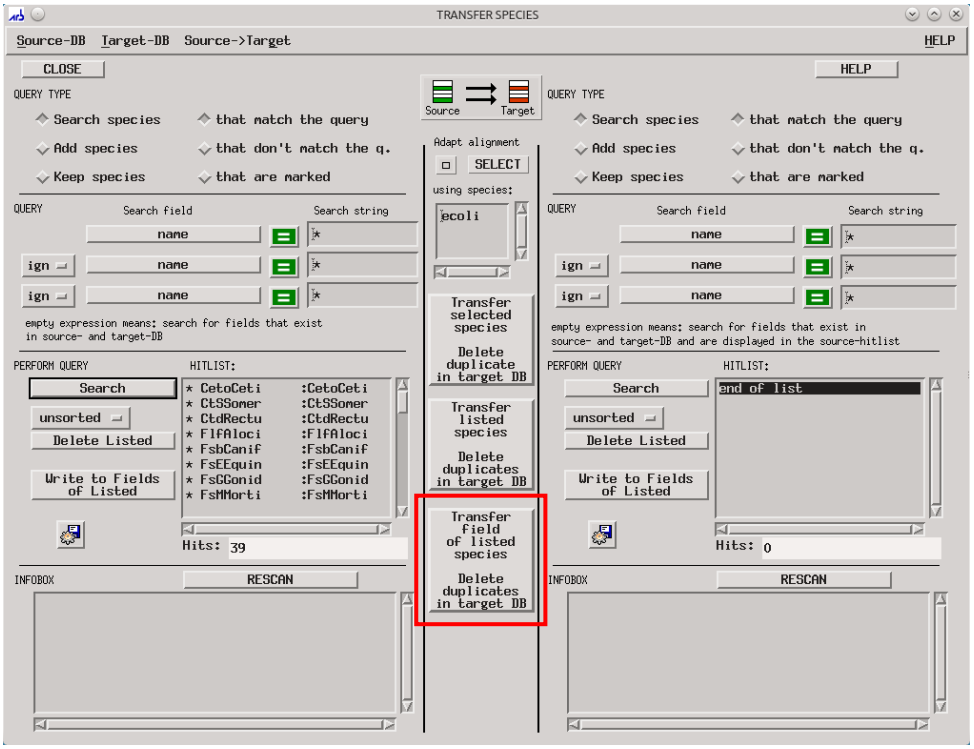
Besides using the built in ClustalW program of ARB it is also possible to run external multiple alignment programs like MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/) or MAFFT (http://mafft.cbrc.jp/alignment/server/). With ARB 6 both have also been integrated into ARB like ClustalW but nevertheless we still explain the process here. To use them externally it is necessary to first export the sequences in Multi-FASTA format, run the external program on the sequences, and afterwards overwrite the (unaligned) sequences in the ARB database by the aligned ones using the `Merge Two ARB Databases` tool!

***Note: Again it is crucial that all special characters are removed as shown in 6.4!***

- Export your sequence as described in 9.2 in MULTI-FASTA format.
- Run the external programs making sure that the sequences are written to a file in Multi-FASTA format.
- Create a new ARB database with the aligned FASTA sequences as described in 5.2.

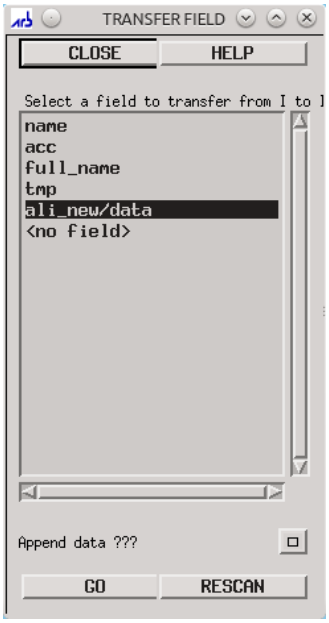
***Note: In this case you have to select the import filter manually to `fasta_wgap.ift` since with the `fasta.ift` filter the gaps introduced by the aligner will be removed again. The `type` and `name` of the alignment should be consistent with sequences in the ARB database where you have exported the sequences from. In the following question box you have to select `Use found names`, since the names are needed for replacing the sequences in the subsequent merge process.***

- save the whole new ARB database with a new name and close ARB
- start ARB and in the intro select `MERGE TWO ARB DATABASES` as shown in 9.3
- in the first following select the database with the aligned sequences imported in FASTA format and in the second window the database containing the unaligned sequences plus additional information from e.g. the GenBank or ENA.
- In the next window click on `Check IDs ...` and activate the `override` button
- close and click on `Transfer Species`
- In the following `Transfer Species` window search for all species in Source-DB (left side).
- All sequences you have aligned should now be shown in the `HITLIST`
- Click on `TRANSFER FIELD OF LISTED IN SPECIES…` (red box)




The `Transfer Species` dialog

- The TRANSFER FIELD window pops up




The TRANSFER FIELD window

- Select `ali_xxx/data` (this is the database field that contains the aligned sequence data) and click on `GO.` According to the `name` field of the database all sequences in Target-DB will be overwritten by the (aligned) sequences in target-DB if `ali_xxx/data` exists in both databases. If the names of the alignments are not synchronized, a second alignment will be generated in the Target-DB, i.e. the original field is not overwritten (you can have multiple alignments in an ARB database in parallel and switch between them via `ARB_MAIN` window → `Sequence` → `Sequence/Alignment Admin`!)
- Click on `CLOSE` in the `TRANSFER SPECIES window`
- Click on `Save whole target DB as ...` in the `ARB_MERGE window` and save the database with a new name
- Go to `FILE` → `QUIT` to close the `ARB_MERGE` window
- Format the alignment and remove the gaps at the end of the sequences (if necessary) as shown in 6.4.1 and 6.4.2.
- Proceed with the reconstruction of phylogenetic trees

# 7 Reconstruction of phylogenetic trees

***Note: Although most of the examples presented here are taken from ribosomal RNA analysis (SILVA SSU databases), all programs in ARB (e.g. ARB_EDIT4, ARB_Parsimony and ARB_Neighbor Joining) will also work with protein sequences. All procedures and commands will be the same!***

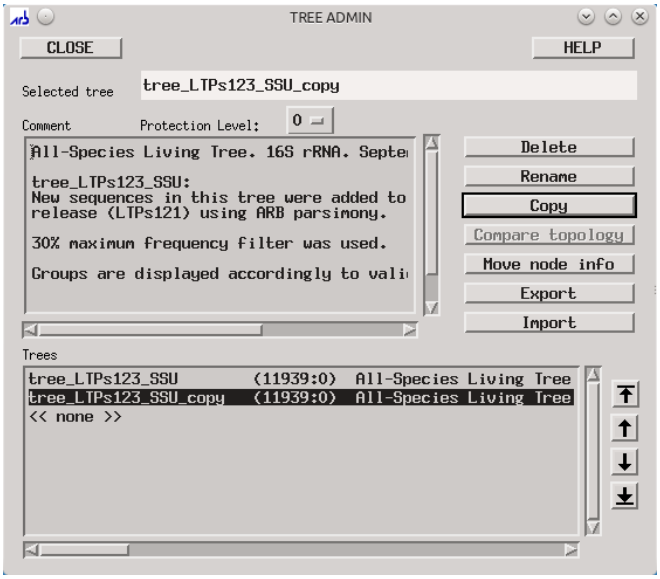
## *7.1 Managing trees in ARB*

You can organize your trees in the ARB database by opening the `TREE ADMIN` window:

`ARB_MAIN` → `Tree` → `Tree admin`

In this window you can do a lot of things with the trees: delete, rename, copy, export, import etc.

Sometimes it is good to make a copy of the tree you want to e.g. add your sequences of interest to, or you want to optimize using different criteria. This gives you a backup in case something goes wrong.




To copy a tree:

- Select the tree you would like to copy from the `Trees` field
- Click on `COPY`
- You will be asked to enter the name of the new tree (ARB will add "tree_" at the beginning)
- Click on `GO`
- The new tree will show up in the `Trees` field
- Click on `CLOSE`

## *7.2 Quick-add sequences to an existing tree with ARB_Parsimony*

- Mark the sequences you want to add to an existing tree with the `Search & Query` tool
- Open `ARB_Parsimony` with `ARB_MAIN` → `Tree` → `Add Species to Existing Tree` → `ARB Parsimony (Quick add marked)`
- The SET PARSIMONY OPTIONS window will pop up

  → Select the tree which you want to add the sequences to

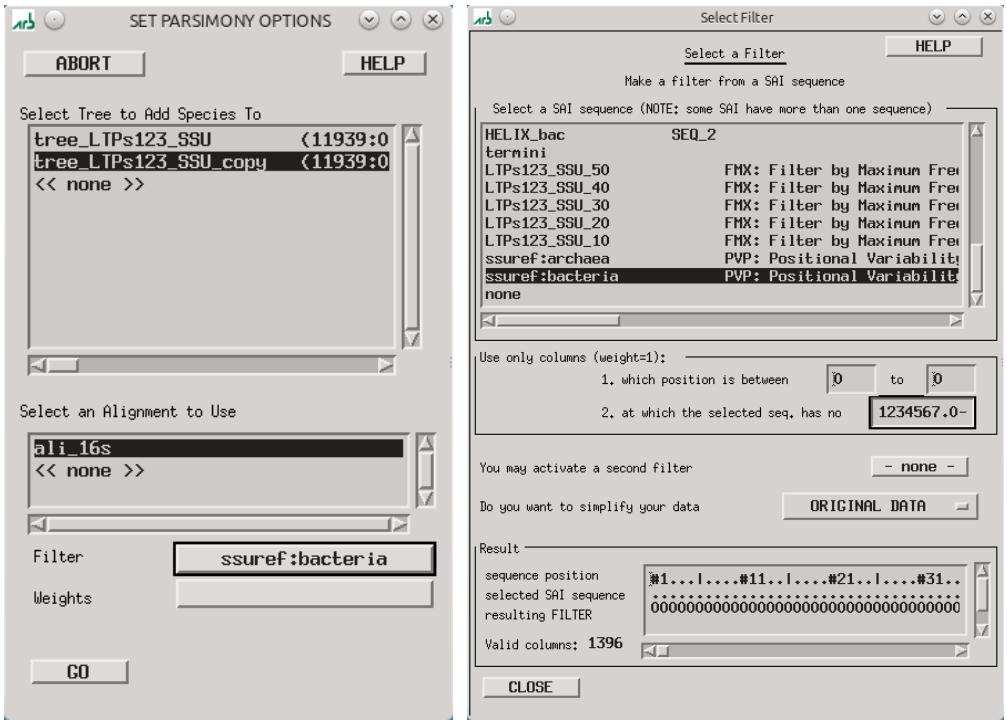
  → Select an appropriate filter for the phylogenetic group. If you don't know the phylogenetic position of your sequence, use a proper `Position Variability by Parsimony` filter and exclude the highly variable positions (1-7 or 8/9) by typing the corresponding numbers in the box 'Use only columns: 2. at which the selected seq. has no (like this: `12345678.0-=`)
- Adjust the filtering of columns by taking into account how many columns are left; this is shown at `Result` → `Valid columns`

- To get rid of vector and primer sequences, you can use a second filter like the `termini` filter which cuts off all colums outside the 16S rRNA gene (to also exclude the primer sequences you have to e.g. take the `termini` filter and shorten it by hand using `ARB_EDIT4`).

→ Click on `CLOSE`

→ Click on `GO`

After the calculations are finished, the new sequences will appear in the selected tree. If you have highlighted one of your sequences in the `Search and Query` window, you can use the `Jump` button ( Jump ) to go to it quickly in the tree.

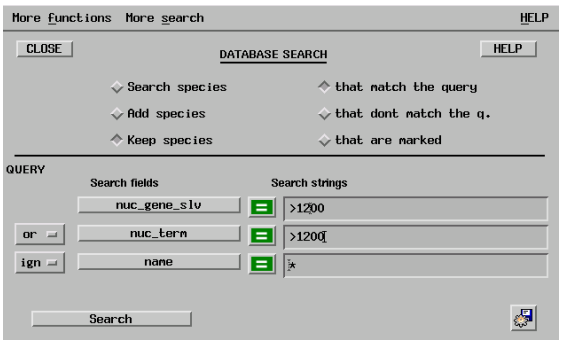



Selection of filter(s)

### *7.3 Calculation of filters by base frequency*

Marking the set of sequences:

- Mark all sequences of the phylogenetic group you want to investigate in detail in the tree or by using the `Search and Query` tool
- If you have marked sequences in the tree, go to the `Search and Query` tool and get them listed by `Search species that are marked`
- Now use the `Keep species that match the query` function to keep only nearly full length sequences in case you have supplemented your e.g. SILVA Ref database with partial sequences (for 16S rRNA "search" in the fields `nuc_gene_slv`, and `nuc_term` using the string `>1200` and combine it by `or`)




***Note:*** If `nuc_term` is not available e.g. if you have obtained the sequences by yourself, you have to count the values for the field `nuc_term` first by using the functions in the `Search and Query` tool to modify the fields of listed species as described in 4.1.5.

Calculate the filter:

- `ARB_MAIN` window → `SAI` → `Create SAI using …` → `Filter by base frequency`

  A new window pops up (showing the alignment of your marked sequences - but not editable).

