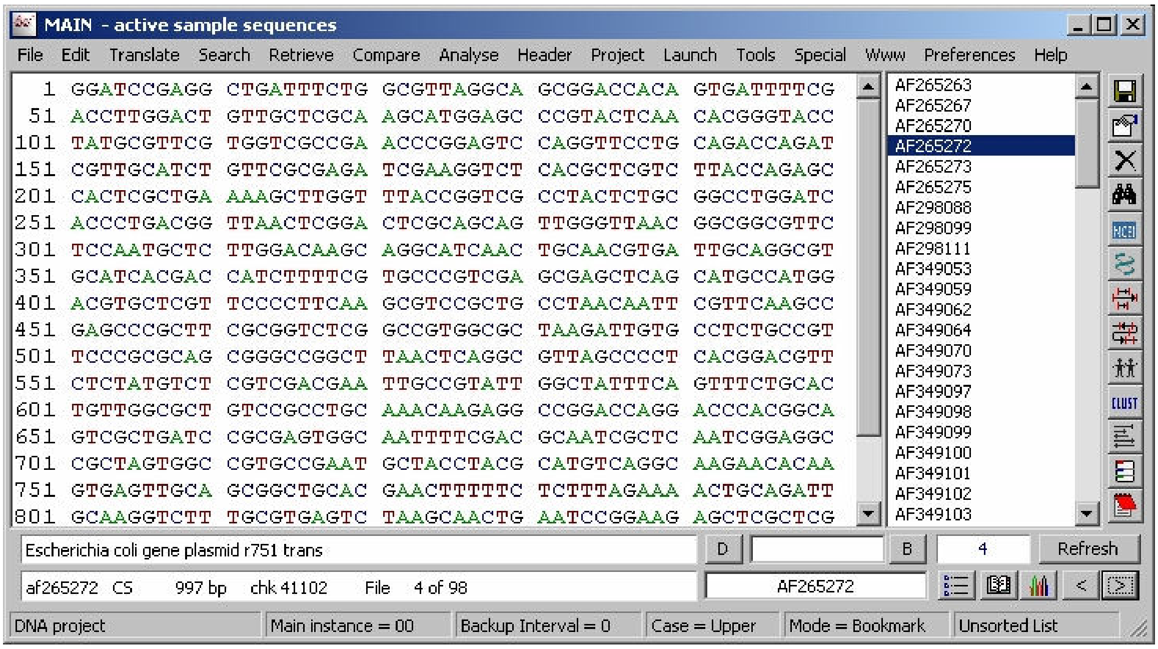
SEQ tools

...a win32 software suite for analysis and handling
of nucleotide and protein sequences

Portable Document Format (pdf) of Online Manual



```
  1 GGATCCGAGG CTGATTTCTG GCGTTAGGCA GCGGACCACA GTGATTTTCG
 51 ACCTTGGACT GTTGCTCGCA AGCATGGAGC CCGTACTCAA CACGGGTACC
101 TATGCGTTCG TGGTCGCCGA ACCCGGAGTC CAGGTTCCTG CAGACCAGAT
151 CGTTGCATCT GTTCGCGAGA TCGAAGGTCT CACGCTCGTC TTACCAGAGC
201 CACTCGCTGA AAAGCTTGGT TTACCGGTCG CCTACTCTGC GGCCTGGATC
251 ACCCTGACGG TTAACTCGGA CTCGCAGCAG TTGGGTTAAC GGCGGCGTTC
301 TCCAATGCTC TTGGACAAGC AGGCATCAAC TGCAACGTGA TTGCAGGCGT
351 GCATCACGAC CATCTTTTCG TGCCCGTCGA GCGAGCTCAG CATGCCATGG
401 ACGTGCTCGT TCCCCTTCAA GCGTCCGCTG CCTAACAATT CGTTCAAGCC
451 GAGCCCGCTT CGCGGTCTCG GCCGTGGCGC TAAGATTGTG CCTCTGCCGT
501 TCCCGCGCAG CGGGCCGGCT TAACTCAGGC GTTAGCCCCT CACGGACGTT
551 CTCTATGTCT CGTCGACGAA TTGCCGTATT GGCTATTTCA GTTTCTGCAC
601 TGTTGGCGCT GTCCGCCTGC AAACAAGAGG CCGGACCAGG ACCCACGGCA
651 GTCGCTGATC CGCGAGTGGC AATTTTCGAC GCAATCGCTC AATCGGAGGC
701 CGCTAGTGGC CGTGCCGAAT GCTACCTACG CATGTCAGGC AAGAACACAA
751 GTGAGTTGCA GCGGCTGCAC GAACTTTTTC TCTTTAGAAA ACTGCAGATT
801 GCAAGGTCTT TGCGTGAGTC TAAGCAACTG AATCCGGAAG AGCTCGCTCG
```

```
AF265263
AF265267
AF265270
AF265272
AF265273
AF265275
AF298088
AF298099
AF298111
AF349053
AF349059
AF349062
AF349064
AF349070
AF349073
AF349097
AF349098
AF349099
AF349100
AF349101
AF349102
AF349103
```

Escherichia coli gene plasmid r751 trans

4

af265272 CS 997 bp chk 41102 File 4 of 98

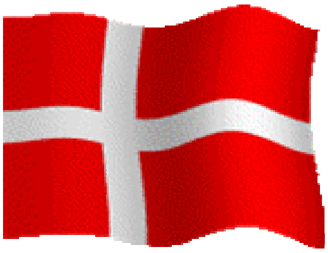
AF265272

DNA project | Main instance = 00 | Backup Interval = 0 | Case = Upper | Mode = Bookmark | Unsorted List

NCBI

works with ncbi

Dr. Soeren W. Rasmussen DSc., PhD., MSc



March 2007

Soeren. W. Rasmussen

education:	DSc., PhD., MSc.
phone:	+45 3616 2259 +45 6170 2259
e-mail:	swr@seqtools.dk
homepage in Danish	www.swr.dk

For more than 30 years I have been engaged in research in the ultra structure of meiotic cells, immunolocalisation of topoisomerase II in meiotic cells, sequencing the yeast genomic, characterizing the *Blumeria graminis* genome by sequencing expressed sequence tags and working in Bioinformatics in a major Danish medical company. Over the past 10 years, a large number of web based services has become available adding analysis possibilities to those provided by an equally large number of stand-alone applications carrying out a number of specific tasks. With an sequence collection of just a few thousand, a major problem is to manage, analyze and compare sequences in a rational fashion - and to keep the analyses updated. For an individual scientist or a small lab without access to a professional data department sequences stored on the hard drive of a PC as single files comprise an almost unmanageable problem. This prompted me to embark on writing the first software package, dnatools, which was intended to serve as a local sequence depository allowing the user to perform a variety of analyses using third party programs and to maintain an updated overview of the data. As of May 1., 2002 I moved from the Carlsberg Laboratory to a position at H. Lundbeck A/S in Bioinformatics. The copyright to dnatools belong to Carlsberg A/S and leaving Carlsberg at the same time put an end to my work with dnatools. The source code, was handed over to Carlsberg A/S and the support of the software package was terminated. Since then I have written a new software package, SEQtools which was published on the internet on October, 2002. The new package includes new versions of many of the functions previously included in DNATools but also a number of new functions related to various aspects of micro array design.

Welcome to SEQtools

In response to requests from several users of seqtools a new version of the program has been included on the download page: Seqtools School Edition intended for classroom teaching in elementary school.

The setup file installs two versions of seqtools: a limited version including only very basic functions and a full version of the seqtools package.

general features

SEQtools 8.3 is a win32 software package for handling and analysis of nucleotide and protein sequences. The program includes a series of trivial functions to help you carry out common operations. In addition SEQtools will assist you with more demanding tasks like unattended batch blast search at NCBI. SEQtools includes advanced facilities for retrieving, storing, handling and listing search results.

special features

Special functions are included for design of micro array gene expression analysis experiments, for expression analyses with the SAGE procedure and for managing small EST projects. Utilities are included for primer design and ordering, renaming files, creating codon usage tables, building local searchable databases, aligning nucleotide and protein sequences, comparing sequences and a lot more...

user interaction

SEQtools is a very responsive software package. User comments and suggestions are highly appreciated and play a key role in keeping the program bug-free and up to date. You can use SEQtools free of charge for as long as you wish if you keep your registration alive by confirming the registration every 60 days.

Visit www.seqtools.dk regularly to stay updated

SEQtools downloads

Seqtools version 8.3 is now available for users with a full license.

No further development will take place in version 8.2 except for correction of critical errors and problems.

setup or update ?

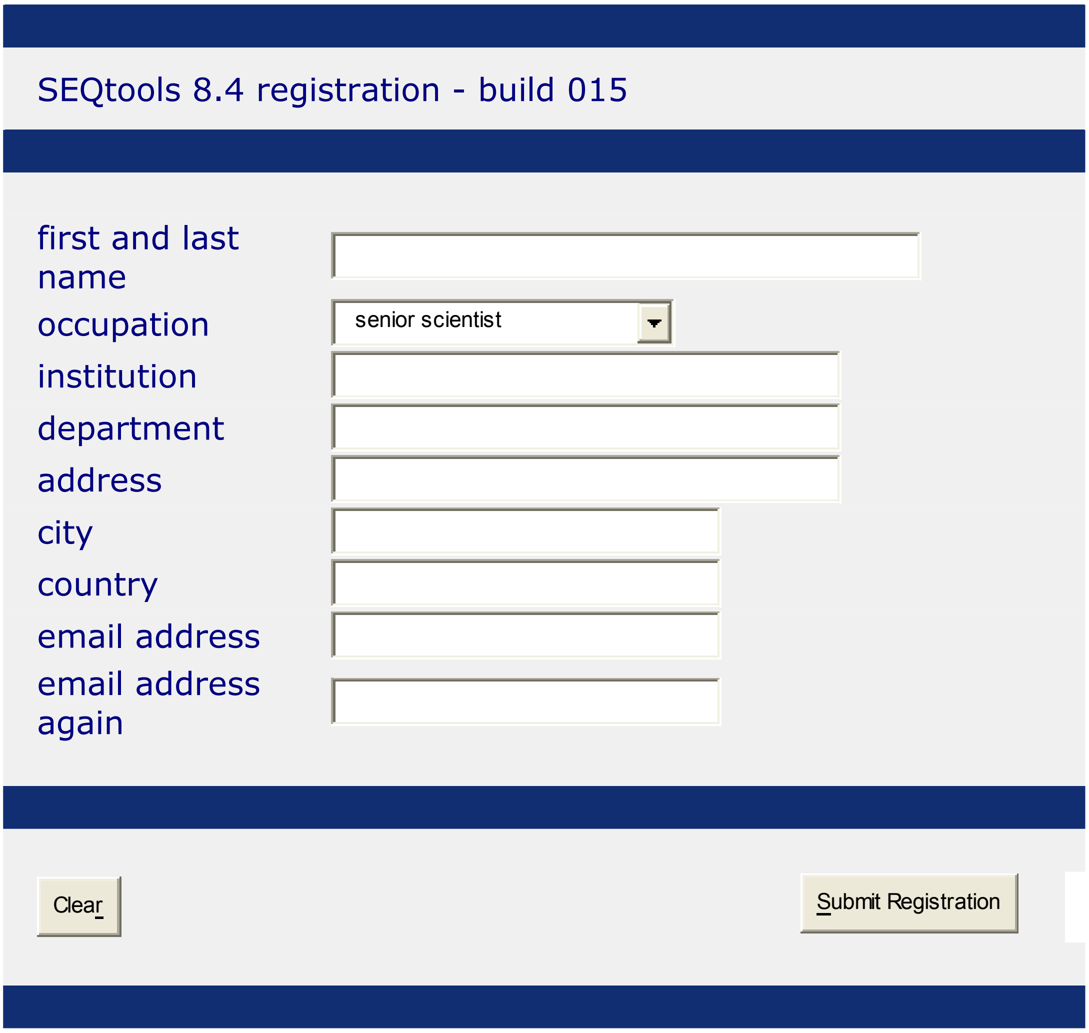
The download page contains files both for installing seqtools for the first time and files for updating an existing installation without performing a full installation of the program. You will also find links to supplementary stand-alone programs which communicate directly with seqtools.

how Do I keep my registration alive ?

You can use SEQtools 8.2 free of charge for as long as you wish if you keep your registration alive by confirming the registration every 60 days.

when can I expect to receive my new license agreement ?

Even though I normally issue licenses the same day I receive your request I may be out of my office and thus unable to renew your license immediately after receiving the request. In case you haven't received your license agreement within a few days please send me an email and explain the situation.



SEQtools 8.4 registration - build 015

first and last name	
occupation	senior scientist
institution	
department	
address	
city	
country	
email address	
email address again	

how do I register ?

To register your free copy of SEQtools for the first time or to extend an existing registration, fill in the form above and click the "Submit Registration" button.

Please note that all registration applications are read by myself. Forms not including the required information are rejected based on the assumption that people not capable of filling out a simple form will not be able cope with this fairly advanced software package.

If you wish to purchase a full license which covers all versions of seqtools, never expires and does not require repeated renewal, please visit the on-line payment page.

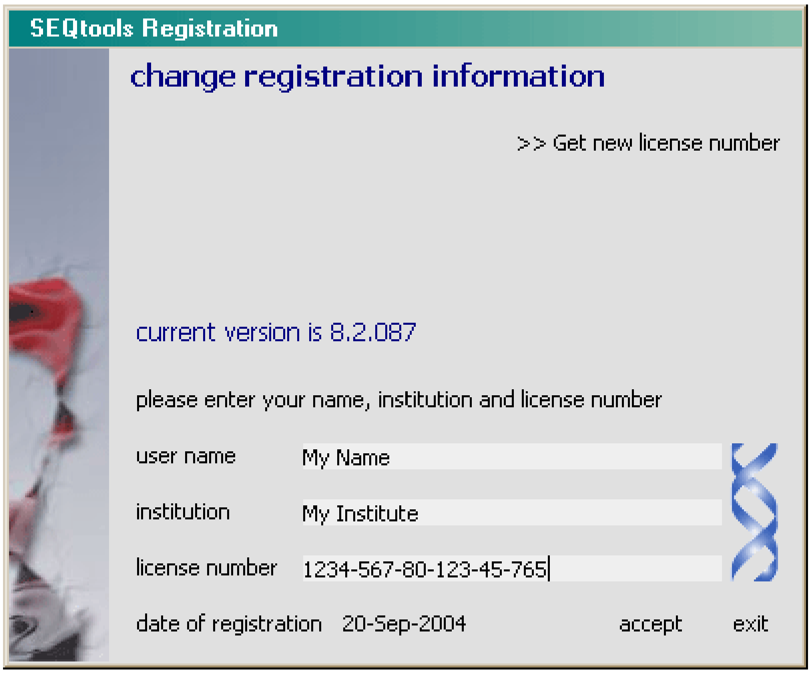
when do I get my license ?

In most cases you will receive a registration agreement from me by email the same day. If you do not receive your registration agreement within a few days the information you have entered in the registration form may have been incomplete or incorrect - or the registration agreement may have been trapped in spam/virus filters on your mail server. In such events send me an email explaining the situation.

enter license key in seqtools?

When you have received the email with your user name and reg-key, start seqtools and open the seqtools registration form (Help/Registration/Enter Registration Information). Type - or even better - copy/paste user name, institution and the registration key in the relevant fields of the form and click accept.

Be careful to enter the registration information EXACTLY as it appears in the registration agreement. Unsuccessful registration in seqtools is in most cases due to misspelled user name and/or incorrect registration key. Note that the user name is both case and space sensitive.



SEQtools Registration

change registration information

current version is 8.2.087

please enter your name, institution and license number

user name My Name

institution My Institute

license number 1234-567-80-123-45-765

date of registration 20-Sep-2004

SEQtools manual

The seqtools web manual is completely rewritten and now provides relevant descriptions of functions and facilities included in the SEQtools program suite.

The context sensitive help build into the program is will be revised in the near future.

viewlets - animated help

Realizing that seqtools may not be the easiest program to get familiar with I have begun writing a series of animated ViewLets describing different aspects of the program. Have a look at the two first Viewlets describing how a project is created and how to perform a batch blast search at NCBI.

Note added October 2005: please note that the viewlets were generated before several updated were implemented.

how do I get help when I'm lost?

The seqtools web manual has been completely rewritten and should be consulted first. When you fail to find information on a specific subject - or the information in the manual is incorrect - don't hesitate to contact me and ask. Or post your question on the discussion board.

Visit www.seqtools.dk regularly to stay updated

1. INTRODUCTION

1.1 about seqtools

SEQtools 8.3 is a win32 software package for handling and analysis of nucleotide and protein sequences. The program includes a series of trivial functions to help you carry out common operations. In addition SEQtools will assist you with more demanding tasks like unattended batch blast search at NCBI. SEQtools includes advanced facilities for retrieving, storing, handling and listing search results.

1.1.1 Special features

Special functions are included for design of micro array gene expression analysis experiments, for expression analyses with the SAGE procedure and for managing small EST projects. Utilities are included for primer design and ordering, renaming files, creating codon usage tables, building local searchable databases, aligning nucleotide and protein sequences, comparing sequences and a lot more. Recently an option to export sequence data to a ms excel spreadsheet has been included.

1.1.2 User interaction

SEQtools is a very responsive software package. User comments and suggestions are highly appreciated and play a key role in keeping the program bug-free and up to date. You can use SEQtools free of charge for as long as you wish if you keep your registration alive by confirming the registration every 60 days.

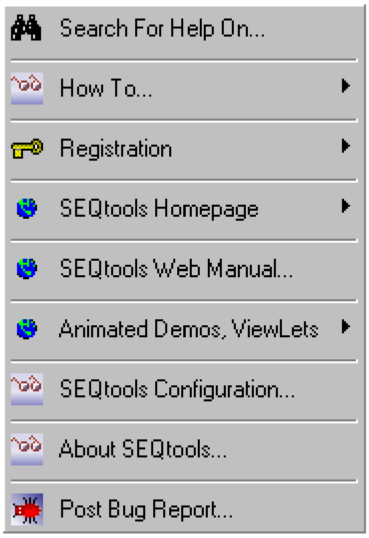
1.2 seqtools help sources

SEQtools does not come with a printed manual. As the whole SEQtools organization consists of a single person it is simply not possibly to maintain the code, the context sensitive help and the web help. Although I try to keep the context sensitive help which is build into the program up-to date, the updating usually lags several revisions behind. Pressing F1 brings up context sensitive help information relating to the currently active program item.

The SEQtools homepage includes a fairly comprehensive manual which is currently being revised to cover the latest changes to the program. I will attempt to maintain this source of help information up-to date with relevant illustrations covering the different topics.

1.3 registration and licenses

You can access the SEQtools registration form either from the program as shown below or by visiting www.seqtools.dk



Providing SEQtools to users free of charge has the dual advantage that users all over the world get free access to a fairly comprehensive software package for sequence handling and analysis. In return I get information about bugs and receive useful user input in the form of suggestions and comments from a large number of users.

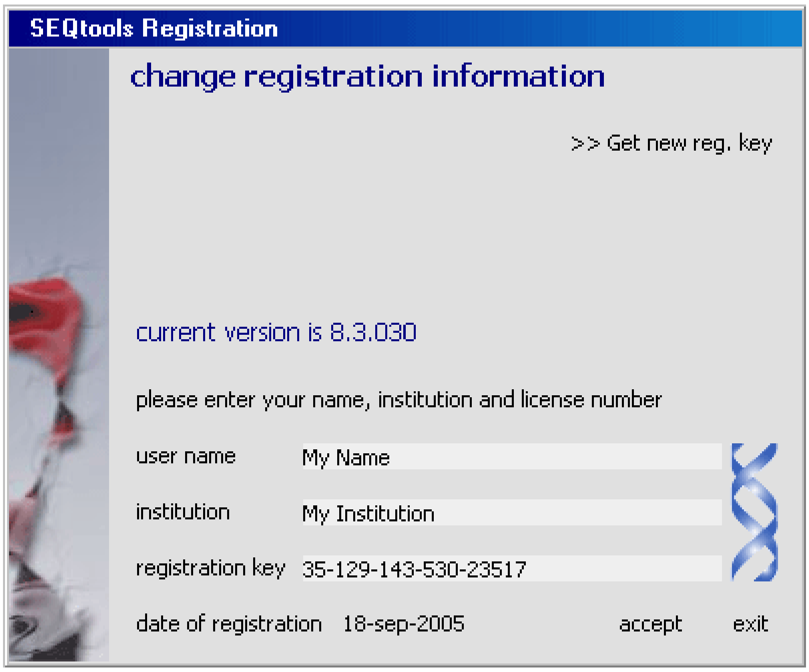
The difficult economic situation of many students and scientist in third world countries is an additional argument for making the use of SEQtools free of charge. The only condition for the free access to SEQtools is that users are requested to register after a testing period of 60 days and there after to keep their registration alive by renewing their license every 60 days.

Old users of SEQtools already know that SEQtools is updated very frequently. Unlike most other authors of software packages I prefer to correct bugs right away and upload the corrected version. This used to create the problem that users often complained about bugs that were already corrected but not yet downloaded on their pc.

Recently I have included an "update-tester" in SEQtools. Every time you start SEQtools it visits the download page to see if new updates are available - and notifies you if there are. You may experience that your license no longer works after upgrading to a newer version of SEQtools.

In this case you just have to renew your license to cover the upgraded version.

The user name and the registration key is entered in the form shown below. Note that this information must be entered *exactly* as in the license agreement. The user name is case and "space" sensitive. Entering incorrect information will terminate SEQtools immediately. You can extended your license at www.seqtools.dk or by sending an email to me.



SEQtools Registration

change registration information

current version is 8.3.030

please enter your name, institution and license number

user name: My Name

institution: My Institution

registration key: 35-129-143-530-23517

date of registration: 18-sep-2005

1.4 user interaction

Seqtools has evolved in close association with its users. Numerous users have contributed significantly to the program by suggesting new functions to be included in the suite and - not least - by testing functions and reporting the result to the author.

As SEQtools is maintained by a small organization there is a very short distance between coding a program revision and the publication of the update. This has the advantage that bug fixes are made available to the users very rapidly, usually the same day the bug is reported.

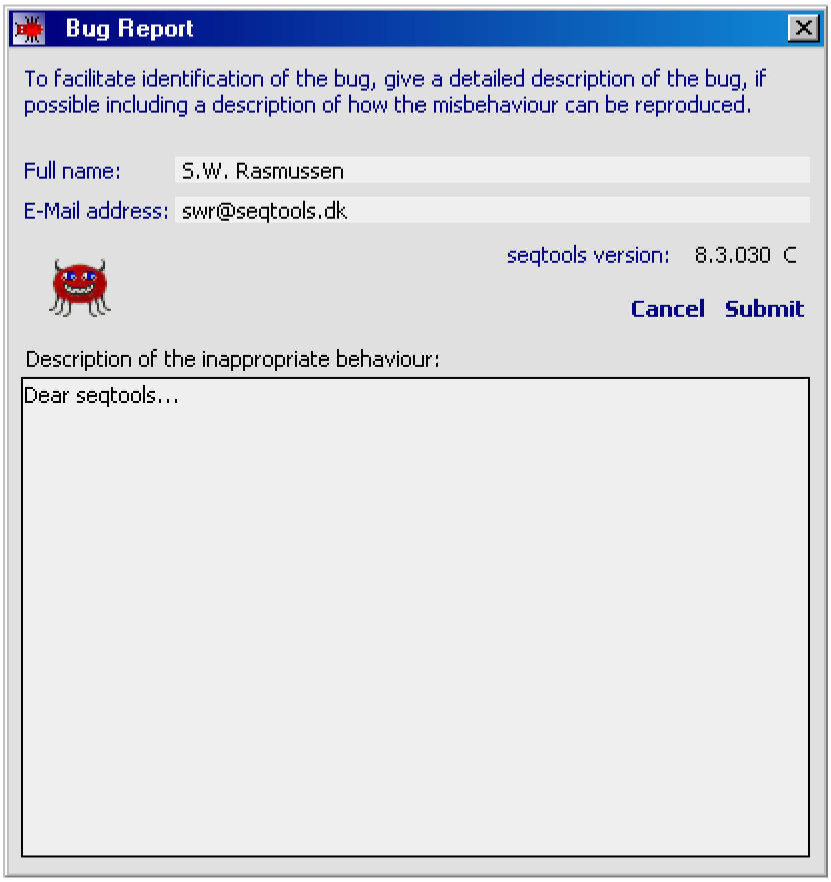
The disadvantage of the frequent revisions is that you need to update the program often. Each time SEQtools is opened it will contact the download page on the web to check if an update is available. If a revision is available you are informed as SEQtools loads.

It is strongly recommended that you update your SEQtools installation when a new update is available. As the auto-update process does not require reinstallation of SEQtools I believe that this is a minor inconvenience to ensure that you always work with a version of SEQtools without known bugs or other problems.

1.5 support - bug reports

You can find the latest additions and corrections to SEQtools on the revision history section of the homepage. As the last resort write an email to me describing the problem (please include as many details as possible) and I will do my best to assist you.

It is also possible to submit a bug report directly from SEQtools. Look under the help menu to load the bug report form.



To facilitate identification of the bug, give a detailed description of the bug, if possible including a description of how the misbehaviour can be reproduced.

Full name: S.W. Rasmussen

E-Mail address: swr@seqtools.dk

seqtools version: 8.3.030 C

Description of the inappropriate behaviour:

Dear seqtools...

2. SEQTOOLS INSTALLATION

2.1 download and setup seqtools ver. 8.3

2.1.1 Complete setup file (including excel.exe)

The complete setup file, *seqtools83ex.msi*, contains in addition to the basic SEQtools application and the auxiliary and emboss components, ms excel. After downloading the installation file (approximately 15 Mb) double-click the seqtools83ex.msi file to install SEQtools.

2.1.2 Seqtools core setup file

The core SEQtools setup file *seqtools83.msi* contains all files necessary to install SEQtools application without spreadsheet support (ms excel not included). After downloading the setup file (approximately 10 Mb) double-click the seqtools83.msi file to run the installation program.

2.1.3 Older versions of Seqtools

Seqtools is developed and tested on Windows 2000 and Windows XP operating systems. Users not using either of these operating systems may not be able to run SEQtools versions 8.2 or 8.3.

Version 8.0 of SEQtools can be installed using, *seqtools80.msi* (8.0.804) and 8.2 using *seqtools82c.msi* (8.2.094).

Note, however, that none of these version are supported. You are of course welcome to contact me in case you run into problems but do not expect too much...

2.1.4 Registration and license

When you install SEQtools for the first time you automatically get a students license which is valid for 60 days. When the students license expires you can extend the license for a new 60 days period by renewing the registration (version 8.0 and 8.2 only). You can continue renewing your license for as long as you wish. Using ver. 8.3 requires a full license.

Note, however, that you cannot upgrade beyond the version number covered by your license. In case you wish to upgrade to a newer version of SEQtools you must first get a new license - even if the 60 day period has not yet expired. You can read more about registration and licenses on the registration page.

2.2 auxiliary components

If you wish to update functions depending on NCBI support, trace file processing and viewing and the functions using emboss programs, you can do this by downloading the self-extracting *auxiliary8.exe* and *emboss8.exe* files and install the components from the SEQtools *Help/SEQtools Configuration* menu as described below.

The auxiliary8.exe file contains the following programs:

2.2.1 Clustal

...is required for multi-sequence alignment. ClustalX is a stand-alone program launched from SEQtools with the selected sequences as parameters. ClustalW is a command line DOS program entirely controlled by the SEQtools user interface. Version 1.83 of clustalx/w is included in the auxiliary file available for downloading. The clustal programs were written by:

- Toby Gibson *EMBL, Heidelberg, Germany.*
- Des Higgins *UCC, Cork, Ireland.*
- Julie Thompson *IGBMC, Strasbourg, France.*
- Francois Jeanmougin *IGBMC, Strasbourg, France.*

2.2.2 Blastall, Formatdb, bl2seq, Blastclust, Fastacmd

...are required for a number of tasks all depending on this collection of utilities made available by NCBI. The tasks include creating and searching local databases with the five blast programs, comparing sequences, performing batch blast searches at Genebank etc. The current version of the NCBI programs is 2.2.11.

2.2.3 Blastcl3

...is required for database searching on Genbank databases at NCBI. In some cases there may be a problem if you are behind a firewall. Consult the NCBI blastcl 3help file and/or your system administrator for advice.

2.2.4 Entrez

...is required for retrieval of sequence records and Medline entries from Entrez

2.2.5 Convert_trace

...is required to extract and import chromatograms generated by the most common auto-sequencers. In addition to convert_trace, the two dll's: read.dll and zlib.dll are necessary for this function. Convert_trace is part of the Staden package.

2.2.6 Chromas

...is required for viewing chromatogram files. Note that versions of chromas earlier than 162 are freeware whilst newer versions require registration and a license fee. Read more about chromas on the chromas website.

2.2.7 t-Coffee

...is required for optimization of sequence alignments generated by clustalw. Note that t-coffee is extremely greedy with respect to RAM resources. If the amount of free RAM is insufficient t-coffee stalls and fails to optimize the alignment. The auxiliary file contains version 1.37 of t-coffee. Read more about t-coffee on the t-coffee website.

2.3 emboss programs

The four Emboss programs interfaced with SEQtools are Windows versions of selected programs from the Emboss package. The programs are from the distribution made available by Andria Blavier and include version 2.7.1-0.7, September 2004 of the package.

2.3.1 Fuzznuc, Fuzzpro, Fuzztran

...are required for nucleic acid pattern search, protein pattern search and protein pattern search after translation - respectively.

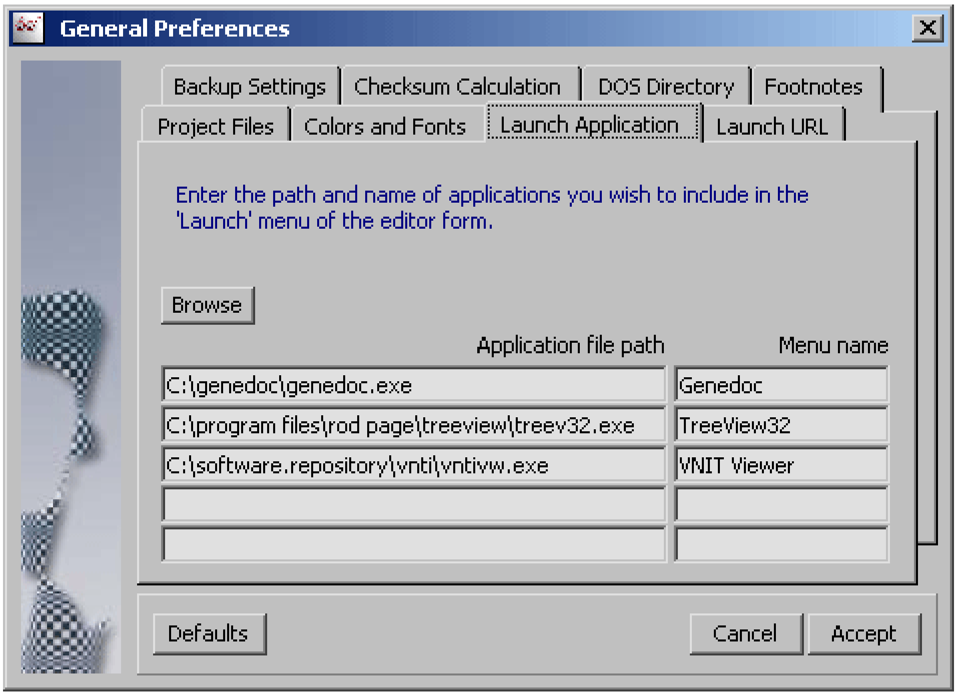
2.3.2 Merger

...is required for merging two overlapping nucleotide sequences.

2.4 interfaced programs

2.4.1 How to associate external programs to seqtools

To associate a 3rd party program to SEQtools open *Preferences/General Preferences/Launch Application*, place the cursor on the first empty line of the list, click *Browse* to find the location of the program you wish to associate and click *Enter* to include the application in the list. As described below GeneDoc and TreeView communicate directly with SEQtools while other programs just opens when their name on the Launch menu is clicked.



Enter the path and name of applications you wish to include in the 'Launch' menu of the editor form.

Application file path	Menu name
C:\genedoc\genedoc.exe	Genedoc
C:\program files\rod page\treeview\treev32.exe	TreeView32
C:\software.repository\vnti\vntivw.exe	VNIT Viewer

2.4.2 GeneDoc

GeneDoc is a powerful editor which allows you to manually edit and add a wide range of attributes to multi-sequence alignments generated by Clustal W. To make GeneDoc available to SEQtools download the program from the GeneDoc homepage , run the setup file and tell SEQtools where to find the genedoc exe-file as described above.

2.4.3 TreeView

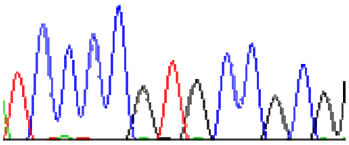
TreeView is a simple program for displaying phylogenies on Windows PCs. It has the following features:

1. - reads many different tree file formats (including NEXUS, PHYLIP, Hennig86, NONA, MEGA, and ClustalW/X)
2. - supports standard the TrueType and Postscript fonts
3. - supports native graphics file format (Windows metafile) for copying and saving
4. - print preview
5. - can print multiple trees per page, and one tree over more than one page
6. - has drag-and-drop facility for easy opening of files
7. - includes access to Web-based online help
8. - includes facilities for editing trees

To make TreeView available to SEQtools download the program from the TreeView homepage, run the setup file and tell SEQtools where to find the exe-file as described above.

2.4.4 Chromas

The interaction between SEQtools and Chromas is entirely handled by SEQtools and does not require that Chromas is included in the *Launch* menu as described above unless of course you wish to have direct access to this application outside the SEQtools control.

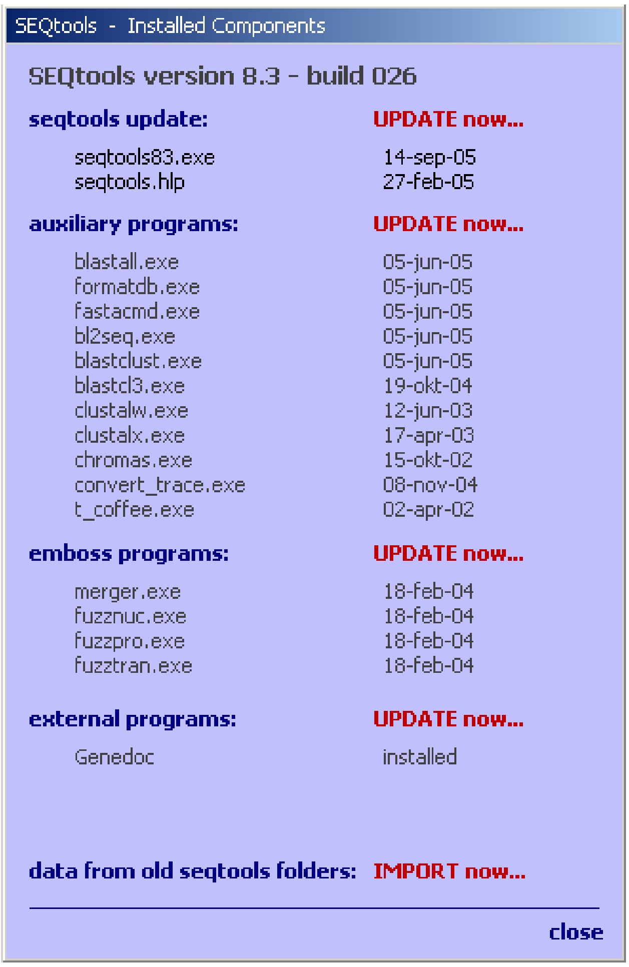


Chromas version 223 is included in the auxiliary8.exe file and the data transfer between SEQtools and Chromas is automatically established when the auxiliary programs are installed. Visit the Chromas homepage to read more

2.5 verify installation

2.5.1 Installed components

This form, *Help/SEQtools Configuration* lists currently installed auxiliary and emboss components and their file dates. Installed external supported are listed in the lower part of the form. If items are missing they can be downloaded and installed by clicking the "*UPDATE now...*" fields of the form.