  → `Config` → `Column Filter`

  Default settings for a common **50% positional conservation filter**:

  `Min. similarity=50%`

  `Max. similarity=100%`

  `'.' if occurs most often => forget whole column`

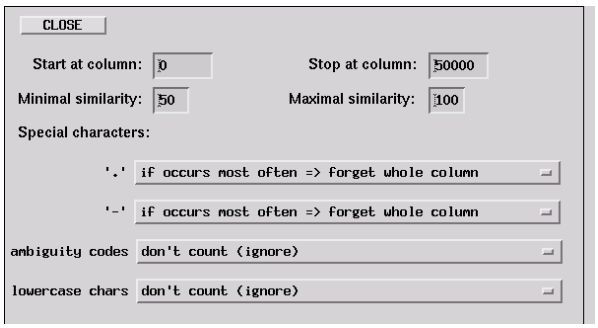
  `'-' if occurs most often => forget whole column`

  `ambiguity codes: don't count`

  `lowercase chars: don't count`

  → `Calculate` → `Column filter`

  → `File` → `Export filter`




- Name of filter: e.g. `XXX_50%`

### *7.4 Maximum parsimony trees:*

#### *7.4.1 ARB maximum parsimony*

If you already have a tree you want to work on (e.g. the general tree delivered with the 16/18S or 23/28S rRNA ARB dataset):

- Make a copy of the tree (e.g. the one with the most sequences) in your dataset
- Remove all partial sequences (e.g. <1200 nt for 16S rRNA) from the tree (not only those from the group you are interested in)
- Mark all remaining sequences in the group you are interested in (can be up to a few thousand sequences)
- Optimize the group of interest by `ARB_Parsimony`:

  `ARB_MAIN` window → `Tree` → `Add species to existing tree` → `ARB Parsimony (interactive)`

  → `Filter: e.g. xxx_50% & termini`

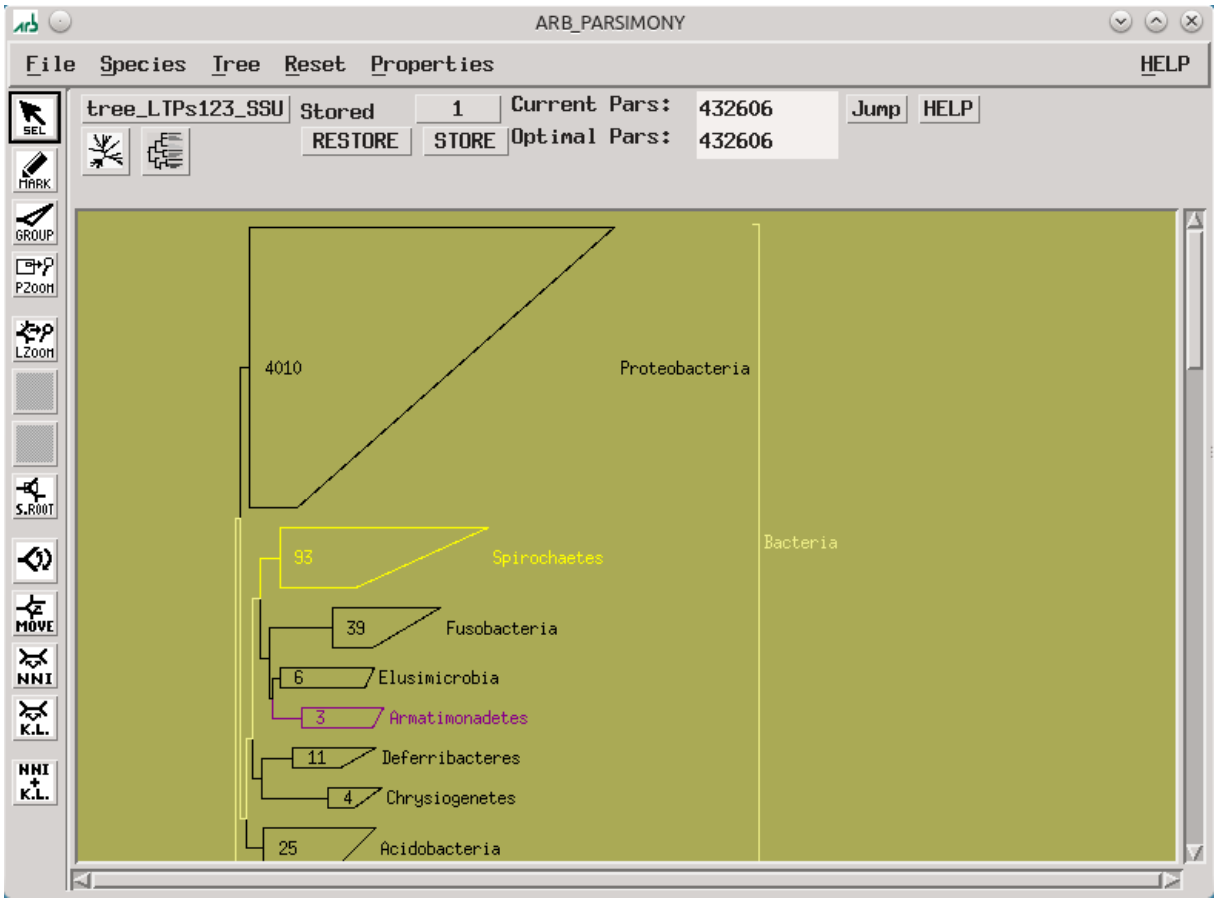
  → Click on `GO`
- The new coloured `ARB_PARSIMONY` window will appear
- Use the `Set root` button (  ) and the grouping function to group everything except the group you are interested in
- Tree optimization:

  → Global and Local Optimization

  → NNI+KL button on left-hand side (until the parsimony value is not decreasing anymore)

  → Calculate Branch Lengths

  → Quit pink window (tree will be saved)
- Repeat these calculations with no or different filters




Example for optimizing the phylogenetic reconstruction of a group of sequences with ARB_PARSIMONY

**If you do not have a tree to work on, or want to reconstruct a tree from scratch, there are two possibilities:**

1 You can make a tree with any other phylogenetic reconstruction method (e.g. `NEIGHBOUR JOINING [ARB_DIST]` or the PHYLIP or ML programs implemented in ARB) and use this tree as a starting point for optimization with the `ARB_PARSIMONY` program.

2 Arbitrarily select three of your species of interest and create an initial tree using neighbour joining or PHYLIP parsimony (does not matter which, because with three sequences only one topology is possible).

Use the `Add Species to Existing Tree` option to add as many sequences as desired to the initial tree.

Optimize the group of species you are interested in as described above.

### *7.4.2 PHYLIP DNA-Parsimony (parsimony version 3.6a3)*

Mark all full length sequences of the group you are interested in (for 16S rRNA >1200nt) with the `Search and Query` tool or in the tree (can be up to several hundred sequences)

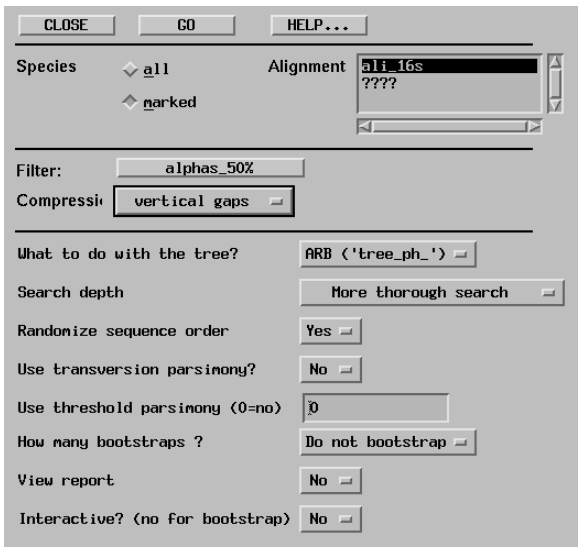
`ARB_MAIN window → Tree → Build tree from sequence data → Maximum Parsimony methods → Phylip DNAPARS`

- The `Phylip DNAPARS` window will appear

| | |
|---|---|
| `Species:` | `marked` |
| `Alignment:` | `your DNA/RNA alignment` |
| `Filter:` | e.g. `xxx_50% & termini` |
| `Compression:` | `vertical gaps` is recommended |
| `What to do with the tree?:` | `ARB ('tree_ph')` exports the tree automatically back to ARB |
| `Search depth:` | `More thorough search` is recommended |
| `Randomize sequence order:` | `Yes` is recommended |
| `Use transversion parsimony:` | `No` is recommended |
| `Use threshold parsimony:` | `0 means no` is recommended |
| `How many bootstraps?` | Do not bootstrap or do at least 100 replicates |
| `View report:` | `Yes` will give you the original PHYLIP tree output as text file |
| `Interactive:` | `Yes` allows you to adjust all parameters with the original PHYLIP menu |

→ Click on `GO`

- A shell window will pop up and show the settings and the tree calculation process. When finished a message like "`Output written to file 'outfile', Tree also written onto file 'outtree' and Press return to close window`" will be shown in the shell window. Now you can press return and close the Parsimony window by clicking on `CLOSE`. The calculated tree is saved and can be selected from the main `ARB_MAIN` window → `Trees` → `Tree admin`
- Repeat this calculation with no or different filters.




PHYLIP DNAPARS options in ARB

- It is recommended to **not use filters with bootstrapping** – it will further reduce the information!

For a detailed description of the PHYLIP DNAPARS options please refer to the PHYLIP manual at: http://evolution.genetics.washington.edu/phylip/doc/dnapars.html

### *7.4.3 PHYLIP Protein-Parsimony (parsimony version 3.6a3)*

Mark all the full length sequences of the group you are interested in with the `Search and Query` tool or in the tree.

`ARB_MAIN window → Tree → Build tree from sequence data → Maximum Parsimony methods → Phylip PROTPARS`

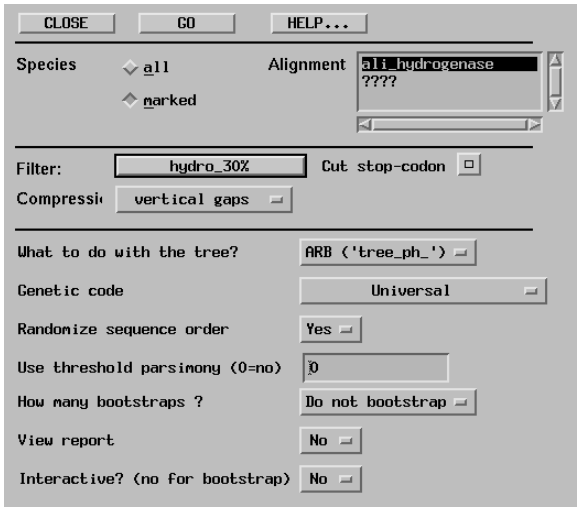
- The grey PHYLIP Parsimony window will appear

| | |
|---|---|
| `Species:` | `marked` |
| `Alignment:` | `your protein alignment` |
| `Filter:` | e.g. `xxx_30% & termini` |
| `Compression:` | `vertical gaps` is recommended |
| `What to do with the tree:?` | `ARB ('tree_ph')` exports the tree automatically back to ARB |
| `Genetic code:` | `Universal` (for rRNA) |
| `Randomize sequence order:` | `Yes` is recommended |
| `Use threshold parsimony:` | `0 means no` is recommended |
| `How many bootstraps?` | Do not bootstrap or do at least 100 replicates |
| `View report:` | `Yes` will give you the original PHYLIP tree output as text file |
| `Interactive:` | `Yes` allows you to adjust all parameters with the original PHYLIP menu |

→ Click on `GO`

A shell window will pop up and show the settings and the tree calculation process. When finished a message like "`Output written to file 'outfile', Tree also written onto file 'outtree' and Press return to close window`" will be shown. Now you can press return and close the Parsimony window by clicking on `CLOSE`. The calculated tree is saved and can be selected from the main `ARB_MAIN` window → `Trees → Tree admin`

- Repeat this calculation with no or different filters.




PHYLIP PROTPARS options in ARB

- It is recommended to **not use filters with bootstrapping** – it will further reduce the information!

For a detailed description of the PHYLIP PROTPARS options please refer to the PHYLIP manual at: http://evolution.genetics.washington.edu/phylip/doc/protpars.html

### *7.5 Distance matrix trees:*

#### *7.5.1 ARB neighbour joining*

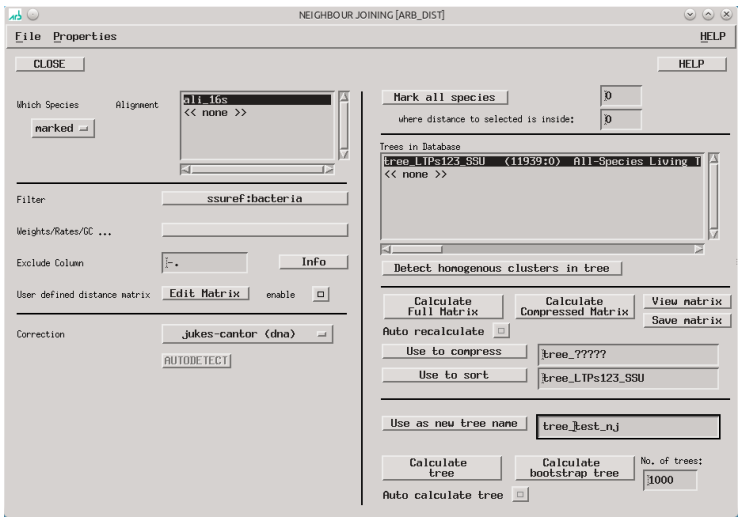
- Mark all the sequences of the group you are interested in (e.g. 16S rRNA >1200nt) with the `Search and Query` tool or in the tree (can be up to a few thousand sequences)

  `ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Distance matrix methods` → `Distance Matrix + ARB NJ`
- The `NEIGHBOR JOINING` [ARB_DIST] window will appear

| | |
|---|---|
| `Which Species:` | `marked` |
| `Alignment:` | `your DNA/RNA alignment` |
| `Filter:` | e.g. `xxx_50% & termini` |
| `Correction:` | `jukes-cantor` |
| `Use to compress:` | only interesting if you want to calculate compressed similarity matrices – ignore here |
| `Use to sort:` | does only matter for the calculation of similarity matrices – ignore possible error messages |
| `Use as new tree name:` | give your new tree a name |

  → `Calculate tree`
- If you like you can also calculate bootstrap values by setting the numbers of trees and a click on `Calculate bootstrap tree.` It is recommended to **not use filters with bootstrapping**!

  Please note that the ARB NJ bootstrap algorithm is a bit special. It will remove sequences from the tree calculation that can not unambiguously be placed in the tree.
- When finished, quit window (tree is saved). The calculated tree can be selected from the main `ARB_MAIN` window → `Trees` → `Tree admin`
- Repeat these calculations: with no or different filters; with e.g. Felsenstein correction.




ARB Neighbour Joining

#### *7.5.1.1 Calculation of a similarity matrix with ARB Neighbour joining*

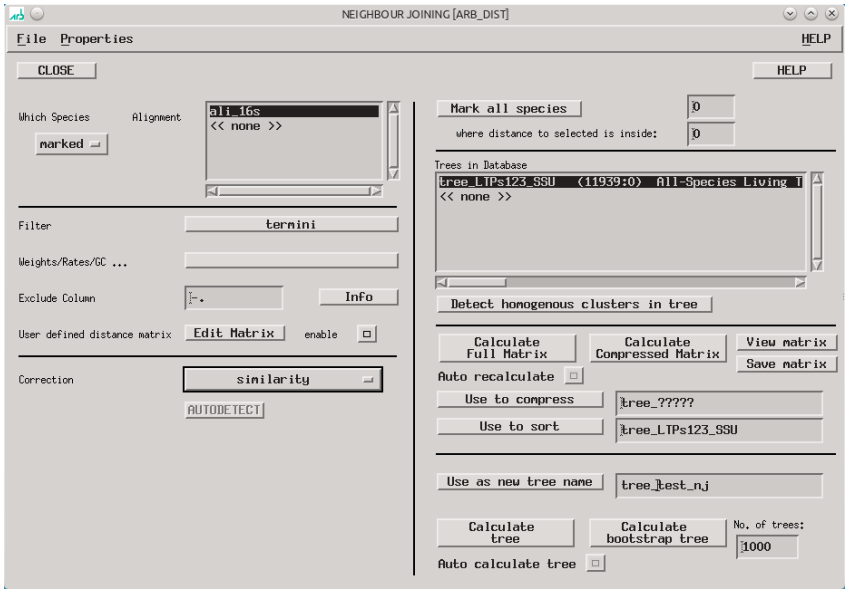
- Mark all the sequences you are interested in with the `Search and Query` tool

  `ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Distance Matrix + ARB NJ`

- The `NEIGHBOR JOINING` [ARB_DIST] window will appear

| | |
|---|---|
| `Species:` | `marked` |
| `Alignment:` | `your DNA/RNA alignment` |
| `Filter:` | only to adjust the sequences and cut off primers with something like `termini` |
| `Correction:` | `similarity` |
| `Use to compress:` | `The matrix will be compressed according to the groupings in the selected tree` |
| `Use to sort:` | When a tree is selected the matrix will be sorted according to the order in the tree |

→ Click on `VIEW MATRIX`




ARB Neighbour Joining, settings for similarity matrix

### *7.5.2 Distance matrix trees with PHYLIP version 3.6a3*

**There are two possibilities for the calculations:**

1. Mark the full length sequences you are interested in with the `Search and Query` tool

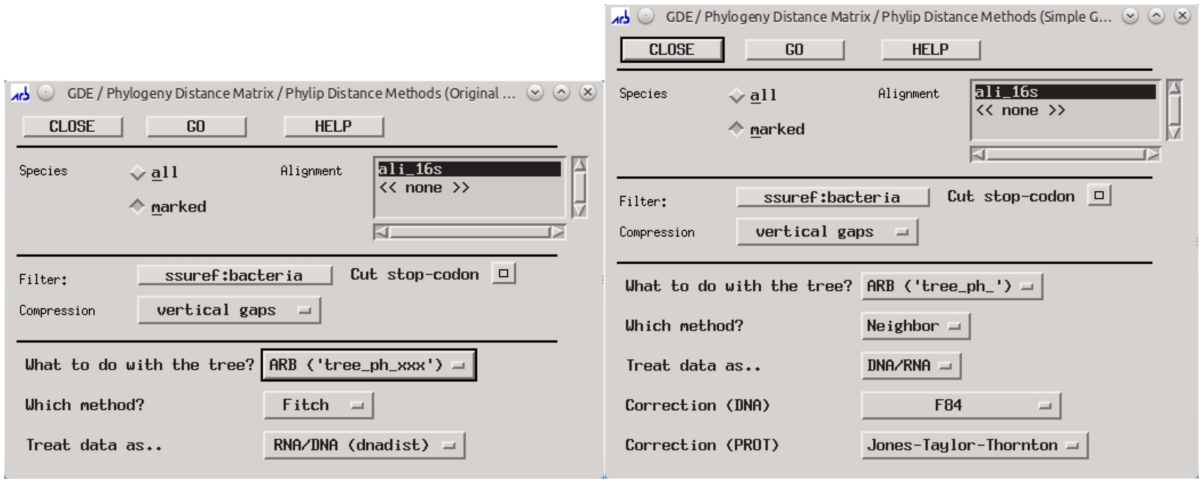
`ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Distance matrix methods` → `Phylip Distance Methods (Original Phylip, Interactive)`

In this case you will get the original command line menus from DNADIST/PROTDIST.

2. Mark the full length sequences you are interested in with `Search and Query`

`ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Distance matrix methods` → `Phylip Distance Methods (Simple GUI Based Interface)`

In this case you can adjust some of the parameters of DNADIST/PRODIST from the GUI (**G**raphical **U**ser **I**nterface)

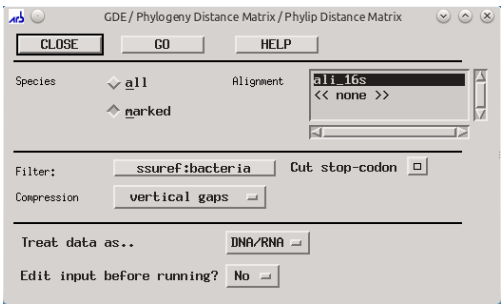



PHYLIP distance programs in ARB (left: interactive or right: simple GUI based)

For a detailed description of the PHYLIP options please refer to the PHYLIP manual at:
http://evolution.genetics.washington.edu/phylip/doc/dnadist.html
http://evolution.genetics.washington.edu/phylip/doc/protdist.html

#### *7.5.2.1 Calculation of a distance matrix with PHYLIP version 3.6a3 (DNADIST/ PROTDIST)*

- Mark the full length sequences you are interested with the `Search and Query` tool or in the tree `ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Phylip Distance Matrix`
- Make your initial settings
- You will get the original command line menus from DNADIST/PROTDIST.
- When the program is finished the matrix will be shown in a text editor.




For a detailed description of the PHYLIP options please refer to the PHYLIP manual (see links above).

### *7.6 Maximum likelihood trees:*

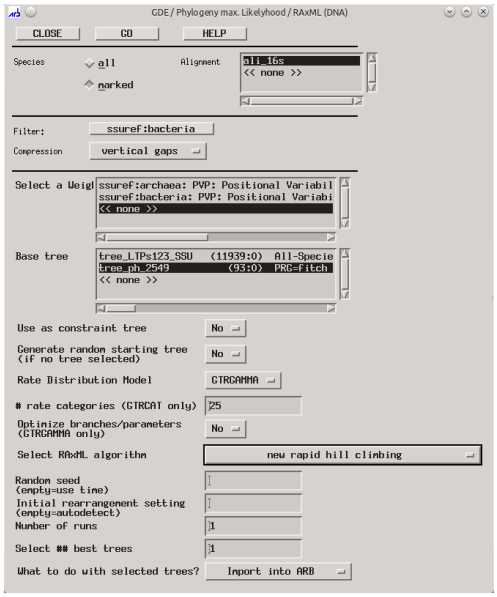
#### *7.6.1 RAxML (DNA) V 7.7.2*

- Mark all full length sequences of the group you are interested in (16S rRNA >1200nt) with the `Search and Query` tool or in the tree

  `ARB_MAIN window → Tree → Build tree from sequence data → Maximum Likelihood methods → RAxML (DNA)`
- The `RAxML (DNA)` window will appear

| | |
|---|---|
| `Species:` | `marked` |
| `Alignment:` | `your DNA/RNA alignment` |
| `Filter:` | `xxx_50% & termini` |
| `Compression:` | `vertical gaps` is recommended |
| `Select a Weighting Mask:` | `none` |
| `Base tree:` | `none` |
| `Rate Distribution Model:` | `GTRGAMMA` is recommended |
| `What to do with selected trees?:` | `Import into ARB` |

  → Click on `GO`
- Bootstrapping can be performed by selecting "`rapid bootstrap analysis`" in the "`Select RAxML algorithm`" drop down menu. The "`number of runs`" should be set to at least 100.
- A shell window will pop up and show the tree calculation process. When the calculation is finished a message like "`Press return to close window`" will be shown. Now you can press return and close the RAxML window by clicking on `CLOSE` The calculated tree is saved and can be selected from the main `ARB_MAIN` window → `Trees` → `Tree admin`
- Repeat calculations with no or different filters




The RAxML (DNA) window in ARB

For more information, especially on the additional settings which are not addressed above, please refer to: http://sco.h-its.org/exelixis/web/software/raxml/

However, for the beginning we recommend to use the default settings provided by ARB. **Better invest your time in optimization of your alignment and comparison of results from different treeing methods, then in fine-tuning of the algorithm's settings (this applies to all methods!)**

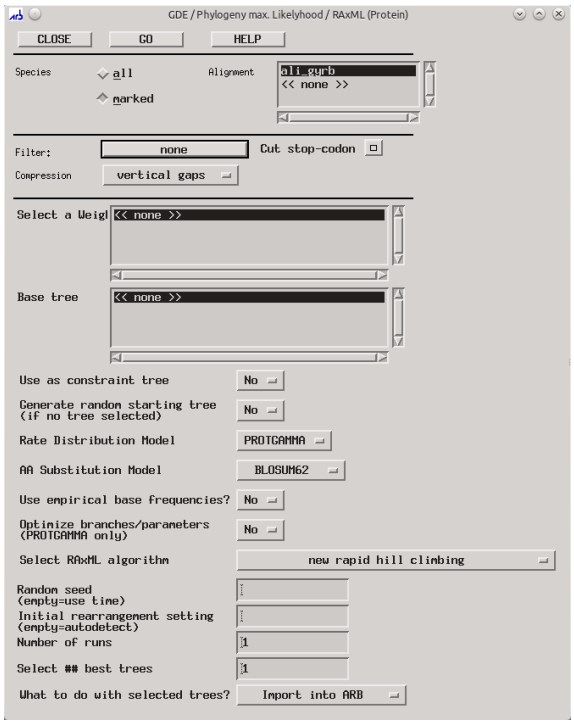
### *7.6.2 RAxML (Protein) V 7.7.2*

- Mark all full length sequences of the group you are interested in (16S rRNA >1200nt) with the `Search and Query` tool or in the tree (max. around 200, depending on the speed of your machine)

  `ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Maximum Likelihood methods` → `RAxML (Protein)`
- The `RAxML (Protein)` window will appear

| | |
|---|---|
| `Species:` | `marked` |
| `Alignment:` | `your protein alignment` |
| `Filter:` | `xxx_30%` |
| `Compression:` | `vertical gaps` is recommended |
| `Select a Weighting Mask:` | `none` |
| `Base tree:` | `none` |
| `Rate Distribution Model:` | `PROTGAMMA` is recommended |
| `AA Substitution Model:` | `BLOSUM 62` |
| `What to do with selected trees?:` | `Import into ARB` |

  → Click on `GO`
- Bootstrapping can be performed by selecting "`rapid bootstrap analysis`" in the "`Select RAxML algorithm`" drop down menu. The "`number of runs`" should be set to at least 100.
- A shell window will pop up and show the tree calculation process. When the calculation is finished a message like "`Press return to close window`" will be shown. Now you can press return and close the RAxML window by clicking on `CLOSE` The calculated tree is saved and can be selected from the main `ARB_MAIN` window → `Trees` → `Tree admin`
- Repeat calculations with no or different filters




The RAxML (Protein) window in ARB

For more information, especially on the additional settings which are not addressed above, please refer to: http://sco.h-its.org/exelixis/web/software/raxml/

### *7.6.3 PHYML (DNA) V2.4.5*

- Mark all full length sequences of the group you are interested in with the `Search and Query` tool or in the tree
  `ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Maximum Likelihood methods` → `PHYML (DNA)`
- The `PHYML (DNA)` window will appear

  `Species:` `marked`

  `Alignment:` `your DNA/RNA alignment`

  `Filter:` e.g. `xxx_50% & termini`

  `Compression:` `vertical gaps` is recommended

  `What to do with the tree?:` `ARB ('tree_phyml')exports the tree automatically back to ARB`

  `Nuc. Substitution model:` `Chose one of the models for DNA:`

  For DNA sequences, the default choice is HKY (Hasegawa et al., 1985). This model is analogous to K80 (Kimura, 1980), but allows for different base frequencies. The other models are JC69 (Jukes and Cantor, 1969), F81 (Felsenstein, 1981), F84 (Felsenstein, 1989), TN93 (Tamura and Nei, 1993) and GTR (e.g., Lanave et al. 1984, Tavaré 1986, Rodriguez et al. 1990). More information about models can be found in Swofford on page 434 and in the diagram below.

  `Base frequency estimates:` `ML/empirical,` can only be applied to models that allow unequal base frequencies like GTR or HKY!

  `Ts/tv ratio:` `fixed/estimated,` can only be applied to models that allow different substitution types like HKY, K2P or TN93

  `Proportion of invariable sites:` `fixed/estimated (slower)`

  `Interactive:` `Yes` allows you to adjust all parameters with the original PHYML command line menu, which shows up after clicking on GO. You can also use this to make bootstrapped trees. Just type in B 100 and press Enter (the value should show up in the command line menu). To start the analysis type Y and press Enter.

  → Click on `GO`

- A shell window will pop up and show the settings and the tree calculation process. When the calculation is finished a message like "`Tree tree_phyml_XXX read into database" will show up and "Press return to close window"` will be shown in the shell window. Now you can press return and close the PHYML window by clicking on `CLOSE` The calculated tree is saved and can be selected from the main `ARB_MAIN` window → `Trees` → `Tree admin`
- Repeat this calculation with no or different filters.

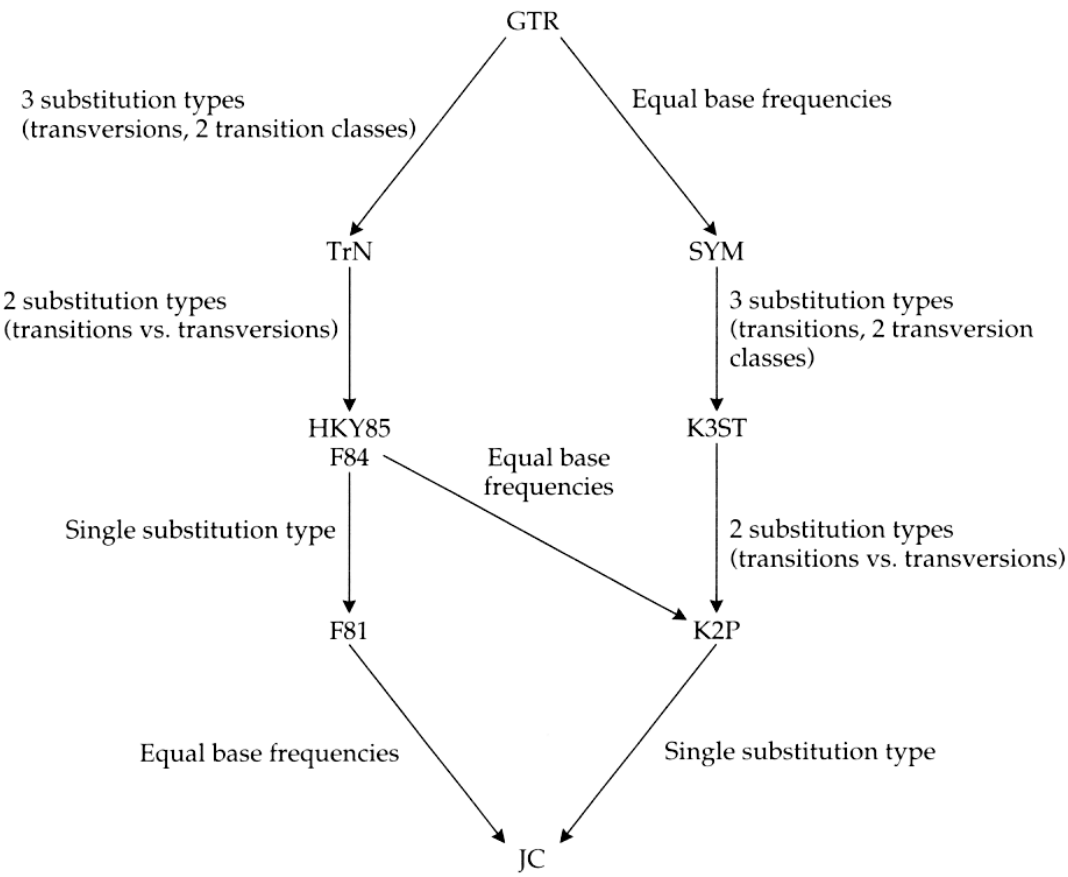
- Bootstrap analysis can be performed using the "`Interactive?: Yes`". → Click on `GO`. A shell window will pop up and you have to type in "`b, 100, y, return`" After this the menu has changed, now showing "`B Non parametric bootstrap analysis yes (100 replicates)`". Type in "`y and return`" to start the analysis. After the calculation seems to be finished press "`return`" again.

```
- PHYML v2.4.5 -

Settings for this run:

 D                                Data type (DNA/AA)  DNA
 I        Input sequences interleaved (or sequential)  interleaved
 S                        Analyze multiple data sets  no
 B                  Non parametric bootstrap analysis  no
 M                   Model of nucleotide substitution  HKY
 E            Base frequency estimates (empirical/ML)  empirical
 T                     Ts/tv ratio (fixed/estimated)  fixed (ts/tv = 4.00)
 V   Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
 R         One category of substitution rate (yes/no)  yes
 U                        Input tree (BIONJ/user tree)  BIONJ
 O                              Optimise tree topology  yes
 L     Last optimisation step on numerical parameters  yes


b
100
y
```

```
- PHYML v2.4.5 -

Settings for this run:

 D                                Data type (DNA/AA)  DNA
 I        Input sequences interleaved (or sequential)  interleaved
 S                        Analyze multiple data sets  no
 B                  Non parametric bootstrap analysis  yes (100 replicates)
 M                   Model of nucleotide substitution  HKY
 E            Base frequency estimates (empirical/ML)  empirical
 T                     Ts/tv ratio (fixed/estimated)  fixed (ts/tv = 4.00)
 V   Proportion of invariable sites (fixed/estimated)  fixed (p-invar = 0.00)
 R         One category of substitution rate (yes/no)  yes
 U                        Input tree (BIONJ/user tree)  BIONJ
 O                              Optimise tree topology  yes
 L     Last optimisation step on numerical parameters  yes


y
```