SEQtools - Installed Components

SEQtools version 8.3 - build 026

seqtools update:	**UPDATE now...**
seqtools83.exe	14-sep-05
seqtools.hlp	27-feb-05
auxiliary programs:	**UPDATE now...**
blastall.exe	05-jun-05
formatdb.exe	05-jun-05
fastacmd.exe	05-jun-05
bl2seq.exe	05-jun-05
blastclust.exe	05-jun-05
blastcl3.exe	19-okt-04
clustalw.exe	12-jun-03
clustalx.exe	17-apr-03
chromas.exe	15-okt-02
convert_trace.exe	08-nov-04
t_coffee.exe	02-apr-02
emboss programs:	**UPDATE now...**
merger.exe	18-feb-04
fuzznuc.exe	18-feb-04
fuzzpro.exe	18-feb-04
fuzztran.exe	18-feb-04
external programs:	**UPDATE now...**
Genedoc	installed
data from old seqtools folders:	**IMPORT now...**

2.5.2 Updating seqtools and external components

Seqtools looks for new updates when you start the program and notifies you if an update is available. New updates can be downloaded and installed from *Help/SEQtools Configuration*. You have to exit and restart SEQtools in order to install the updates when downloading is completed.

To manually update SEQtools or one of the auxiliary components simply click the relevant *UPDATE now...* field on the form. Look at the download page on the web for more details.

2.5.3 Install/update external programs from seqtools

After installing SEQtools start the program and click *Help/SEQtools Configuration...* to display the current configuration of SEQtools. Then simply click the relevant *UPDATE now...*field to download and install auxiliary (or emboss) programs.

2.5.4 Install/update external programs manually

If this fails (for example because you are behind a firewall) it is possible to carry out the installation/updating of auxiliary8 and emboss8 programs manually:

- - download the auxiliary8.exe and emboss8.exe files from the SEQtools download page
- - unzip the self-extracting files accepting the default destination *c:\~seqtools* suggested by winzip
- - start SEQtools and accept to install auxiliary and emboss files from a local folder

2.5.5 License and registration

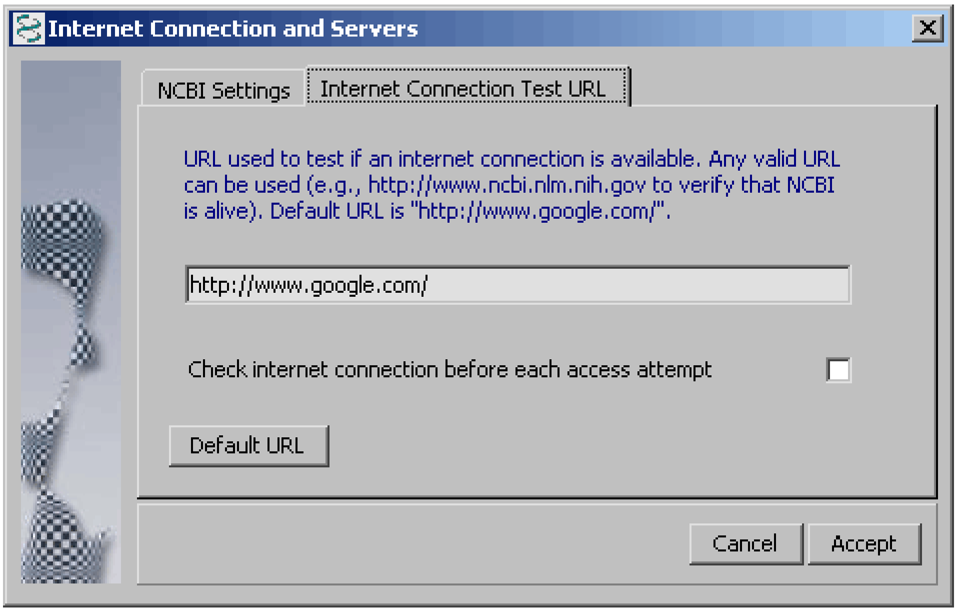
WARNING - Replacing your current version of SEQtools with an update *not* covered by your license will cause SEQtools to stop working. In this event, use the registration form to get a new registration key allowing you to use the update.

2.5.6 Importing old SEQtools data

In case you want SEQtools to search for existing data and components on your hard disk from a previous installation, simply click the *import user data* field. This will start a search/import facility copying old data to your current data folders, default location: *c:\windows folder\ST8_TEMP\*

2.6 internet access

When SEQtools loads it checks whether or not a live connection to the Internet can be detected. This check involves an attempt to connect to the URL specified in the form shown below. The default URL is *www.google.com* but can be altered if you prefer a different URL.



URL used to test if an internet connection is available. Any valid URL can be used (e.g., http://www.ncbi.nlm.nih.gov to verify that NCBI is alive). Default URL is "http://www.google.com/".

http://www.google.com/

Check internet connection before each access attempt □

2.7 system requirements

2.7.1 Computer

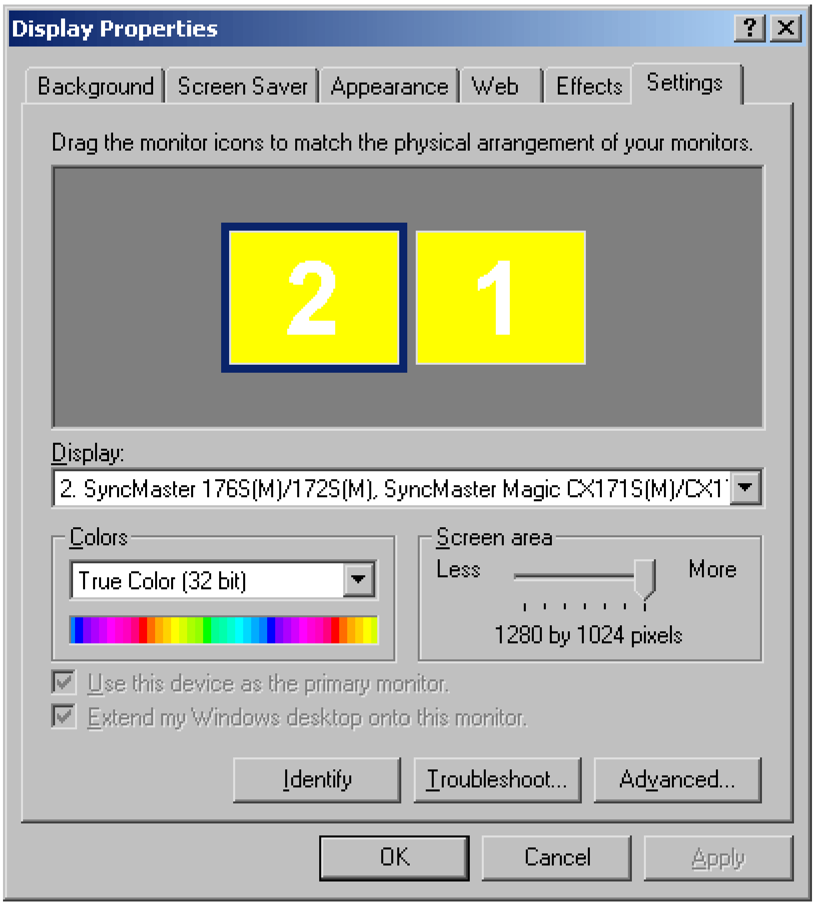
Seqtools is designed to run on 32bit Windows based PC's and it is not possible to run the program on Macs, Unix, Linux and other non-Windows operating systems. However, according to some users SEQtools runs fine under *Virtual PC* on Macintosh.

Obviously the program runs more efficiently the more powerful the computer is. A Pentium3 600 Mhz with 256 Mb RAM works well for small projects, i.e., 1-500 sequences each of 500 bp. If you intend to work with larger projects, it is a good idea to add more RAM to your PC. Currently we are using a dual processor 2.8 Ghz Pentium IV equipped with 2 Gb RAM.

If you have problems running SEQtools on your PC let me try to help you. Write an email with as many details as possible describing the nature of the problem.

2.7.2 Dual monitor display

With a large number of forms SEQtools benefits from computers set up to use two monitors. It is a major advantage when analyzing sequences with SEQtools to split the tasks on two separate monitors. Both Win2000 and WindowsXP operating systems handle this smoothly without requiring additional drivers - except an extra screen card.



Drag the monitor icons to match the physical arrangement of your monitors.

Display:
2. SyncMaster 176S(M)/172S(M), SyncMaster Magic CX171S(M)/CX1

Colors: True Color (32 bit)

Screen area: 1280 by 1024 pixels

☑ Use this device as the primary monitor.

☑ Extend my Windows desktop onto this monitor.

2.8 firewalls

Firewalls installed on local servers may in some cases prevent SEQtools from communicating with external resources such as the NCBI blast server. While I will try to assist you as much as I can, it often turns out that such problems are very difficult to solve. The NCBI blastcl3 help file may be helpful in setting up the communication between SEQtools and NCBI programs through a firewall.

3. SEQTOOLS FEATURES

This page contains a number of general topics which could not conveniently be included under any of menu item title captions. In many cases more detailed/supplementary descriptions are found in one or more of the following pages.

3.1 organization of the user manual

3.1.1 Introduction

This major revision of the seqtools manual comprise a complete reorganization and rewriting of most topics of the manual including new screen shots of all seqtools forms. A long time has passed since the first version of the seqtools user manual was written. Since then a number of minor revisions have been made to the user manual in an attempt to cover new additions and modifications to the program. However, despite these efforts the application and its documentation now has diverged to an extent where major parts of the manual described features no longer relevant - and failed to mention important additions to seqtools.

As it is not nearly as interesting to write documentation as it is to build new facilities for the application this major revision has been postponed for a long time. The current manual was written February 2005 and hopefully will last for some time.

3.1.2 Organization of the manual

Apart from the first three sections (1. Introduction, 2. Installation, 3. Features) and the last section (16. Primer) of the manual, the description of the various seqtools functions and facilities strictly follows the menu structure of the main editor form (section 3.5 below). This may not be the most optimal arrangement for the user, but hopefully makes it easier for me to keep the manual up-to-date in the future.

3.1.3 How to use the manual

Access to topics covered by the manual is by menu item caption of the main seqtools editor form. This retrieves in most cases a single page containing descriptions of all sub-topics included under the main topic. In some cases additional pages were necessary to cover special items which could not conveniently be contained on a single page.

The disadvantage of this organization is that finding documentation to items not immediately identifiable by the menu or sub-menu caption is difficult. In such cases the context sensitive help may help guiding you towards the relevant section of the user manual.

3.1.4 User comments

In case you find that this manual insufficient you are welcome to contact me with criticism and preferably with constructive suggestions for improvements.

3.2 the dos folder

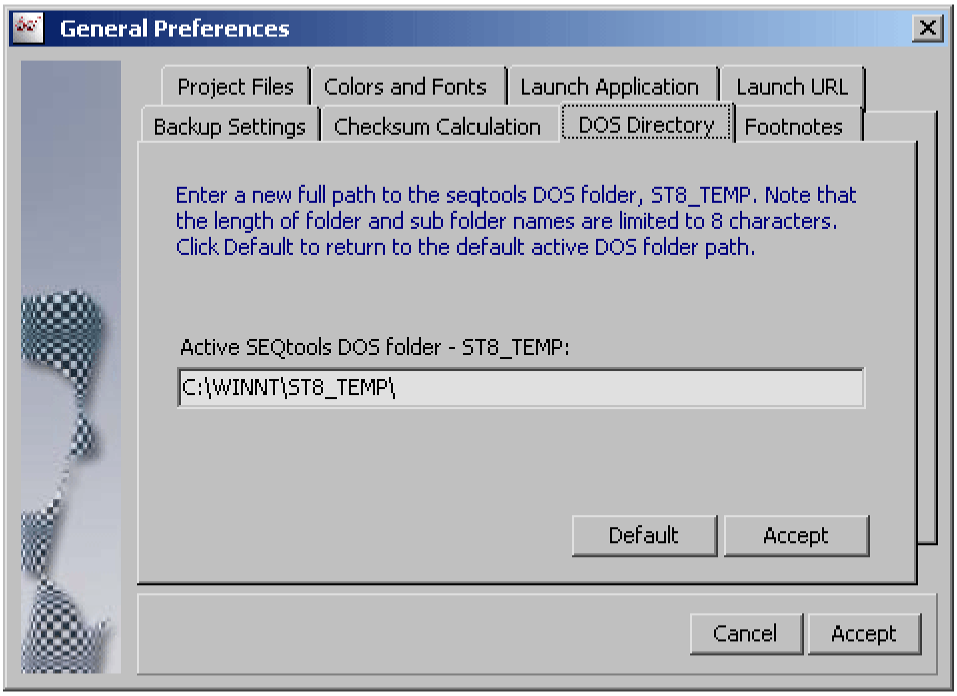
A number of SEQtools functions uses command line dos programs. To avoid problems with the length of file paths (many dos programs are unable to handle file paths unless they follow the old 8+3 syntax) all such programs and associated components reside in a special SEQtools folder on the c drive under the folder containing the operating system (WINNT, Windows): *C:\WindowsFolder\ST8_TEMP*.

When SEQtools starts it checks whether all necessary external components are available in this folder. If components are missing the user is warned and encouraged to download the missing components. The components are contained in two self extracting compressed files, *auxiliary8.exe* and *emboss8.exe*.

When you install the full SEQtools packages you automatically install also these components. New updates of the auxiliary and emboss programs and be downloaded and installed without re-installing SEQtools. Use the functions under the *Help/SEQtools Configuration* menu to perform this task.

3.2.1 Moving the dos folder to a new location

It is possible - but not recommended - to move the SEQtools dos folder to a different location. If you prefer the dos folder to be located in a different location use the *Preferences/General Preferences/DOS Directory* to choose a new location. Click *Accept* to copy the entire content of the ST8_TEMP folder to the new location. Note that the new path must follow the standard dos syntax (8+3) to pass the verification routine before the new path is accepted.



Enter a new full path to the seqtools DOS folder, ST8_TEMP. Note that the length of folder and sub folder names are limited to 8 characters. Click Default to return to the default active DOS folder path.

Active SEQtools DOS folder - ST8_TEMP:

C:\WINNT\ST8_TEMP\

3.2.2 Components located in the dos folder

The following sub-folders and files must present in the SEQtools dos folder:

\ST8_TEMP\data\*.*	Contains the 26 NCBI data files required by the different NCBI programs
\ST8_TEMP\DB\*.*	Contains local databases created by formatdb. Each local database consist of 5 files all with the sane name but with different extensions
\ST8_TEMP\EMBOSS\...	Includes two sub-folders: *acd* containing four acd files and *data* containing five data files required by emboss programs
\ST8_TEMP\TMP\*.*	Contains temporary files created by different SEQtools functions. The TMP folder is cleared when SEQtools closes
\ST8_TEMP*.*	Contains executables and dll's for a number of components used by SEQtools

3.3 command line options

SEQtools creates and saves a specific ini-file for each instance of the program. This implies that you can create pre-defined instances of the program for different sequence types. Note that you must create a new icon on your desk
top with the instance parameter (/I=NN) *before* you open the SEQtools

instance to set the preferences for the instance.

Proceed as follows: Create a new SEQtools icon on your desktop. *Right-click* the icon to display the Windows pop-up menu. *Left-click* the *Properties* line of the pop-up menu and edit the load path for the program as described below. Then launch the SEQtools instance, set the preferences and exit SEQtools to save the ini-file associated with the new instance.

valid command line parameters:

1. **SEQtools instance number** (/I= (00 - 99)
2. **full path to sequence file to load when SEQtools opens**

examples:

set project type

```
c:\app.folder\seqtools83.exe /I=00  (main instance, default)
c:\app.folder\seqtools83.exe /I=01  (nucleotide project)
c:\app.folder\seqtools83.exe /I=02  (protein project)
c:\app.folder\seqtools83.exe /I=03  (primer project)
```

load specified file

```
c:\app.folder\seqtools83.exe c:\mydir\myfolder\my_sequence.seq /I=05
c:\app.folder\seqtools83.exe c:\mydir\myfolder\my_project.fms /I=10
c:\app.folder\seqtools83.exe c:\mydir\myfolder\my_protein.seq /I=15
c:\app.folder\seqtools83.exe c:\mydir\myfolder\my_primer.seq /I=20
c:\app.folder\seqtools83.exe c:\mydir\myfolder\my_project.plp /I=25
```

3.4 data files

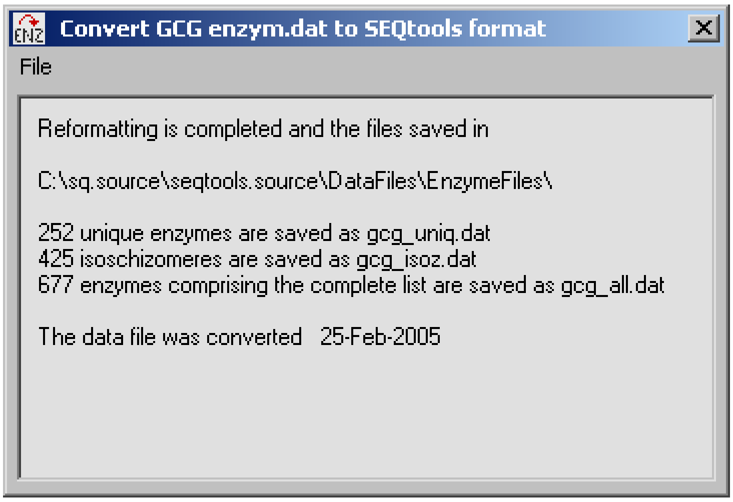
Seqtools uses two types of data files: restriction enzyme data files and codon usage table files. When SEQtools is installed four restriction enzyme files and four codon usage files are included in the installation. The Data files are located in the main application folder in the *...\Program Files\seqtools 8.3\DataFiles\EnzymeFiles\* and the *...\Program Files\seqtools 8.3\DataFiles\CodonFiles\* sub-folders. Seqtools uses its own file format and both file types must thus be processed before they can be used in the program as described below.

3.4.1 Restriction enzyme data files

Updated restriction enzyme data files can be downloaded from ReBase. In addition to enzyme data files, the ReBase homepage contains a very useful search function which allows you to search their data base with the name of an enzyme or with a recognition pattern. Visit the ReBase homepage to download the restriction enzyme data file in GCG format.

3.4.2 Convert gcg data files to seqtools format?

Seqtools uses a slightly different enzyme data file format than the GCG program so it is necessary to use *Tools/Conversion Functions/Convert GCG Restriction Enzyme File...* to convert the file format so that the data file can be used by SEQtools as illustrated by the three screen shots below:



Reformatting is completed and the files saved in

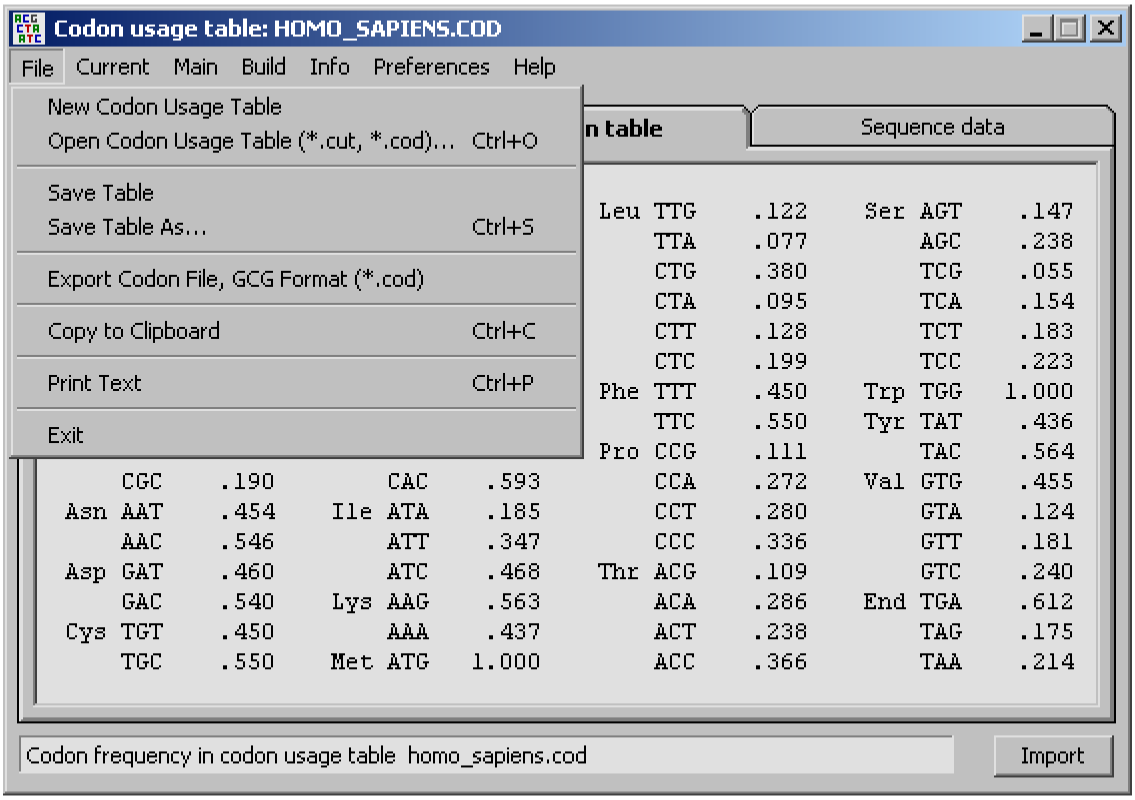
C:\sq.source\seqtools.source\DataFiles\EnzymeFiles\

252 unique enzymes are saved as gcg_uniq.dat
425 isoschizomeres are saved as gcg_isoz.dat
677 enzymes comprising the complete list are saved as gcg_all.dat

The data file was converted   25-Feb-2005

3.4.3 Codon usage tables

Codon usage tables can be obtained from a number of sources for example from the Japanese Kazusa DNA Research Institute/Codon Usage Database. Remember to specify a GCG like style. The easiest way is to save the table directly from the Internet browser window as a plain text file with the extension *.cod in the folder *...\Program Files\seqtools 8.3\DataFiles\CodonFiles*\mycodons.cod. Note that some browsers adds a *.txt extension to the file in addition to the *.cod extension you typed (...myfile.cod.txt). To avoid this enclose the filename+extension in quotes before saving from the browser.



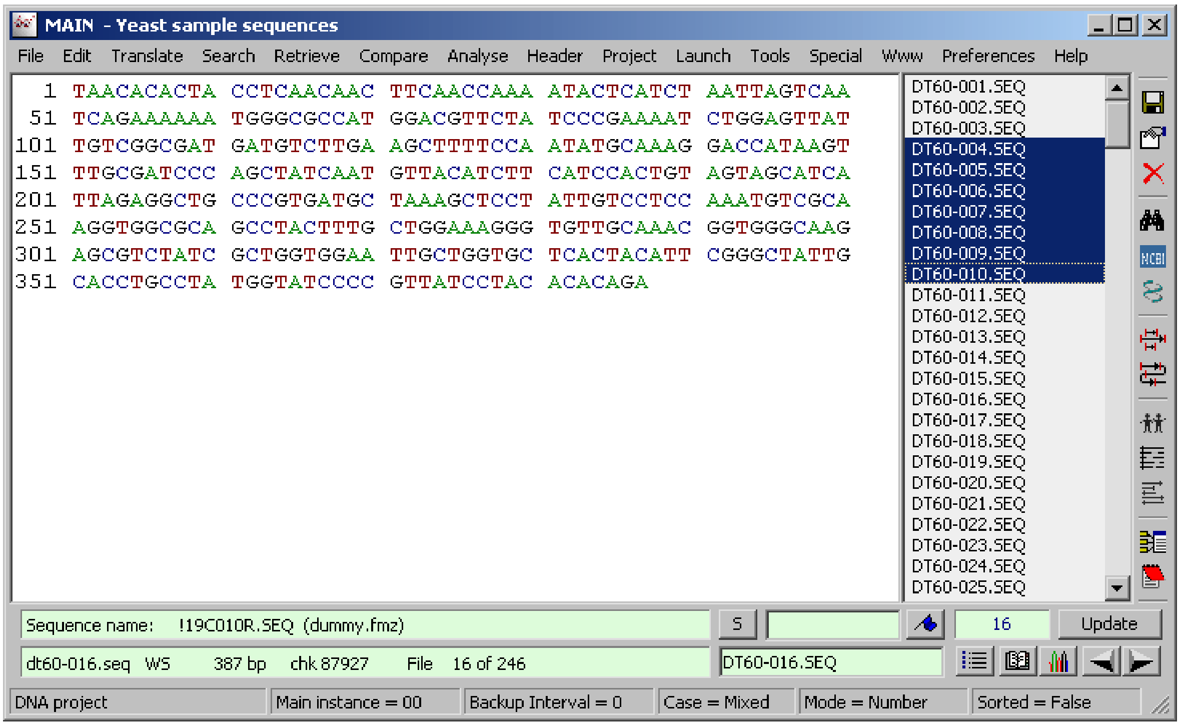
```
                                      Leu TTG   .122    Ser AGT   .147
                                          TTA   .077        AGC   .238
                                          CTG   .380        TCG   .055
                                          CTA   .095        TCA   .154
                                          CTT   .128        TCT   .183
                                          CTC   .199        TCC   .223
                                      Phe TTT   .450    Trp TGG  1.000
                                          TTC   .550    Tyr TAT   .436
                                      Pro CCG   .111        TAC   .564
    CGC   .190        CAC   .593          CCA   .272    Val GTG   .455
Asn AAT   .454    Ile ATA   .185          CCT   .280        GTA   .124
    AAC   .546        ATT   .347          CCC   .336        GTT   .181
Asp GAT   .460        ATC   .468      Thr ACG   .109        GTC   .240
    GAC   .540    Lys AAG   .563          ACA   .286    End TGA   .612
Cys TGT   .450        AAA   .437          ACT   .238        TAG   .175
    TGC   .550    Met ATG  1.000          ACC   .366        TAA   .214
```

Codon frequency in codon usage table homo_sapiens.cod

3.5 the main seqtools editor

Below is a screen shot of the main SEQtools editor. The form includes of a sequence panel, a sequence list (right clicking the sequence list toggles between a sorted list, a project order list. Pressing <F5> lists the matches from a local blast search). In the lower part are two info fields, a goto/bookmark field, an editable sequence name field and command buttons for an extended sequence list, the sequence header, chromatogram display and navigation buttons. The *Update* button reformats the sequence after editing.

Parking the cursor over the upper info field and holding down the right mouse button retrieves blast information for the displayed sequence (if the information is available). The vertical panel to the right contains shortcuts to a number of commonly used functions.


MAIN - Yeast sample sequences
File Edit Translate Search Retrieve Compare Analyse Header Project Launch Tools Special Www Preferences Help
S
Update

```
  1 TAACACACTA CCTCAACAAC TTCAACCAAA ATACTCATCT AATTAGTCAA
 51 TCAGAAAAAA TGGGCGCCAT GGACGTTCTA TCCCGAAAAT CTGGAGTTAT
101 TGTCGGCGAT GATGTCTTGA AGCTTTTCCA ATATGCAAAG GACCATAAGT
151 TTGCGATCCC AGCTATCAAT GTTACATCTT CATCCACTGT AGTAGCATCA
201 TTAGAGGCTG CCCGTGATGC TAAAGCTCCT ATTGTCCTCC AAATGTCGCA
251 AGGTGGCGCA GCCTACTTTG CTGGAAAGGG TGTTGCAAAC GGTGGGCAAG
301 AGCGTCTATC GCTGGTGGAA TTGCTGGTGC TCACTACATT CGGGCTATTG
351 CACCTGCCTA TGGTATCCCC GTTATCCTAC ACACAGA
```

DT60-001.SEQ
DT60-002.SEQ
DT60-003.SEQ
DT60-004.SEQ
DT60-005.SEQ
DT60-006.SEQ
DT60-007.SEQ
DT60-008.SEQ
DT60-009.SEQ
DT60-010.SEQ
DT60-011.SEQ
DT60-012.SEQ
DT60-013.SEQ
DT60-014.SEQ
DT60-015.SEQ
DT60-016.SEQ
DT60-017.SEQ
DT60-018.SEQ
DT60-019.SEQ
DT60-020.SEQ
DT60-021.SEQ
DT60-022.SEQ
DT60-023.SEQ
DT60-024.SEQ
DT60-025.SEQ

Sequence name: !19C010R.SEQ (dummy.fmz) | 16

dt60-016.seq WS 387 bp chk 87927 File 16 of 246 | DT60-016.SEQ

DNA project | Main instance = 00 | Backup Interval = 0 | Case = Mixed | Mode = Number | Sorted = False

3.6 SEQTOOLS project types

Before you create a new SEQtools project you need to decide which type of sequences you wish the project to contain. In cases where you load a project which is previously generated and saved from SEQtools, the SEQtools auto-detects the project type from the *first* sequences in the selection and sets editor options accordingly.

It is not possible to mix nucleotide, protein or primer sequences in the same project. If you which to work with different sequence types simultaneously, open separate instances of SEQtools - one for each sequence type and use copy/paste to transfer sequences *of the same type* between the separate instances of SEQtools.

3.6.1 Nucleotide / trace projects

This project is restricted to include nucleotide sequences. If protein sequences are generated by translation of nucleotide sequences the protein sequences *do not* become part of the project when the project is saved.

Extracted trace files (chromatograms from auto sequencers) require a nucleotide project. If you create a new project exclusively consisting of trace files SEQtools auto-detects the project type and create a trace project. A trace project is similar to a normal nucleotide project. You can add more normal sequences and new trace files to a trace project and save the entire mixed project by one of the four methods described below.

The original trace file is *not* modified by being loaded and saved from SEQtools. Instead an association/link is created between the extracted, normal SEQtools version of the trace sequence and the original trace file. Provided that the path to original trace file is not changed the chromatogram can be retrieved and displayed by clicking the trace icon on the main editor form.

If you attempt to load a non-nucleotide sequence into a nucleotide project you are warned before SEQtools cancels the load operation.

3.6.2 Protein projects

This project type is limited to protein sequences. If you attempt to load a non-protein sequence into a protein project you are warned before SEQtools cancels the load operation.

The project type is auto-detected by SEQtools based on the *first* sequence in the load selection or a multi-sequence file.

3.6.3 Primer projects

This project type only holds primer sequences. If you attempt to load a non-primer sequence into a primer project you are warned before SEQtools cancels the load operation.

3.6.4 Conversion of projects

It is possible to convert primer projects to nucleotide projects and vice versa. This option is useful if you for example want to perform a blast search at Genbank with a collection of primer sequences.

Note, however, that due to the different structure of sequence and primer headers converting a nucleotide project to a primer project - *and saving the project as a primer project* will lead to irreversible loss of all information contained in the original sequence headers.

The fact that several symbols (brackets, IUB symbols) which are allowed for primer sequences but not in normal nucleotide sequences implies that the conversion option should be used with caution, especially when converting normal sequences to primers and primers with degenerate positions to sequences.

3.7 Working with projects

3.7.1 Create projects

A SEQtools project is automatically created when you load a collection of sequences into the program. This can either be done by navigating to a specific folder and selecting one or more sequence file, by loading a multi-sequence file or by creating an empty sequence file and entering the sequence by manually typing or by copy/paste.

3.7.2 Modify projects

It is possible to add more sequences to an existing project by importing multiple single sequences, multi-sequence files or by creating new sequences manually. New sequences added to a project are appended to the sequences already included in the project.

Sequences can also be removed from the project by selecting the sequences to be removed from the project in the sequence list and click the remove icon. Note that removed sequences remain on the hard drive, i.e., are NOT deleted but only excluded from the project.

With this facility it is for example possible to perform a database search with all sequences contained in a given project - and remove sequences with matches worse than a specified expect value.

3.7.3 Save/export projects

Sequences can be saved/exported in three different ways:

- - as single sequences,
- - as a multi-sequence file in fasta or SEQtools format
- - as a so called pfp file which is a list containing the full path from which the sequences in the project were imported or
- - a psp file which also consist of a path list, but in this case the save-path for all project files.

The latter option is not enabled until the project is saved as single files.

Note for the pfp and psp save methods that changing the physical location of the sequence files on the hard disk *after* the pfp and psp file are generated will prevent these sequence files from being loaded from the pfp and psp path-list file.

3.8 About sequence names

3.8.1 Normal sequence name - Most of the functions related to handling multiple sequences in SEQtools were developed during a small EST project carried out at the Carlsberg Laboratory. The purpose of the project was aimed at obtaining information about the *Blumeria* (mildew) genome and gene expression to better understand the interaction between the obligate plant parasite and its host, barley.

All the clones from the cDNA libraries used in the analysis were sequenced twice, with an F (forward) primer and with a R (reverse) primer. The 5' sequences of the insert were used for database searching for homology in public data bases while the 3' (polyA) sequences were used to create links to SAGE profiles generated from the same developmental stages.

The insert lengths of the cDNA libraries were rather short (only very few were full length ORFs) which turned out to be an advantage when searching the international databases. It also implied that in many cases the F and R sequence overlapped and could be replaced by the merged, complete sequence of a particular insert. This feature of the libraries allowed us to replace the F and R sequences by their merged sequence which both improved the quality of the sequence and reduced the number of sequences in the Blumeria database.

In cases where the F and R sequence of a clone/insert did not overlap, i.e., where sequence information was not available to link F and R sequences from the same clone, the file name was used instead as a link between corresponding F and R sequences of the same insert. Obviously this requires that files / clones must be named consistently as described below.

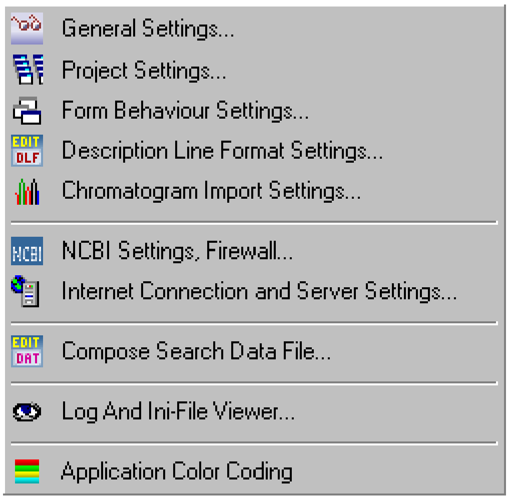
In order to keep track of the F and R sequences originating from the same insert/clone, all sequences were named using *-F, -R* and *-M* to indicate the 5', 3' and merged sequence.

3.8.2 Long sequence name - Loading a new sequence with a long, non-DOS, file name into SEQtools automatically transfers the long file name into the *Long name* variable of SEQtools.

For new sequences which have not previously been formatted by SEQtools, a *Long name* is automatically created consisting of the file name followed by the number symbol (#) and a random 8-digit number (e.g. C00018-F #47382957). The Long Sequence name cannot be changes by the user.

3.9 Setting user preferences

There is a number of options for the user to customize the appearance and behavior of SEQtools through extensive preference facilities.


General Settings...
Project Settings...
Form Behaviour Settings...
Description Line Format Settings...
Chromatogram Import Settings...
NCBI Settings, Firewall...
Internet Connection and Server Settings...
Compose Search Data File...
Log And Ini-File Viewer...
Application Color Coding

These options are described in details under the *Preference* menu item. At this point it suffice to briefly mention which aspects of SEQtools behavior that are adjustable by preference settings.

General settings	Project files, Colors and fonts, Launch applications, Launch URLs, Backup settings, Checksum calculation, DOS folder location, Footnotes.
Project settings	Trace file folder, Global timeout, Project blast settings, Project title, User data, Sequence format, Color patterns, Header
Form behavior settings	Main editor, Header forms, Blast forms, Compare forms, Tools, Translate, Primer forms, Special function forms
Description line format	General settings, Left-trim lines, Right-trim lines, Replace lines
Chromatogram import settings	Basecallers, Preset options, Trimming, N-threshold, Gap-quality
NCBI inifile settings and editor	Inifile settings for blast searches
Internet connection and servers	NCBI settings, Internet connection test URL
Compose search data file	Predefined groups, User defined groups
Log and Ini-file viewer:	Ini-files for multiple instances of SEQtools, Log-file for several batch functions

Application color coding	Assign color schemes to multiple instances of SEQtools

3.10 sequence annotation

3.10.1 Auto-annotation

Seqtools includes various methods of auto-annotating sequences. The most powerful ones are the batch blast functions which allows you to perform unattended blast searches at NCBI/Genbank with a large number of sequences. Depending of your pc you can load 10 - 20.000 sequences into a single project and perform batch blast search on them all. Seqtools stores this information in the sequence header associated with each sequence.