The intuitive workflow for performing bootstrap analysis with PhyML

For a detailed description of the PHYML options please refer to the PHYML manual at: http://atgc.lirmm.fr/phyml/




The different substitution models for DNA/RNA (taken from Swofford, page 434)

### *7.6.4 PHYML (Amino acids) V2.4.5*

- Mark all full length sequences of the group you are interested in with the `Search and Query` tool or in the tree
  `ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Maximum Likelihood methods` → `PHYML (Amino Acids)`
- The grey PHYML window will appear

  `Species:` marked
  `Alignment:` your protein alignment
  `Filter:` e.g. `xxx_30%`
  `Compression:` `vertical gaps` is recommended
  `What to do with the tree?:` `ARB ('tree_phyml')exports the tree automatically back to ARB`
  `AA substitution model: Chose one of the substitution models for Aminoacids:`
  For amino-acid sequences, the default choice is JTT (Jones, Taylor and Thornton, 1992). The other models are Dayhoff (Dayhoff et al., 1978), mtREV (as implemented in Yang's PAML), WAG (Whelan and Goldman, 2001) and DCMut (Kosiol and Goldman, 2005),RtREV (Dimmic et al.), CpREV (Adachi et al., 2000) VT (Muller and Vingron, 2000), Blosum62 (Henikoff anf Henikoff, 1992) and MtMam (Cao, 1998). For detailed information see Felsenstein page 222.
  `Proportion of invariable sites: fixed/estimated (slower)`
  `Interactive: Yes` allows you to adjust all parameters with the original PHYML command line menu, which shows up after clicking on GO. You can also use this to make bootstrapped trees. Just type in B 100 and press Enter (the value should show up in the command line menu). To start the analysis type Y and press Enter.

  → Click on `GO`
- To calculate bootstrap trees please refer to PHYML (DNA) chapter above.
- A shell window will pop up and show the settings and the tree calculation process. When the calculation is finished a message like "`Tree tree_phyml_XXX read into database" will show up and "Press return to close window"` will be shown in the shell window. Now you can press return and close the PHYML window by clicking on `CLOSE` The calculated tree is saved and can be selected from the main `ARB_MAIN` window → `Trees` → `Tree admin`
- Repeat this calculation with no or different filters.

For a detailed description of the PHYML options please refer to the PHYML manual at: http://atgc.lirmm.fr/phyml/

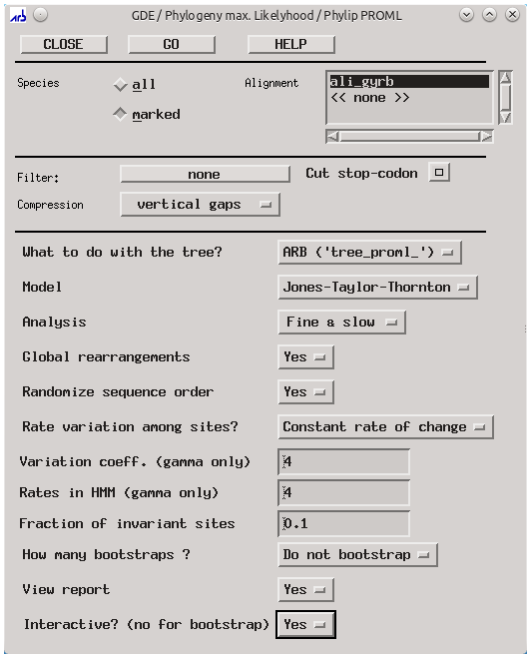
### *7.6.5 Phylip PROML V3.6a3*

- Mark all full length sequences of the group you are interested in with the `Search and Query` tool or in the tree

  `ARB_MAIN` window → `Tree` → `Build tree from sequence data` → `Maximum Likelihood methods` → `PHYLIP PROML`
- The grey PHYLIP PROML window will appear

| | |
|---|---|
| `Species:` | `marked` |
| `Alignment:` | `your protein alignment` |
| `Filter:` | e.g. `xxx_30% & termini` |
| `Compression:` | `vertical gaps` is recommended |
| `What to do with the tree?:` | `ARB ('tree_ph')` exports the tree automatically back to ARB |
| `Analysis:` | `Fine and Slow` is recommended |
| `Global rearrangements` | `Yes` is recommended (slow!) |
| `Randomize sequence order:` | `Yes` is recommended |
| `How many bootstraps?` | `Do not bootstrap` or do at least 100 replicates |
| `View report:` | `Yes` will give you the original PHYLIP tree output as text file |
| `Interactive:` | `Yes` allows you to adjust all parameters with the original PHYLIP menu |

→ Click on `GO`




- A shell window will pop up and show the settings and the tree calculation process. When the calculation is finished a message like "`Output written to file 'outfile', Tree also written onto file 'outtree' and Press return to close window`" will be shown in the shell window. Now you can press return and close the PROML window by clicking on `CLOSE` The calculated tree is saved and can be selected from the main `ARB_MAIN` window → `Trees` → `Tree admin`
- Repeat this calculation with no or different filters.

For a detailed description of the PHYLIP PROML options please refer to the PHYLIP manual at: http://evolution.genetics.washington.edu/phylip/doc/proml.html

### *7.7 Calculating trees with PHYML (on command line) V3.0*

PHYML V3.0 is a simple, fast and accurate algorithm to estimate large phylogenies by Maximum Likelihood. The corresponding paper has been published be Guindon and Gascuel, 2003 in the Journal of Systems Biology, 52(5):696-704. The source code, binaries, and a webserver can be found at http://www.atgc-montpellier.fr/phyml. Although the tool PHYML is also directly accessible in ARB, for large scale phylogenetic reconstructions it is still reasonable to start the program from the command line. You can easily export your alignment in a format that can be handled by PHYML (Phylip format) and after the calculations are done you can import the tree into ARB. To export your sequences just follow the steps described in 9.2.

After you have exported your sequences in the Phylip format and you have installed the PHYML program on your computer you can start the tree reconstruction process from your installation folder by typing:

```
./phyml
```

In this case the program will ask you for the filename and you will get a PHYLIP like menu where you can change the settings by typing its corresponding characters or you provide the settings directly on the command line, e.g.:

```
./phyml -i <input file> -q -d aa -m JTT -c 4 -a e
```

A detailed description of the command-line interface of PhyML 3.0 you find at
http://www.atgc-montpellier.fr/phyml/usersguide.php?type=command

After the calculation has been finished several files are produced.

- Rename the xxx.xx_phyml_tree.txt file to xxx.tree
- start ARB and go to `ARB_MAIN` window → `Tree` → `Tree Admin` → `Import` to import the tree

### *7.8 Calculating trees with RaxML (on command line) V8.0.x*

RAxML is a very powerful, fast and accurate algorithm to estimate large phylogenies by Maximum Likelihood. The first corresponding paper has been published by Alexandros Stamatakis, 2006 in *Bioinformatics* 22(21):2688-2690. The source code, binaries, and a webserver (plus detailed documentation) can be found at http://sco.h-its.org/exelixis/web/software/raxml/. Although RAxML is directly accessible in ARB, for large scale phylogenetic reconstructions it is still reasonable to start the program from the command line. You can easily export your alignment in a format that can be handled by PHYML (Phylip format) or FASTA and after the calculations are done you can import the tree into ARB. To export your sequences just follow the steps described in 9.2.
After you have exported your sequences and you have installed the RAxML program on your computer you can start the tree reconstruction process (easy & fast way) by typing:

raxmlHPC -m GTRCAT -s test_100.phylip -n test_100

where test_100.phylip is the input file and test_100 will be used to name several output files

There are two additional versions of the program available raxmlHPC-PTHREADS and raxmlHPC-MPI. For multi-core systems (with 2 or more processors) raxml-PTHREADS is recommended because it speeds up the calculation process significantly by using all processors in parallel. The respective command would look like this:

raxmlHPC-PTHREADS -T 4 -m GTRCAT -s test_100.phylip -n test_100

The MPI version is for cluster computing by spreading bootstrapping to several machines. It needs OpenMPI installed – please ask you system administrator for details.

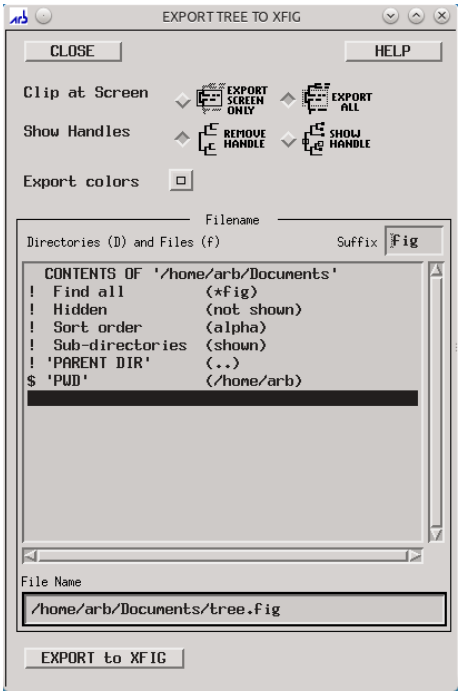
After the calculation has been finished several files are produced.

- Rename the RAxML_result.xx file to xxx.tree
- start ARB and go to `ARB_MAIN` window → `Tree` → `Tree Admin` → `Import` to import the tree

## 7.9 Exporting trees from ARB to external programs

### 7.9.1 Exporting trees in the EMF format via XFIG

A common file exchange format for graphics is represented by the Enhanced Metafile (EMF). ARB uses an external tool called XFIG to create this kind of files:




- ARB_MAIN window → Tree → Export tree to XFIG

  The EXPORT TREE TO XFIG window will appear

  EXPORT ALL

  REMOVE HANDLE

  EXPORT to XFIG
- Dismiss Error Message

(If you can't see the tree → enlarge window by using the button in the upper right-hand corner)

→ File (hold mouse button!) → Export...

→ Language: EMF

→ Output File: enter file name e.g. **tree.emf**

→ The tree file is being saved in the directory you have specified

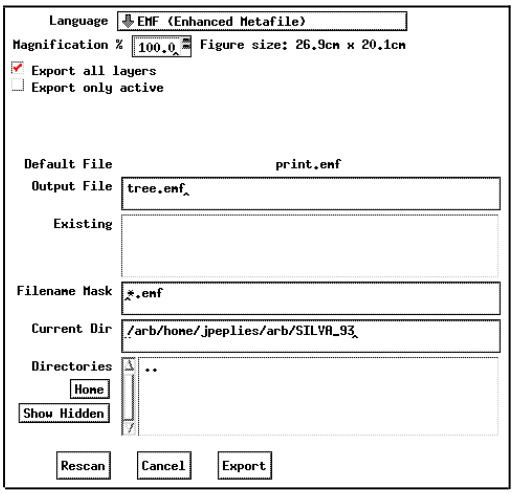
→ File → Exit (to exit XFIG)

If you prefer to finalize your trees using Windows-based tools like PowerPoint, go on like this:

- Transfer the file to your Windows machine
- If you can't see the file extensions under Windows go to My Computer → View → Options → View → Unmark Hide file extensions for known file types → OK
- Start Microsoft PowerPoint

  → Insert → Picture → From File... → select file **tree.emf**

  (The tree might be imported in a much too large size thus you need to adjust the size to the page → setting the zoom factor to 10% helps)
- Double-click on the tree and convert the picture to a Microsoft Office drawing. Now you can edit the tree with respect to the text and line settings etc.




Remark: The main advantage of (the very old) XFIG is the various export formats provided by this tool. For example, you can easily create a PDF file out of your tree to share preliminary results with your colleagues. However, if you are not really happy with the quality of the EMFs created by XFIG, please refer to the next chapter which is introducing a powerful alternative.

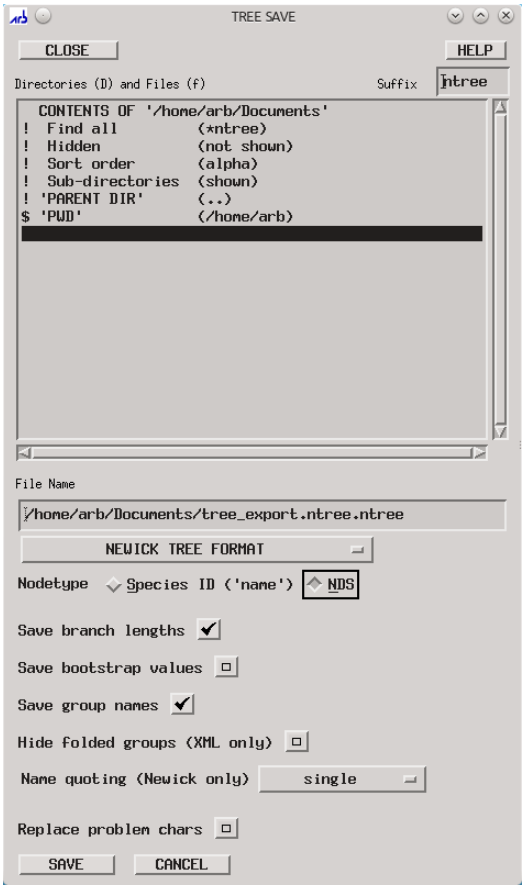
### *7.9.2 Exporting trees in the NEWICK format*

If you export your trees in the NEWICK format, you have the possibility to import them into external tools which allow you to completely reformat your tree outside of ARB. Sometimes this is the faster option for finalizing trees for later publication. One useful option is represented by the tool FigTree which can be obtained for usage in Windows, Linux and Mac OS X environments from http://tree.bio.ed.ac.uk/software/figtree. It is “a graphical viewer of phylogenetic trees and a program for producing publication-ready figures”.

You can directly import the trees to FigTree in the NEWICK format and beautify/finalize them according to your requirements. Finally, you can export the optimized tree view from FigTree also in the EMF format to e.g. add brackets highlighting sequence clusters in e.g. PowerPoint (this is not possible with FigTree). The advantage of this workflow is the much better quality of the EMFs compared to the XFIG solution. We strongly recommend this option to create publication-quality tree-based figures.