Seqtools contains advanced *facilities* for handling and displaying this information. It is possible to select a particular blast search and list or display this information excluding/hiding results from other search results stored in the sequence header. As all information is stored in the RAM memory of your pc processing large amount of sequences requires quite a lot of RAM.

Information stored in sequence headers can be search in a number of ways making this a very flexible system. You can read more about this in the special *Header* section of this manual

3.10.2 User annotation

It is possible to manually enter your own comments and copy/paste external information into sequence headers. This, however, must be done sequence-by-sequence.

3.11 batch operations

One of the strong features of SEQtools is the facilities for performing batch operations. A number of tasks such as changing sequence names and performing blast searches locally or on databases at Genbank can be performed without user intervention. Some users have reported successful batch analysis of as many as 30,000 est sequences in a single job running over several days.

The fact that you can launch several concurrent instances of seqtools makes it possible for example to run large blast search jobs at genbank while performing other analyses with a different instance of seqtools.

You can even run parallel batch search jobs at genbank with separate instances of seqtools. Seqtools only uses very few pc resources for processing and storing search results as they arrive from ncbi. The auto-save function of the batch blast function reduces the risk of loosing data in case of pc crashes during a search job.

3.12 seqtools file types

Seqtools uses a number of different file types, some of them for saving various types of data others for importing data. The table below lists the extension of file types recognized or created by SEQtools.

ALN, PIR, PHY, MSF	Output files from sequence alignment with Clustalw.
DAT, SDF, GCG	Restriction enzyme and user created search data files.
PLP, PSP	Project path files, used to store the full paths for all files in a project for reloading the complete project or a sub-group of the project.
FOF	File of files. Includes a list including the names of all sequences included in the project.
TXT, RTF, LST, RPT, LOG, TAB	Various ASCII files containing sequence lists, reports, logs etc.
SEQ, DNA, PRO	General extensions for DNA or protein sequence files.
CUT, COD	Codon usage tables, SEQtools and GCG format.
FMS, FMZ, TMS, MSF, DMS, FAS, FSA, GB, LGF, GBK, GCG, MBL, FMS	Various types of multi-sequence files.
B!!, BA!, BAK	Backup files from timed project auto-backups.
TPL, ESF	Template and complete submission file for transfer of EST sequences to Genbank.
STF, PTF, DTF, MTF, SMF, CGI, TDT	Extensions used in SAGE related functions.
SGD, MCA	Extensions used for files created by EST clustering functions.
OOF, COF, MSG	Primer mail order files.
BMP, WMF	Image files.
MTP, MPF, IGF, IMG	Microtiter plate index, Micro array project

file, Imagene GeneID file

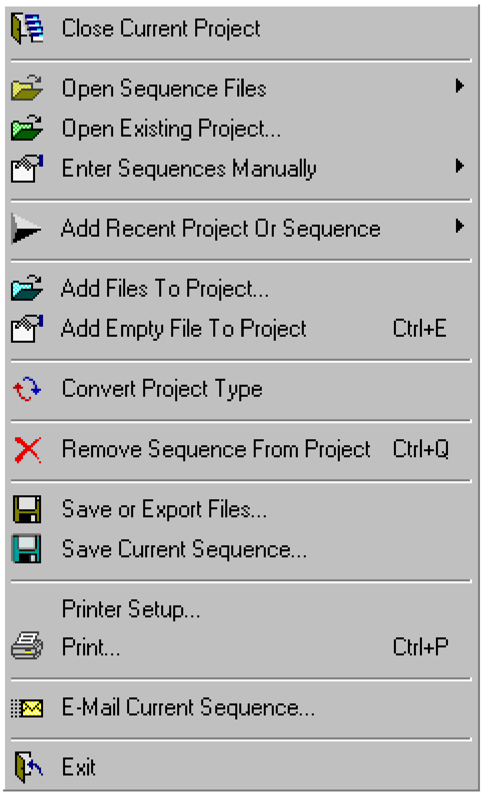
3.13 Application files and folders created and managed by seqtools

\windows\NCBI.ini	Ini-file for blast programs.
\app\ST8##.INI	Ini-file for instance ## of SEQtools. Contains all user preferences for instance ##. Each instance (maximum number of open instances of SEQtools is 99) has its own set of preferences.
\app\ST8_instances.dat	Seqtools session dat-file. Keeps track of open instances of SEQtools.
\app\BackupData\	Contains timed backup files for open / active projects. Each instance of SEQtools has its own timed backup file.
\app\DataFiles\CodonFiles\*.*	Codon usage tables.
\app\DataFiles\EnzymeFiles\*.*	Contains all restriction enzyme data files.
\app\DataFiles\genbank_databases.dat	Contains a list of available Genbank databases for advanced batch database searching at Genbank.
\app\UserData\NNN\*.*	Auto-generated default folders for storing various data: _array _blast _cluster _database _default _genbank _multiseq _primer _protein _psgfiles _sage

4. SEQTOOLS MENUS

This page contains a brief summary of the functions grouped under each of the 15 menu titles for *nucleotide projects*. The menu layout is slightly different for *protein projects* and more so for *primer projects*. Some menu items are not displayed when seqtools is opened for handling protein and primer sequences. The manual contains a separate page, *primer functions*, describing specific facilities related to primer design and ordering. The differences in menu layout are smaller for protein projects and are not treated separately.

4.1 file menu



The *File* menu contains a number of facilities for creating and modifying projects. With these functions you can select sequences to be included in a project, add more sequences to an existing project or remove selected sequences from a project.

Sequences can either be loaded as collections of single sequence files, as multi-sequence files or a mixture of both. Seqtools examines each file to be loaded to see if it contains a single sequence or is a multi-sequence file.

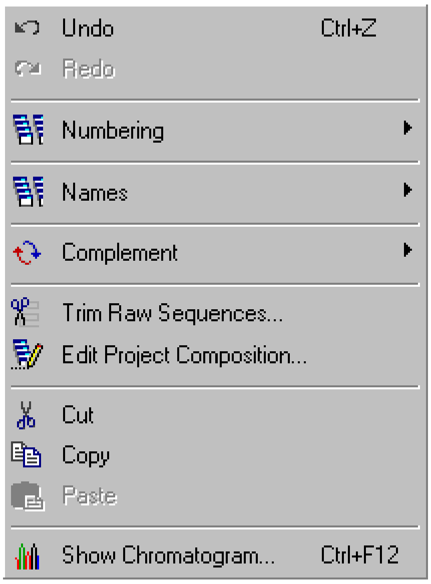
If you need to enter sequences manually or by copy/paste it is necessary first to create an empty file to hold the sequence.

Multi-sequence files can either contain the complete sequence and annotation for each sequence in the multi-sequence file or be a list (a plp-file / psp-file) of file-paths to each sequence file.

In the first case all project files must be located in the same folder while in case of plp-files / psp-files the file paths can point to sequence files located in different folders.

The save and export facilities allow you to save / export sequence files in the most common sequence formats.

4.2 edit menu

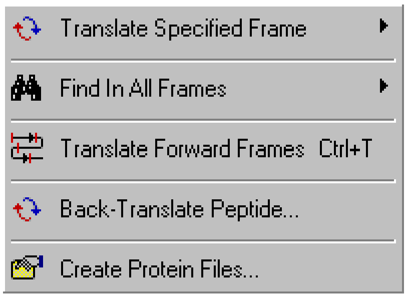


The *Edit* menu includes functions related to sequence editing in the broadest sense.

With these functions you can edit sequence names and numbering, remove vector parts of raw sequences generated by auto-sequencers (Trim Raw Sequences...), convert one or more sequences to their complementary sequence and remove sequences from the project based on sequence quality (Edit Project Composition...).

Most of the functions for sequence editing apply to nucleotide sequences and are not visible when seqtools is in *protein mode*.

4.3 translate menu

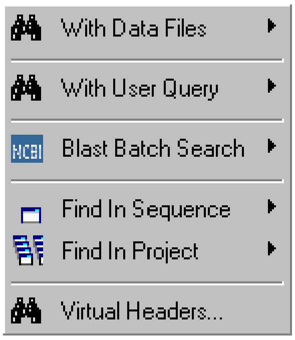

Translate Specified Frame
Find In All Frames
Translate Forward Frames Ctrl+T
Back-Translate Peptide...
Create Protein Files...

The *Translate* menu contains a number of options for translating nucleotide sequences into protein. In addition it allows you to rapidly find the longest open reading frame or the longest stretch without stop codons in an unknown nucleotide sequence.

If a protein sequence is displayed you can back-translate it into a nucleotide sequence if you provide information about the expected codon usage in the form of a codon usage table either retrieved from a web resource or created by yourself.

As for the Edit menu above, most of the *Translate* functions apply to nucleotide sequences and are not visible when seqtools is in *protein mode*.

4.4 search menu



The *Search* menu includes a number of advanced options for searching with your nucleotide or protein sequences.

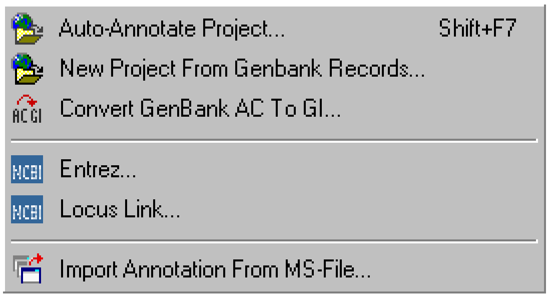
The functions range from trivial searching your sequence with a query string to unattended batch blast searching all sequences in the project against Genbank or against a local sequence database created by yourself.

In addition you may look for repeats, introns and similar/identical sequences in the project.

Some of these functions uses programs included in the Emboss collection others depend on the NCBI program collection.

Batch searching Genbank requires an Internet connection.

4.5 retrieve menu

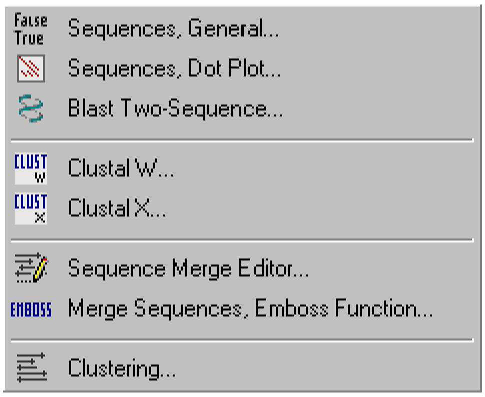

Auto-Annotate Project... Shift+F7
New Project From Genbank Records...
Convert GenBank AC To GI...
Entrez...
Locus Link...
Import Annotation From MS-File...

The *Retrieve* menu includes various functions for annotating your sequences and for auto-editing already existing annotation.

With these functions you can retrieve the complete annotation from Genbank if your sequences are only identified by their Genbank accession number.

Or you can automatically create a new project consisting of the Genbank sequences with the best match from a blast search on Genbank with your own sequences.

4.6 compare menu



The *Compare* menu comprises several functions for comparison and/or alignment of two or more sequences.

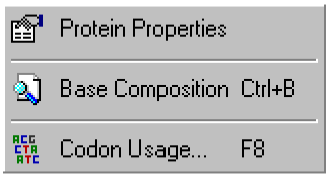
The functions include two-sequence comparison, multi-sequence alignment with

ClustalW or ClustalX, the former option with optional post-processing of the alignment with t-coffee.

The *Compare* menu furthermore contains two functions for multi-sequence merging, one based on the Emboss program the other a simple multi-sequence editor.

Finally, this menu includes functions for sequence clustering based on different methods.

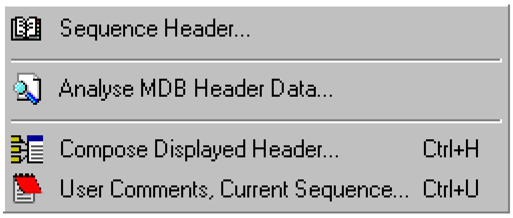
4.7 analyze menu



The *Analyze* menu lists three functions which enables you to view basic properties of a displayed protein sequence, the base composition of a nucleotide sequence and the codon usage of a nucleotide sequence.

The latter function furthermore enables you to create a new codon usage table and to include the codon usage of the currently displayed nucleotide sequence in this - or an already existing - codon usage table.

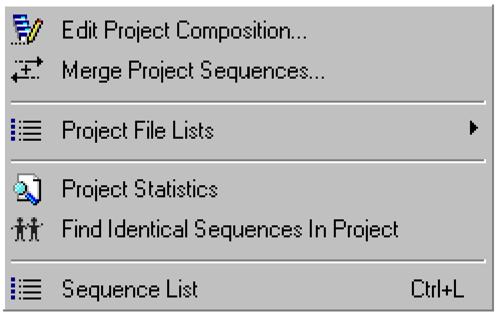
4.8 header menu


Sequence Header...
Analyse MDB Header Data...
Compose Displayed Header... Ctrl+H
User Comments, Current Sequence... Ctrl+U

Different options related to handling sequence annotation is collected under the *Header* menu. This includes display of the items currently selected as the *virtual* header, options for displaying the result of local multi-database searches and a form enabling you to enter your personal comments to the sequence.

Finally, the "Compose Displayed Header..." option allows you to select items of the complete annotation to be included in the *virtual* header.

4.9 project menu

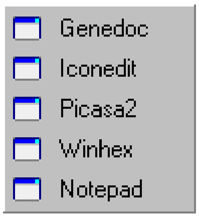

Edit Project Composition...
Merge Project Sequences...
Project File Lists
Project Statistics
Find Identical Sequences In Project
Sequence List Ctrl+L

The *Project* menu includes functions and options related to the handling of the sequences contained in the current project.

With these functions you can create lists of sequence names and file paths, merge overlapping forward and reverse sequences from the same insert, calculate project statistics, find (and remove) duplicate sequences in the project (irrespectively of the sequence names) and display a list of the sequences contained in the project.

The latter function is quite elaborate enabling you to include selected information from the current *virtual* with extensive options for formatting the displayed sequence list.

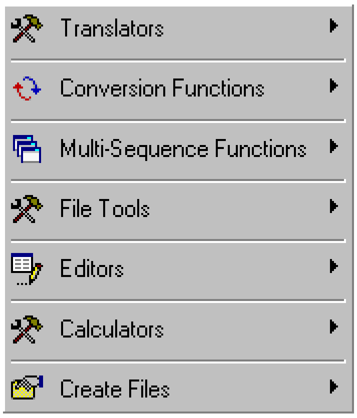
4.10 launch menu



The *Launch* menu contains up-to five third party programs selected from the "General Preferences...".

Two external programs (GeneDoc and TreeView) are recognized by seqtools and can be accessed from the ClustalW alignment result form. T-coffee mentioned above under the *Compare* menu is also accessed from the ClustalW result form but do not have a user interface. This program is included in the seqtools installation and need not be installed separately via "General Preferences...".

4.11 tools menu



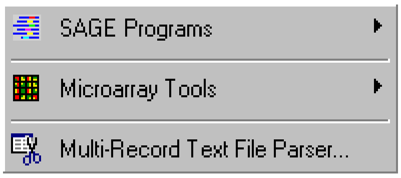
Various functions and facilities have been collected under the *Tools* menu. These include a codon-amino acid translator, a IUB symbol translator and three converters (GCG restriction enzyme file -> seqtools format, project -> search database, Genbank accession numbers -> GI numbers).

The menu also includes several tools for multi-sequence handling (building of local databases, batch-editing of sequence annotation, building new projects from Genbank records). Among the "File Tools" are various facilities for viewing, searching and creating different file types.

The "Editors" include options to customize restriction enzyme search datafiles and to compose/edit FastA definition lines for multi-sequence files.

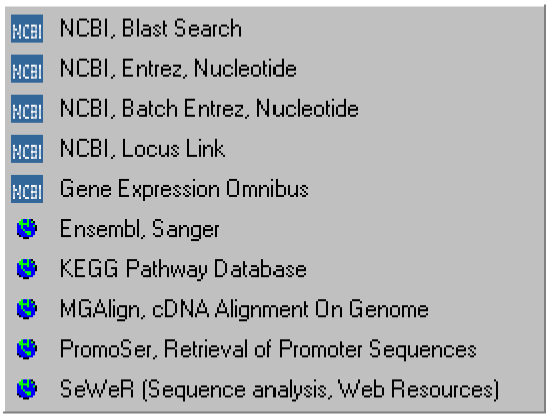
Finally, the "Create Files" item covers functions for submission of EST sequences to Genbank and a multi-sequence annotation parser.

4.12 special menu



The *Special* menu contains a number of functions and facilities for handling and processing data for SAGE (serial analysis of gene expression) and for oligo-nucleotide based micro-array analysis of gene expression.

4.13 www menu

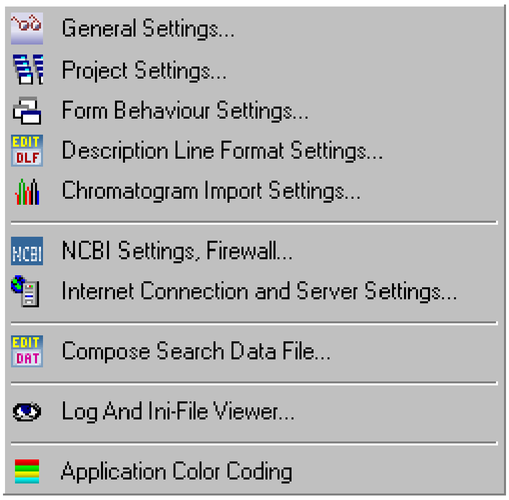

NCBI, Blast Search
NCBI, Entrez, Nucleotide
NCBI, Batch Entrez, Nucleotide
NCBI, Locus Link
Gene Expression Omnibus
Ensembl, Sanger
KEGG Pathway Database
MGAlign, cDNA Alignment On Genome
PromoSer, Retrieval of Promoter Sequences
SeWeR (Sequence analysis, Web Resources)

The *WWW* menu is a list of URL to selected resources on the Internet.

You can customize the list from the "General Preferences..." form. The list holds up-to ten URL's.

Genbank resources are indicated by NCBI-icons and others by globe-icons.

4.14 preferences menu


General Settings...
Project Settings...
Form Behaviour Settings...
Description Line Format Settings...
Chromatogram Import Settings...
NCBI Settings, Firewall...
Internet Connection and Server Settings...
Compose Search Data File...
Log And Ini-File Viewer...
Application Color Coding

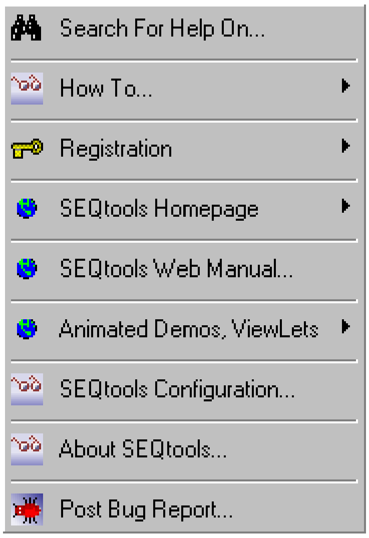
The *Preferences* menu includes a series of forms containing options for customizing the behavior and appearance of seqtools. Most of the menu items are self-explanatory, other more obscure. Among the latter category is the "Form Behavior Settings..." which enables you to decide if a given seqtools form should always stay on-top of other form on your desktop.

In some cases the "NCBI Settings, Firewall..." are important to establish an Internet connection through a local firewall.

The "Application Color Coding" allow you to color code different instances (one instance in *primer mode*, a second in *DNA mode* and perhaps a third instance running in *protein mode*) of seqtools running simultaneously on your pc to facilitate identifying each instance of seqtools.

It is highly advisable to invest in a second monitor if you are using seqtools regularly, especially when you run several instances of the program simultaneously.

4.15 help menu



The *Help* menu contains both different help items, options for registering seqtools (entering the registration key in the program as well as on-line renewing your registration) and a form listing the current seqtools configuration (file dates and installed auxiliary components) and program update options.

You also find a form for reporting bugs on the help menu.

The "Animated Demos, Viewlets..." menu item only contains two animated sequences illustrating program basics. It has been my intension for a long time to write more viewlets describing other aspects of seqtools. It is, however, quite time consuming to produce viewlets so you may have to keep waiting for more animations.

4.1 FILE MENU

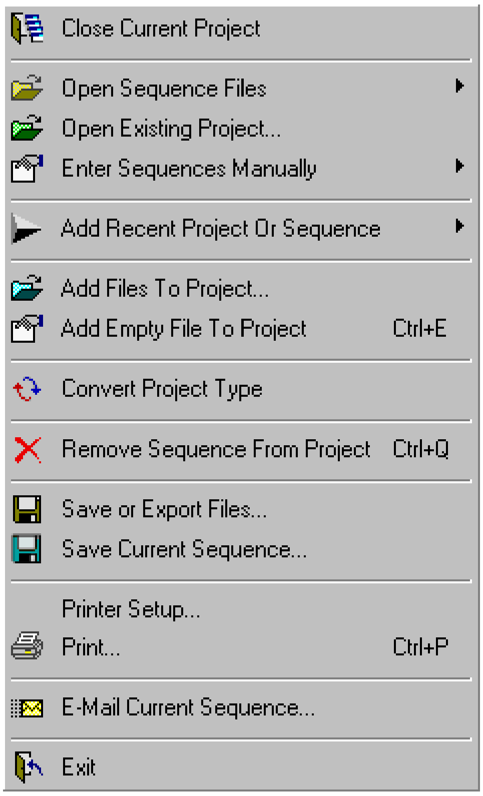
4.1.1 about files and projects

A project in SEQtools is simply a collection of one or more sequences of the same type (*nucleotide*, *protein* or *primer*). It is *not* possible to include different sequence types in the same project. If you wish to create a project from more that one sequence file, all files to be loaded must be located in the *same* folder.

You can *add* more sequence files to an *existing* project from other folders. In most cases SEQtools will auto-detect both the file type (nucleotide, protein or primer), sequence format (SEQtools, embl, fasta, genbank, etc.) and file format (single, trace, multi-sequence - or a mixture of the three) and create the project from the selected files without your intervention.

Saving a project is most conveniently done by using the standard SEQtools multi-sequence format which saves all sequences in the project in a single file (with or without compressing the file).

The file menu contains the following menu items (described in more detail in separate sections below):

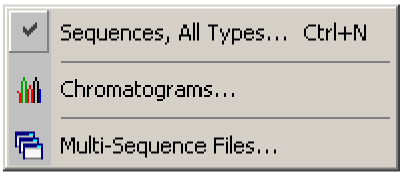


4.1.2 close current project

SEQtools issues a warning before closing the current project offering to save the sequences. Closing a project *without* saving the data will cause irreversible loss of editorial changes to the sequences as well as all information added to the sequence headers.

4.1.3 open sequence files

Sequence files to be included in a project can be selected in different ways as indicated in the screenshot of the *Open Sequence Files* menu shown below.

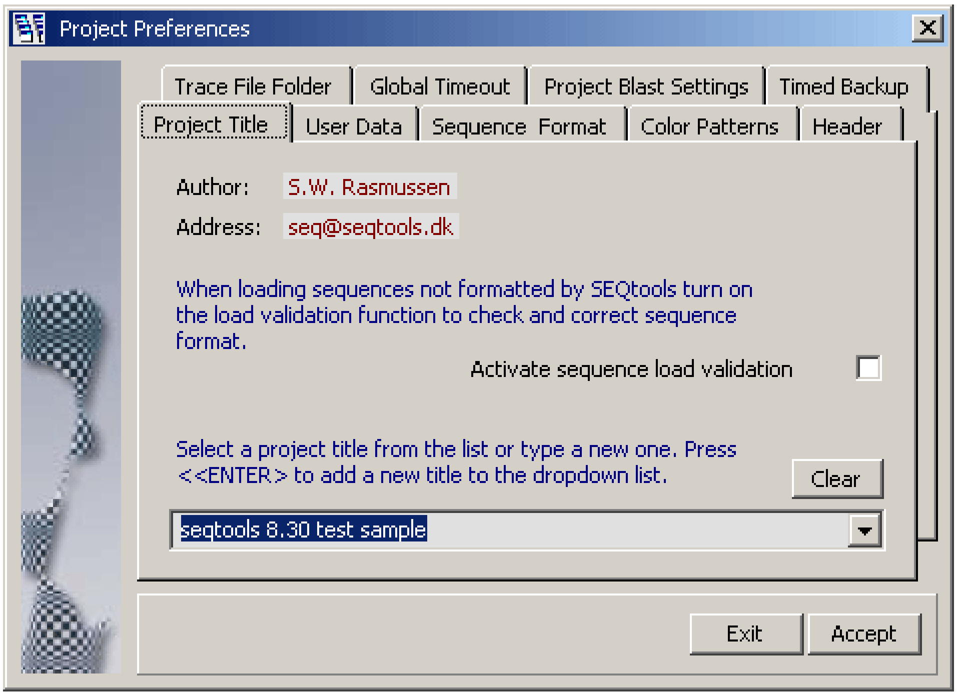


SEQtools attempts to determine sequence *type* and *format* and *file format* before loading the data into a new project. In most cases this does not require user intervention *provided all sequences to be loaded are of the same type* (nucleotide, primer or protein).

The project *type* (nucleotide, primer or protein) is determined by the *first* sequence loaded. If a sequence of a different *type* is encountered a warning is issued and loading is interrupted.

SEQtools recognizes and loads four sequence *formats* either as single sequence files or as collections of sequences in multi-sequence files: *SEQtools*, *EMBL*, *Genbank* and *Fasta*

Before the file selection form is loaded the *Project Preferences* form is opened to enable you to give the project a title and to set various parameters for the new project.



Author: S.W. Rasmussen

Address: seq@seqtools.dk

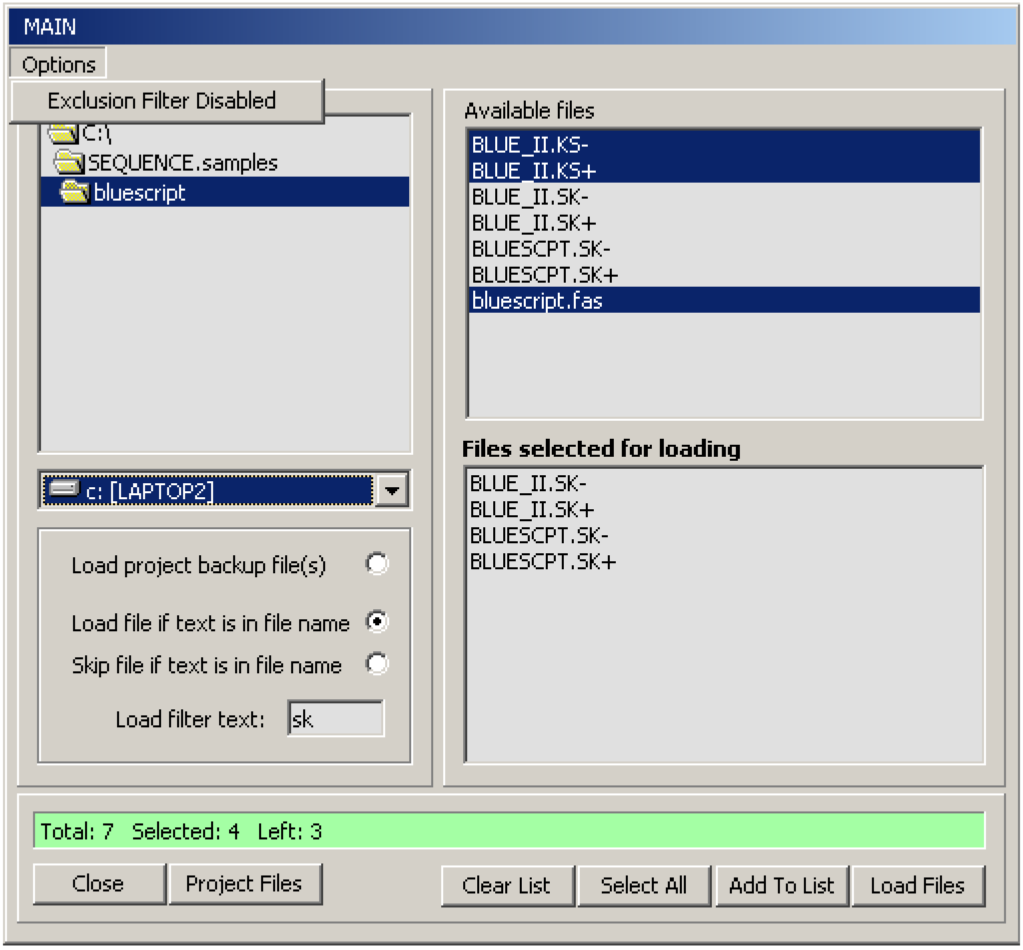
When loading sequences not formatted by SEQtools turn on the load validation function to check and correct sequence format.

Activate sequence load validation □

Select a project title from the list or type a new one. Press <<ENTER> to add a new title to the dropdown list.

seqtools 8.30 test sample

The *File Selection* form is used to select the sequence files for the project. A *drive* list box and a *file* list box allows you to navigate between drives and directories to locate the sequence files you wish to include in the project. The top file list contains all files in the selected directory. The bottom file list shows the files currently selected for loading.



C:\
SEQUENCE.samples
bluescript

c: [LAPTOP2]

Load project backup file(s) ○
Load file if text is in file name ●
Skip file if text is in file name ○
Load filter text: sk

Available files
BLUE_II.KS-
BLUE_II.KS+
BLUE_II.SK-
BLUE_II.SK+
BLUESCPT.SK-
BLUESCPT.SK+
bluescript.fas

Files selected for loading
BLUE_II.SK-
BLUE_II.SK+
BLUESCPT.SK-
BLUESCPT.SK+

Total: 7 Selected: 4 Left: 3

Files are selected from the directory file list by pointing or dragging the mouse pointer to highlight one or more file names. A discontinuous series of files is created by holding down the <CTRL> key while clicking the filenames to be included in the project. Clicking the *Add Files* command button activates the selection. File names can be removed from the list of selected file names by clicking the file name.

Files with the following extensions (cab, log, fof, exe, ini, sys, com, hlp, bat, oof, cof, msg, cut, cod, lst, zip, dat, qscore.fasta, gap_qscore.fasta) *cannot* be selected and loaded into a project unless the *Options/File Exclusion Enabled/Disabled* option is set to *File Exclusion Disabled*.

It is possible to add a case-insensitive *filter* to the selection by typing characters in the text field. Only files which *include* or *do not include* - depending on the selected option - these characters in their file names will be selected/deselected when the *Add To List* command button is clicked.

When the auto-backup option is active (*Preferences/Project Settings/Timed Backup*) a complete backup of all sequences and sequence headers of the project is saved - at the specified time interval - to a multi-sequence file (*.fms) located in the main application folder (normally *c:\SEQtools 8.3\BackupData\*). If you need to load a backup copy of a previous project select the *Load project backup file(s)* option on the load form to set the path to this folder and load the *.fms multi-sequence file into a new project.

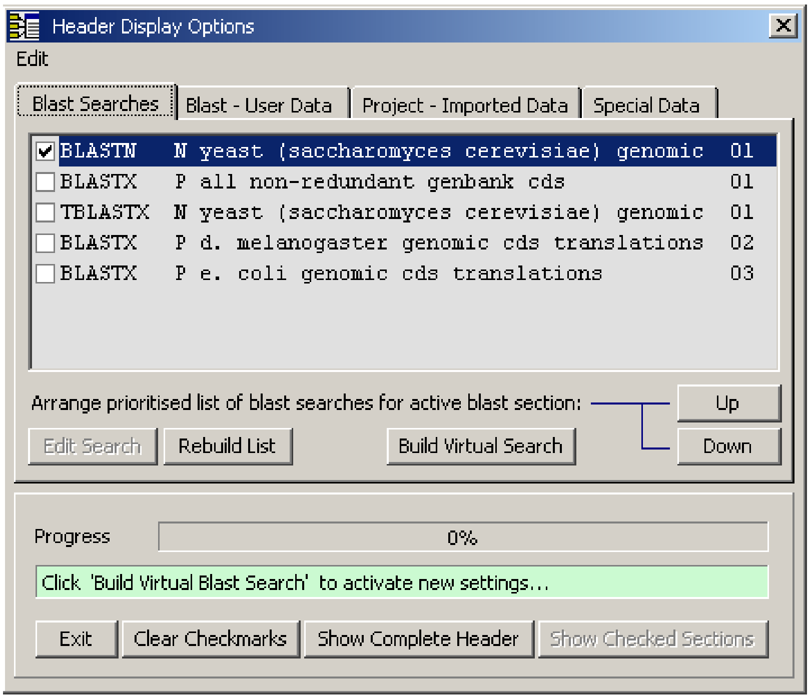
If you are loading more than 300 sequences into a project, SEQtools offers to turn off the timed backup function. This function is often not required for large projects and turning it off saves resources for processing other functions.

When selection is completed, clicking the *Load Files* command button causes the selected files to be loaded into the specified project. It is *not* possible to select the same file twice nor is it possible to select files from different directories when a new project is created. Additional files can be added to the project later.

If you already know that the sequences to be loaded are contained in a multi-sequence file (SEQtools, Genbank or Fasta format) just select the *Multi-Sequence Files…* menu item. This opens a standard Windows file dialog box for selecting the multi-sequence file. The file selection form is not loaded in this case.

It is possible to select and load a mixture of normal single files and multi-sequence files.

When sequence loading is completed and a new project created SEQtools displays a summary of the annotation (primarily a list of blast search results) available for the loaded sequences. This is described in more detail under *4.8 Header* menu and its sub-items.



```
☑ BLASTN    N yeast (saccharomyces cerevisiae) genomic  01
□ BLASTX    P all non-redundant genbank cds             01
□ TBLASTX   N yeast (saccharomyces cerevisiae) genomic  01
□ BLASTX    P d. melanogaster genomic cds translations  02
□ BLASTX    P e. coli genomic cds translations          03
```

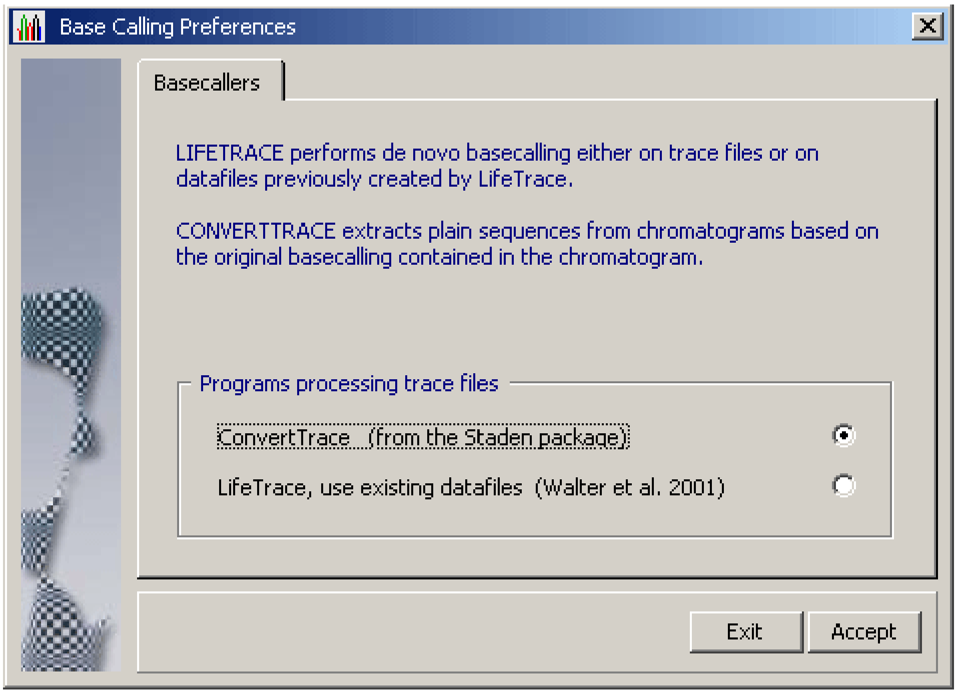
Arrange prioritised list of blast searches for active blast section:

Progress 0%

Click 'Build Virtual Blast Search' to activate new settings...

4.1.4 basecalling chromatograms

SEQtools auto-detects if the file to be loaded is a chromatogram produced by an automated sequencer. Extraction of the plain DNA sequence from the trace file is, by default, carried out by the *convert_trace* program from the Staden package while viewing the traces is done by *Chromas* (see screenshot below).



Base Calling Preferences

Basecallers

LIFETRACE performs de novo basecalling either on trace files or on datafiles previously created by LifeTrace.

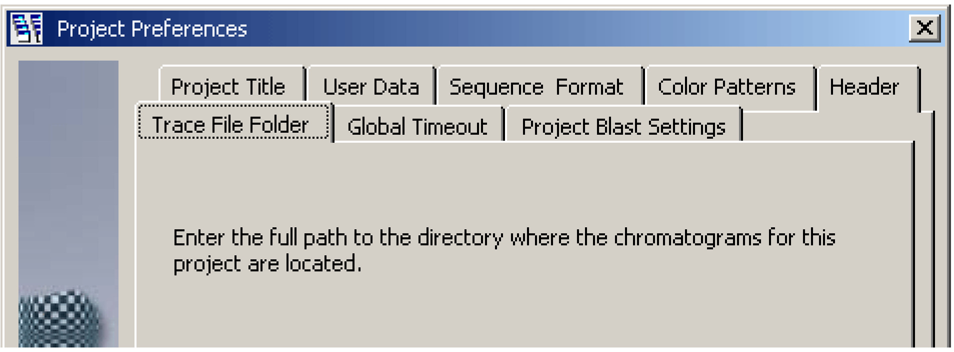
CONVERTTRACE extracts plain sequences from chromatograms based on the original basecalling contained in the chromatogram.

Programs processing trace files

ConvertTrace (from the Staden package) ●

LifeTrace, use existing datafiles (Walter et al. 2001) ○

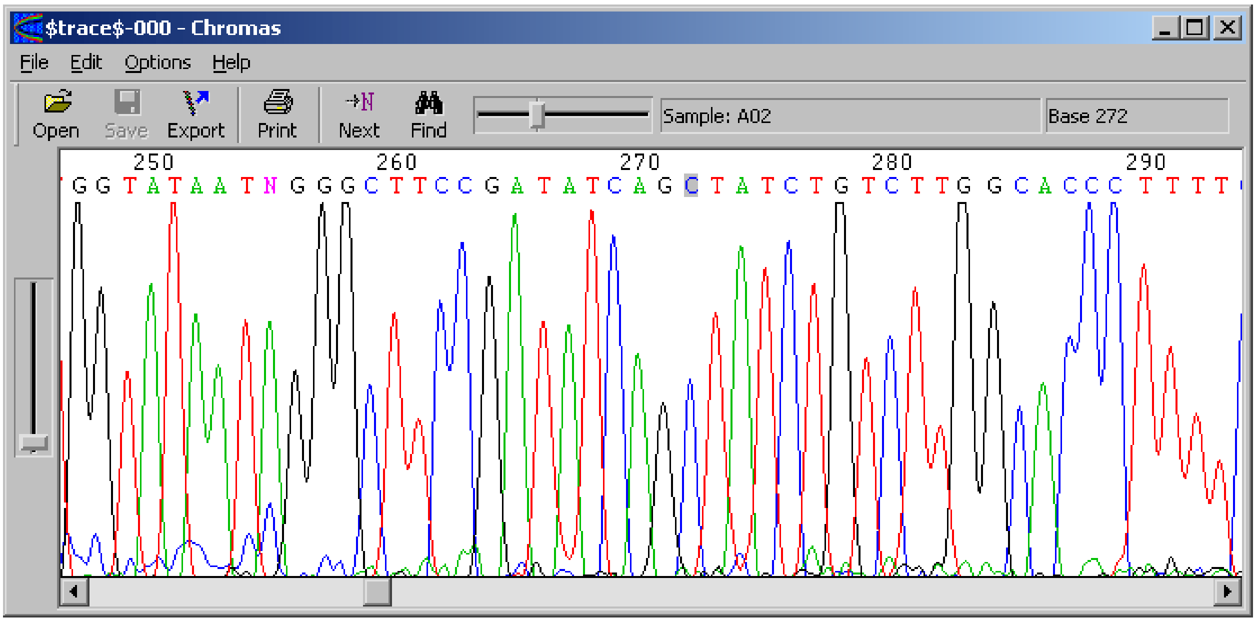
The link between the extracted sequence and the chromatogram is the *Long Filename* of the sequence *and* the path to the trace file folder set in *Preferences/Project Settings/Trace File Folder*.



Project Preferences

Enter the full path to the directory where the chromatograms for this project are located.

Provided this association is intact the chromatogram can be retrieved later and viewed with the *Chromas* program.



To maintain this connection it is important that the long sequence name is not changed in SEQtools. If you alter the long file name for a sequence, the link is broken and can only be re-established if you enter the name of the trace file corresponding to the SEQtools sequence again.

If you want to check a certain position in your sequence against the chromatogram, highlight the region in the main SEQtools editor and press CTRL+C to copy the region to the clipboard. The highlighted region in the sequence is colored blue to facilitate locating it.

In *Chromas*, click *Edit/Find*... to display the search form. Press CTRL+V to paste the selected region of your sequence into the search form of *Chromas* and click *Find*. SEQtools removes spaces, CR, LF, and numbers from the selected region, so it does not matter if your selection spans two lines.

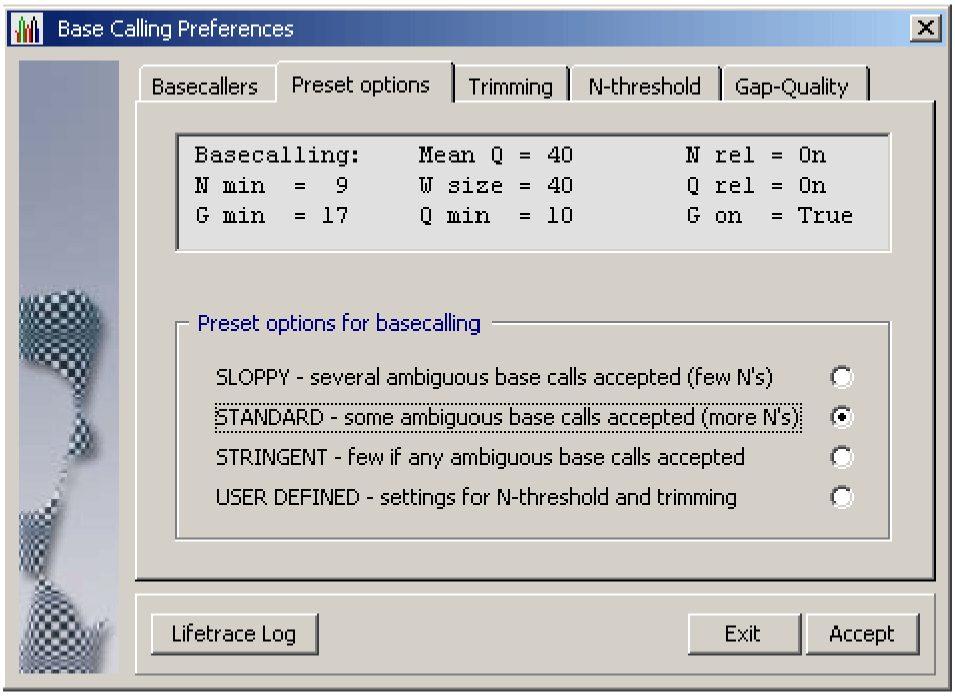
The advantage of keeping SEQtools formatted sequences and the original trace files separate is that all SEQtools functions, including automated annotation for example generated by blast searching can be maintained in the sequence headers.

4.1.4.1 convert_trace

4.1.4.1 Convert_Trace is the default program used by SEQtools to extract plain nucleotide information from chromatogram files. The extracted nucleotide sequence is generated by the basecalling performed by the application which created the chromatogram and does *not* allow the user to modify/adjust the way the basecalling is carried out.

4.1.4.2 lifetrace

4.1.4.2 LifeTrace on the other hand is a stand-alone basecaller which uses information included in the chromatogram to perform *de-novo* basecalling utilizing its own algorithm for calling bases.



```
Basecalling:    Mean Q = 40        N rel = On
N min  =  9     W size = 40        Q rel = On
G min  = 17     Q min  = 10        G on  = True
```

Preset options for basecalling

SLOPPY - several ambiguous base calls accepted (few N's) ○

STANDARD - some ambiguous base calls accepted (more N's) ●

STRINGENT - few if any ambiguous base calls accepted ○

USER DEFINED - settings for N-threshold and trimming ○

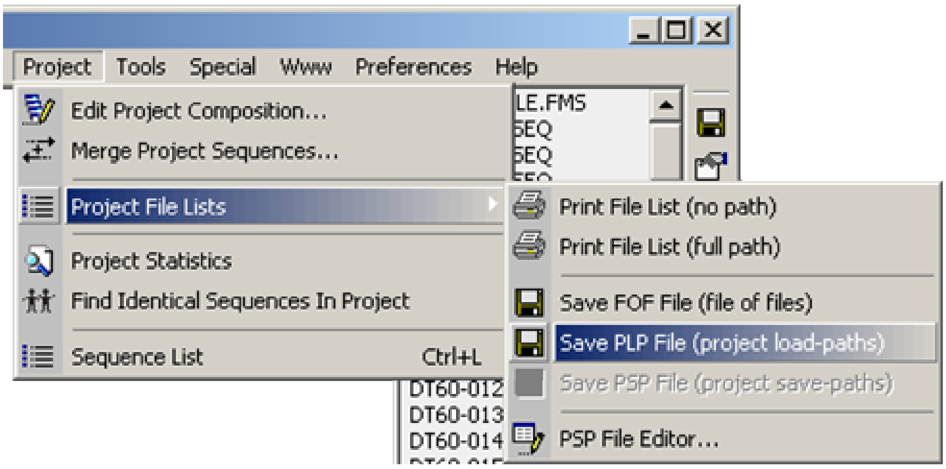
LifeTrace runs on Linux/Unix systems and requires a more complex setup than *convert_trace*. In brief: Sequences must be copied to a Linux/Unix computer running *LifeTrace* to generate the data files used by SEQtools to post-process the basecalling. The advantage is that the user has full control over the basecalling operation as well as of the post-processing by SEQtools. Take a look at the preferences form above to get an impression of the options available when *LifeTrace* is used for basecalling/extraction of the nucleotide sequence from a chromatogram.

LifeTrace is particularly effective when applied to MegaBACE capillary sequencing machines. A detailed description of the *LifeTrace* /SEQtools setup and interaction and the command line arguments are given on separate pages of this manual.

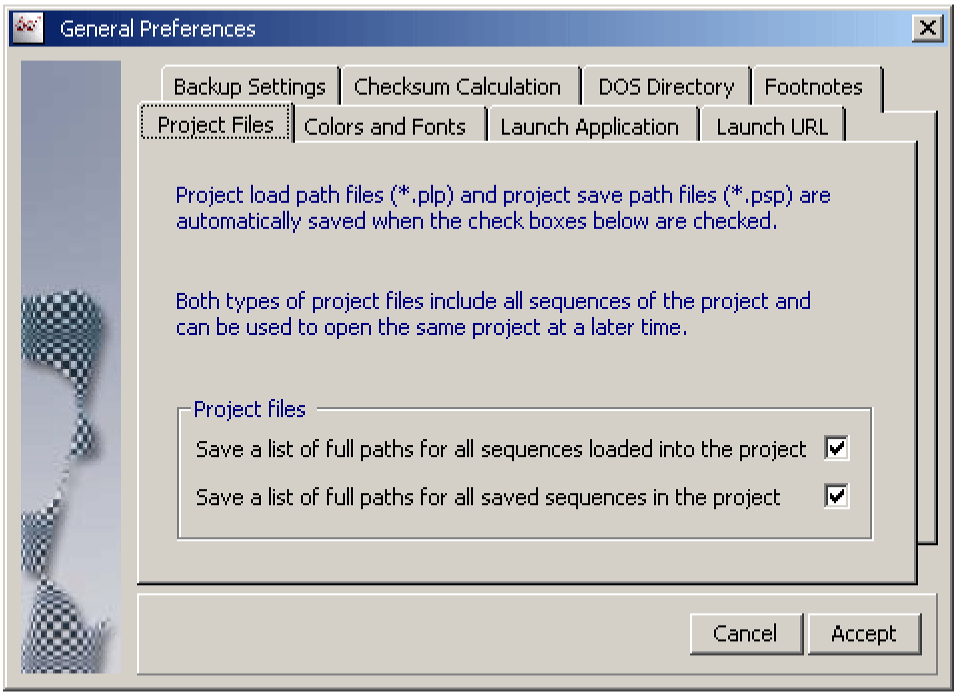
4.1.5 open existing project

If a *.psp (project save paths) or a *.plp (project load paths) for a project exists it is possible to re-open the project from the *Open Existing Project* menu. The *.psp and *.plp files are lists of full paths to all sequence files included in the project. The files may be located in different directories and can be single or multi-sequence files - or a mixture of the two types.

The *.plp and *.psp files can be saved by clicking the *Project/Project File Lists* as shown by the screenshot below.


Project
Tools
Special
Www
Preferences
Help
Edit Project Composition...
Merge Project Sequences...
Project File Lists
Project Statistics
Find Identical Sequences In Project
Sequence List Ctrl+L
Print File List (no path)
Print File List (full path)
Save FOF File (file of files)
Save PLP File (project load-paths)
Save PSP File (project save-paths)
PSP File Editor...
LE.FMS
SEQ
SEQ
DT60-012
DT60-013
DT60-014

The *.plp file is auto-generated when the project is created while the *.psp file is auto-built/re-built each time the project is saved. This option is enabled in *Preferences/General Settings/Project Files*



Project load path files (*.plp) and project save path files (*.psp) are automatically saved when the check boxes below are checked.

Both types of project files include all sequences of the project and can be used to open the same project at a later time.

Project files

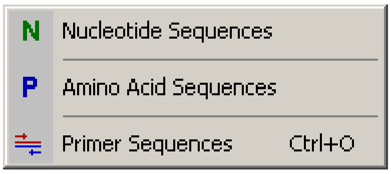
Save a list of full paths for all sequences loaded into the project ☑

Save a list of full paths for all saved sequences in the project ☑

4.1.6 enter sequences manually

In case you wish to enter sequences manually either by typing the sequence or by copy/paste from other applications or from additional instances of SEQtools you need to tell SEQtools which *type* (nucleotide, primer or protein) of sequences you intend to include in the project. When you choose this option, SEQtools sets the project type and opens an empty file ready for receiving the new sequence.

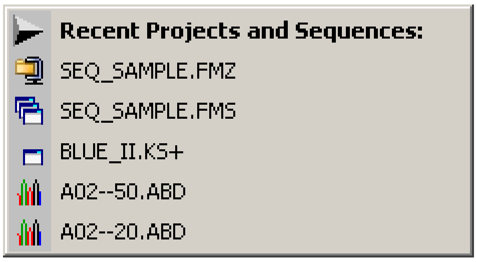
Each additional sequence requires that you first create a new, empty, page (see below) to hold the sequence before you start typing or copy/paste. Remember that a project can only hold one *type* of sequence



4.1.7 load recent project or sequence

SEQtools stores the last 20 opened sequence files (single and multi-sequence) in the *Open Recent Project or Sequence* list for easy loading of often accessed files. It is only possible to select and load *one* file from the list at a time. Note that this list may include sequence files belonging to different sequence *types*.

The different sequence file *formats* are indicated by different icons. To clear the list of recently opened files, click the title line of the list.



4.1.8 add more files to a project

Once a project is created more sequence files can be added to the project using the load form described in sections 4.1.3.. and 4.1.5. Note, however, that using the 4.1.3 sub-menu will close the current project and create a *new* SEQtools project while the *Add Files To Project...* add the selected files to the existing project.

Apart from this difference the load form works exactly in the way described in section 4.1.3.

It is also possible to add more sequences to the project using the *Add Recent Project Or Sequence*

While adding sequence files to the project SEQtools warns you if you load sequences with filenames already present in the project. If you choose to override the warning and accept multiple files with identical names, SEQtools will modify the filenames of such files if the project is saved as single sequence files in order to avoid overwriting the first saved file with subsequent sequence files with the same name.

Notice that the file *type* (nucleotide, primer or protein) of files to be added to an existing project must be of the same type as the files in the project.

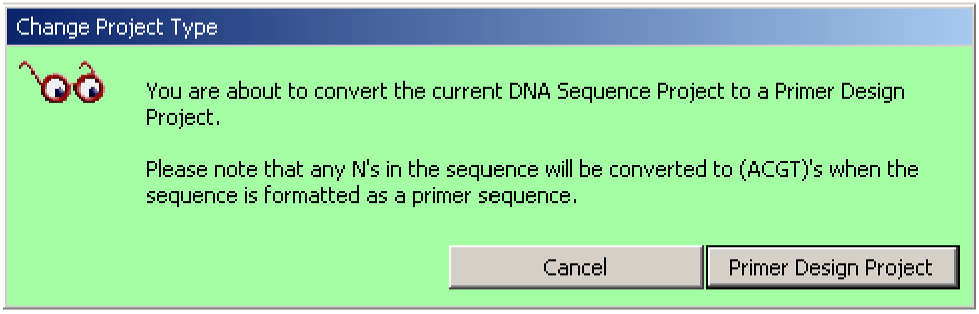
Sequences loaded with this function are *appended* to the list of sequences already in the project.

4.1.9 add an empty file to a project

Before you can add sequences to an existing project by typing the sequence or by copy/pasting the sequence from a different source you must first add an empty page to the project to hold the sequence. Click *Add Empty File To Project* to append an empty page to the end of an existing project.

4.1.10 convert project type

Occasionally it is convenient to be able to perform a blast search on Genbank databases with oligonucleotides designed for microarrays. This can most easily be done by loading the oligonucleotides into a primer project in SEQtools and subsequently convert the project to a nucleotide project. This function *Convert Project Type* enables you to convert primer projects to nucleotide projects and *vice versa*.



You are about to convert the current DNA Sequence Project to a Primer Design Project.

Please note that any N's in the sequence will be converted to (ACGT)'s when the sequence is formatted as a primer sequence.

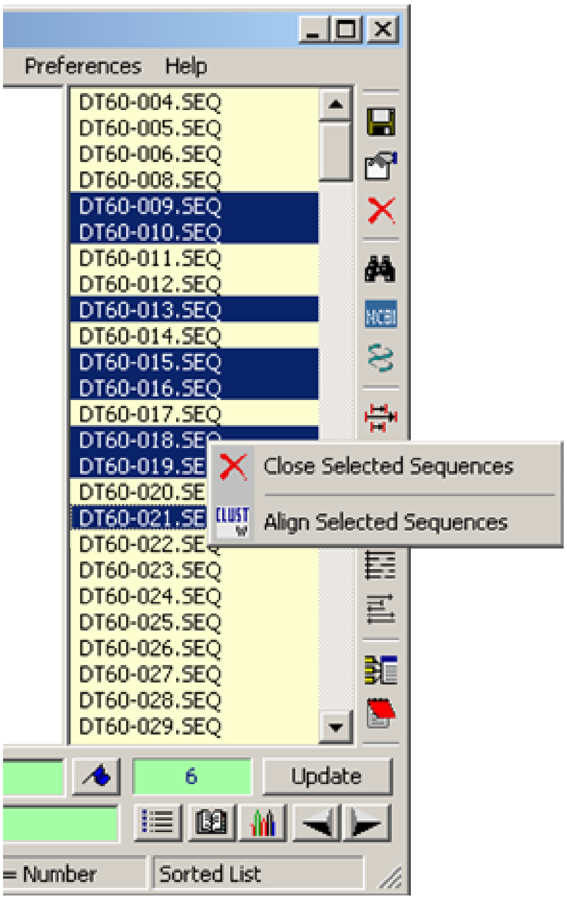
Important note: Converting a nucleotide project to a primer project will irreversibly remove all information stored in sequence headers due to the different design of the header structure of the two project types in SEQtools.

4.1.11 remove sequence from project

To remove a *single* sequences from a project simply highlight the sequence to be removed in the sequence list and click *Remove Sequence From Project*. The removed sequence is *not* removed from the hard disk, just no longer a member of the project.

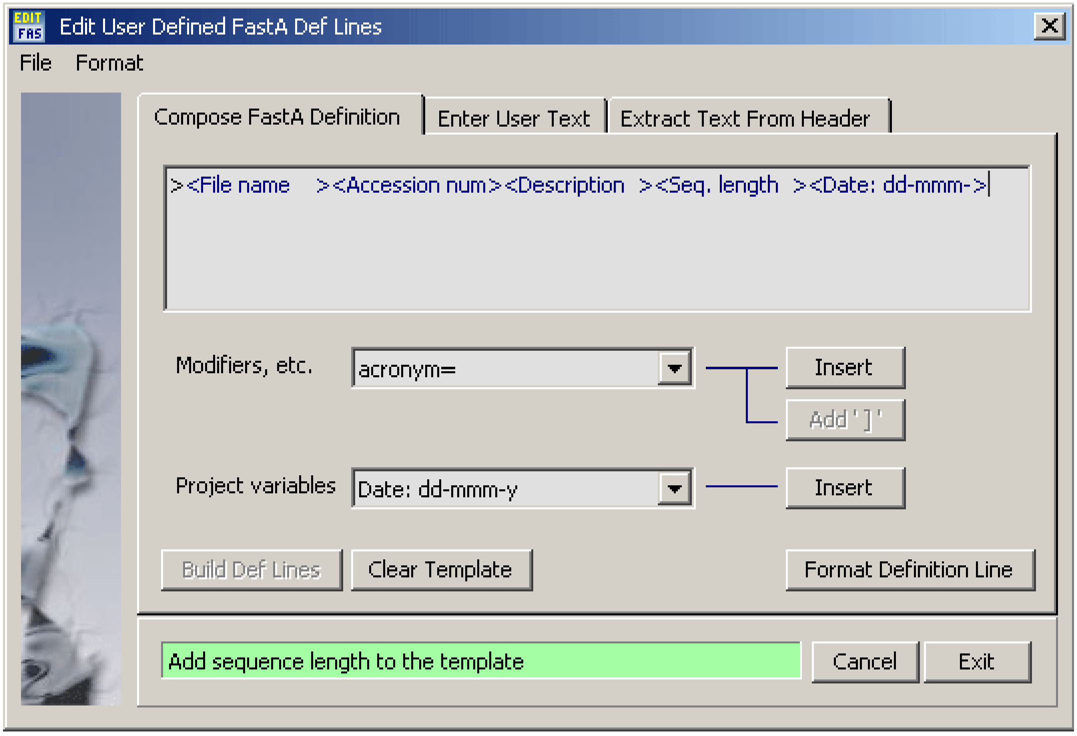
To remove a selection of sequences from a project proceed as follows: Hold down <CTRL> while clicking the sequences to be removed.

<Shift+Right-Click> on the sequence list to open the pop-up menu. Select *Close Selected Sequences* to remove the selected sequences from the project. Again, the sequences are *not* deleted from the hard disk but only removed from the project.



4.1.12 save project / export files

This function *File/Export Formats* formats the sequence and its header so that they can be loaded into other nucleotide and protein analysis programs. There is a special function which allow you to customize the single line header - the *Definition Line* - used in Fasta format.



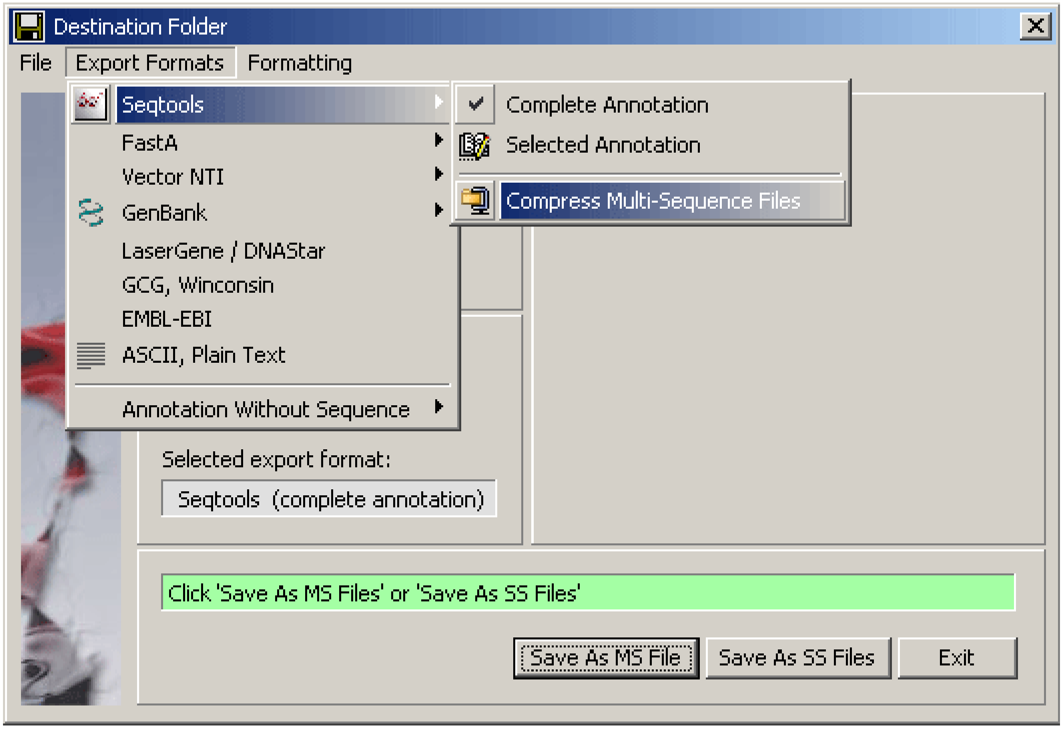
><File name ><Accession num><Description ><Seq. length ><Date: dd-mmm->

Modifiers, etc. acronym=

Project variables Date: dd-mmm-y

Add sequence length to the template

The different save/export formats supported by SEQtools are shown in the screenshot of the save/export form. Additional options are available for several of the export formats. Among these is an option for *compressing* multi-sequence SEQtools files which facilitates loading the file into a SEQtools project and saves disk space.

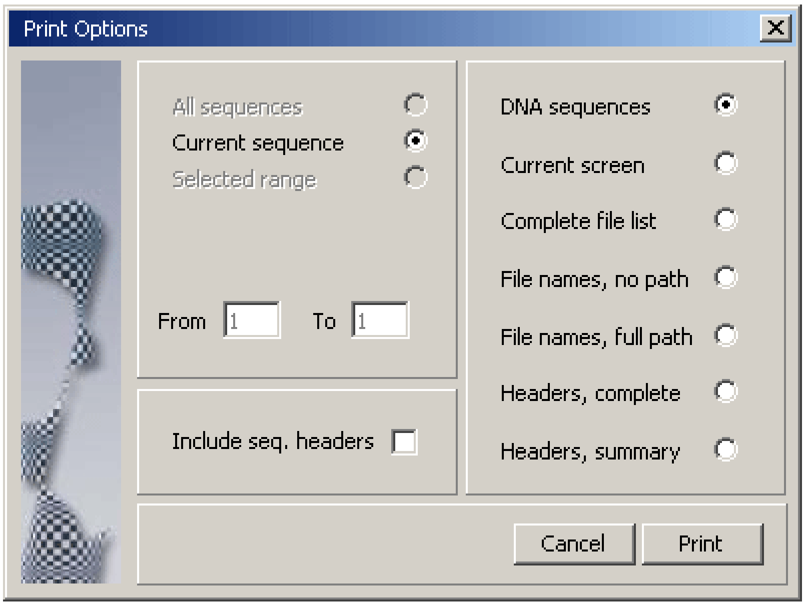


Selected export format: Seqtools (complete annotation)

Click 'Save As MS Files' or 'Save As SS Files'

4.1.13 print project

Printing projects is usually not a relevant option. In most cases the amount of data included in a project makes printing meaningless. As a consequence the printing facilities in have not been revised for a long time and may not work as indicated on the print form. Users in need for more sophisticated printing options are welcome to contact me for an update of the print functions. Till then I intend to leave things as they are...



All sequences ○
Current sequence ●
Selected range ○

From 1 To 1

Include seq. headers ☐

DNA sequences ●
Current screen ○
Complete file list ○
File names, no path ○
File names, full path ○
Headers, complete ○
Headers, summary ○

4.1.14 e-mail current sequence

With this function you can send the currently displayed plain sequence by e-mail with an attached comment. In case you need to send the entire project the sequences must be saved in a multi-sequence file and e-mailed as an attachment using the standard e-mail Windows program.

4.1.15 close project and exit

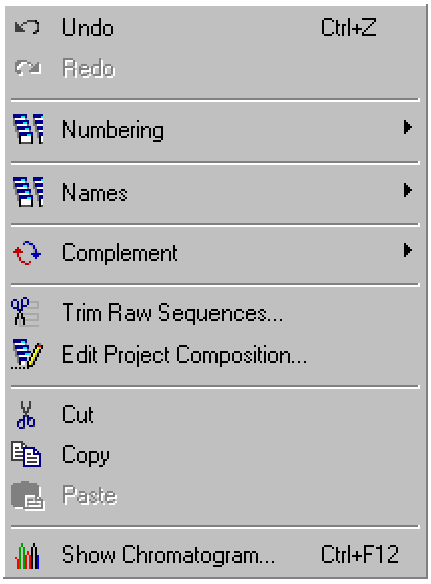
Before SEQtools closes the user is advised - twice - to save the project. Keep in mind that SEQtools keeps all project data in RAM until the project is saved. Closing SEQtools without saving the project will lead to irreversible loss of all data of the project.

Note that large batch blast search jobs - which may last several days - includes an option to auto-save the project every time a specified number of searches has been performed. This reduces the risk of data loss (in case of power failure for example) while the batch searching is running. See section 4.4 of the manual for a more detailed description of this option.

4.2 EDIT MENU

4.2.1 about functions for editing

Under the *Edit* menu is collected several functions all directed towards batch editing sequences and their names. Some are straight forward others more complex. Below each menu item is explained in some detail.

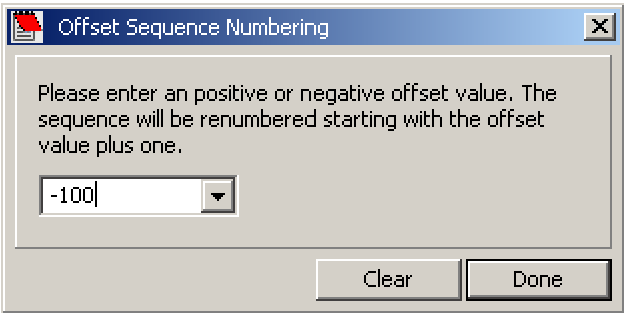


4.2.2 undo / redo changes

These options allow you to undo editorial changes. Note that changes are *not* recorded until you press the *Update* command button in the main editor.

4.2.3 numbering sequence residues

Simply enter a positive or negative value to offset the sequence numbering. Enter a zero to get the normal numbering back.



Please enter an positive or negative offset value. The sequence will be renumbered starting with the offset value plus one.

-100

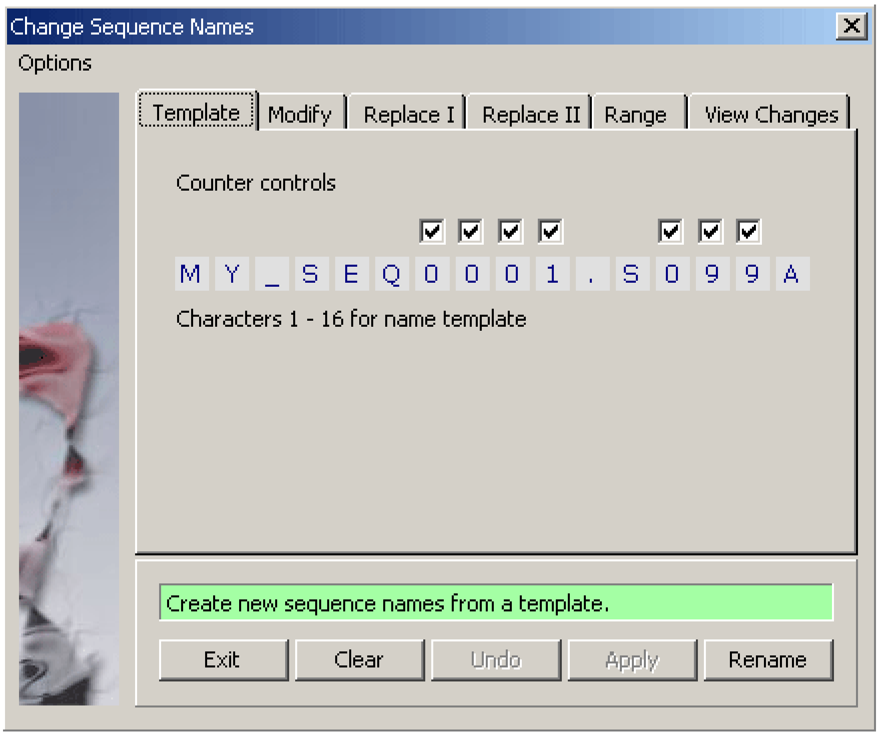
4.2.4 renaming sequences

Manual editing of individual file names can be performed by clicking the field displaying the current sequence name on the main editor form. Editing the names of individual files should be done *after* batch-renaming all files of the project.

Batch-renaming will irreversibly eliminate any changes previously made to the names of individual files. The options for batch editing sequence names are quite complex allowing you to change/edit/customize names in almost any way you can imagine. On the last tab of this form you can inspect the changes *before* you implement them by clicking the *Apply* command button..

4.2.1.1 Change names - With this function entirely new sequence names can be generated based on a template of 16 characters (the maximum length of sequence names in SEQtools). Type the characters you wish including numerical characters. In the latter case a check box appears above the character field. Putting a check mark in one or more of the check boxes creates a *counter* which will increment by one per sequence.

The example below includes two counters, a 4-digit and a 3-digit counter. See the result of the renaming operation below. Note that at least one of the counters must be able to hold the total number of sequences in the project.



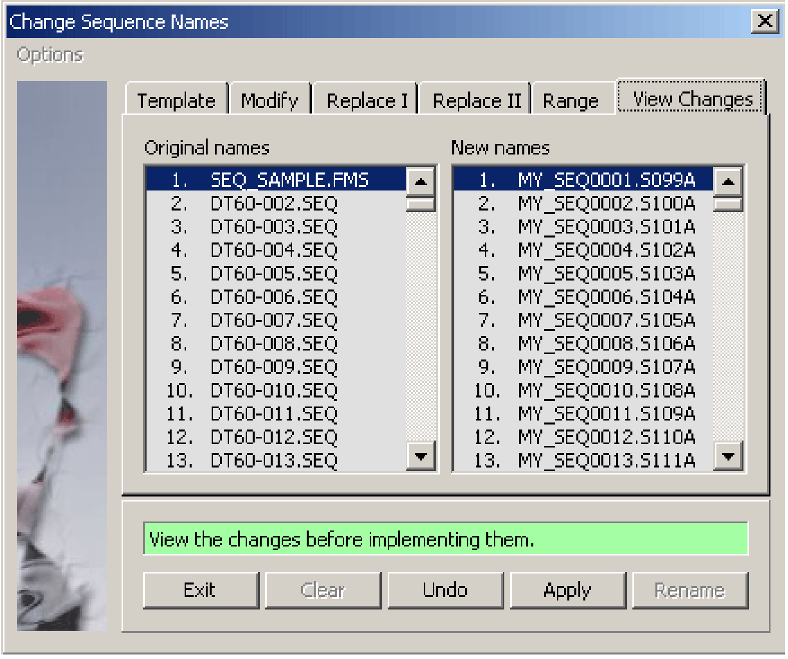
Counter controls

						☑	☑	☑	☑			☑	☑	☑	
M	Y	_	S	E	Q	0	0	0	1	.	S	0	9	9	A

Characters 1 - 16 for name template

Create new sequence names from a template.