How to export trees in the NEWICK format from ARB:

- `ARB_MAIN` window → `Tree` → `Tree admin`
- The `TREE ADMIN` window will appear; select a tree and click on `Export`
- The `TREE SAVE` window will appear; make your settings:
  - → `File Name:` enter an output file name
  - → Select `NEWICK TREE FORMAT`
  - → `Nodetype: NDS` is recommended
  - → Make your additional settings
- Press `SAVE` to write the NEWICK file to the directory you have selected in the upper box




You can now directly open the file with FigTree.

# 8 Probe functions

## *8.1 Probe design*

- Mark all sequences of the group you are interested in with the `Search and Query` tool or in the tree or directly in a tree
- `ARB_MAIN` window → `Probes` → `Design Probes`
- The `PROBE DESIGN` window will appear

  → Select an appropriate PT_Server
- Enter parameters for probe design:

  **`Length of output:`** how many suggestions `Probe_Design` should show

  **`Max. non group hits:`** max. number of sequences/species that can be targeted by a probe without mismatches outside your species or group of interest – relaxes the specificity of the probe design

  **`Max. hairpin bonds:`** not implemented

  **`Min group hits (%):`** if you have marked a group of sequences/species, `Probe_Design` is allowed to suggest probes which do not cover the complete group – relaxes the sensitivity of the probe design

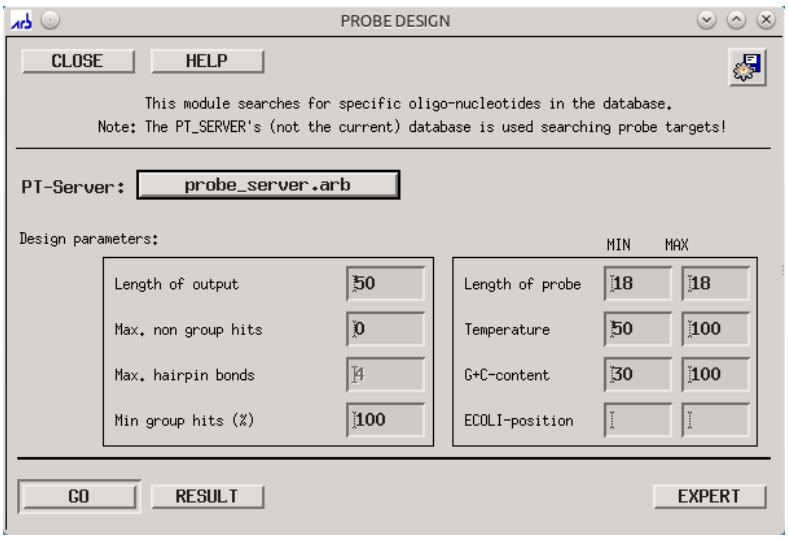
  **`Length of probe:`** the length of the probes in bases

  **`Temperature:`** set an allowed range of melting temperatures. The theoretical melting temperature is calculated by the 4+2 rule (GC pair 4°C, AT pair 2°C)

  **`G+C content:`** set an allowed range of G+C contents.

  **`ECOLI-position:`** if you would like to force `Probe_Design` to only report probes in a certain region of the sequence/species, you can adjust it here. **Note:** You have to write in the absolute alignment positions!!

→ Click on `GO`




`Probe_Design` parameters window

- ARB will generate a sorted list of possible probes
- The following information is given in the output list:

  **`Probe design Parameters:`** a summary of the selected parameters

  **`Target:`** target sequence

  **`le:`** probe length

  **`apos:`** absolute probe position, probes are grouped as A, B, C, etc. according to target site, with the “best” probe first. Overlapping probes are shown as e.g. A + (bases) or A – (bases), relative to the first probe in the group. Best is relative, `Probe_Design` takes into account the theoretical melting temperature and the specificity – see `Decrease T by n*.3C`

  **`ecol:`** probe position relative to the *E. coli* alignment

  **`qual:`** probe quality (specificity) indicator (optimum at 20)

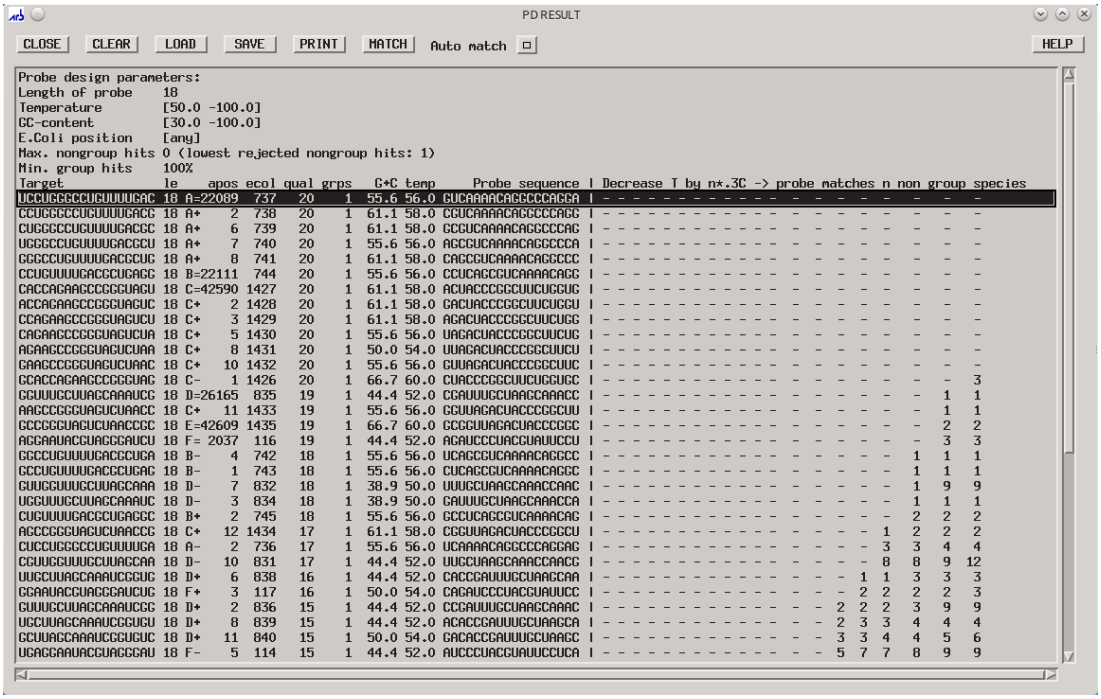
  **`grps:`** total number of sequences covered by the probe

  **`G+C:`** GC content of the probe

  **`temp:`** theoretical melting temperature (4GC+2AT)

  **`Probe sequence:`** probe sequence

  **`Decrease T by n*.3C:`** gives you information about the theoretical specificity of the probe. By decreasing the optimal hybridization temperature x times 0.3 C° (x=the sum of columns) the indicated number of additional non-target sequences in the database would theoretically hybridize with the probe.




List of possible probes generated by `Probe_Design`

- The selection of a probe in the output list (by clicking) automatically transfers the target sequence to the `Probe_Match` tool and to the `Probe` field in the `ARB_EDIT4` window

### *8.2 Probe match*

- ARB_MAIN window → Probes → Match Probes
- The PROBE MATCH window will appear

  **Target string:** the reverse complement sequence of your probe. If you have done Probe_Design first, you can just select a probe from the output list – the sequence will be automatically transferred. Here you can also type in manually a probe you like to check against the database.

  **Used PT Server:** the PT_Server you like to check your probe against

  **Accepted mismatches:** here you can adjust if also sequences with mismatches should be reported

  **Use weighted mismatches:** ✓

  Briefly: The program takes into account the position and the quality of the mismatch; e.g. G and edge mismatches are down weighted (please have a look in the EXPERT menu to learn how the weighted mismatches are calculated)

  **Check complement too:** ✓

  **Mark in database:** ✓

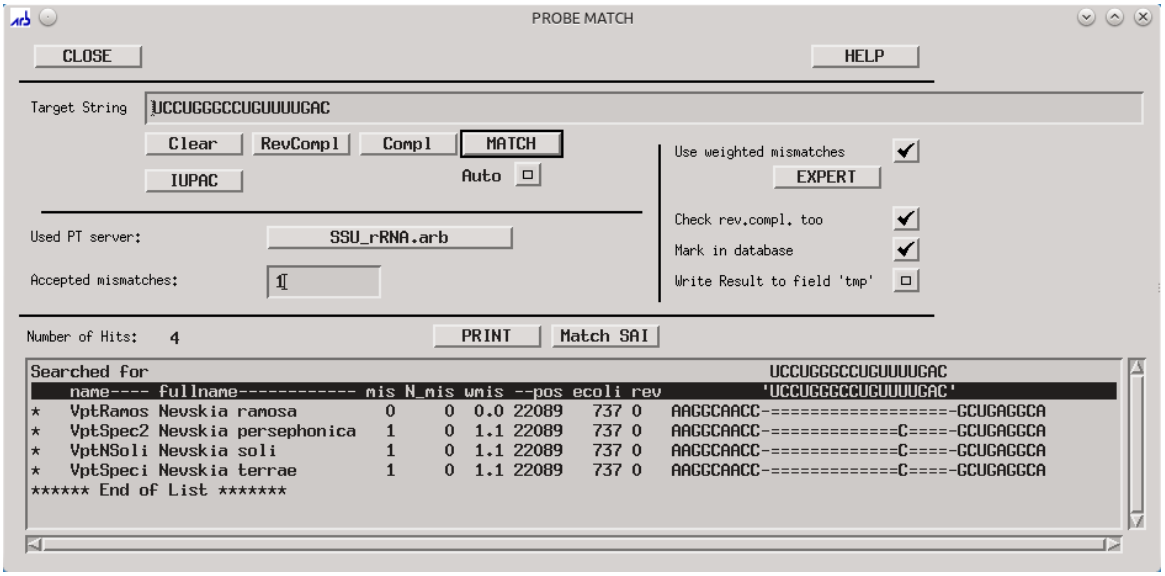
  **Write results to field 'tmp':** gives all the sequences which are listed in the Hitlist a tag in the tmp field

  **Print:** opens a menu where you can adjust parameters for printing, you will also get a small preview

→ Click on MATCH

- You will see the following output:

| | |
|---|---|
| **Searched for:** | the target (reverse complement of the probe) sequence |
| **name:** | the short name (ARB internal identifier) of the sequence |
| **fullname:** | the corresponding full name |
| **mis:** | absolute number of mismatches |
| **N_mis:** | number of N (ambiguous) bases |
| **mis:** | number of mismatches if weighted |
| **pos:** | absolute start position of the probe in the corresponding alignment |
| **ecoli:** | *E. coli* start position of the probe |
| **rev:** | 0=normal match; 1=reverse complement match |
| **in the last row** | you see the probe corresponding target sequence and some adjacent bases.<br>= means perfect match<br>AGCU means strong mismatch (if weighted)<br>agct means weak mismatch (if weighted) |




The `Probe_Match` window

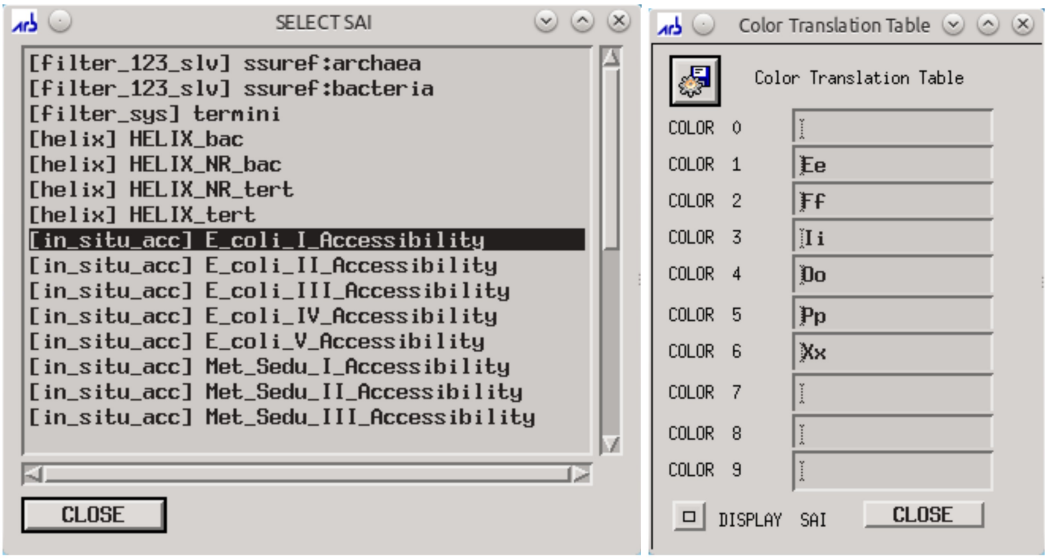
### *8.2.1 Match SAI (e.g., visualisation of target site accessibility)*

The `match_SAI` function allows you to map any kind of sequence associated information (helix structure, filters etc.) colour-coded or in clear text on a region of the target sequence matched by a probe. For details please refer to the paper by Kumar et al., BMC Bioinformatics 2005, 6:61.

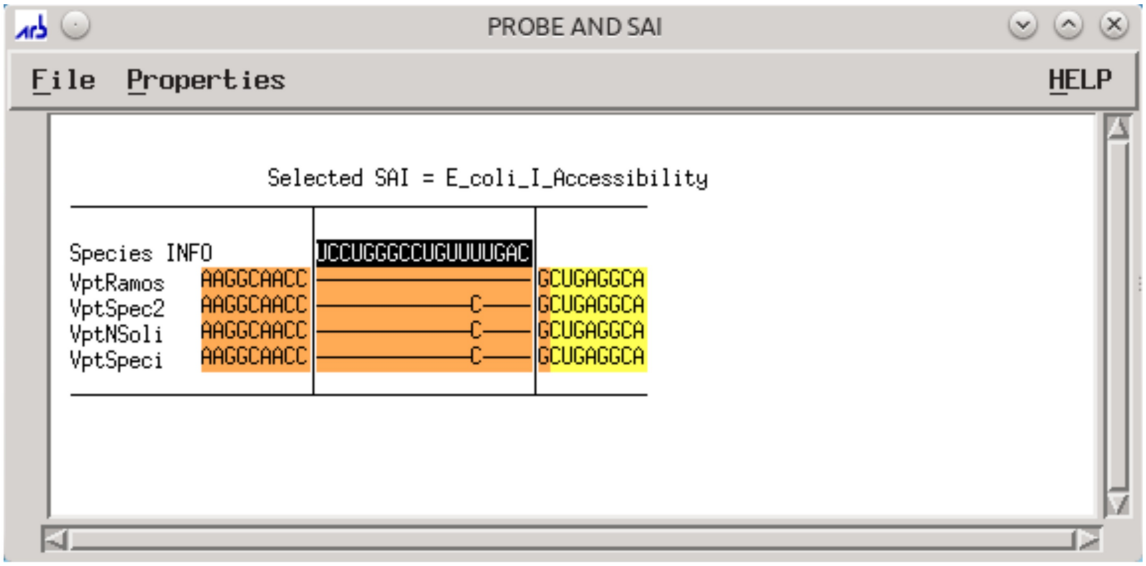
If you design probes for using them with single cell fluorescence *in situ* hybridisation (FISH), you can map the *in situ* accessibility of the 16S rRNA as described by Behrens et al., in Applied an Environmental Microbiology 2003, Vol. 69, no. 3, p. 1748 on your region of interest.