4.2.4.2 View renamed sequences - View the changed sequence names on the panel to the right before implementing the names by pressing the *Apply* command button.



Original names		New names	
1.	SEQ_SAMPLE.FMS	1.	MY_SEQ0001.S099A
2.	DT60-002.SEQ	2.	MY_SEQ0002.S100A
3.	DT60-003.SEQ	3.	MY_SEQ0003.S101A
4.	DT60-004.SEQ	4.	MY_SEQ0004.S102A
5.	DT60-005.SEQ	5.	MY_SEQ0005.S103A
6.	DT60-006.SEQ	6.	MY_SEQ0006.S104A
7.	DT60-007.SEQ	7.	MY_SEQ0007.S105A
8.	DT60-008.SEQ	8.	MY_SEQ0008.S106A
9.	DT60-009.SEQ	9.	MY_SEQ0009.S107A
10.	DT60-010.SEQ	10.	MY_SEQ0010.S108A
11.	DT60-011.SEQ	11.	MY_SEQ0011.S109A
12.	DT60-012.SEQ	12.	MY_SEQ0012.S110A
13.	DT60-013.SEQ	13.	MY_SEQ0013.S111A

View the changes before implementing them.

4.2.4.3 Modify sequence names - This function makes it possible to make complex changes to parts of the file names without affecting other parts of the names. The function initially separates the sequence name into the *title* and *extension* and treats the two components of the file name independently.

With this function, characters can be replaced or removed inside the name. Addition or replacement can be made from the left or from the right of the two parts of the name. The last tab on the form lists the original and the new names of all files of the project. With this function parts of the old file names can be preserved while unwanted characters can be removed. The new file names are validated and renaming disrupted if the renaming results in duplicate file names.

Clicking the *Apply* command button activates the renaming of all files of the project according to the settings of the options and text. If the renaming operation generates duplicate file names, the operation is interrupted and the remaining original names are preserved.

Clicking *Undo* eliminates all changes to the file names of the project. This does not affect changes made to the sequences and their headers.

The *Close* command button closes the window preserving the current changes as listed in the new names combo box. To cancel without renaming, reset the sequence names *before* closing.

File name characters - This text field can hold up to 8 characters which can be added to or inserted into the current file names as selected by the options buttons.

Extension characters - This text field can hold up to 3 characters which can be added to or inserted into the current file extension as selected by the options buttons.

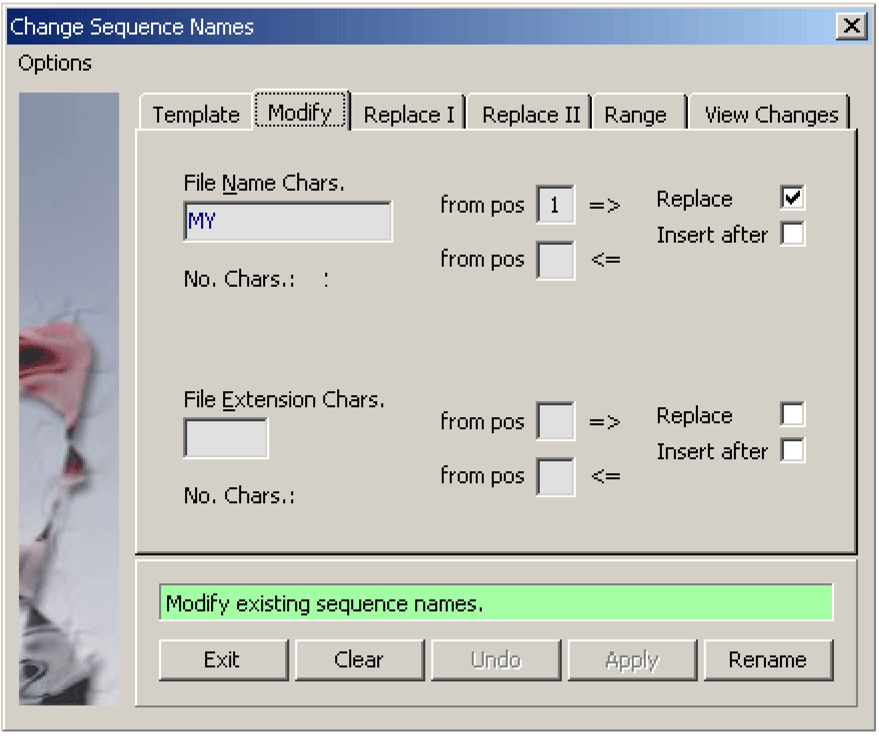
Position fields - The values entered in these fields give the position of insertion or replacements from left or right of names and extensions.

Add / Insert - This option causes the characters in the *text/extension* fields to be added/inserted into the file names/extensions at the position from the left/right as set by the two position fields. Inserting spaces into file names/extensions has no effect on the file names or extensions.

If the number of characters to be added causes the length of the name plus extension to exceed a total length of 16 characters the excess characters are truncated from the left or right end of the names and extensions.

Replace - This option causes the characters in the text/extension fields to *replace* the same number of characters from the left or right of the file names/extensions as set in the position fields. Replacing characters with spaces deletes the characters from names/extensions.

Increment - If the text boxes *only* contain numerical characters a check box appears which, when checked, causes the increment of the value in the text boxes (increment is one per sequence of the project).



File Name Chars. MY

from pos 1 => Replace ☑

Insert after □

from pos <=

No. Chars.: :

File Extension Chars.

from pos => Replace □

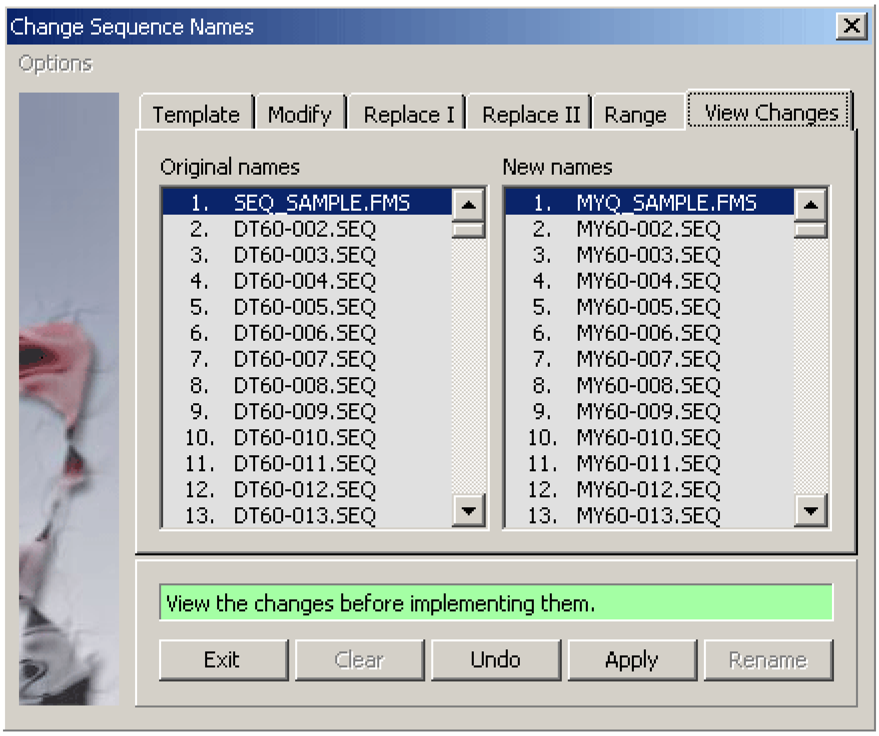
Insert after □

from pos <=

No. Chars.:

Modify existing sequence names.

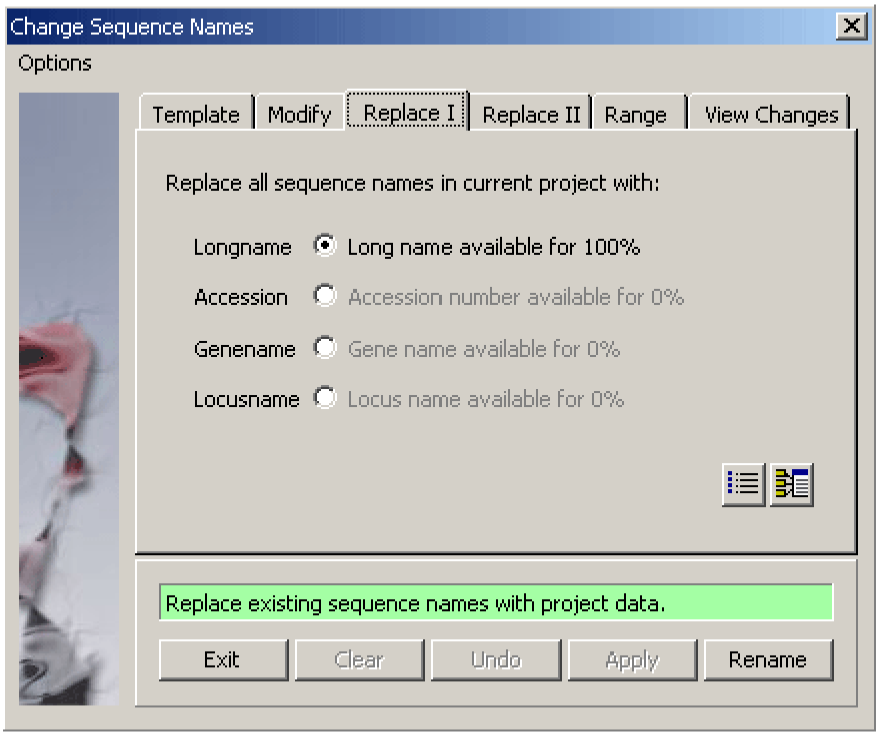
4.2.4.4 View modified names - View the changed sequence names on the panel to the right before implementing the names by pressing the *Apply* command button.



Original names	New names
1. SEQ_SAMPLE.FMS	1. MYQ_SAMPLE.FMS
2. DT60-002.SEQ	2. MY60-002.SEQ
3. DT60-003.SEQ	3. MY60-003.SEQ
4. DT60-004.SEQ	4. MY60-004.SEQ
5. DT60-005.SEQ	5. MY60-005.SEQ
6. DT60-006.SEQ	6. MY60-006.SEQ
7. DT60-007.SEQ	7. MY60-007.SEQ
8. DT60-008.SEQ	8. MY60-008.SEQ
9. DT60-009.SEQ	9. MY60-009.SEQ
10. DT60-010.SEQ	10. MY60-010.SEQ
11. DT60-011.SEQ	11. MY60-011.SEQ
12. DT60-012.SEQ	12. MY60-012.SEQ
13. DT60-013.SEQ	13. MY60-013.SEQ

View the changes before implementing them.

4.2.4.5 Replace I - Batch replace sequence names with one of the enabled categories on this tab. Disabled options imply that the relevant information is not available for *all* sequences of the project.



Replace all sequence names in current project with:

Longname ● Long name available for 100%

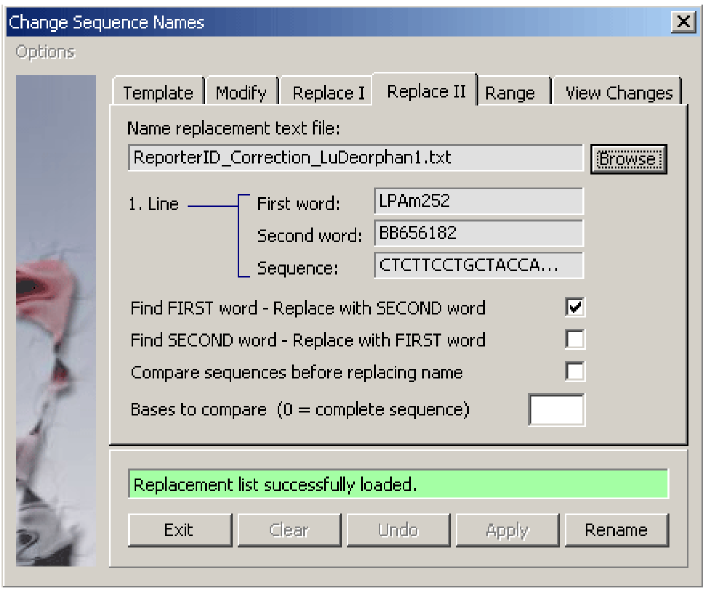
Accession ○ Accession number available for 0%

Genename ○ Gene name available for 0%

Locusname ○ Locus name available for 0%

Replace existing sequence names with project data.

4.2.4.6 Replace II - Complex function to replace *project* sequence names with the names for the *same* sequence but contained in a text file with different annotation. Eg. sample: Assume you have an annotated project and a fasta file with the *same* sequences. With this function you can replace the project sequence name with the first or the second word of the fasta definition line. Before replacement takes place the two sequences are compared and only identical sequences will be renamed.



Name replacement text file: ReporterID_Correction_LuDeorphan1.txt

1. Line
- First word: LPAm252
- Second word: BB656182
- Sequence: CTCTTCCTGCTACCA...

Find FIRST word - Replace with SECOND word ☑

Find SECOND word - Replace with FIRST word □

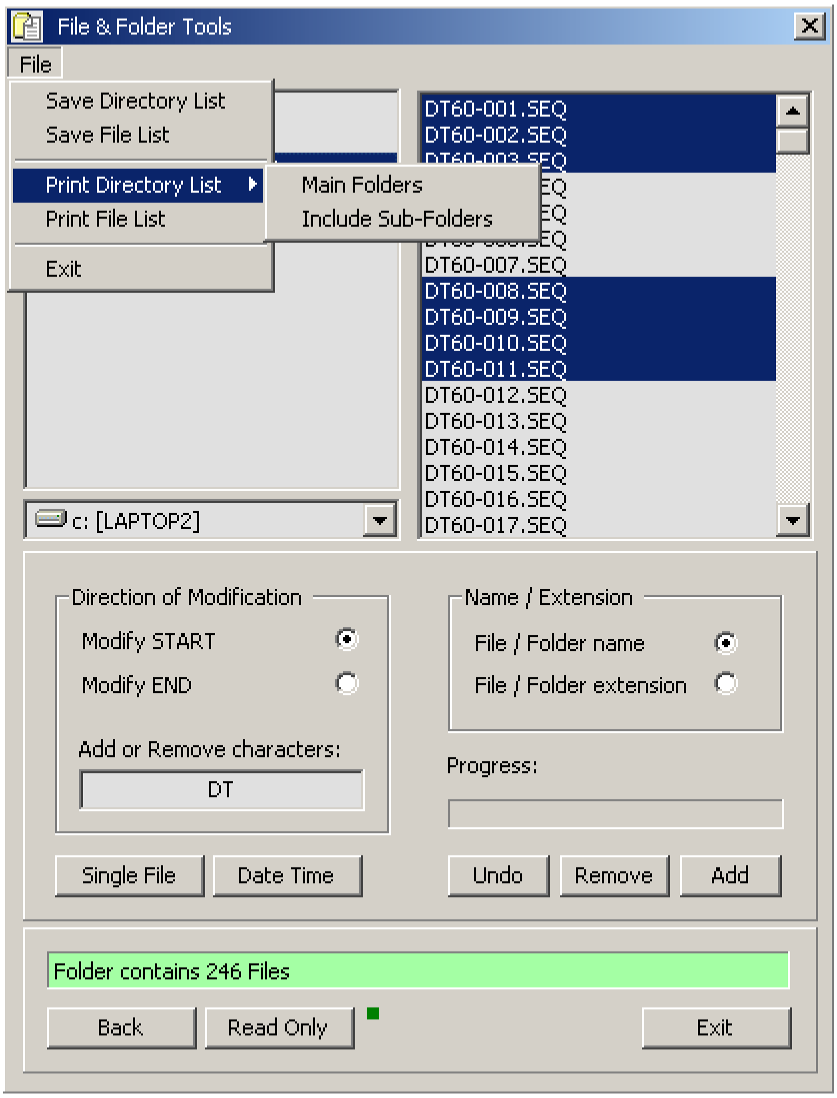
Compare sequences before replacing name □

Bases to compare (0 = complete sequence)

Replacement list successfully loaded.

4.2.4.7 File and Folder Tools - This small program enables you to carry out a number of operations on file and folder names. You can edit the file titles and extensions, change file dates, print and save file and folder lists etc.

The program is very useful if - for example - you want to print out an index of the content of a CD or change all file dates to the current date.



DT60-001.SEQ
DT60-002.SEQ
DT60-007.SEQ
DT60-008.SEQ
DT60-009.SEQ
DT60-010.SEQ
DT60-011.SEQ
DT60-012.SEQ
DT60-013.SEQ
DT60-014.SEQ
DT60-015.SEQ
DT60-016.SEQ
DT60-017.SEQ

Direction of Modification
Modify START ●
Modify END ○
Add or Remove characters: DT

Name / Extension
File / Folder name ●
File / Folder extension ○

Progress:

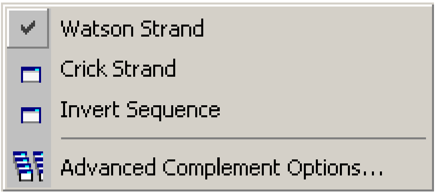
Folder contains 246 Files

4.2.5 complementing sequences

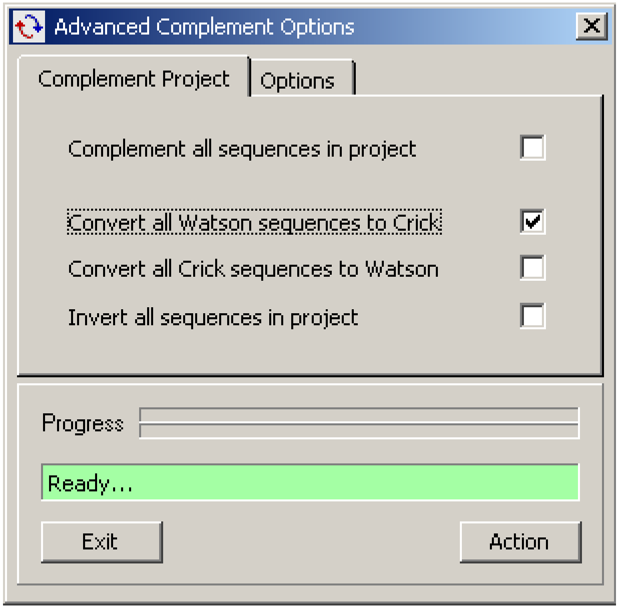
The *Watson / Crick* options generates the complementary DNA sequence and displays it with the 5' end to the left. *Invert* sequence inverts the current sequence and should be used with caution. The function is useful when copying sequences written 3' to 5'. In all other cases , i.e. with sequences written 5' to 3' create the inverted sequence will have no relationship to the original sequence.

Note - The information describing the orientation of the DNA sequence is saved with the file and retrieved when the file is loaded.

In the sequence lists the following codes are used to indicate the orientation of the sequence: *WS* - Watson strand, *CS* - Crick strand, *WI* - Inverted Watson strand and *CI* - Inverted Crick strand. In cases where orientation information cannot be retrieved or is incomplete, ?'s replaces one or both orientation characters.



The *Complement / Invert* operation can be performed on the entire project by using the *Advanced Complement Options*, a batch version of the above functions.



Complement all sequences in project □

Convert all Watson sequences to Crick ☑

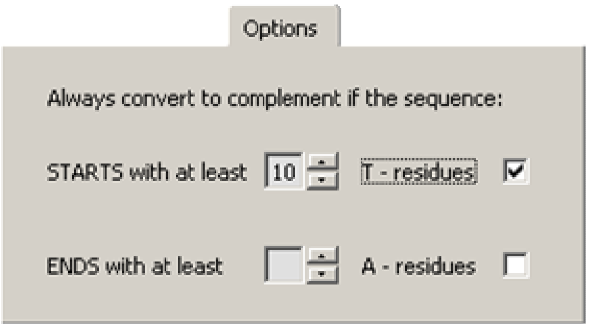
Convert all Crick sequences to Watson □

Invert all sequences in project □

Progress

Ready...

In case you only wish to batch complement polyA sequences set the minimum number of A's / T's for complementation to be performed.



Always convert to complement if the sequence:

STARTS with at least 10 T - residues ☑

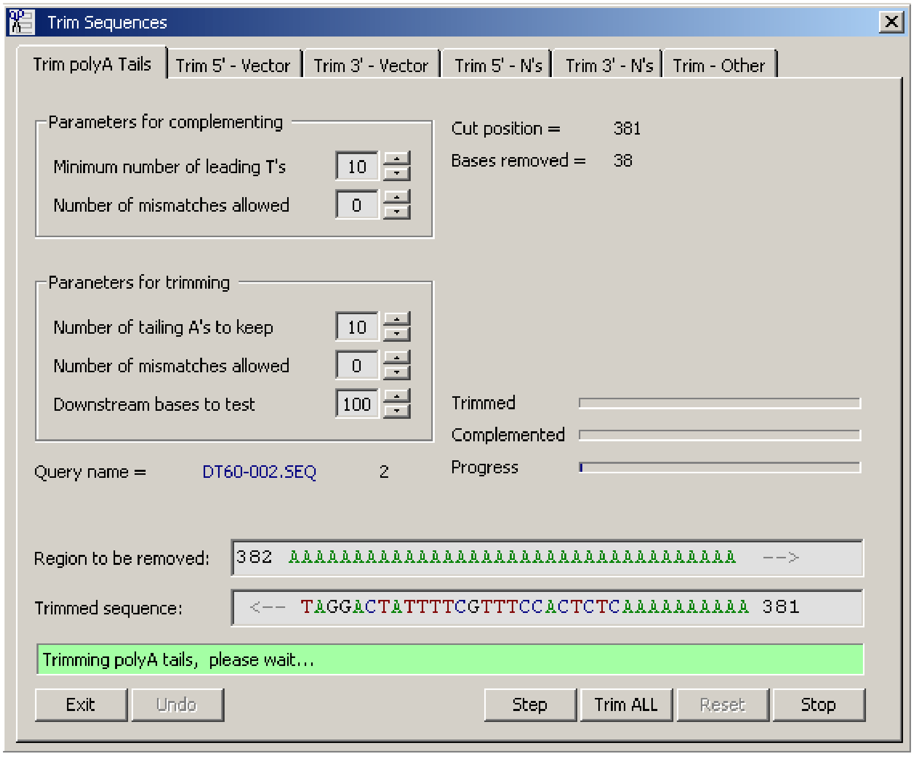
ENDS with at least A - residues □

4.2.6 trim raw sequences

This form includes five utilities for processing raw sequence data. All functions allow you either to process sequences one-at-a-time in *step mode* or to launch *auto-trimming*. While auto-trimming is running the operation can be paused and the user taking over continuing stepwise. All functions also include undo and reset buttons letting you reset trimmed sequences contained in the project to the state prior to a trimming operation. To save resources, you have the option of turning the *undo* function off before opening the form. In this case, the undo button is not shown.

4.2.6.1 Remove PolyA Tails - This function is designed for removing all bases upstream of a leading polyT region. In EST sequencing from the 3' end all inserts normally contain stretch of T corresponding to the polyA tail of the cDNA clone.

In situations where the sequencing primer position is very close to the start of the insert, the upstream vector part of the sequence is often biased by dye terminator signals and is not recognized by a comparison with the sequence of the vector. This function only considers the T's of the and thus trims correctly, also in cases where upstream vector sequence is ambiguous.
The options and the output of the function is illustrated by the screen dump below. In cases where you wish to reduce the length of leading T stretches, this can be done by entering the maximum number of T's to retain after trimming.



Parameters for complementing

Minimum number of leading T's 10

Number of mismatches allowed 0

Paraneters for trimming

Number of tailing A's to keep 10

Number of mismatches allowed 0

Downstream bases to test 100

Cut position = 381

Bases removed = 38

Query name = DT60-002.SEQ 2

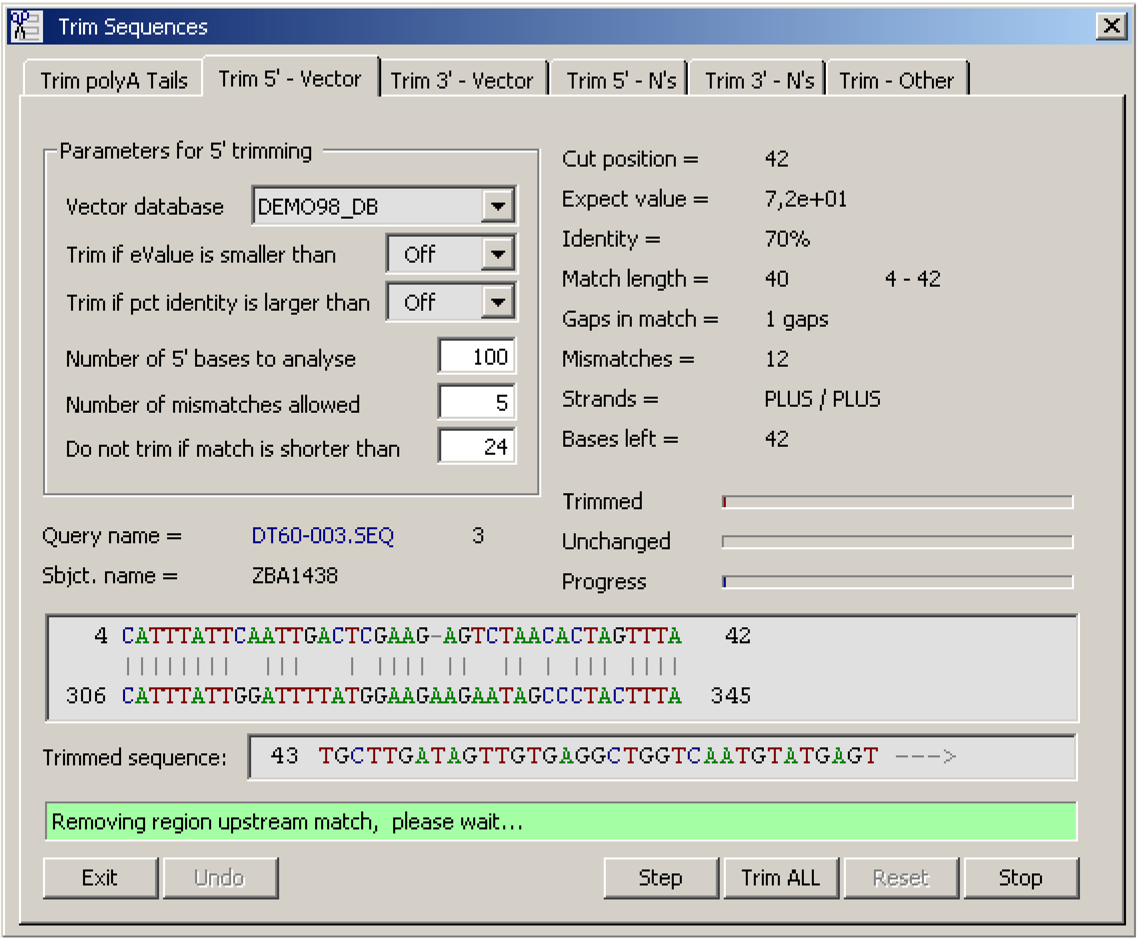
Region to be removed: 382 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -->

Trimmed sequence: <-- TAGGACTATTTTCGTTTCCACTCTCAAAAAAAAAA 381

Trimming polyA tails, please wait...

4.2.6.2 Remove Vector Sequence - Based on a database containing the sequence of the cloning vector(s) this function performs a blastn search, evaluates the result and trims the sequence if the selected criteria are met. The matching region and the start of the sequence after trimming are displayed in the two fields if you use the step option otherwise the main editor form is hidden to avoid using resources on updating and displaying the sequence. The settings as well as an example of the output is shown below.

Please note that this function require that a local vector database is already created. Use the functions for creating local databases if a suitable vector database is not available.



Parameters for 5' trimming

Vector database: DEMO98_DB

Trim if eValue is smaller than: Off

Trim if pct identity is larger than: Off

Number of 5' bases to analyse: 100

Number of mismatches allowed: 5

Do not trim if match is shorter than: 24

Cut position = 42

Expect value = 7,2e+01

Identity = 70%

Match length = 40 4 - 42

Gaps in match = 1 gaps

Mismatches = 12

Strands = PLUS / PLUS

Bases left = 42

Query name = DT60-003.SEQ 3

Sbjct. name = ZBA1438

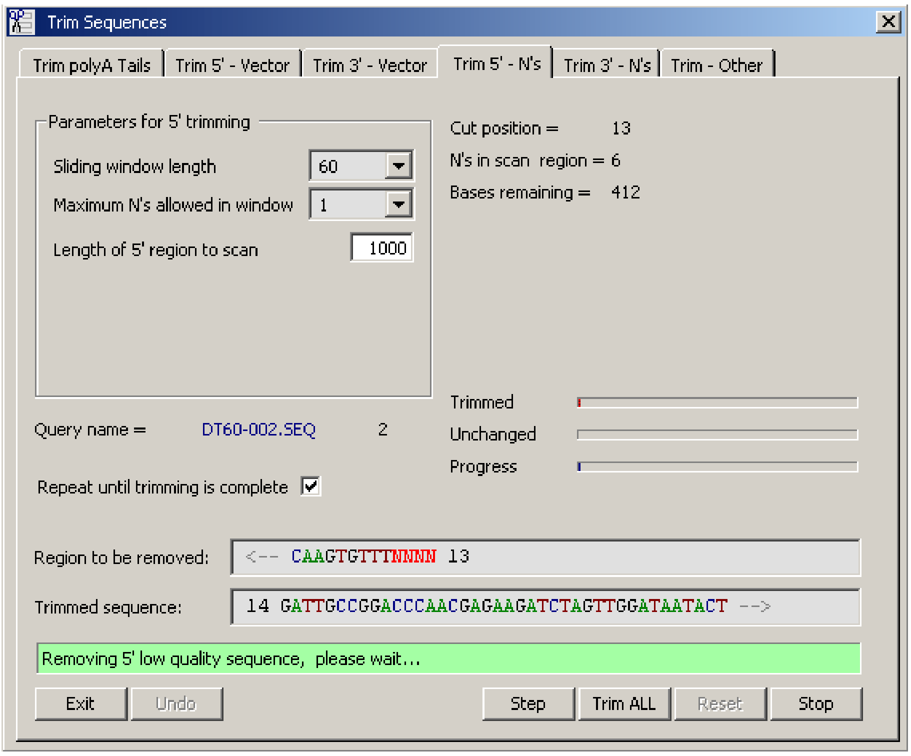
```
  4 CATTTATTCAATTGACTCGAAG-AGTCTAACACTAGTTTA   42
    ||||||||  |||   | |||| ||  || | ||| ||||
306 CATTTATTGGATTTTATGGAAGAAGAATAGCCCTACTTTA  345
```

Trimmed sequence: 43 TGCTTGATAGTTGTGAGGCTGGTCAATGTATGAGT --->

Removing region upstream match, please wait...

4.2.6.3 Remove Low Quality Sequence - After removal of vector sequence, low quality sequence regions can be automatically removed from the 5' and 3' ends of the raw sequences. The function determines the number of N's in a window sliding from the start/end of the sequence. The first time a window-sized region is encountered which meets the selected criteria, trimming occurs at the most upstream/downstream position of the window. By default trimming is repeated until all low quality regions are removed.

This function for removing low quality sequence is less accurate - but is much simpler to use - than the function included with the basecalling facility exploiting the external basecaller*LifeTrace*.



Sliding window length 60

Maximum N's allowed in window 1

Length of 5' region to scan 1000

Cut position = 13

N's in scan region = 6

Bases remaining = 412

Query name = DT60-002.SEQ 2

Repeat until trimming is complete ☑

Region to be removed: <-- CAAGTGTTTNNNNN 13

Trimmed sequence: 14 GATTGCCGGACCCAACGAGAAGATCTAGTTGGATAATACT -->

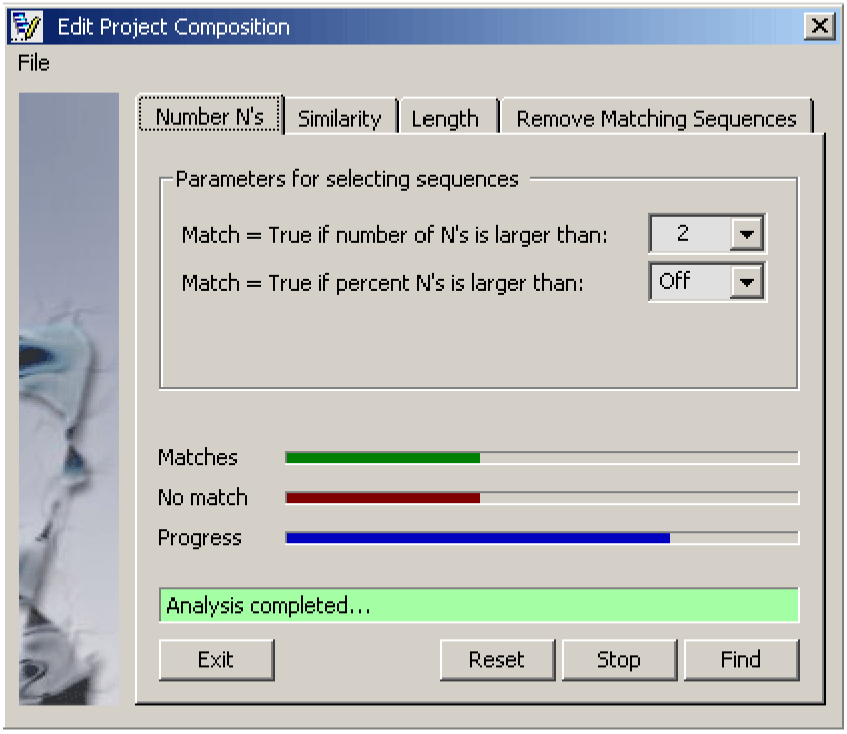
Removing 5' low quality sequence, please wait...

4.2.6.4 Simple trimming - This function (not illustrated) allows you to either cut the sequences at fixed 5' and 3' positions or to enter a 5' and 3' string which must *exactly* match the sequence for trimming to occur. Cutting occurs at the first position downstream of the 5' string and at the first position upstream of the 3' string. If a perfect match is not found, no cutting occurs.

4.2.7 edit project composition

With the functions on this form you can edit the composition of the current project by removing specific sequence groups such as low quality sequences (with a large number of N's), sequences with significant match to vector sequences etc.

4.2.7.1 Remove low quality sequences - Enter either maximum number or percentage of N' accepted in a sequence and click the *Find* command button. The function will analyze the project and display the result in the results tab, *Remove Matching Sequences*. Each sequence is labeled *True* or *False* indicating whether or not the specified criteria were met.



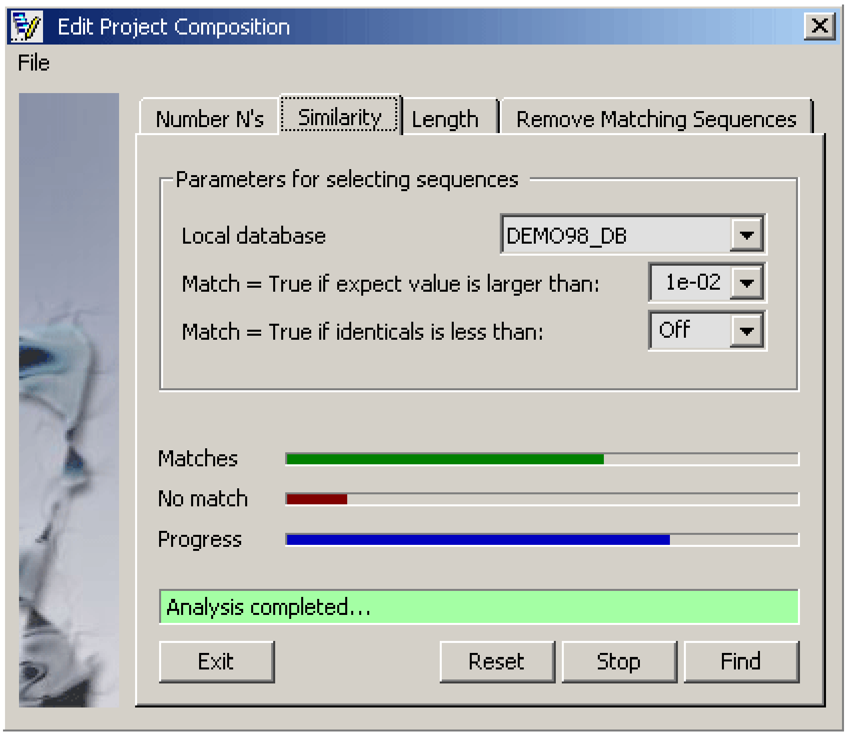
Parameters for selecting sequences

Match = True if number of N's is larger than: 2

Match = True if percent N's is larger than: Off

Analysis completed...