To do this you have to go to the `Properties` menu of the `PROBE AND SAI` window (press the `Match SAI` button of the `PROBE MATCH` window to get there) and:

- **`Select Display Field:`** this will change the database field which is shown for each sequence. Normally, this will be `full_name`
- **`Select SAI:`** take one from the group of `[in_situ_acc]` depending on your organism (delivered with the latest ARB/SILVA database release)
- **`Define Color Translations:`** every character in the corresponding SAI has to have a colour assigned – if you change something you have to save afterwards by clicking on the disk symbol. If you tick **`Display SAI`** you will also see how the SAI looks like.
- You can adjust the colours and fonts in **`the Set Colors and Fonts`** menu. This has to be done twice – here and also in the `ARB_EDIT4 Colors and Fonts` menu, since they are not synchronized. If you do not synchronize them by hand you will get different colours between the `PROBE AND SAI` window, `ARB_EDIT4` and `Sec_Edit.`




The `SELECT SAI` menu and the `Colour Translation Table`




The `PROBE AND SAI` window after adjusting the translation table and colours. The accessibility of the 16 rRNA region is shown according to the colour code published in Behrens et al.

### *8.3 ARB_EDIT4 pattern search function (check a probe in the alignment):*

- Open `ARB_EDIT4` via `ARB_MAIN` window → `Sequence` → `Edit Sequences` → `Using marked species and tree` or by clicking on the button  in the ARB main window
- Open the `pattern matching functions` of the Editor by clicking on  in the upper lefthand corner, next to the red quit symbol.
- The selected probe can be found in the field `Probe`
- To display the probe in the alignment you have to tick  the box after the  button
- Use the `LAST` and `NEXT` buttons to jump to the probe position
- The probe target region is highlighted (here yellow); mismatches can also be displayed (here pink)

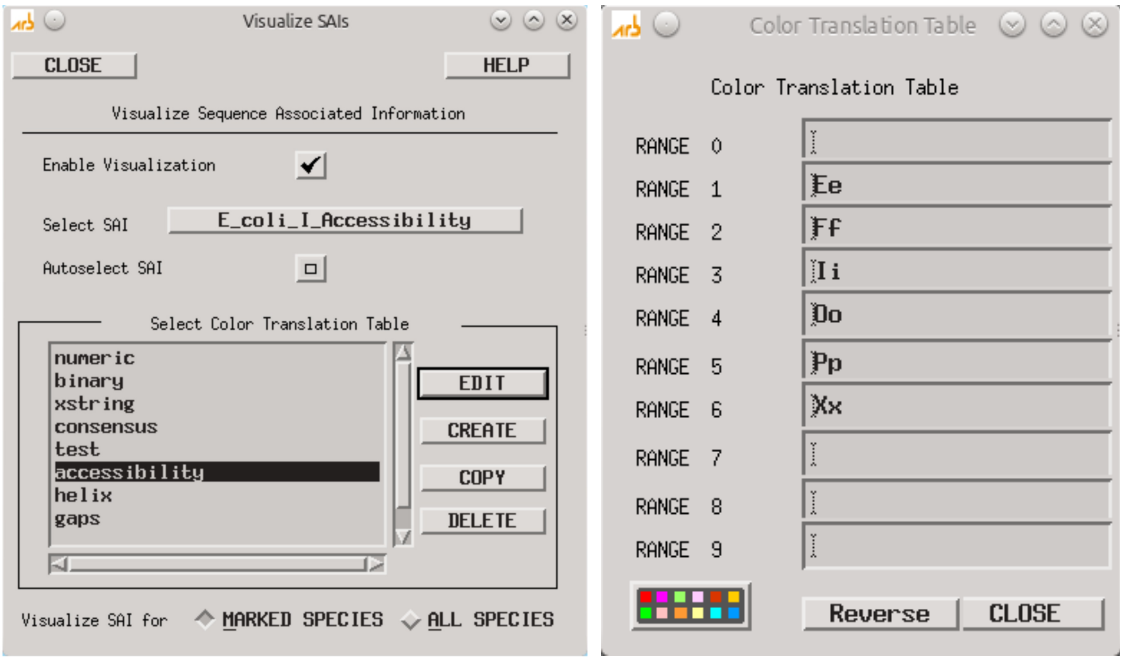

ARB_EDIT4 with a probe matching a certain region shown in grey

### *8.3.1 Display SAI (e.g., visualisation of target site accessibility)*

- To display the SAIs e.g. *in situ* accessibility for FISH go to → `View` → `Visualize SAIs and tick 'Enable Visualization'`

  → `Select SAI` (take one from [`in_situ_acc`] depending on your organism)

  → `Select a color translation table` – for visualizing accessibility you have to create a new one.

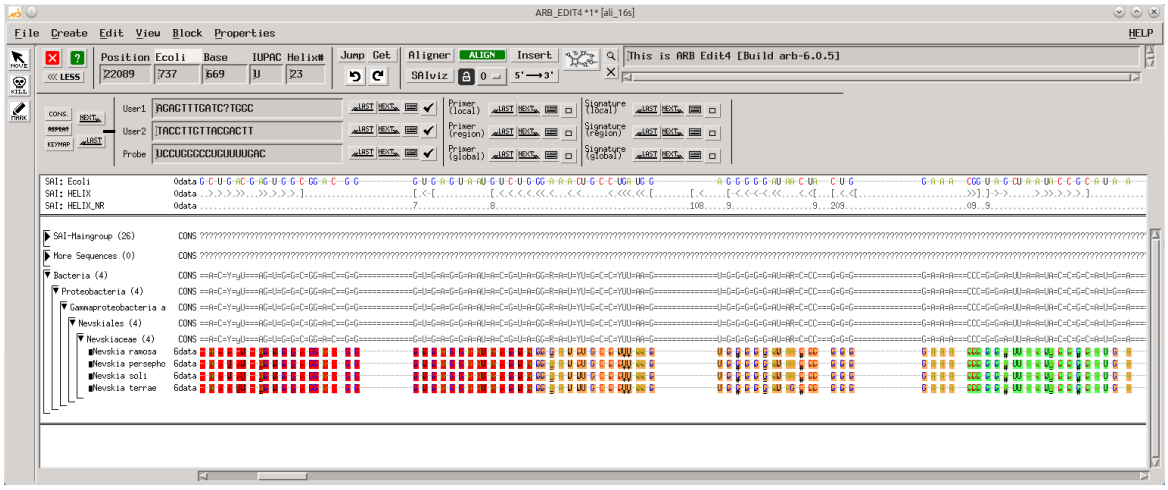



Example for "visualize_SAI" and the color translation table for "accessibility"

**Remember,** that you have to adjust the colors (Range 1-6) in the → `Properties` → `Change colors and Fonts menu of ARB_EDIT4!`

**Example** of the Colors mapping according to Behrens et al: *in situ* accesibility for FISH probes:

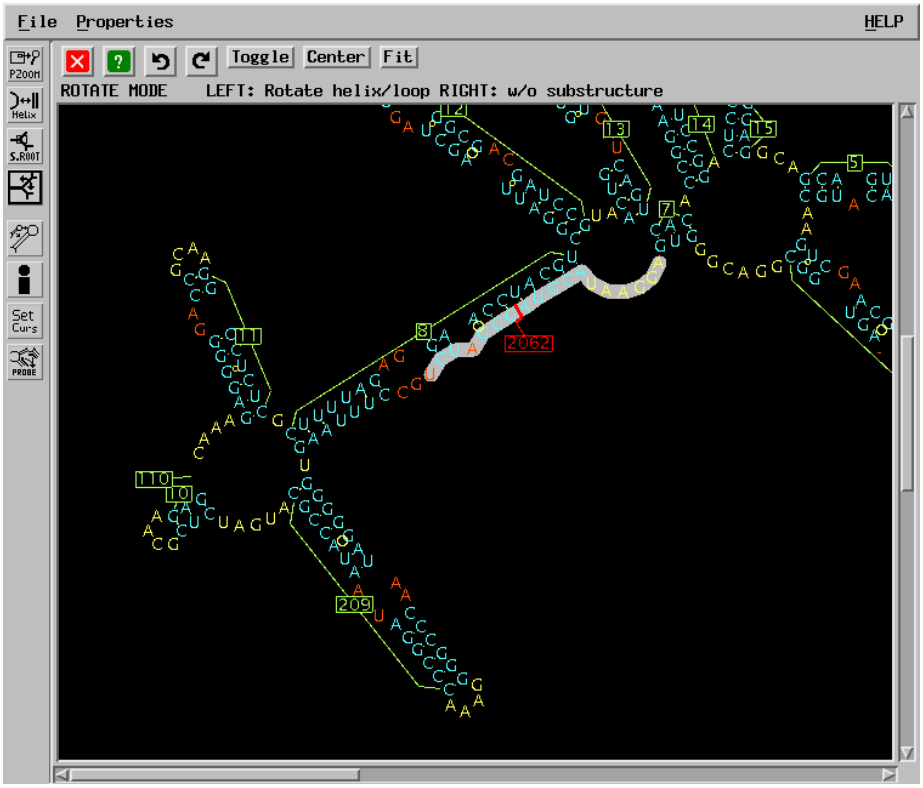
In ARB_EDIT4 it will look like this:




ARB_EDIT4 with “SAI_visualization” enabled. Grey is the probe, pink the mismatches, yellow and orange the accessibility of the target region according to Behrens et al.: yellow class III, orange class II

### *8.4 Secondary structure editor*

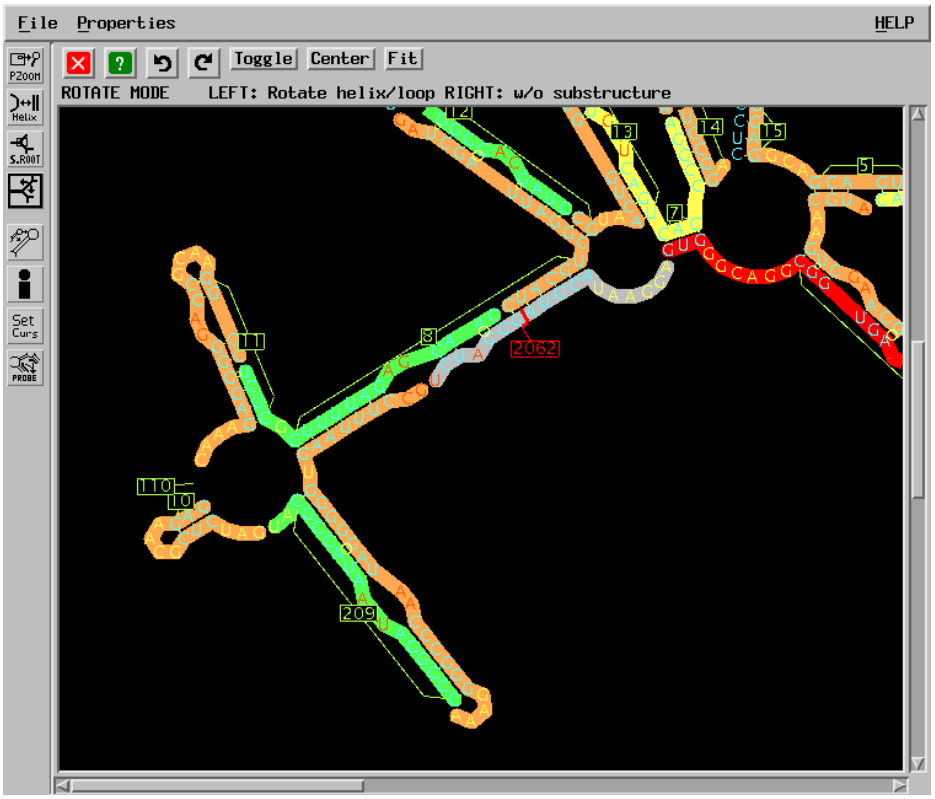
- To check the secondary structure of the probe target position (for ribosomal RNA) you can open the secondary structure editor implemented in ARB: click on the button  in the `ARB_EDIT4` window
- The probe is automatically highlighted in the secondary structure editor.
- The red line shows the current cursor position in the Aligner.




Secondary structure Editor in ARB_EDIT4

- The *in situ* accessibility can be visualized by → `Properties` → `Display Options`

-> `Tick the Visualize_SAI button`




*In situ* accessibility visualization in “Sec_Edit”. The colours correspond to the colours in the Editor

### *8.5 Multiple probes*

The multiple probe function in ARB does not calculate probes de novo, this has to be done with the `Probe_Design` module! see 8.1.

Let's assume you are interested in designing probes to cover a certain group of sequences – in our example it's the *Pirellula* group. In an ideal world you would just mark all members of the group and go to the `Probe_Design` tool, adjust your parameters so that there are no outgroup hits allowed and all sequences you have selected have to be covered by the suggested probes. If you do this you will get an empty `Probe_Design` window with the message "There are no results". This is not the fault of ARB, but of molecular evolution, which means there is no conserved e.g. 18mer across all the group of sequences you marked available. Now you can play around to see if you might get results with a 17mer or 16mer or relax your specificity by allowing outgroup hits. Sometimes this works, but with the increasing amount of sequences in the databases most probably it won't.

The smarter way is to find a reasonable combination of probes which cover the whole group when they are mixed in a hybridization experiment. To support this, ARB has implemented the `Multiple-Probe calculation tool`. But, before you start with the calculation of two- or three-probe combinations you have to generate a set of possible probes for the species or group of sequences you are interested in. In general you perform this as described in 8.1 but to get a list of probes you have to relax the sensitivity by decreasing gradually the value in `Min group hits (%)`.The goal of the procedure is to get a reasonable list of probes which target several subgroups of the group of interest (here *Pirellula*). Test it by selecting some of the probes and do a `Probe_Match` as described in 8.2. If all probes are only targeting a single subgroup of your sequences of interest you can also design probes for the different subgroups individually and merge the results afterwards in the `Multi_Probe tool`.

If you have a reasonable list or lists in the `Probe_Design` result window you have to **save** them for use with the `Multi_Probe tool`!

- `PD RESULT` window → click on SAVE


- the grey save box will appear → select a directory and name for your probelist and click on save
- `ARB_MAIN` window → `Probes` → `Calculate Multi-Probes`
- The grey `MULTI_PROBE` window will appear

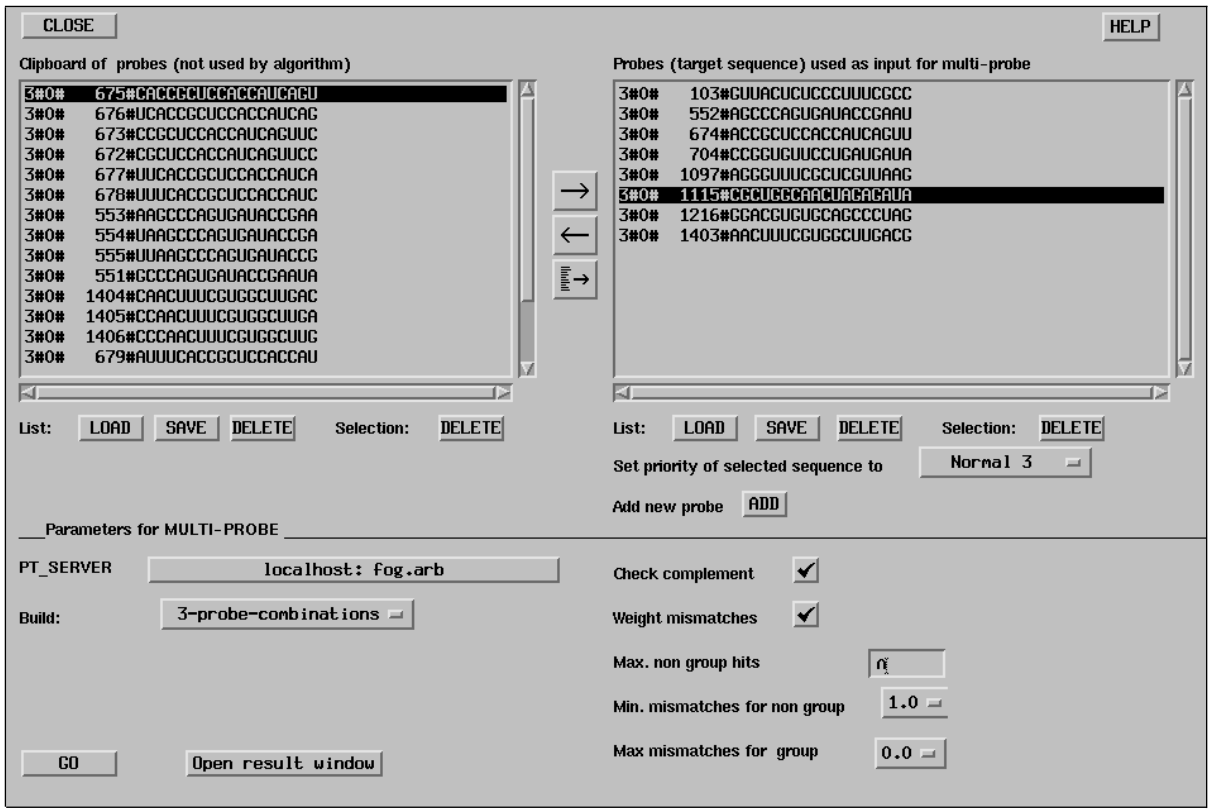
  → you will see two windows; the left one is called `Clipboard of probes`, the right one is called `Probes`.
- Load your probe-list(s) by clicking on `LOAD` below the left window. You will see a list of probes and their Ecoli-positions. You can add more probes from different lists by repeating the procedure. Select the probes that should be used for the calculation of multi-probes using the `arrow` buttons. With → and ← you can add and remove probes to the input list, with ⇥ you can move

all probes from the left side to the input list. The input list can also be saved, loaded etc. You can add a probe manually by typing in the target sequence after clicking on the ADD button.

The probes in the input list should have different specificities to cover in combination the complete group of sequences you are interested in. This can be done by, e.g., selecting probes from different regions for the input list.

- Select an appropriate PT_SERVER (the one which you have used for `Probe_Design`)
- **`Build:`** Select if ARB should calculate `2,3 or more probe-combinations`

| | |
|---|---|
| `Check complement` | ✓ |
| `Weight mismatches` | ✓ |
| `Max. non group hits` | `0` |
| `Min. mismatches for non group` | `1.0` |
| `Max mismatches for group` | `0.0` |

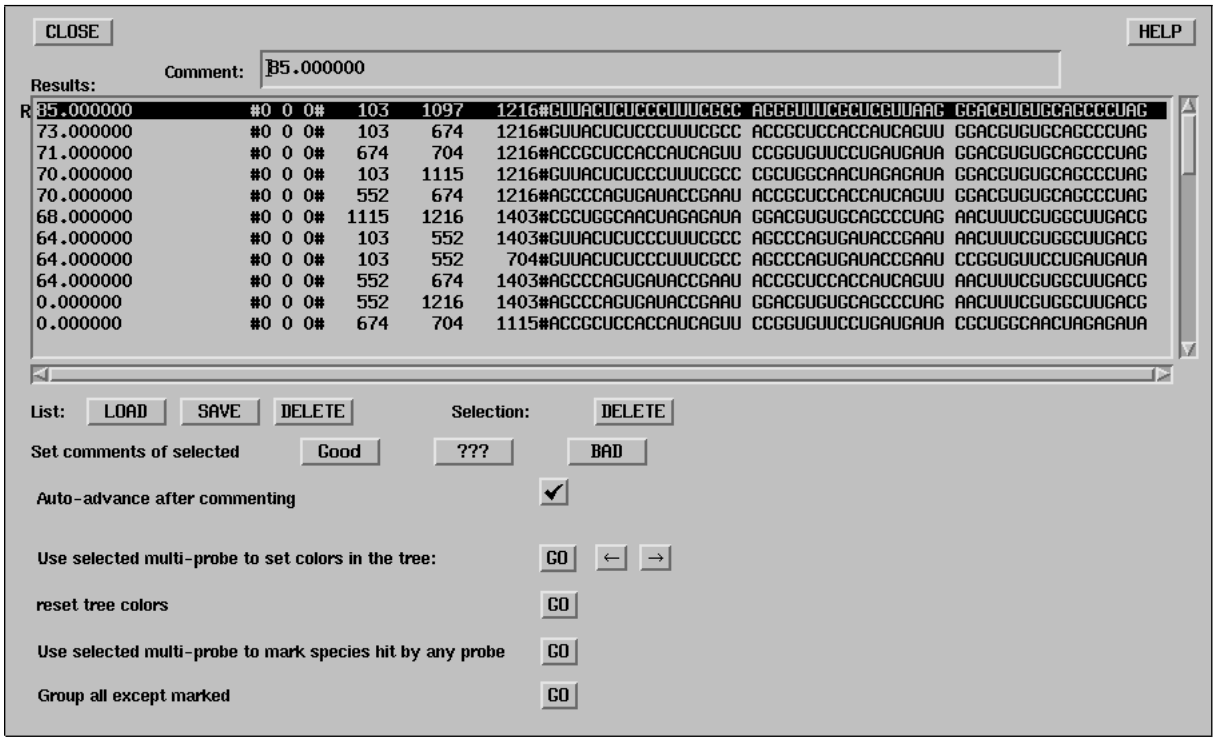



The "`Multi_Probe`" window with selected probes

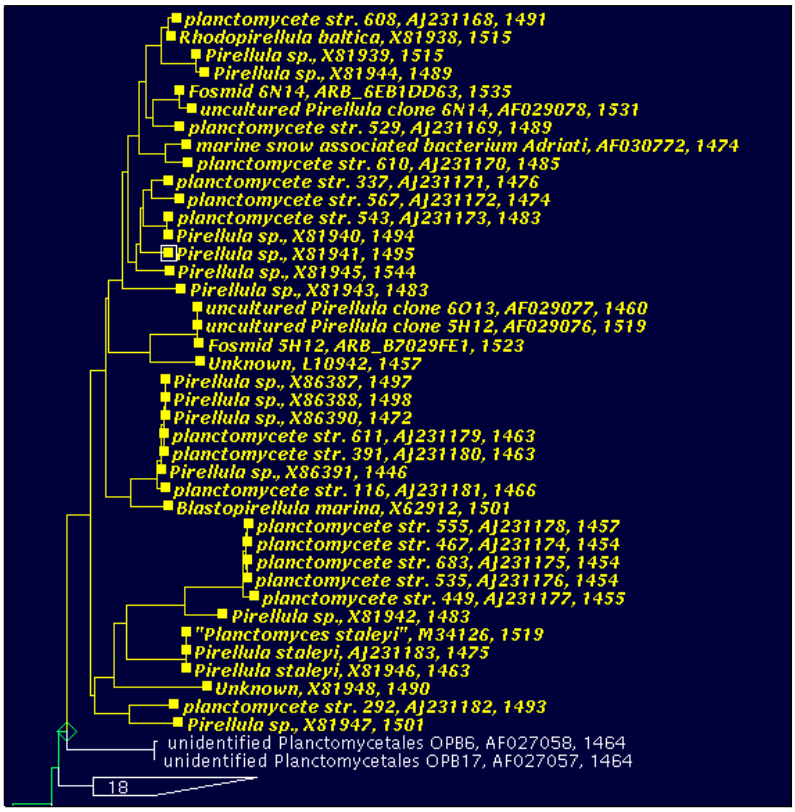
→ click on `GO`

- ARB will calculate all combinations and score them. **It is important that the sequences (here the `Pirellula` group) you are interested in are marked – and only these!!**
- the grey `Multi_Probe` combination results window will pop up, showing the score (the higher the better), the probe positions and the sequences.
- you can `LOAD`, SAVE and `DELETE` the list or single combinations and give the selected combination a comment like good or bad

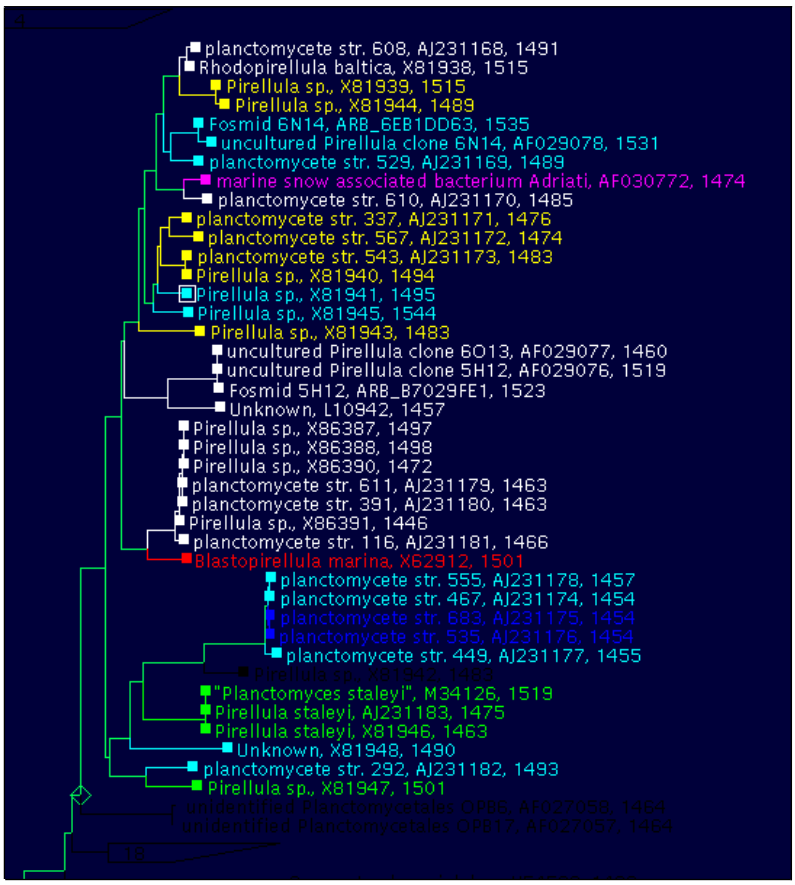
- to visualize the sequences that are covered by your probe combination click on: `Use selected multi-probe to set colours in tree GO`. By clicking on the arrow buttons you can select the next or previous probe combination for visualization.
- the three probes are shown in red, green and blue or if two or more probes hit the same sequence in a combination of the corresponding colour n the tree (e.g. red and green = yellow, red, green and blue = white). You can adjust the colours in `ARB_MAIN` window → `Properties` → `Tree settings` → `Tree colours`




The "`Multi_Probe`" result window




All selected sequences for `Probe_Design` and `Multi_Probe` design




Visualization of the probe combinations in the tree

***Description:*** The three probe combination selected in the example covers - with the exception of one sequence (black) - the complete group of interest. Most of the sequences are targeted by at least two probes (yellow, magenta, cyan), some of them are targeted by three probes (white).

***Note:*** The same procedure can be used if species or groups of sequences have to be targeted by at least two probes in different regions of the rRNA according to the multiple-probe approach. This is necessary to be sure that the hybridization results obtained really correspond to the sequences (organisms) the probe was designed for. With the incredible diversity in the environment, a single probe might give you a nice signal based on a cross reaction with an unknown organism currently not covered by the database. If two probes or even three probes give you the same results it will enhance significantly the probability of being true-positive!

# 9 Additional features in ARB

### *9.1 Generating unique IDs for the sequences (species) in the database*

***Note: This function assures that all ARB internal identifiers in the database (and stored in a field called "name"!) are unique and follow the rules of ARB! This is essential to assure consistency of your database.***

**Description:** The uniqueness of the ID (name) is primarily guaranteed by the accession number (acc field). ARB uses the public or if not available an ARB internal temporary accession number (in all cases the content of the `acc` field) to build the ARB unique database identifier (you find it in the `name` field). Also the content of the `full_name` field is involved in the naming but for better understanding rather ignore this fact. Identical content of the `acc` field should always lead to the same ARB ID (name), independent from the `full_name` field content. With the December 2007 release of ARB, the system allows to add an additional field which is taken into account to generate the ID (name). This was necessary because of the increasing amount of genome sequences that have been made available. rRNA or protein sequences from genomes are normally identified by a single accession number only, and the different genes or regions on a genome or metagenome sequence are unambiguously identified by their respective start positions on the sequence. Because initially ARB IDs (names) were based on the accession numbers alone, this lead to a huge amount of false duplicates indicated by .1, .2 etc.. To solve this issue the user is now allowed to tell the system to use an additional field for the generation of ARB names. For nearly all applications it is recommended to use the `start` field. The additional field is database specific.

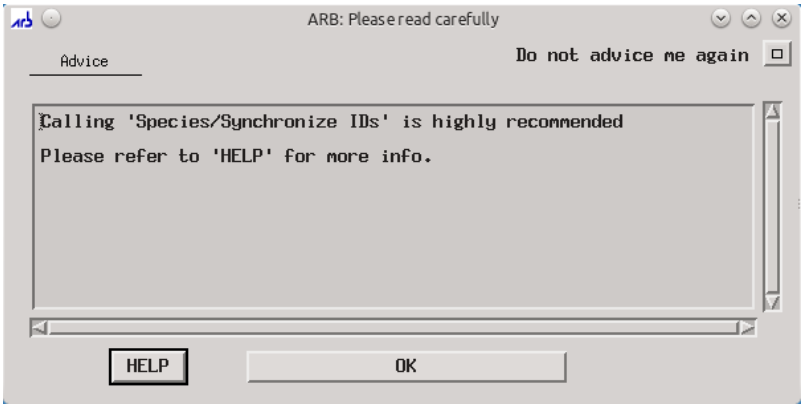
To set the additional field:

- `ARB_MAIN` window → `Tools` → `Name server admin (IDs)`

`The Name Server Admin Windows appears`


→ add “start” to the `Additional species ID field`

- after closing, the following window will pop up indicating that generating new names is now highly recommended




Proceed to generate new names:

- `ARB_MAIN` window → `Species` → `Synchronize IDs`
  → Click on `GO`

***Note:*** The "name server" file which stores the existing IDs is machine (computer) specific! Thus, if you obtain a dataset from someone who is working on a different computer and you want to merge or import sequences from this dataset into your own database, use the `Synchronize IDs` function to synchronise the unique IDs (names) of the two databases. For all SILVA databases the `start` position has been added by default as the additional field for name generation.
**If renaming fails with a corresponding message, try to delete the file "names_start.dat" (the name server file) and repeat the procedure (you find the file in the folder $ARBHOME/lib/nas where $ARBHOME is the path to the directory of your ARB installation).**

### *9.2 Exporting sequences*

For exporting sequences to foreign formats (FASTA, PHYML, RAxML, MrBayes) …

- Mark sequences to be exported
- `ARB_MAIN` window → `File` → `Export` → `Export to external format`
- The `ARB EXPORT` window appears
  → Select a format: e.g. `fasta.eft (for the simple FASTA format)`
  → Select `Export marked` to get only the marked sequences
  → Select `Filter` if you want to apply a filter to remove e.g. highly variable positions. In this case only the valid positions will be exported

→ Select `Compress` no to get an output with the full alignment or `vertical gaps` to remove all gaps that are not necessary for this subset of sequences or `all gaps` to get the unaligned sequences.