4.2.7.2 Similarity analysis - With this function each sequence in the project is compared to the selected local database. Running the function with the set parameters then divides the sequences contained in the project into a *True* and a *False* group. Either group can subsequently be removed from the project.



Parameters for selecting sequences

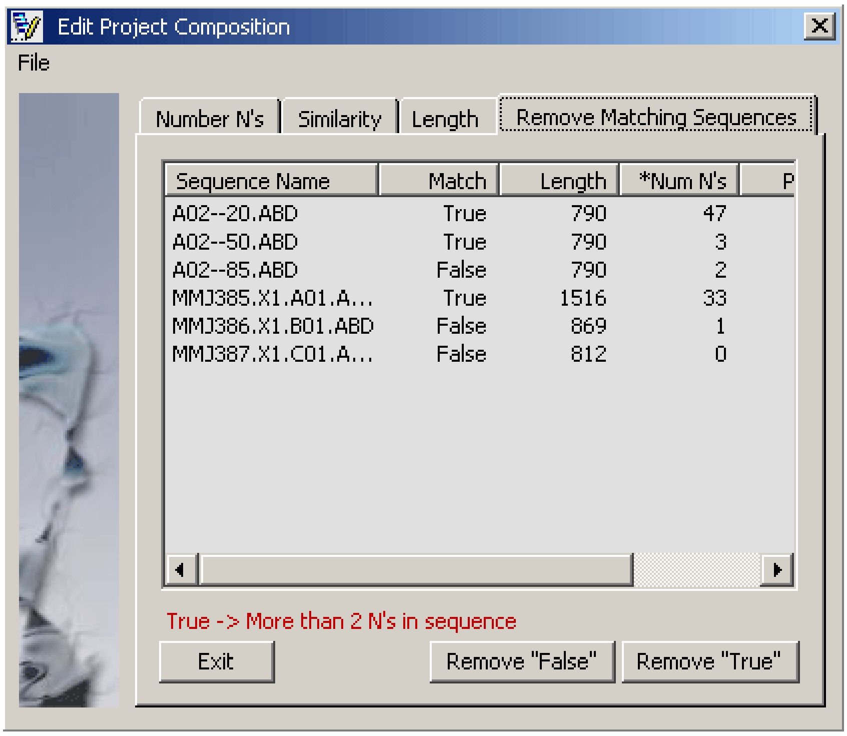
Local database: DEMO98_DB

Match = True if expect value is larger than: 1e-02

Match = True if identicals is less than: Off

Analysis completed...

4.2.7.3 Remove short/long sequences - The last function simply measures sequence length and splits the project sequences into two groups depending on the set length cutoff.



Sequence Name	Match	Length	*Num N's	P
A02--20.ABD	True	790	47	
A02--50.ABD	True	790	3	
A02--85.ABD	False	790	2	
MMJ385.X1.A01.A...	True	1516	33	
MMJ386.X1.B01.ABD	False	869	1	
MMJ387.X1.C01.A...	False	812	0	

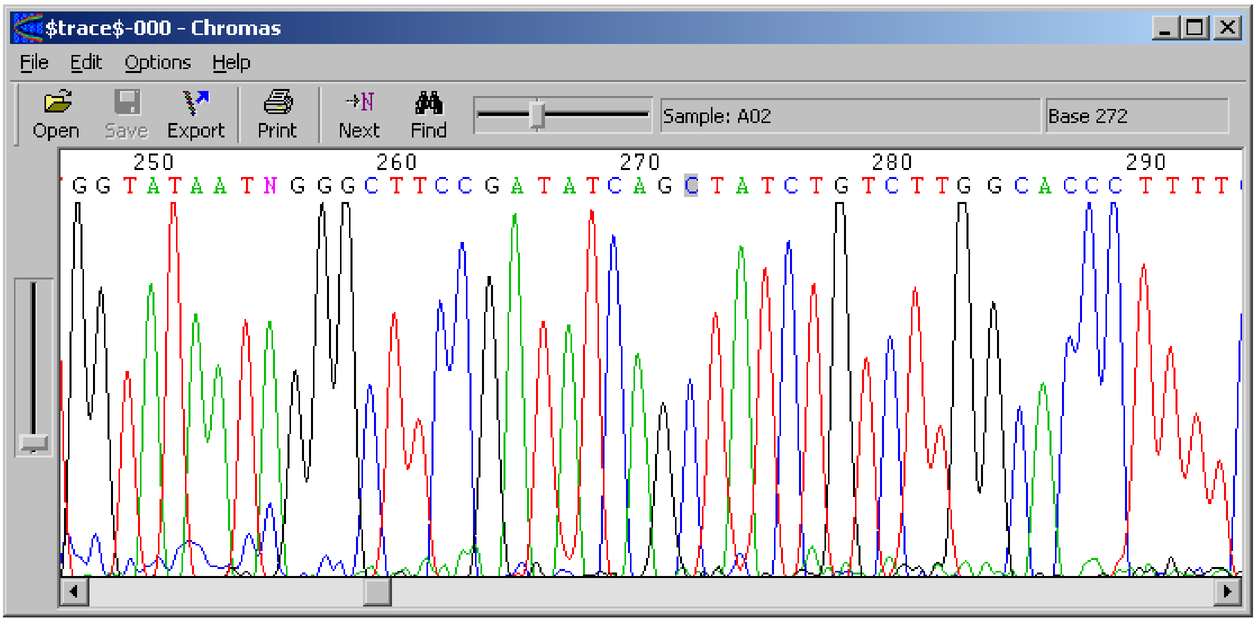
True -> More than 2 N's in sequence

4.2.8 cut / copy / paste

Trivial Windows functions for moving sequences from one instance of SEQtools to another, importing sequences etc. Click *Update* to format the an imported sequence.

4.2.9 show chromatogram

Viewing and editing chromatograms is performed by the external program. Chromas runs completely independent of SEQtools except for opening trace files from within SEQtools. Read more about chromatograms and the association between the project sequence and the chromatogram under the *File* and *Preferences* menus.



Sample: A02

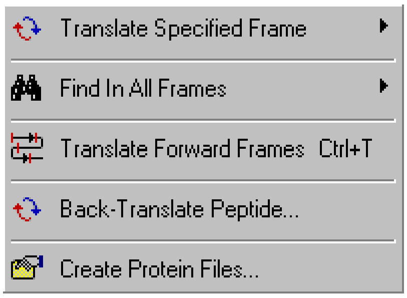
Base 272

```
 250                 260                 270                 280                 290
GGTATAATNGGGCTTCCGATATCAGCTATCTGTCTTGGCACCCTTTT
```

4.3 TRANSLATE MENU

4.3.1 about translating nucleotide sequences

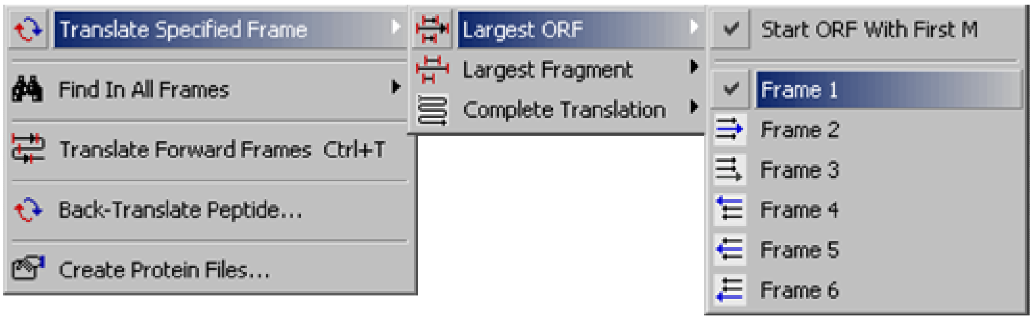
SEQtools includes several options for translating nucleotide sequences as well as for back-translating protein sequences. With the *Find In All Frames* is possible to identify the longest ORF in a nucleotide sequence. The *Translate Specific Frame* you can isolate (Copy/Paste) the isolated protein sequence). The *Translate Forward Frames* provide a link between the translated sequence and the underlying nucleotide sequence. Finally it possible to batch translate all nucleotide sequences contained in the project with the *Create Protein Files* function.


Translate Specified Frame
Find In All Frames
Translate Forward Frames Ctrl+T
Back-Translate Peptide...
Create Protein Files...

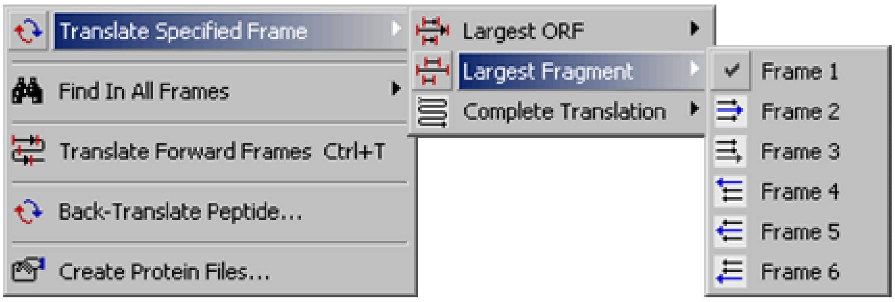
4.3.2 translate in specified frame

This function enables you to translate a nucleotide sequence in the specified reading frame. You have the option to display either the longest ORF, the longest fragment or a complete translation in the specified frame.

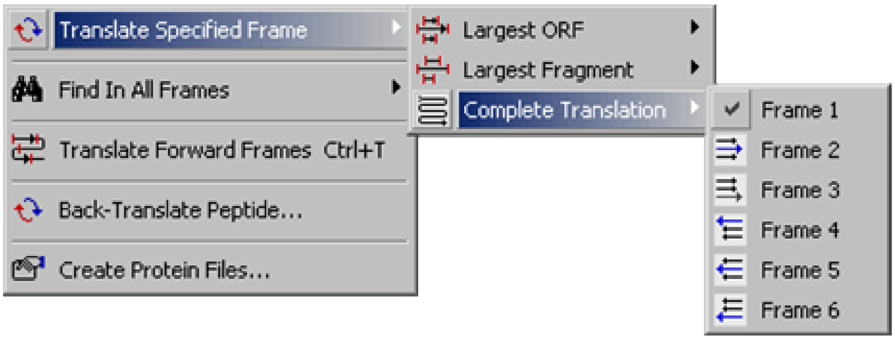
4.3.2.1 Largest ORF submenu.



4.3.2.2 Largest Fragment submenu.



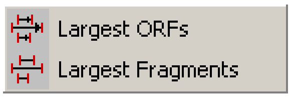
4.3.2.3 Complete Translation submenu.



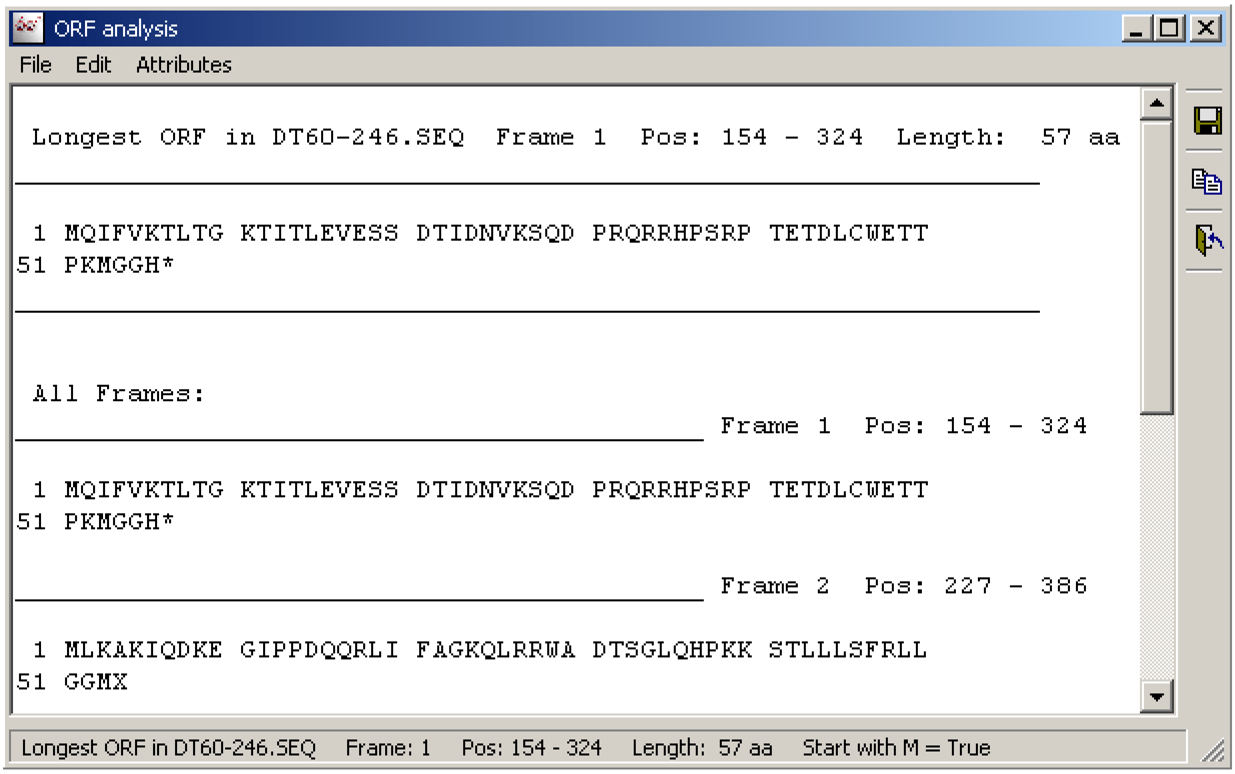
4.3.3 find in all frames

With this function it is possible easily to identify the longest ORF (open reading framing) or fragment (protein region without stop codons) in an unknown nucleotide sequence. The result is displayed in text form which lists the longest orfs/fragments in all six reading frames and the longest of them all.

With information you can re-translate the longest ORF/fragment with the *Translate Specific Frame* function described above to isolate the protein sequence.



The Result form listing ORF's or fragments in all six reading frames of the nucleotide sequence.



```
 Longest ORF in DT60-246.SEQ  Frame 1  Pos: 154 - 324  Length:  57 aa
______________________________________________________________________

 1 MQIFVKTLTG KTITLEVESS DTIDNVKSQD PRQRRHPSRP TETDLCWETT
51 PKMGGH*
______________________________________________________________________


 All Frames:
_______________________________________________  Frame 1  Pos: 154 - 324

 1 MQIFVKTLTG KTITLEVESS DTIDNVKSQD PRQRRHPSRP TETDLCWETT
51 PKMGGH*

_______________________________________________  Frame 2  Pos: 227 - 386

 1 MLKAKIQDKE GIPPDQQRLI FAGKQLRRWA DTSGLQHPKK STLLLSFRLL
51 GGMX
```

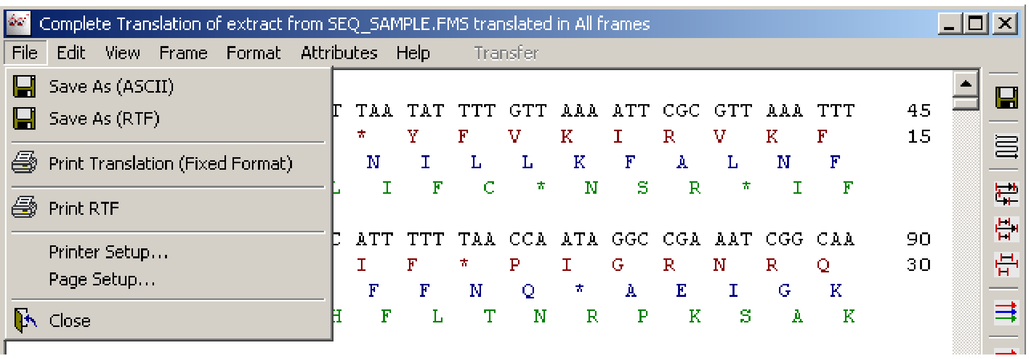
Longest ORF in DT60-246.SEQ   Frame: 1   Pos: 154 - 324   Length: 57 aa   Start with M = True

4.3.4 translate forward frames

The *Translate Forward Frames* displays the translation of the current DNA sequence or an extract thereof in each of the three *forward* reading frames or in all three forward reading frames simultaneously.

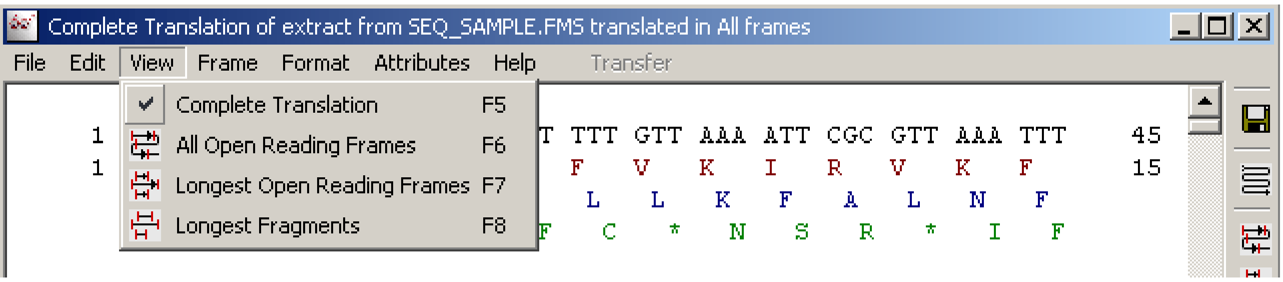
The line numbers correspond to the coordinates of the extracted sequence region. Stop codons are denoted by stars and uncertain (codons including one or more N's) amino acids by X's. The format of the DNA sequence is independent of the selected format in the sequence editor form with block length of 3 and line length of 60 bp.

4.3.4.1 File menu - contains save and prints options for the translated sequence.



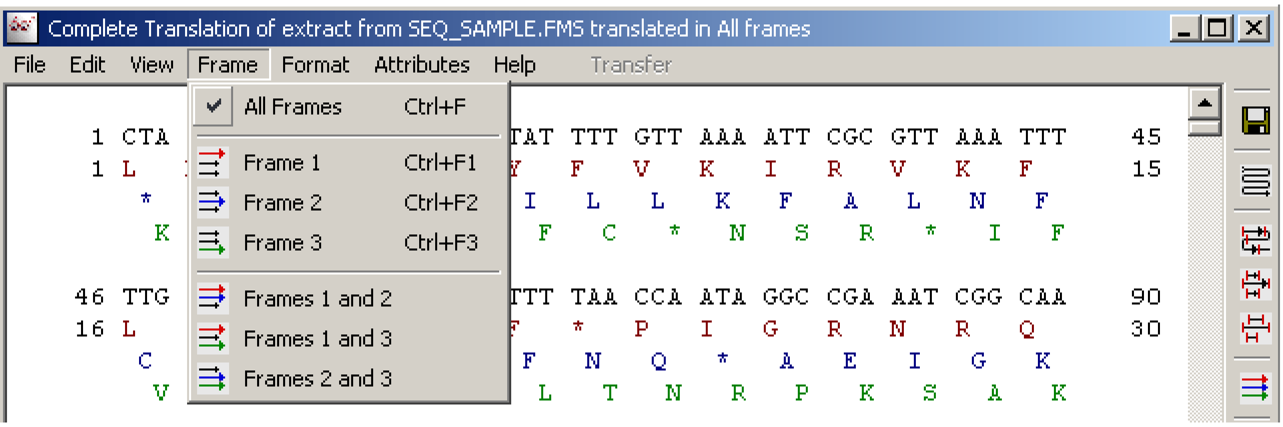
```
T TAA TAT TTT GTT AAA ATT CGC GTT AAA TTT    45
  *   Y   F   V   K   I   R   V   K   F      15
   N   I   L   L   K   F   A   L   N   F
L   I   F   C   *   N   S   R   *   I   F

C ATT TTT TAA CCA ATA GGC CGA AAT CGG CAA    90
  I   F   *   P   I   G   R   N   R   Q      30
   F   F   N   Q   *   A   E   I   G   K
H   F   L   T   N   R   P   K   S   A   K
```

4.3.4.2 View menu - includes the available translate options.


Complete Translation of extract from SEQ_SAMPLE.FMS translated in All frames
File
Edit
View
Frame
Format
Attributes
Help
Transfer
✓ Complete Translation F5
All Open Reading Frames F6
Longest Open Reading Frames F7
Longest Fragments F8

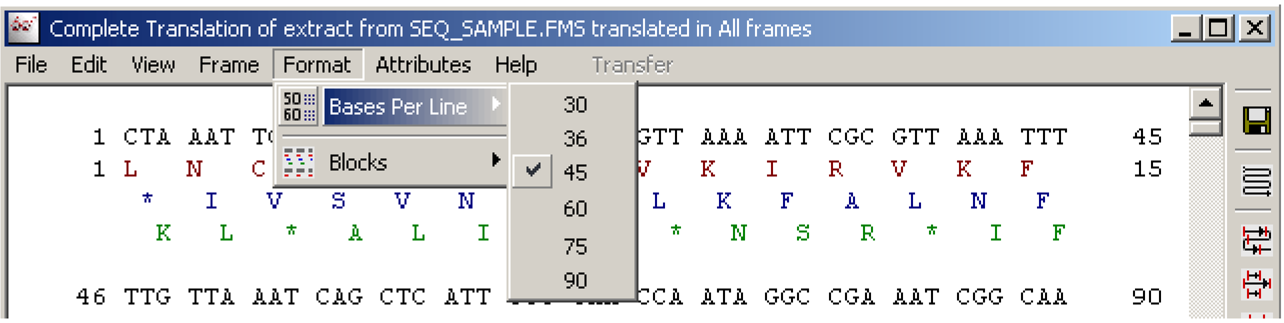
```
  1   T TTT GTT AAA ATT CGC GTT AAA TTT    45
  1     F   V   K   I   R   V   K   F      15
          L   L   K   F   A   L   N   F
      F   C   *   N   S   R   *   I   F
```

4.3.4.3 Frame menu - selects the reading frame (forward only) for the translation.



```
  1 CTA      TAT TTT GTT AAA ATT CGC GTT AAA TTT    45
  1 L        Y   F   V   K   I   R   V   K   F      15
     *        I   L   L   K   F   A   L   N   F
      K        F   C   *   N   S   R   *   I   F

 46 TTG      TTT TAA CCA ATA GGC CGA AAT CGG CAA    90
 16 L        F   *   P   I   G   R   N   R   Q      30
     C        F   N   Q   *   A   E   I   G   K
      V        L   T   N   R   P   K   S   A   K
```

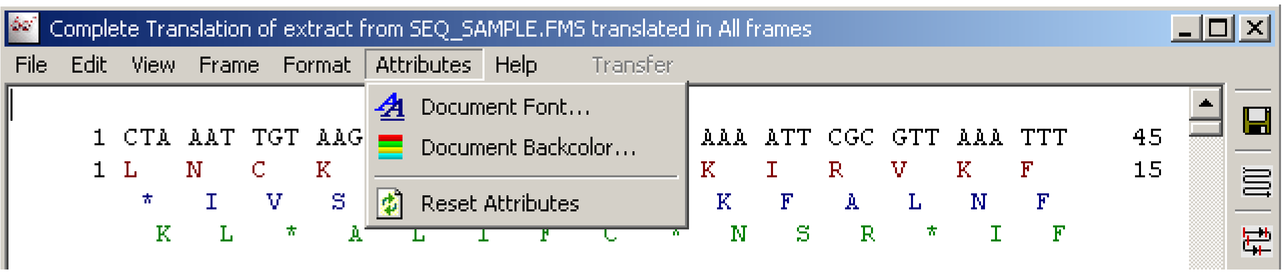
4.3.4.4 Format menu - allow you to select line length and whether or not to divide the sequence in blocks of 10 residues.



```
  1 CTA AAT T(      GTT AAA ATT CGC GTT AAA TTT    45
  1 L   N   C       V   K   I   R   V   K   F      15
     *   I   V   S   V   N    L   K   F   A   L   N   F
      K   L   *   A   L   I    *   N   S   R   *   I   F

 46 TTG TTA AAT CAG CTC ATT      CCA ATA GGC CGA AAT CGG CAA    90
```

4.3.4.5 Attributes menu - contains simple options for annotating the

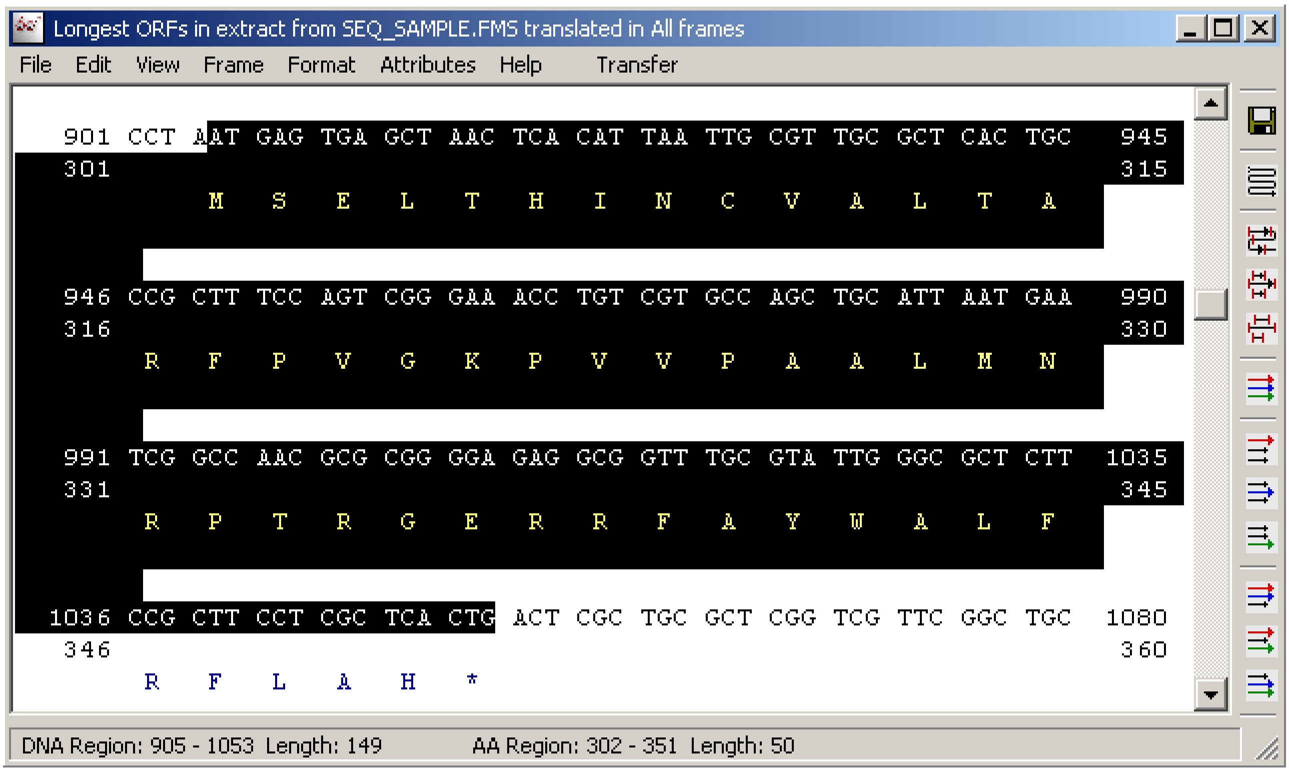
translated nucleotide sequence.



```
   1 CTA AAT TGT AAG                     AAA ATT CGC GTT AAA TTT    45
   1 L   N   C   K                       K   I   R   V   K   F      15
      *   I   V   S                        K   F   A   L   N   F
       K   L   *   A   L   I   F   C   *    N   S   R   *   I   F
```

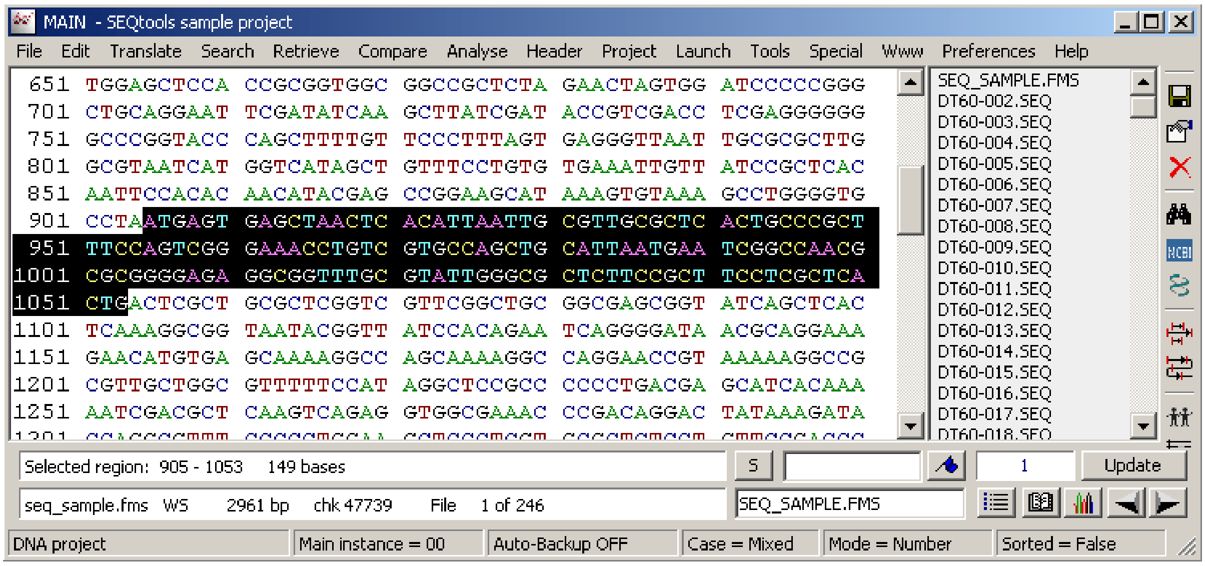
4.3.4.6 Transfer

4.3.4.6 Transfer - it is possible with the transfer options to highlight a nucleotide region (for example corresponding to an interesting portion of the translation and - by clicking *Transfer* - to transfer the highlights to the normal sequence edition (see below).


Longest ORFs in extract from SEQ_SAMPLE.FMS translated in All frames
File
Edit
View
Frame
Format
Attributes
Help
Transfer

```
  901 CCT AAT GAG TGA GCT AAC TCA CAT TAA TTG CGT TGC GCT CAC TGC   945
  301                                                              315
          M   S   E   L   T   H   I   N   C   V   A   L   T   A

  946 CCG CTT TCC AGT CGG GAA ACC TGT CGT GCC AGC TGC ATT AAT GAA   990
  316                                                              330
        R   F   P   V   G   K   P   V   V   P   A   A   L   M   N

  991 TCG GCC AAC GCG CGG GGA GAG GCG GTT TGC GTA TTG GGC GCT CTT  1035
  331                                                              345
        R   P   T   R   G   E   R   R   F   A   Y   W   A   L   F

 1036 CCG CTT CCT CGC TCA CTG ACT CGC TGC GCT CGG TCG TTC GGC TGC  1080
  346                                                              360
        R   F   L   A   H   *

 DNA Region: 905 - 1053  Length: 149        AA Region: 302 - 351  Length: 50
```

Highlights corresponding to the selected region in the *Forward Frame Translation* form.



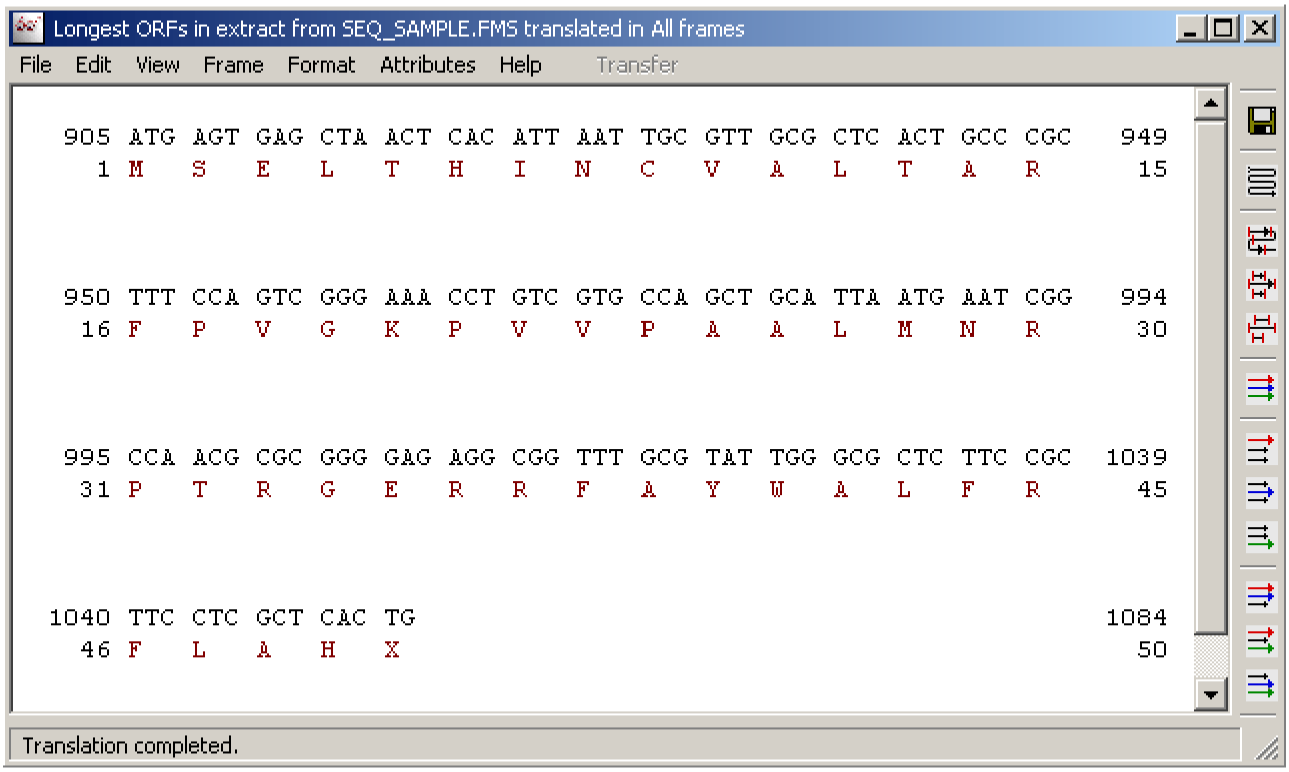
```
 651 TGGAGCTCCA CCGCGGTGGC GGCCGCTCTA GAACTAGTGG ATCCCCCGGG
 701 CTGCAGGAAT TCGATATCAA GCTTATCGAT ACCGTCGACC TCGAGGGGGG
 751 GCCCGGTACC CAGCTTTTGT TCCCTTTAGT GAGGGTTAAT TGCGCGCTTG
 801 GCGTAATCAT GGTCATAGCT GTTTCCTGTG TGAAATTGTT ATCCGCTCAC
 851 AATTCCACAC AACATACGAG CCGGAAGCAT AAAGTGTAAA GCCTGGGGTG
 901 CCTAATGAGT GAGCTAACTC ACATTAATTG CGTTGCGCTC ACTGCCCGCT
 951 TTCCAGTCGG GAAACCTGTC GTGCCAGCTG CATTAATGAA TCGGCCAACG
1001 CGCGGGGAGA GGCGGTTTGC GTATTGGGCG CTCTTCCGCT TCCTCGCTCA
1051 CTGACTCGCT GCGCTCGGTC GTTCGGCTGC GGCGAGCGGT ATCAGCTCAC
1101 TCAAAGGCGG TAATACGGTT ATCCACAGAA TCAGGGGATA ACGCAGGAAA
1151 GAACATGTGA GCAAAAGGCC AGCAAAAGGC CAGGAACCGT AAAAAGGCCG
1201 CGTTGCTGGC GTTTTTCCAT AGGCTCCGCC CCCCTGACGA GCATCACAAA
1251 AATCGACGCT CAAGTCAGAG GTGGCGAAAC CCGACAGGAC TATAAAGATA
```

Selected region: 905 - 1053   149 bases

seq_sample.fms   WS   2961 bp   chk 47739   File   1 of 246   SEQ_SAMPLE.FMS

DNA project | Main instance = 00 | Auto-Backup OFF | Case = Mixed | Mode = Number | Sorted = False

Re-translating the highlighted nucleotide region in the normal sequence editor displays the translation with the translated nucleotides displayed above the protein sequence.

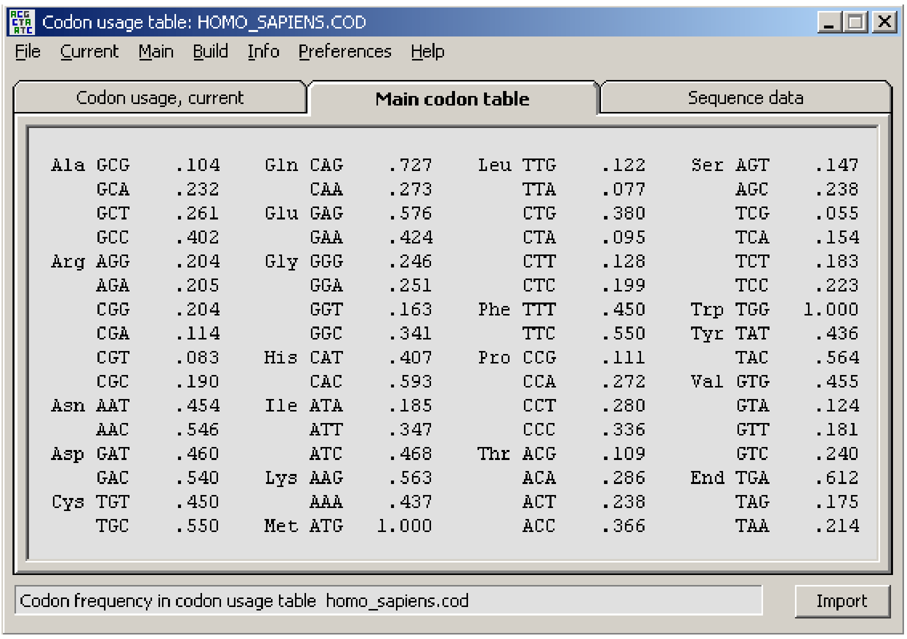

Longest ORFs in extract from SEQ_SAMPLE.FMS translated in All frames
File Edit View Frame Format Attributes Help Transfer

```
 905 ATG AGT GAG CTA ACT CAC ATT AAT TGC GTT GCG CTC ACT GCC CGC   949
   1 M   S   E   L   T   H   I   N   C   V   A   L   T   A   R      15

 950 TTT CCA GTC GGG AAA CCT GTC GTG CCA GCT GCA TTA ATG AAT CGG   994
  16 F   P   V   G   K   P   V   V   P   A   A   L   M   N   R      30

 995 CCA ACG CGC GGG GAG AGG CGG TTT GCG TAT TGG GCG CTC TTC CGC  1039
  31 P   T   R   G   E   R   R   F   A   Y   W   A   L   F   R      45

1040 TTC CTC GCT CAC TG                                            1084
  46 F   L   A   H   X                                              50
```

Translation completed.

4.3.5 back-translate protein sequence

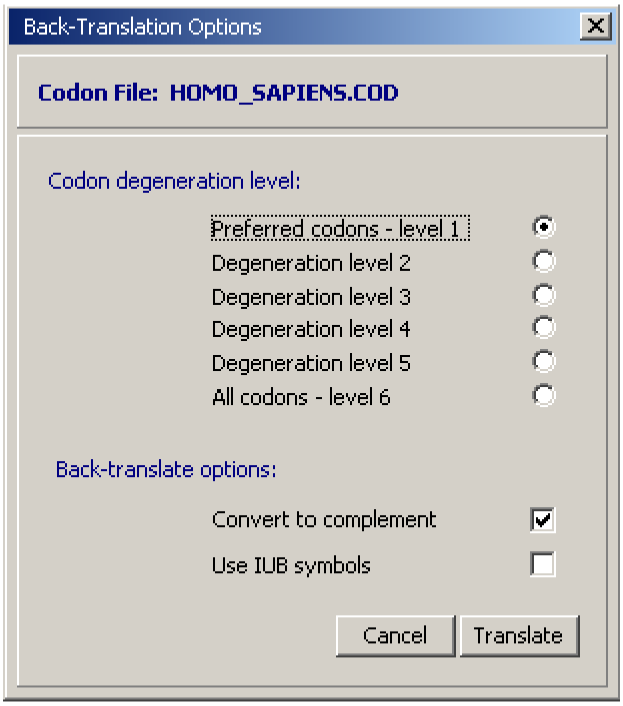
Back-translating protein sequences is useful when designing sequencing primers. When a *protein* sequence is displayed in the normal sequence editor selecting the *Back-Translate* option prompts you to select/load a codon usage data file to supply information about frequently used codons (codon usage) for the particular organism/protein.


Codon usage table: HOMO_SAPIENS.COD
File Current Main Build Info Preferences Help
Codon usage, current
Main codon table
Sequence data
Import

```
Ala GCG   .104   Gln CAG   .727   Leu TTG   .122   Ser AGT   .147
    GCA   .232       CAA   .273       TTA   .077       AGC   .238
    GCT   .261   Glu GAG   .576       CTG   .380       TCG   .055
    GCC   .402       GAA   .424       CTA   .095       TCA   .154
Arg AGG   .204   Gly GGG   .246       CTT   .128       TCT   .183
    AGA   .205       GGA   .251       CTC   .199       TCC   .223
    CGG   .204       GGT   .163   Phe TTT   .450   Trp TGG  1.000
    CGA   .114       GGC   .341       TTC   .550   Tyr TAT   .436
    CGT   .083   His CAT   .407   Pro CCG   .111       TAC   .564
    CGC   .190       CAC   .593       CCA   .272   Val GTG   .455
Asn AAT   .454   Ile ATA   .185       CCT   .280       GTA   .124
    AAC   .546       ATT   .347       CCC   .336       GTT   .181
Asp GAT   .460       ATC   .468   Thr ACG   .109       GTC   .240
    GAC   .540   Lys AAG   .563       ACA   .286   End TGA   .612
Cys TGT   .450       AAA   .437       ACT   .238       TAG   .175
    TGC   .550   Met ATG  1.000       ACC   .366       TAA   .214
```

Codon frequency in codon usage table homo_sapiens.cod

When a codon usage data file is successfully loaded into SEQtools the form below is displayed to enable you to select a degeneration level. Choosing level 1 will result in a primer *without* degenerate positions while level 6 will cover *all possible* degenerate base combinations. The cost in the latter case of course is few primers in the mixture with the correct base sequence exactly matching the nucleotide sequence.



Codon File: HOMO_SAPIENS.COD

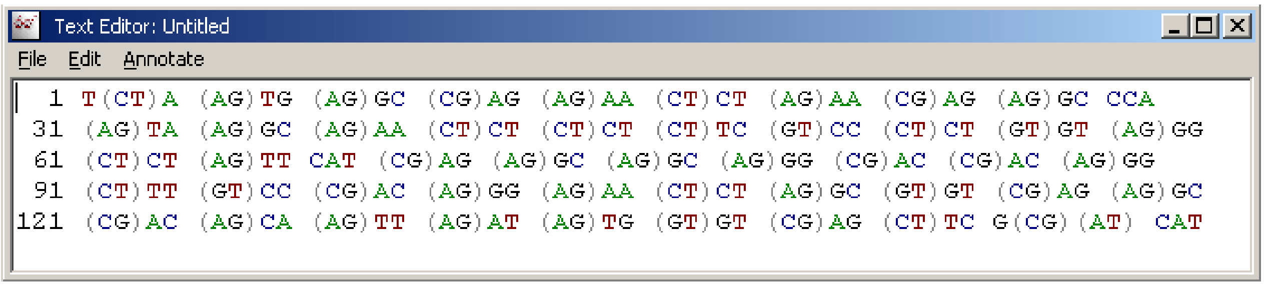
Codon degeneration level:

- Preferred codons - level 1 ●
- Degeneration level 2 ○
- Degeneration level 3 ○
- Degeneration level 4 ○
- Degeneration level 5 ○
- All codons - level 6 ○

Back-translate options:

- Convert to complement ☑
- Use IUB symbols □

The primer sequence after back-translation is displayed in a simple text form. You must then copy/paste the primer sequence into a separate instance of SEQtools opened for handling primers.



```
  1 T(CT)A  (AG)TG  (AG)GC  (CG)AG  (AG)AA  (CT)CT  (AG)AA  (CG)AG  (AG)GC  CCA
 31 (AG)TA  (AG)GC  (AG)AA  (CT)CT  (CT)CT  (CT)TC  (GT)CC  (CT)CT  (GT)GT  (AG)GG
 61 (CT)CT  (AG)TT  CAT  (CG)AG  (AG)GC  (AG)GC  (AG)GG  (CG)AC  (CG)AC  (AG)GG
 91 (CT)TT  (GT)CC  (CG)AC  (AG)GG  (AG)AA  (CT)CT  (AG)GC  (GT)GT  (CG)AG  (AG)GC
121 (CG)AC  (AG)CA  (AG)TT  (AG)AT  (AG)TG  (GT)GT  (CG)AG  (CT)TC  G(CG)(AT)  CAT
```

4.3.6 create protein files

This utility is designed to assist you in the analysis of short EST sequences (expressed sequence tags) in cases where functional identification by data base searching has failed and the correct reading frame thus is unknown.

The utility translates *all* nucleotide sequences of the current project in the selected reading frame(s) and saves each protein sequence in a separate file. The extracted protein sequences can then be searched for example against the Prosite data base of protein motifs, or other data bases including protein signatures.

Translation options:

Complete sequences - the complete translation including X 's and stops.

Largest fragments - largest contiguous amino acid region without stops N-terminal regions; regions starting with a M and ending at the first downstream stop C-terminal regions; regions from the start of the sequence to the first stop

Frame options - (**1**) All reading frames. (**2**) The 3 forward reading frames (A, B, C). (**3**) The 3 reverse reading frames (D, E, F).

Filter option - allows you to disregard protein sequences shorter then the selected minimum length.

Protein file names - The protein file names are constructed by adding _N to the file names of the DNA sequences, where N denotes the reading frame (1-6, or # for all reading frames in the same file). The two characters can be added in one of four ways: (**1**) By replacing the extension of the DNA sequence file name with _N. (**2**) By adding _N to the leftmost six characters of the file name. (**3**) By adding _N to the rightmost six characters of the file name. (**4**) By adding _N to the middle six characters of the file name.

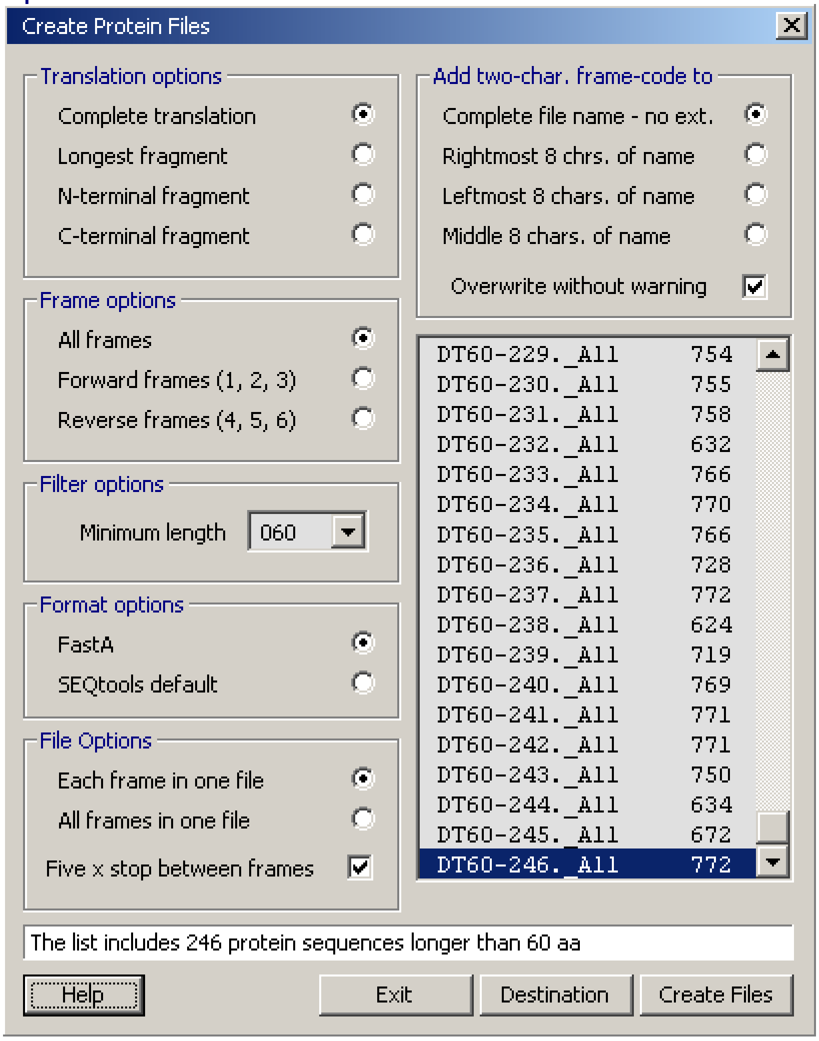
In the latter three cases, the protein file names will lack an extension.

When the protein files are build, the selected file names are validated to avoid duplicate file names. If the selected naming method yields duplicate names, the building is arrested and the used advised to select another method of generating protein file names.

In cases where *none* of the available four methods yields unique protein file names, the original DNA sequence files must be renamed.

File format - The protein files can be saved in either Fasta or GCG format. Each protein file includes a header giving the sequence name, the reading frame and the length of the protein sequence. The protein sequences are broken into lines of 50 characters without line numbering.

Save options - The protein files can either be saved in separate files or in one file per DNA sequence. If the latter option is selected, a 5 x stop separator is inserted between each reading frame if the check box for this option is checked.



Create Protein Files

Translation options
- Complete translation ●
- Longest fragment ○
- N-terminal fragment ○
- C-terminal fragment ○

Frame options
- All frames ●
- Forward frames (1, 2, 3) ○
- Reverse frames (4, 5, 6) ○

Filter options
- Minimum length 060

Format options
- FastA ●
- SEQtools default ○

File Options
- Each frame in one file ●
- All frames in one file ○
- Five x stop between frames ☑

Add two-char. frame-code to
- Complete file name - no ext. ●
- Rightmost 8 chrs. of name ○
- Leftmost 8 chars. of name ○
- Middle 8 chars. of name ○
- Overwrite without warning ☑

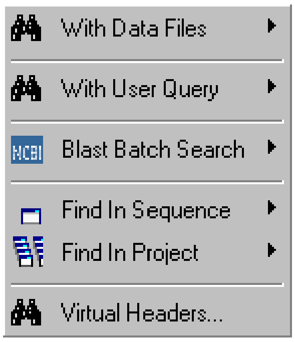
```
DT60-229._All     754
DT60-230._All     755
DT60-231._All     758
DT60-232._All     632
DT60-233._All     766
DT60-234._All     770
DT60-235._All     766
DT60-236._All     728
DT60-237._All     772
DT60-238._All     624
DT60-239._All     719
DT60-240._All     769
DT60-241._All     771
DT60-242._All     771
DT60-243._All     750
DT60-244._All     634
DT60-245._All     672
DT60-246._All     772
```

The list includes 246 protein sequences longer than 60 aa

4.4 SEARCH MENU

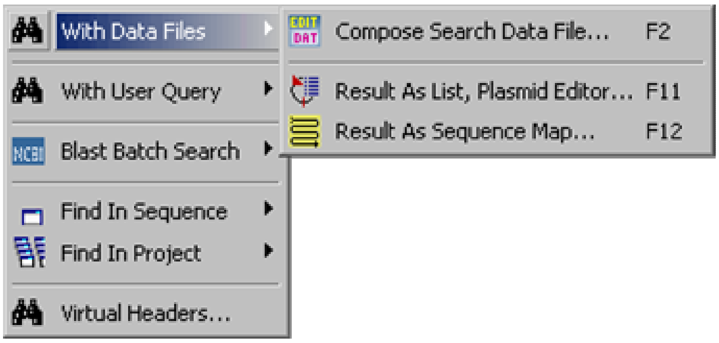
4.4.1 about searching

The search options in SEQtools are quite extensive. The brief descriptions below are primarily intended to give you an overview of the various options for searching sequences and their annotation. It is recommended that you look through the different sections below to learn which options are available. And then experiment to find the most convenient way to use the different functions.

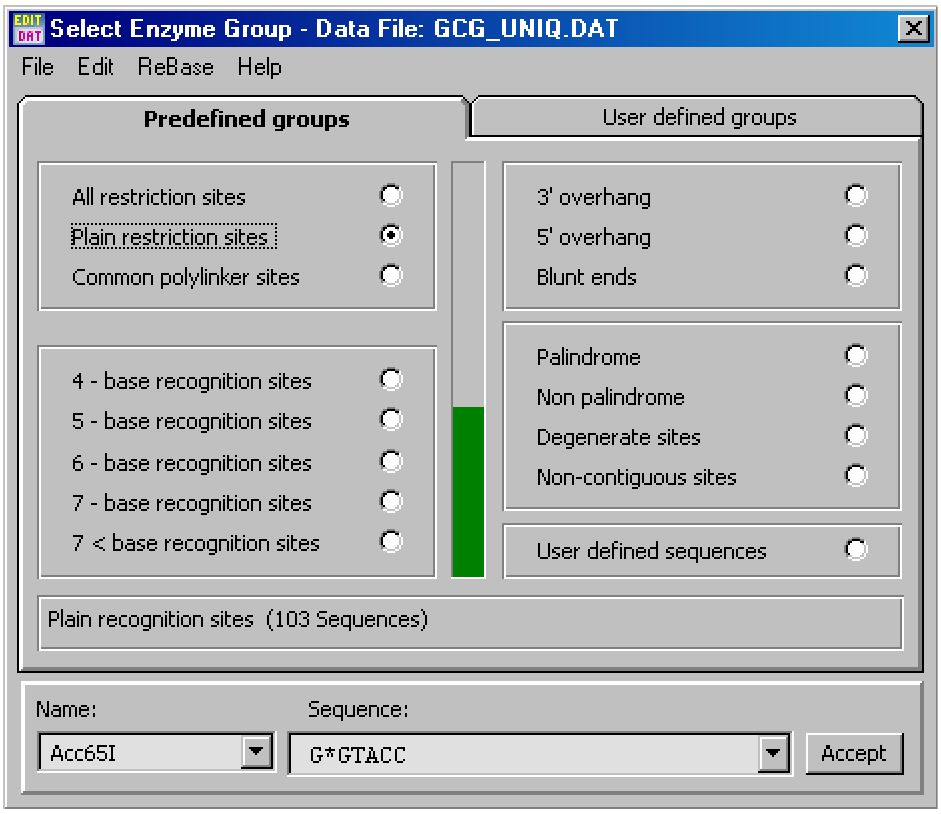


4.4.2 search with data files

This function enables you to sear a single sequence with a collection of restriction enzyme sequences. Use *Compose Search Data File* to build the group of enzymes you want to include in the search data file. Note that you can convert an entire project of primer sequences into a search datafile and use this datafile in exactly the same way as proper restriction enzyme datafiles.



4.4.2.1 Compose restriction enzyme data file - This form contains various options for building a custom designed restriction enzyme datafile. Use the setting *User defined sequences* if you wish to search with a search datafile containing primer sequences (or other user designed search strings).


Select Enzyme Group - Data File: GCG_UNIQ.DAT

File Edit ReBase Help

Predefined groups | User defined groups

All restriction sites ○
Plain restriction sites ●
Common polylinker sites ○

4 - base recognition sites ○
5 - base recognition sites ○
6 - base recognition sites ○
7 - base recognition sites ○
7 < base recognition sites ○

3' overhang ○
5' overhang ○
Blunt ends ○

Palindrome ○
Non palindrome ○
Degenerate sites ○
Non-contiguous sites ○

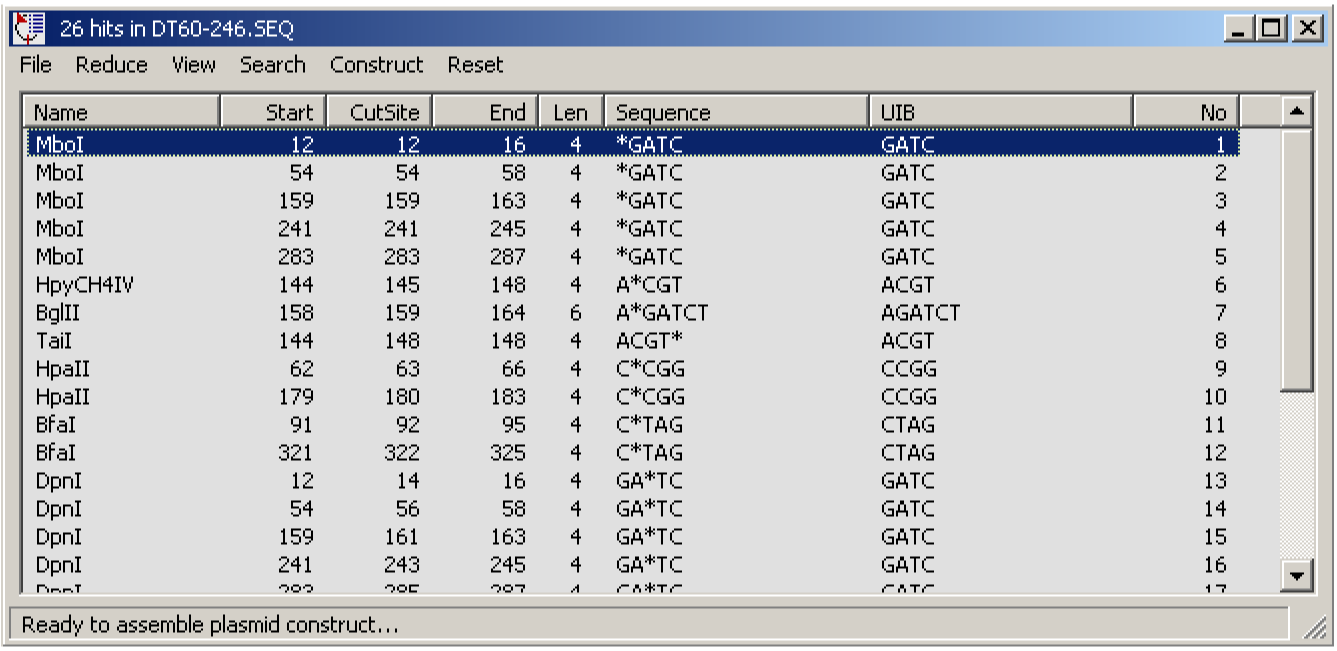
User defined sequences ○

Plain recognition sites (103 Sequences)

Name: Acc65I
Sequence: G*GTACC
Accept

4.4.2.2 Restriction enzyme search, Plasmid editor - The results form shown below displays the result of the datafile search. the form includes various options for filtering the list of matches (remove multi-cutting enzymes for example). Clicking a match line highlights the match in the original sequence.

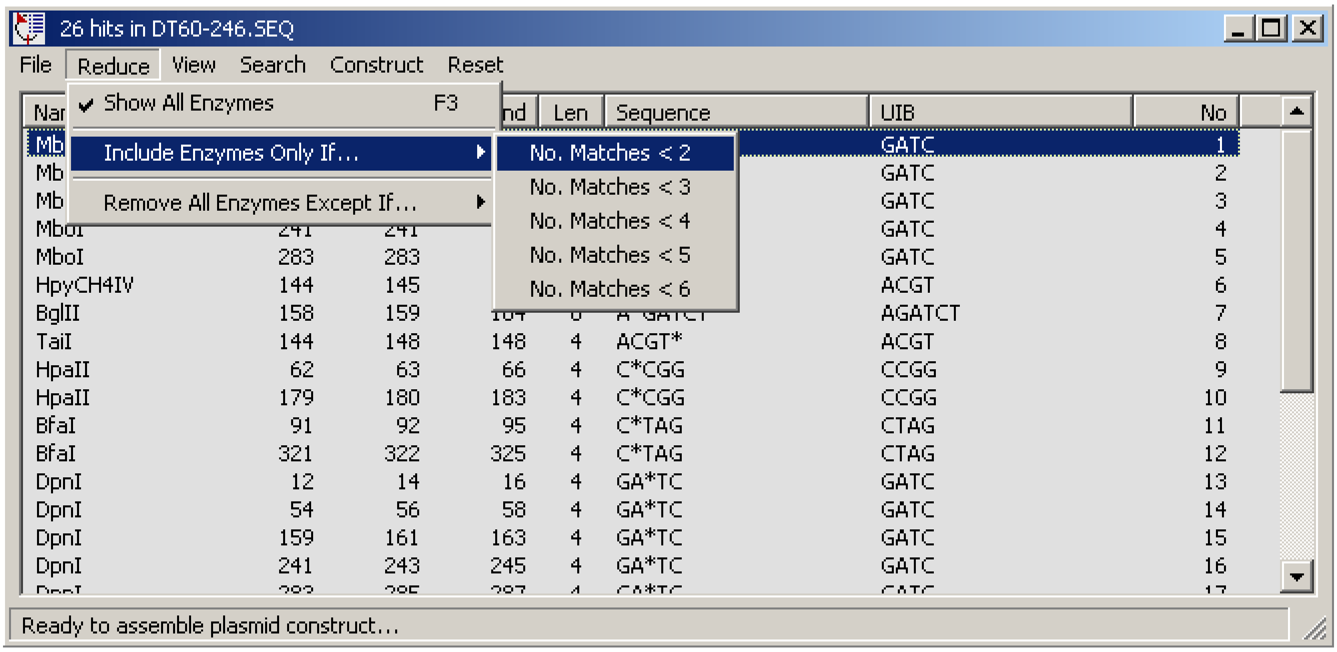
This form also contains a simple function for assembling simple plasmid constructs. This function is described in the following screenshots.



Name	Start	CutSite	End	Len	Sequence	UIB	No
MboI	12	12	16	4	*GATC	GATC	1
MboI	54	54	58	4	*GATC	GATC	2
MboI	159	159	163	4	*GATC	GATC	3
MboI	241	241	245	4	*GATC	GATC	4
MboI	283	283	287	4	*GATC	GATC	5
HpyCH4IV	144	145	148	4	A*CGT	ACGT	6
BglII	158	159	164	6	A*GATCT	AGATCT	7
TaiI	144	148	148	4	ACGT*	ACGT	8
HpaII	62	63	66	4	C*CGG	CCGG	9
HpaII	179	180	183	4	C*CGG	CCGG	10
BfaI	91	92	95	4	C*TAG	CTAG	11
BfaI	321	322	325	4	C*TAG	CTAG	12
DpnI	12	14	16	4	GA*TC	GATC	13
DpnI	54	56	58	4	GA*TC	GATC	14
DpnI	159	161	163	4	GA*TC	GATC	15
DpnI	241	243	245	4	GA*TC	GATC	16
DpnI	283	285	287	4	GA*TC	GATC	17

Ready to assemble plasmid construct...

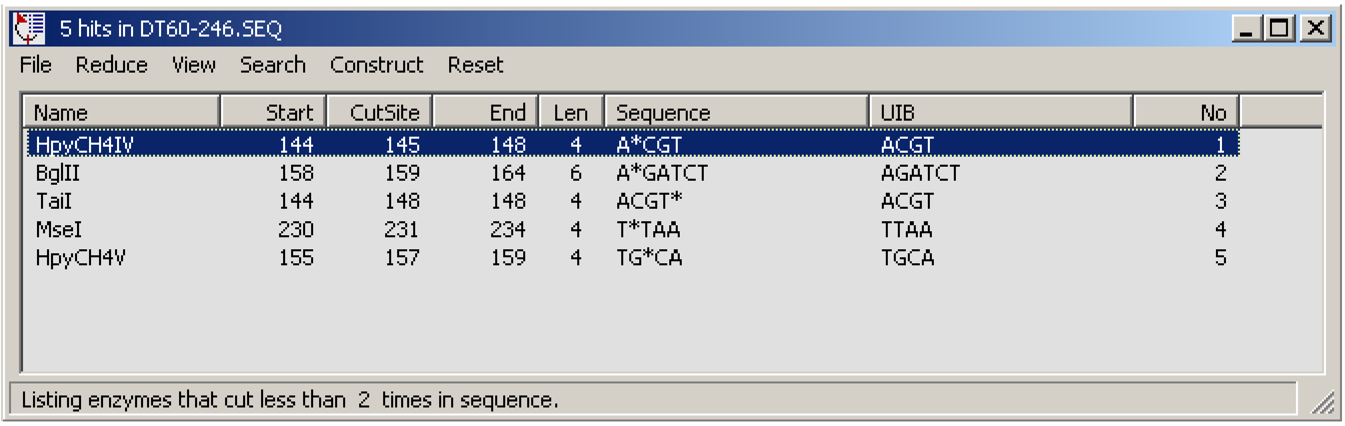
Options under the *Reduce* menu.



Nar			nd	Len	Sequence	UIB	No
Mb						GATC	1
Mb						GATC	2
Mb						GATC	3
MboI	241	241				GATC	4
MboI	283	283				GATC	5
HpyCH4IV	144	145				ACGT	6
BglII	158	159	164	6	A*GATCT	AGATCT	7
TaiI	144	148	148	4	ACGT*	ACGT	8
HpaII	62	63	66	4	C*CGG	CCGG	9
HpaII	179	180	183	4	C*CGG	CCGG	10
BfaI	91	92	95	4	C*TAG	CTAG	11
BfaI	321	322	325	4	C*TAG	CTAG	12
DpnI	12	14	16	4	GA*TC	GATC	13
DpnI	54	56	58	4	GA*TC	GATC	14
DpnI	159	161	163	4	GA*TC	GATC	15
DpnI	241	243	245	4	GA*TC	GATC	16
DpnI	283	285	287	4	GA*TC	GATC	17

Ready to assemble plasmid construct...

The *match list* after removal of restriction enzymes which cut more than once.



Name	Start	CutSite	End	Len	Sequence	UIB	No
HpyCH4IV	144	145	148	4	A*CGT	ACGT	1
BglII	158	159	164	6	A*GATCT	AGATCT	2
TaiI	144	148	148	4	ACGT*	ACGT	3
MseI	230	231	234	4	T*TAA	TTAA	4
HpyCH4V	155	157	159	4	TG*CA	TGCA	5

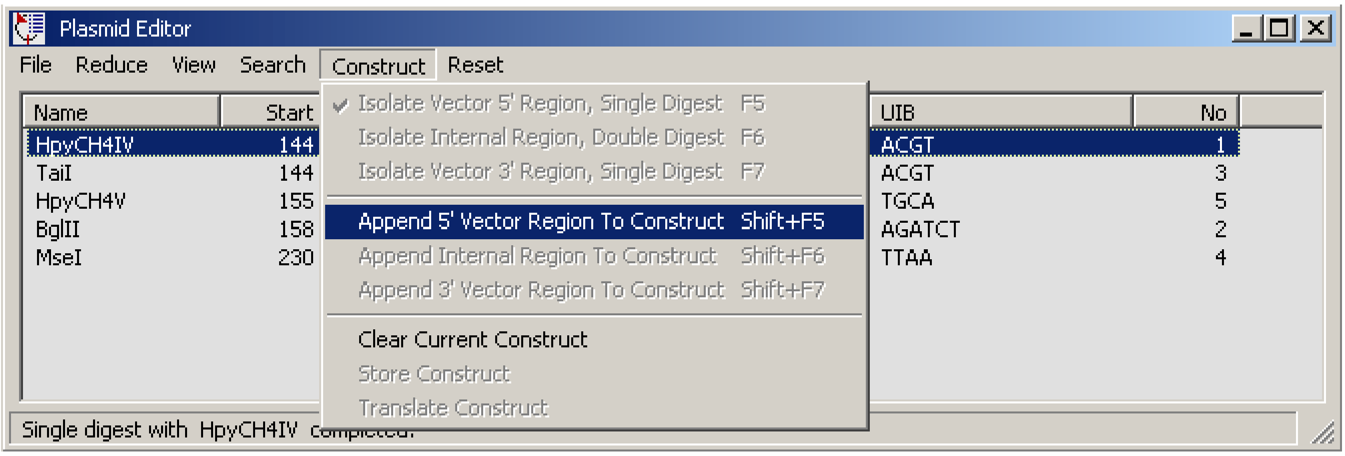
Listing enzymes that cut less than 2 times in sequence.

Proceeding through the steps of the *Construct* menu allows you to digest your sequence with the specified restriction enzymes and to isolate the segment in a separate text form. This facility can be used to build simple plasmid constructs: The operation comprises 3 steps:

(1) Navigate to the vector sequence in the project and isolate of 5' vector arm by a single cut.
(2) Navigate to the sequence to provide the insert and isolate the insert sequence by a double cut.
(3) Navigate back to the vector and and isolate the 3' vector arm by a single cut.

Navigating to a different sequence in the project automatically updates the list of matches for the selected search data file. Before a sequence segment is inserted into the text form, the overhangs are checked for compatibility and the result of the check displayed in an info message.

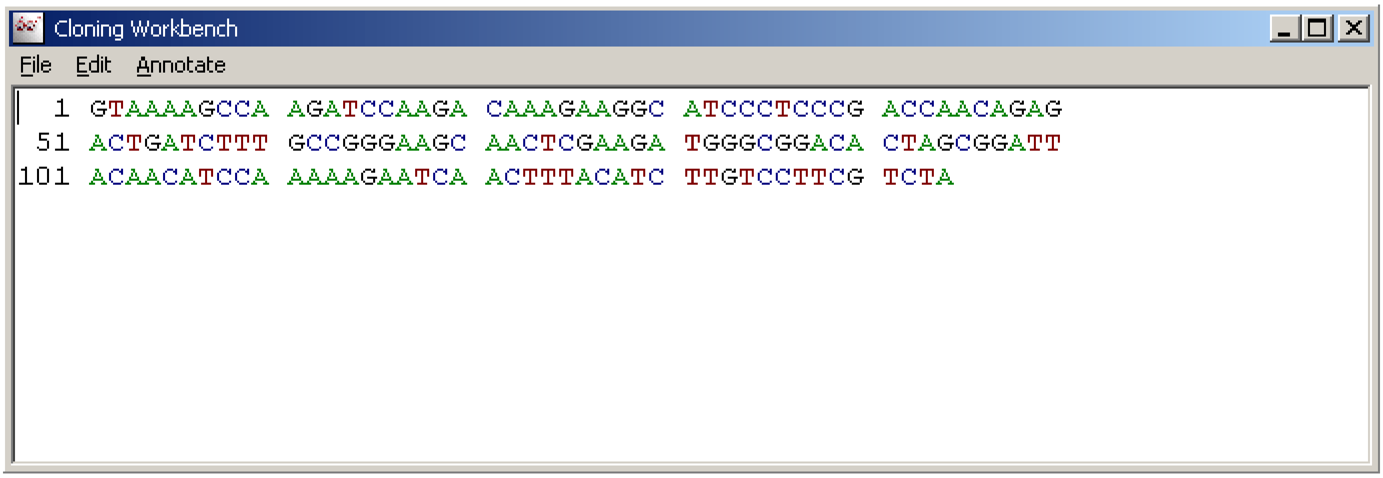
In the example shown below the 5' end of the construct is created by digesting the vector with enzyme *HpyCH4IV*. Clicking *Append 5' Vector Region To Construct* transfers the isolated segment to a text editor.