→ Choose an output file name (by default, file will be named `noname` and saved in the directory where you have started ARB)

→ `GO`

***Note:*** Sequences will be exported using the name field as the identifier in the FASTA format. To change this (or to add other fields to the header), you have to modify the corresponding export filter. You find the export filters in $ARBHOME/lib/export

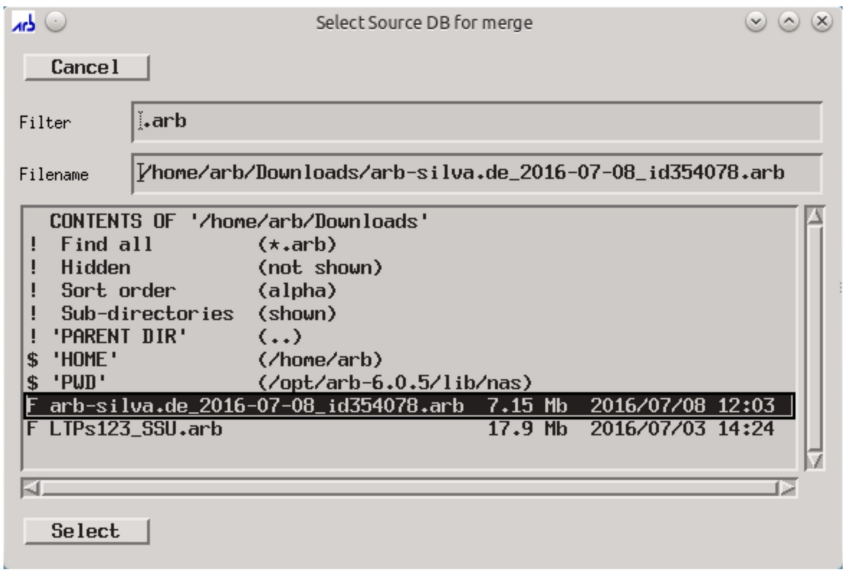
To export sequences to run **PHYML** or **RAxML** choose `phylip` as the output format.

### *9.3 Merging two ARB databases (move data from source to destination database)*

- `ARB_INTRO` window (just after starting ARB) → `MERGE TWO ARB DATABASES`


- `The Select Source DB for merge` window appears

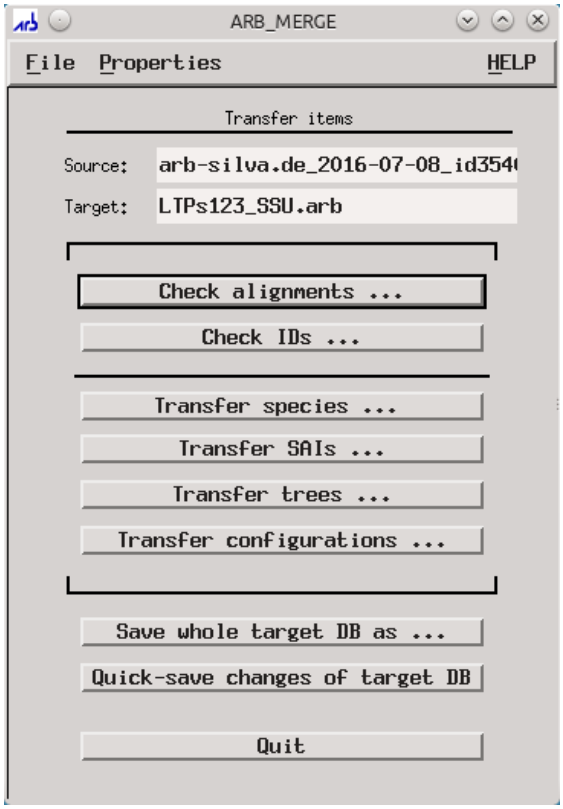
→ Select your source ARB database in the browser and click on `Select`




The `Select Source DB for merge` window

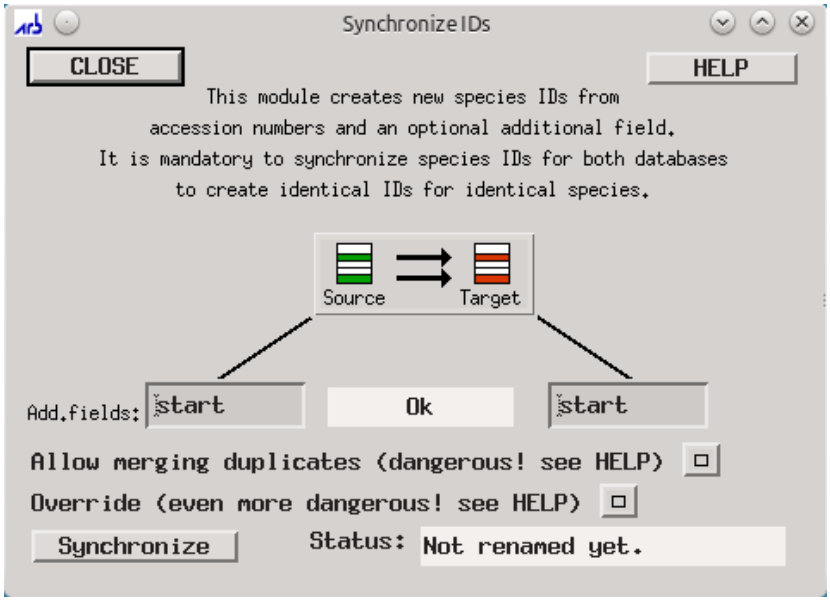
→ Select your destination ARB database in the browser of the following `Select Destination DB for merge` window and click on `Select`

- The ARB_MERGE window appears




- Before a transfer of species (complete entries) or fields is allowed, ARB forces you to generate new names for both databases. Click on → `Check IDs ...`

The `Synchronize IDs` window appears




Make sure that both databases use the same additional fields for name generation, for details see 9.1.

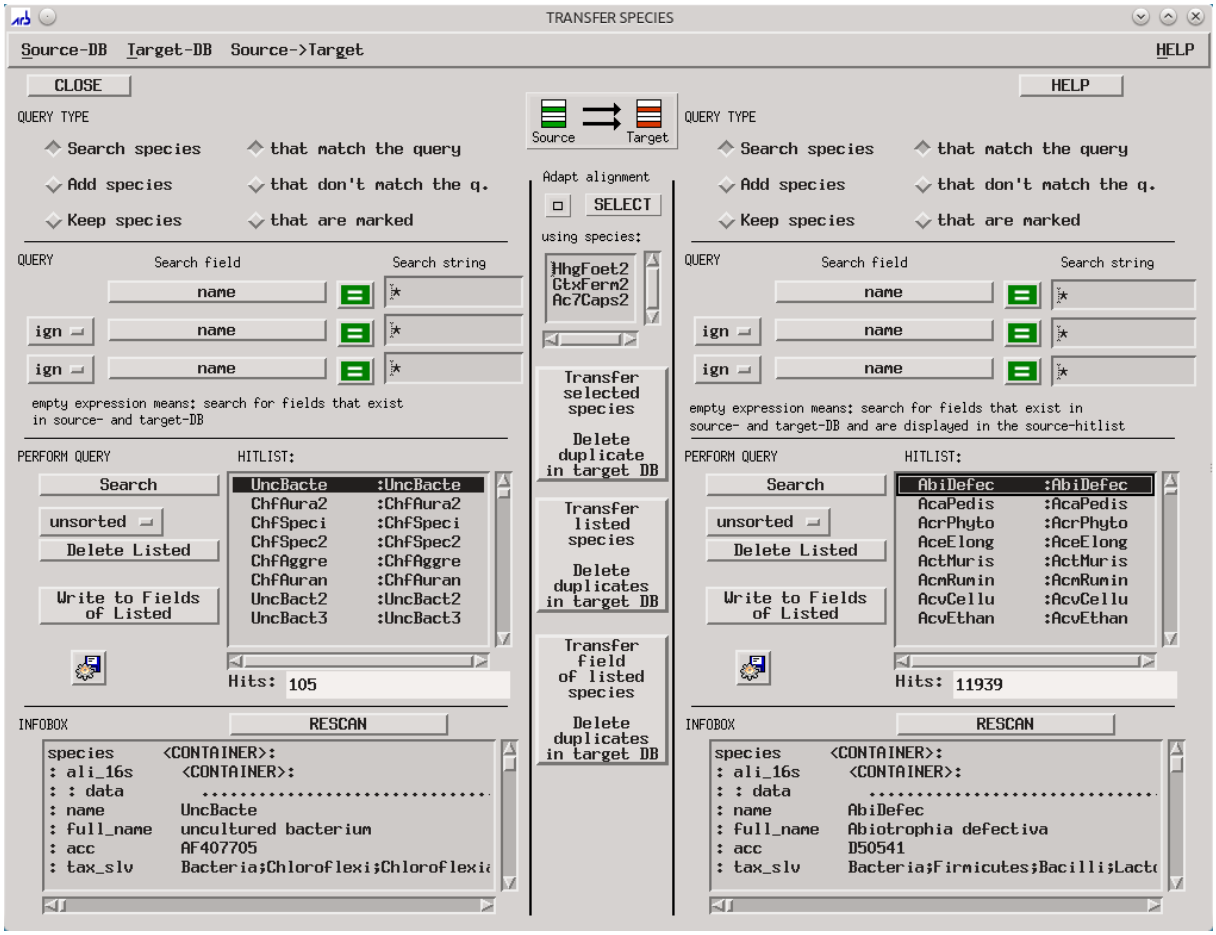
→ `Synchronize`

If no duplicates have been found the Status box shows OK and you can close the window and proceed (see next page).
In case duplicates are found and indicated in the yellow status box, you have to think why this is the case. Reasons could be a missing additional field like the start position – see 9.1, or the existence of real duplicates because of duplicated entries in your database (same accession numbers). If this is the case you should try to resolve this.
You can merge two databases containing duplicates by activating the box `Allow merging duplicates!`
In addition, you also have the option to merge databases without former renaming! Just activate the box `Override` and close the window. For some operations this is necessary like reimporting of aligned sequence information after using an external alignment tool instead of the ARB internal aligner.

***Important Note:*** ***If you are going to merge an old rRNA database with an SILVA database, the correct use of the name server is essential. Because old ARB databases do not contain neither the start field as an additional identifier for the name server nor the start field at all it is recommended to add the start field to the database and name server before starting the merge process. The start field can be created for all database entries in the*** **`ARB_MAIN`** ***window*** ( Species ( Database fields admin ( Create Fields ... . After this is done, list all entries of the database in the Search and Query menu and write 1 to the start field using Write to Fields of Listed. This solution is not perfect, since not all sequences start at position 1, but a practical 99% accurate solution.

( Select a transfer option in the ARB_MERGE window ( Transfer Species ... (alternative options: Transfer SAIs ..., Transfer Trees ..., Transfer Configurations ...)

- The `TRANSFER SPECIES` window appears




The `TRANSFER SPECIES` window

- Essentially this window offers the same functions as the Search and Query tool for both databases: Bring the species to be transferred to the HITLIST on the left.

**Note:** The combination of `Search species that don't match the query` with no search string in the search field name shows all the sequences in the HITLIST which are different between Source-DB and Target-DB.

→ `TRANSFER listed species - Delete duplicates in target DB`

→ `CLOSE`

- In `ARB_MERGE` window → `Save whole target DB as ...`
- Finally: → `Quit`

In case you get the following message “`Key 'XY' exists, but has different type`” the type of the corresponding field in your source database differs from the type in the destination database. To be able to merge sequences you have to adjust the field type. Open your source database in ARB and toggle the expert mode first (`ARB_MAIN` → `Properties`) to access the `Convert fields` … option under `ARB_MAIN` window → `Species` → `Database fields admin` (see also 5.2).

***Note:*** The option `Transfer field of listed species` allows you to transfer only a specific field of the listed species between two databases.

In our example we have supplemented a SILVA type strain dataset of 11.939 sequence entries (Target-DB) with 105 environmental sequences affiliated with a selected genus which we downloaded from the SILVA webpage in .arb format (Source-DB).

# 10 Recommended readings

**Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer.** 2004. ARB: a software environment for sequence data. Nucleic Acid Res. 32:1363-1371.

*The paper describing ARB.*

**Pruesse, E., C. Quast, K. Knittel, B. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner.** 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acid Res. 35:7188-7196.

*The paper describing the SILVA database project.*

**Peplies, J., R. Kottmann, W. Ludwig, and F. O. Glöckner.** 2008. A standard operating procedure for phylogenetic inference (SOPPI) using (rRNA) marker genes. Syst. Appl. Microbiol. 31:251-257.

*A recommended workflow for phylogenetic sequence analysis which reflects our philosophy.*

**Ludwig, W., and H. P. Klenk.** 2001. A phylogenetic backbone and taxonomic framework for prokaryotic systematics, p. 49-65. In D. R. Boone and R. W. Castenholz (ed.), The Archaea and the deeply branching and phototrophic Bacteria, vol. 1. Springer-Verlag, New York.

*A good overview over pylogenetic tree reconstruction and the philosophy behind.*

**Hall, B. G.** 2001. Phylogenetic trees made easy, a how-to manual for molecular biologists. Sinauer Associates, Inc., Sunderland, Massachusetts.

*The book gives a quick overview of the currently used phylogenetic reconstruction methods. It was originally written based on the PAUP program. If you do not want to go into the details of phylogenetic treeing this book is highly recommended.*

**Swofford, D. L., G. J. Olsen, P. J. Waddel, and D. M. Hillis.** 1996. Phylogenetic Inference, p. 407-514. *In* D. M. Hillis, C. Moritz, and B. K. Marble (ed.), Molecular Systematics, second ed. Sinauer Associates, Inc., Sunderland, Massachusetts.

*Compact and comprehensive overview, a must for advanced users.*

**Felsenstein, J.** 2004. Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, Massachusetts.

The book about phylogenetic reconstruction – use it to fill up the gaps left by Swofford.

**Behrens, S., C. Rühland, J. Inacio, H. Huber, A. Fonseca, I. Spencer-Martins, B. M. Fuchs, and R. Amann.** 2003. In situ accessibility of small-subunit rRNA of members of the domains *Bacteria, Archaea*, and *Eucarya* to Cy3-labeled oligonucleotide probes. Appl. Environ. Microbiol. **69:**1748-1758.

*The in situ accessibility paper.*

**Kumar, Y., R. Westram, S. Behrens, B. Fuchs, F. O. Glöckner, R. Amann, H. Meier, and W. Ludwig.** 2005. Graphical representation of ribosomal RNA probe accessibility data using ARB software package. BMC Bioinformatics **6:**61.

*Describes the new visualisation functions for probe accessibility and other sequence associated information in ARB.*