Plasmid Editor
File Reduce View Search Construct Reset
✓ Isolate Vector 5' Region, Single Digest F5
Isolate Internal Region, Double Digest F6
Isolate Vector 3' Region, Single Digest F7
Append 5' Vector Region To Construct Shift+F5
Append Internal Region To Construct Shift+F6
Append 3' Vector Region To Construct Shift+F7
Clear Current Construct
Store Construct
Translate Construct

Name	Start	UIB	No
HpyCH4IV	144	ACGT	1
TaiI	144	ACGT	3
HpyCH4V	155	TGCA	5
BglII	158	AGATCT	2
MseI	230	TTAA	4

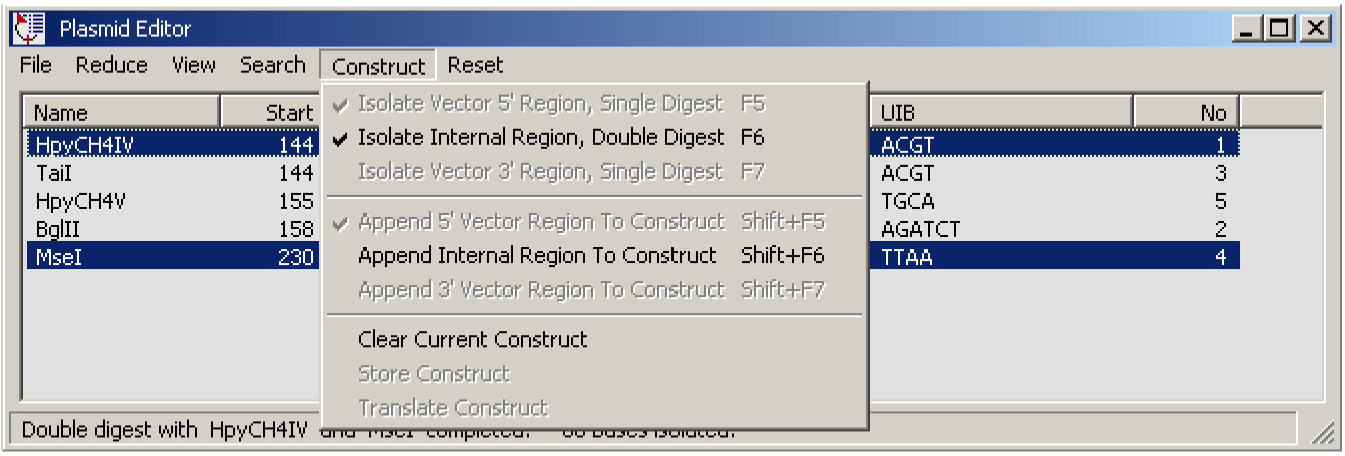
Single digest with HpyCH4IV completed.

The 5' segment of the vector sequence after digesting with enzyme *HpyCH4IV* copied to the text form.



```
  1 GTAAAAGCCA AGATCCAAGA CAAAGAAGGC ATCCCTCCCG ACCAACAGAG
 51 ACTGATCTTT GCCGGGAAGC AACTCGAAGA TGGGCGGACA CTAGCGGATT
101 ACAACATCCA AAAAGAATCA ACTTTACATC TTGTCCTTCG TCTA
```

Isolating the insert by a double digest - hold down the <CTRL> key while clicking the second enzyme in the match list. Note that this option is *only* available *after* the 5' segment has been isolated.

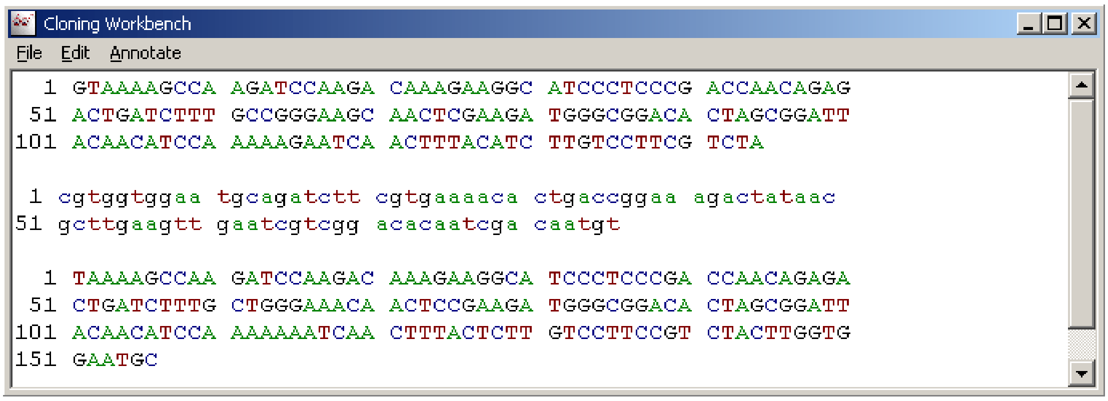

Plasmid Editor
File
Reduce
View
Search
Construct
Reset
Isolate Vector 5' Region, Single Digest F5
Isolate Internal Region, Double Digest F6
Isolate Vector 3' Region, Single Digest F7
Append 5' Vector Region To Construct Shift+F5
Append Internal Region To Construct Shift+F6
Append 3' Vector Region To Construct Shift+F7
Clear Current Construct
Store Construct
Translate Construct

Name	Start
HpyCH4IV	144
TaiI	144
HpyCH4V	155
BglII	158
MseI	230

UIB	No
ACGT	1
ACGT	3
TGCA	5
AGATCT	2
TTAA	4

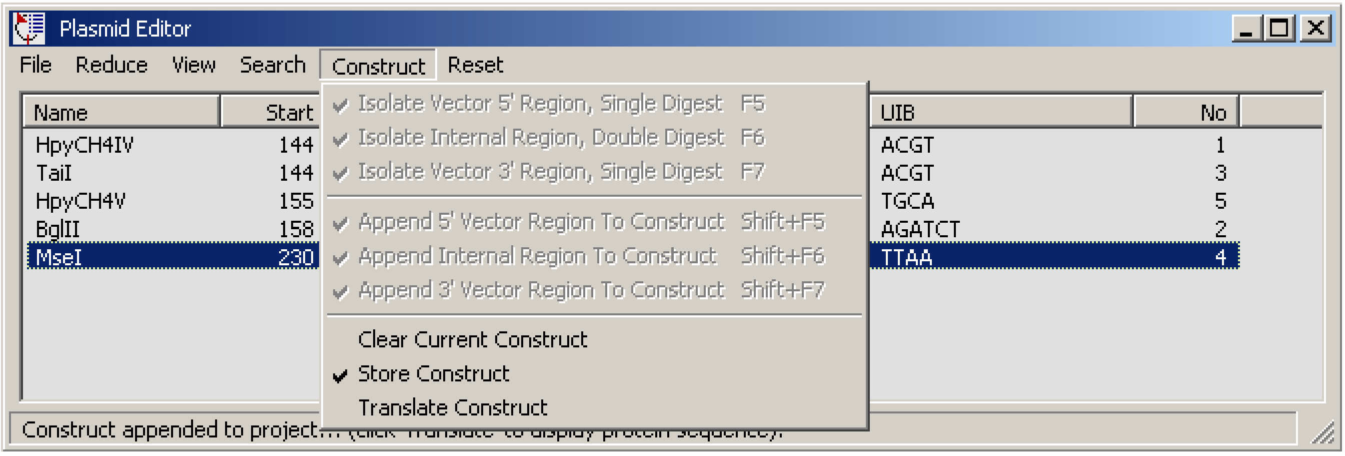
Double digest with HpyCH4IV

Click *Append Internal Region To Construct* to copy the isolated insert sequence to the text form. In the screenshot below the construct has been completed by insertion of the 3' vector arm into the text form.



```
  1 GTAAAAGCCA AGATCCAAGA CAAAGAAGGC ATCCCTCCCG ACCAACAGAG
 51 ACTGATCTTT GCCGGGAAGC AACTCGAAGA TGGGCGGACA CTAGCGGATT
101 ACAACATCCA AAAAGAATCA ACTTTACATC TTGTCCTTCG TCTA

 1 cgtggtggaa tgcagatctt cgtgaaaaca ctgaccggaa agactataac
51 gcttgaagtt gaatcgtcgg acacaatcga caatgt

  1 TAAAAGCCAA GATCCAAGAC AAAGAAGGCA TCCCTCCCGA CCAACAGAGA
 51 CTGATCTTTG CTGGGAAACA ACTCCGAAGA TGGGCGGACA CTAGCGGATT
101 ACAACATCCA AAAAAATCAA CTTTACTCTT GTCCTTCCGT CTACTTGGTG
151 GAATGC
```

The completed plasmid sequence can either be copy/pasted to a second nucleotide instance of SEQtools or appended to the current project as a separate, new sequence file.

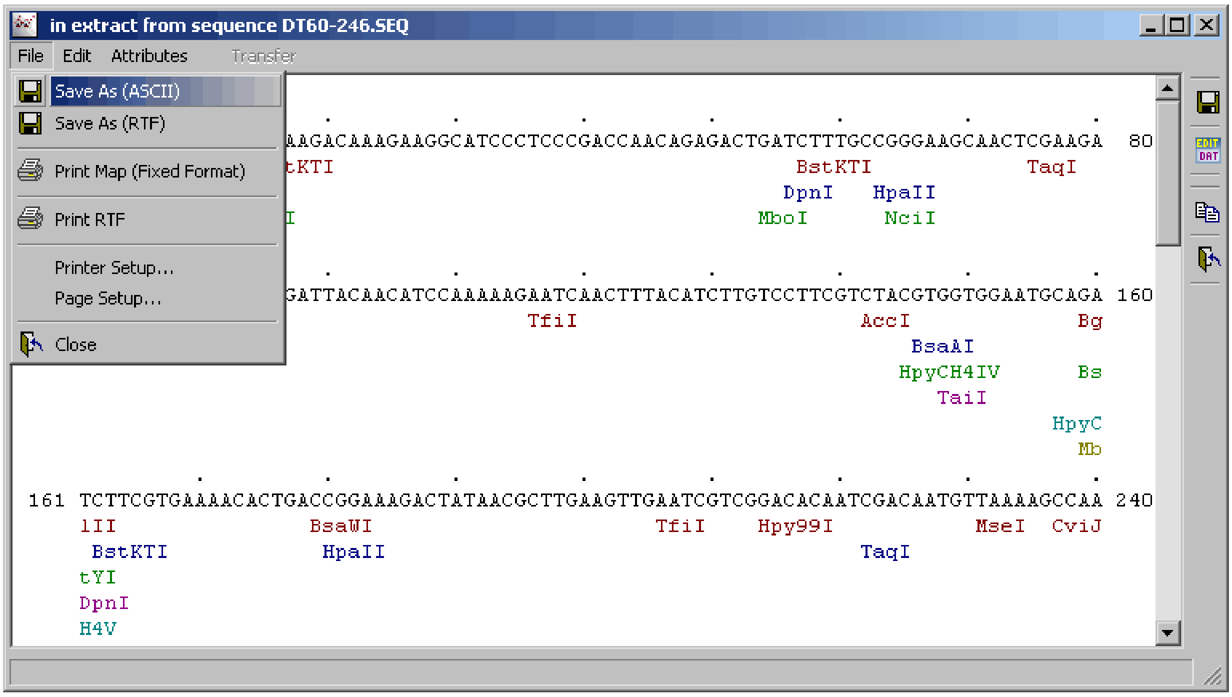

Plasmid Editor
File
Reduce
View
Search
Construct
Reset
Isolate Vector 5' Region, Single Digest F5
Isolate Internal Region, Double Digest F6
Isolate Vector 3' Region, Single Digest F7
Append 5' Vector Region To Construct Shift+F5
Append Internal Region To Construct Shift+F6
Append 3' Vector Region To Construct Shift+F7
Clear Current Construct
✓ Store Construct
Translate Construct

Name	Start
HpyCH4IV	144
TaiI	144
HpyCH4V	155
BglII	158
MseI	230

UIB	No
ACGT	1
ACGT	3
TGCA	5
AGATCT	2
TTAA	4

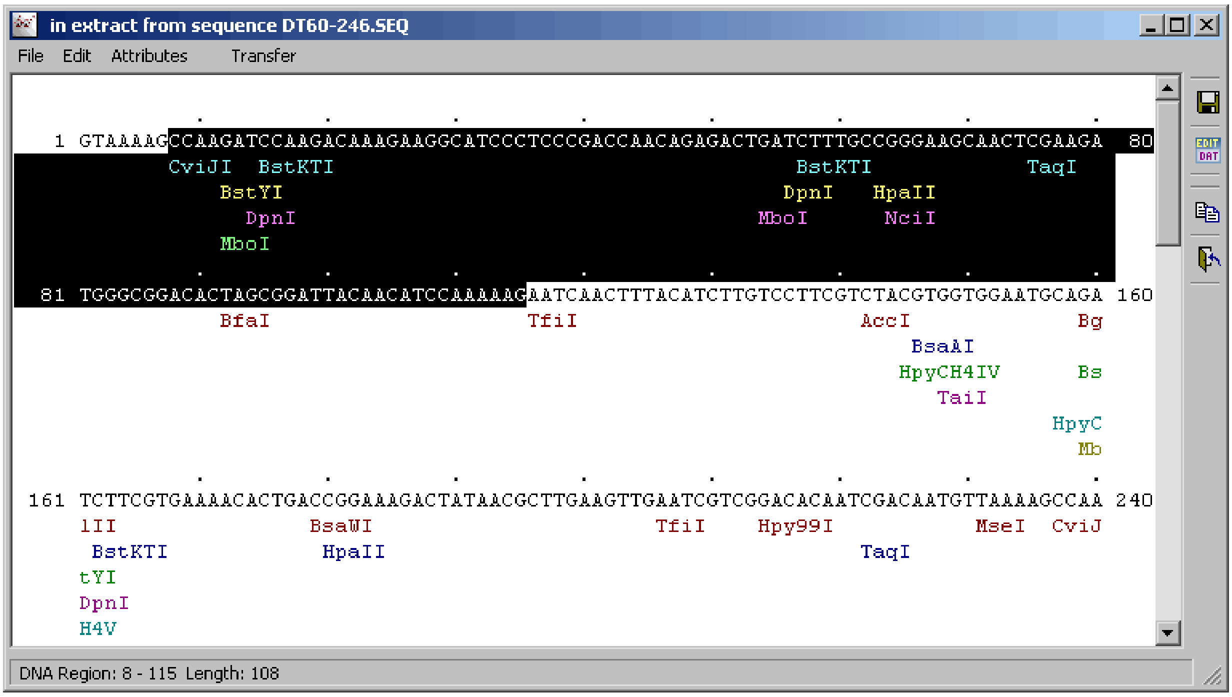
Construct appended to project... (click Translate to display protein sequence).

Restriction map - A restriction map is a second alternative for displaying the result of a datafile search. The first character of the enzyme name marks the cut site.



```
                       .         .         .         .         .         .         .
                    AAGACAAAGAAGGCATCCCTCCCGACCAACAGAGACTGATCTTTGCCGGGAAGCAACTCGAAGA 80
                    tKTI                                    BstKTI            TaqI
                                                           DpnI   HpaII
                    I                                    MboI      NciI

                       .         .         .         .         .         .         .
                    GATTACAACATCCAAAAAGAATCAACTTTACATCTTGTCCTTCGTCTACGTGGTGGAATGCAGA 160
                                       TfiI                      AccI             Bg
                                                                     BsaAI
                                                                    HpyCH4IV      Bs
                                                                       TaiI
                                                                                HpyC
                                                                                  Mb

             .         .         .         .         .         .         .         .
161 TCTTCGTGAAAACACTGACCGGAAAGACTATAACGCTTGAAGTTGAATCGTCGGACACAATCGACAATGTTAAAAGCCAA 240
    lII               BsaWI                      TfiI    Hpy99I           MseI  CviJ
     BstKTI            HpaII                                     TaqI
    tYI
    DpnI
    H4V
```

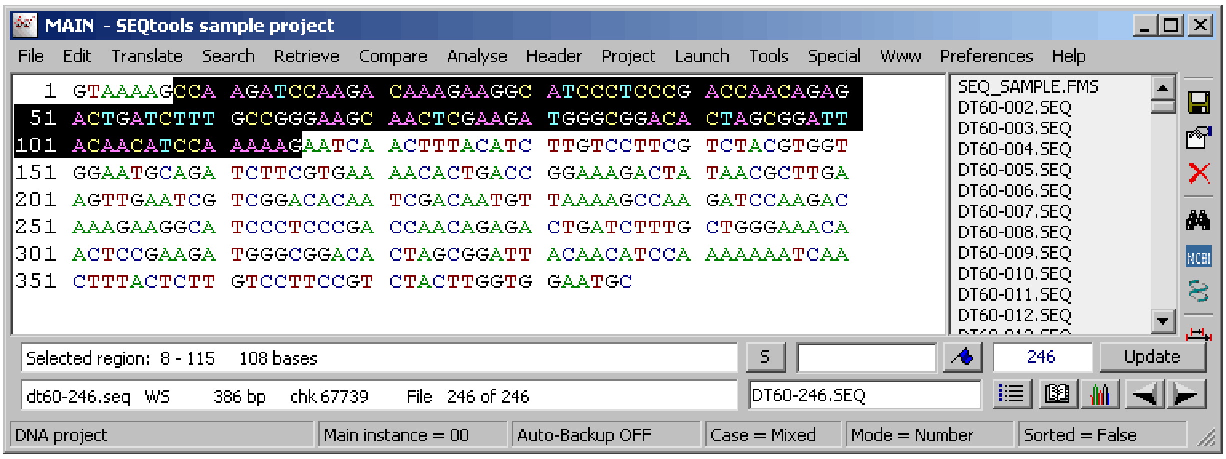
4.4.2.4 Transfer selected region - Highlighting a region of the sequence in the restriction map and clicking *Transfer* closes the restriction map form and transfers the highlights to the sequence displayed in the normal SEQtools editor.



```
             .         .         .         .         .         .         .         .
  1 GTAAAAGCCAAGATCCAAGACAAAGAAGGCATCCCTCCCGACCAACAGAGACTGATCTTTGCCGGGAAGCAACTCGAAGA  80
           CviJI  BstKTI                                    BstKTI            TaqI
               BstYI                                       DpnI   HpaII
                 DpnI                                    MboI      NciI
               MboI
             .         .         .         .         .         .         .         .
 81 TGGGCGGACACTAGCGGATTACAACATCCAAAAAGAATCAACTTTACATCTTGTCCTTCGTCTACGTGGTGGAATGCAGA 160
               BfaI                    TfiI                      AccI             Bg
                                                                     BsaAI
                                                                    HpyCH4IV      Bs
                                                                       TaiI
                                                                                HpyC
                                                                                  Mb
             .         .         .         .         .         .         .         .
161 TCTTCGTGAAAACACTGACCGGAAAGACTATAACGCTTGAAGTTGAATCGTCGGACACAATCGACAATGTTAAAAGCCAA 240
    lII               BsaWI                      TfiI    Hpy99I           MseI  CviJ
     BstKTI            HpaII                                     TaqI
    tYI
    DpnI
    H4V
```

DNA Region: 8 - 115 Length: 108

Highlighted sequence region transferred from the restriction map form.



```
  1 GTAAAAGCCA AGATCCAAGA CAAAGAAGGC ATCCCTCCCG ACCAACAGAG
 51 ACTGATCTTT GCCGGGAAGC AACTCGAAGA TGGGCGGACA CTAGCGGATT
101 ACAACATCCA AAAAGAATCA ACTTTACATC TTGTCCTTCG TCTACGTGGT
151 GGAATGCAGA TCTTCGTGAA AACACTGACC GGAAAGACTA TAACGCTTGA
201 AGTTGAATCG TCGGACACAA TCGACAATGT TAAAAGCCAA GATCCAAGAC
251 AAAGAAGGCA TCCCTCCCGA CCAACAGAGA CTGATCTTTG CTGGGAAACA
301 ACTCCGAAGA TGGGCGGACA CTAGCGGATT ACAACATCCA AAAAAATCAA
351 CTTTACTCTT GTCCTTCCGT CTACTTGGTG GAATGC
```

SEQ_SAMPLE.FMS
DT60-002.SEQ
DT60-003.SEQ
DT60-004.SEQ
DT60-005.SEQ
DT60-006.SEQ
DT60-007.SEQ
DT60-008.SEQ
DT60-009.SEQ
DT60-010.SEQ
DT60-011.SEQ
DT60-012.SEQ

Selected region: 8 - 115 108 bases

246

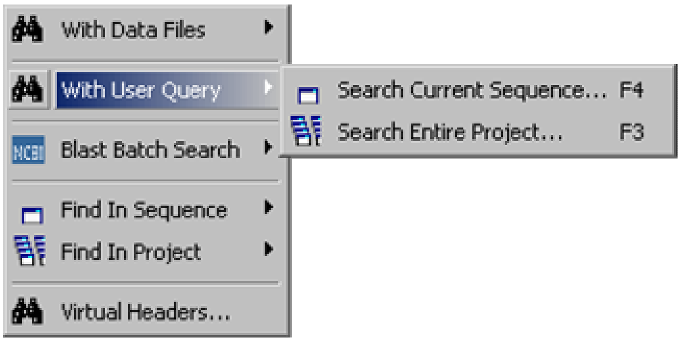
dt60-246.seq WS 386 bp chk 67739 File 246 of 246

DT60-246.SEQ

DNA project | Main instance = 00 | Auto-Backup OFF | Case = Mixed | Mode = Number | Sorted = False

4.4.3 search with user query

Simple search with a sequence query (nucleotide or protein) can be performed either on the currently displayed sequence or on *all* sequences contained in the project. In the latter case the *Search Sequence* form automatically appears when a matching sequence is clicked in *Project Search* mode.



In addition to a plain query string more complex queries can be constructed using the syntax below:

***Syntax*:**

? Any character.
[ ] Any of the characters within the square brackets.
[!] Any characters *other* than those within the square brackets.

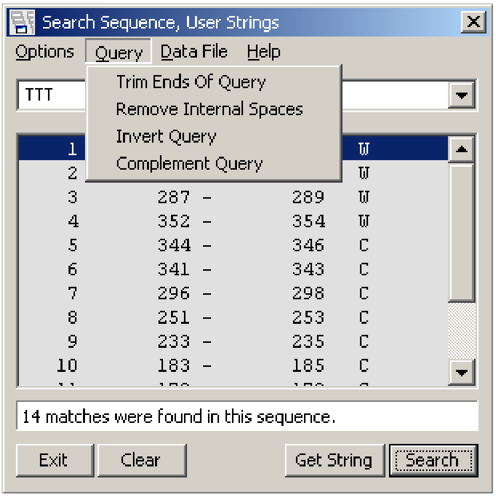
5'/ABCn1-n2/ Between n1 and n2 characters from 5'-end or N-terminal *other* than A, B and C.
/ABCn1-n2/ 3' Between n1 and n2 characters from 3'-end or C-terminal *other* than A, B and C.
/ABCcn1-n2/ Between n1 and n2 characters *other* than A, B and C.

***Examples*:**

Pattern:	**Finds**:	**Does not find**:
ASTS?V	ASTSxV	ASTSV
AST[GHWP]SV	ASTGSV and ASTHSV	ASTNSV and ASTRSV
AST[!GHWP]SV		ASTGSV and ASTHSV
/1-20/AST/4-8/SV	ASTKSV and ASTYSV	
AST/4-8/SV/2-20/	5' xxxASTxxxxSV	5' xxxASTxxxSV
/A1-20/AST/4-8/SV	ASTxxxxSVxxx 3'	ASTxxxxSVx 3'
AST/B4-8/SV/1-20/	5' xxxxASTxxxxxSV ASTxxxxSVxxx 3'	5' xAxxASTxxxxxSV ASTxBxxSVxxx 3'

Where x is any character; 5' and 3' denote the 5'/N-terminal and 3'/C-terminal respectively.

Search current sequence - Result of sequence search. Each line include the start and end of the match as well as the orientation (W=Watson; C=Crick) of the match.

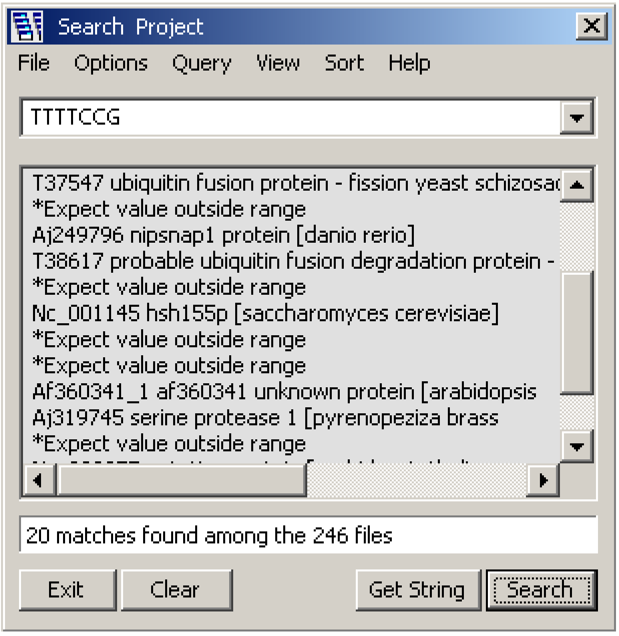


TTT

```
 1                               W
 2                               W
 3      287  -      289   W
 4      352  -      354   W
 5      344  -      346   C
 6      341  -      343   C
 7      296  -      298   C
 8      251  -      253   C
 9      233  -      235   C
10      183  -      185   C
```

14 matches were found in this sequence.

Search entire project - Result of a project search. The *View* option is set to *Descriptions, Virtual Blast Section* listing the matching sequences by their best blast match in the blast search currently selected in the *Compose Header* form as the *Virtual Blast Search*.



TTTTCCG

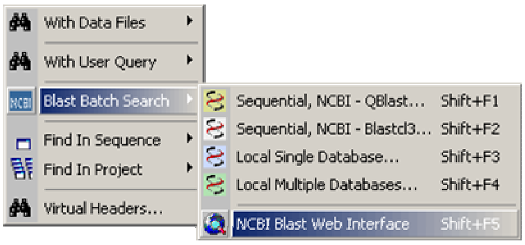
```
T37547 ubiquitin fusion protein - fission yeast schizosac
*Expect value outside range
Aj249796 nipsnap1 protein [danio rerio]
T38617 probable ubiquitin fusion degradation protein -
*Expect value outside range
Nc_001145 hsh155p [saccharomyces cerevisiae]
*Expect value outside range
*Expect value outside range
Af360341_1 af360341 unknown protein [arabidopsis
Aj319745 serine protease 1 [pyrenopeziza brass
*Expect value outside range
```

20 matches found among the 246 files

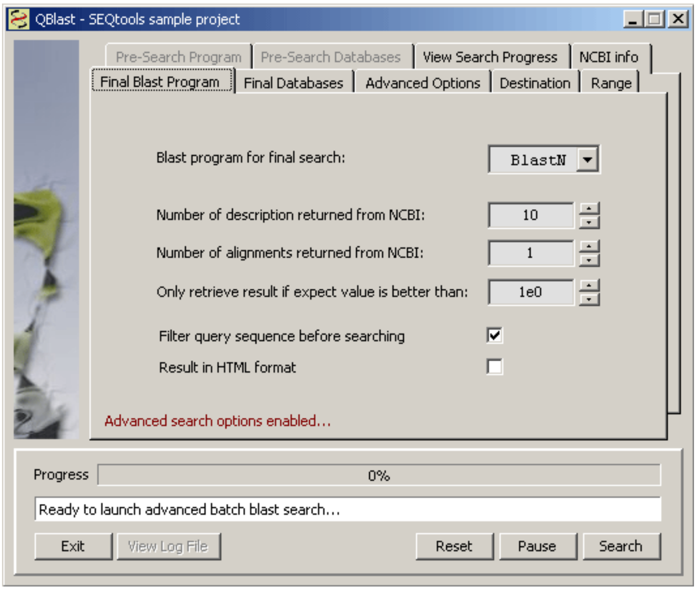
4.4.4 batch blast search

One the very strong features of SEQtools is the *Batch Blast Functions* allowing you to submit some or all sequences of a project to NCBI for homology searching of specified subsections of Genbank.

The Blast functions exist in two almost identical versions in SEQtools: One based on the QBlast scripts the other on the NCBI program *blastcl3.exe*. The first version is a web interface to the blast engine at NCBI while the other is a client/server type of arrangement. In designing both functions a considerable effort has been spent on self-recovery of the functions in case of crashes to ensure that when a batch search job is launched it should run to completion without user intervention. This holds true in nearly all cases, even when the job lasts several days (TBlastX) or includes a large number of sequences (up to 30,000 has been searched successfully). *Results are nice provided you have somewhere to store them in a form that allows you to retrieve them again...* The blast search functions of SEQtools are intimately integrated with the storage/retrieval system of search results. Read more about this under the *Header* menu. Provided your pc is sufficiently powerful you can launch a batch blast job - and continue working (on a different project) in an *different* instance of SEQtools while the blast search runs in the background.


With Data Files
With User Query
Blast Batch Search
Find In Sequence
Find In Project
Virtual Headers...
Sequential, NCBI - QBlast... Shift+F1
Sequential, NCBI - Blastcl3... Shift+F2
Local Single Database... Shift+F3
Local Multiple Databases... Shift+F4
NCBI Blast Web Interface Shift+F5

4.4.4.1 Batch blast at genbank - In most cases the different settings tabs are self-explanatory. Note, however, that you *cannot* import/parse blast results into sequence headers if you choose to get the results as html files. This has to do with the structure of the header/annotation. There is access to the Internet/Entrez at NCBI from within the sequence list displaying search results which to some extent compensates for this by providing an easy link to additional information.



Blast program for final search: BlastN

Number of description returned from NCBI: 10

Number of alignments returned from NCBI: 1

Only retrieve result if expect value is better than: 1e0

☑ Filter query sequence before searching

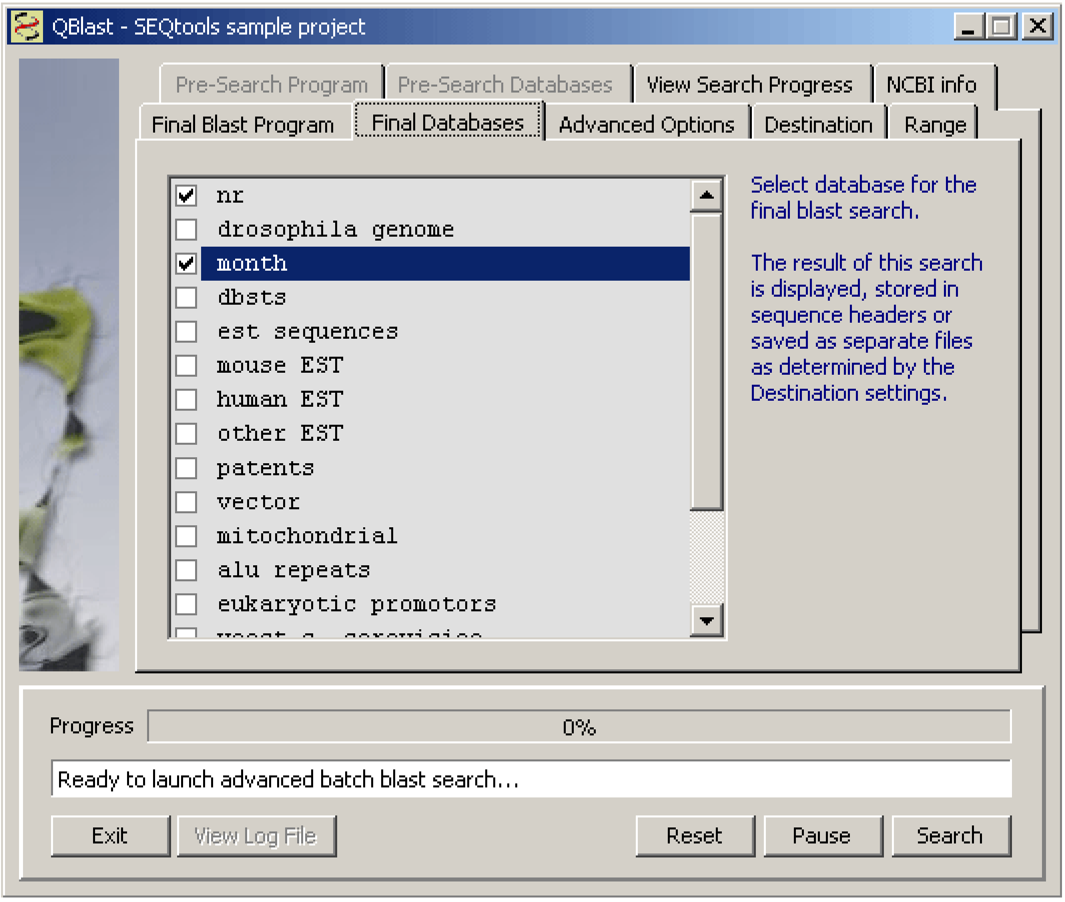
□ Result in HTML format

Advanced search options enabled...

Progress 0%

Ready to launch advanced batch blast search...

The list of available *main* sections of Genbank. The content of the database list (and the available blast programs in the dropdown list above as well) reflects the project type (nucleotide or protein).



☑ nr
□ drosophila genome
☑ month
□ dbsts
□ est sequences
□ mouse EST
□ human EST
□ other EST
□ patents
□ vector
□ mitochondrial
□ alu repeats
□ eukaryotic promotors
□ yeast s. cerevisiae

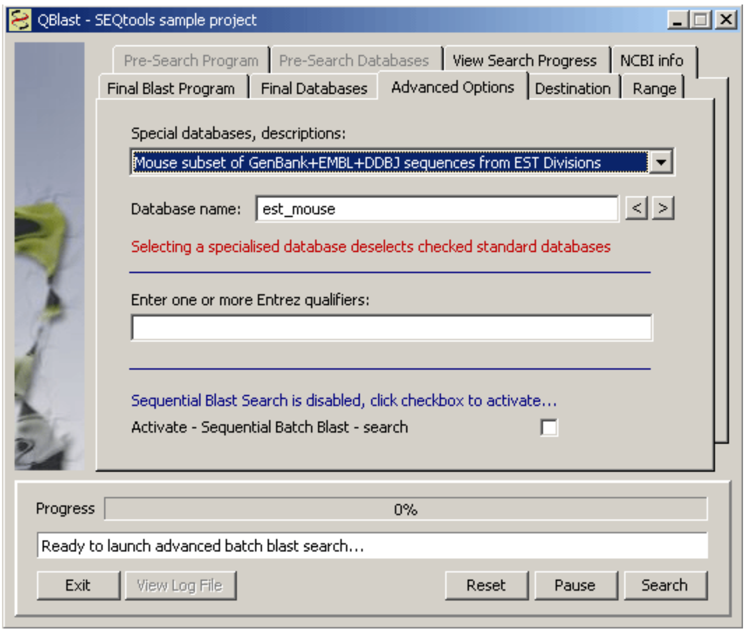
Select database for the final blast search.

The result of this search is displayed, stored in sequence headers or saved as separate files as determined by the Destination settings.

Ready to launch advanced batch blast search...

Under the *Advanced Options* tab is collected additional options for database selection. The list may not be entirely updated, but is the most recent the I could retrieve at NCBI.

Among the advanced options is a checkbox for activating *Sequential Search*. This implies that the function performs *two* sequential blast searches: The first with the project sequences, the second with the best match of the first search. When this option is active two more tabs on the blast form becomes active to allow you to select program and database(s) for the first search.



Special databases, descriptions: Mouse subset of GenBank+EMBL+DDBJ sequences from EST Divisions

Database name: est_mouse

Selecting a specialised database deselects checked standard databases

Enter one or more Entrez qualifiers:

Sequential Blast Search is disabled, click checkbox to activate...

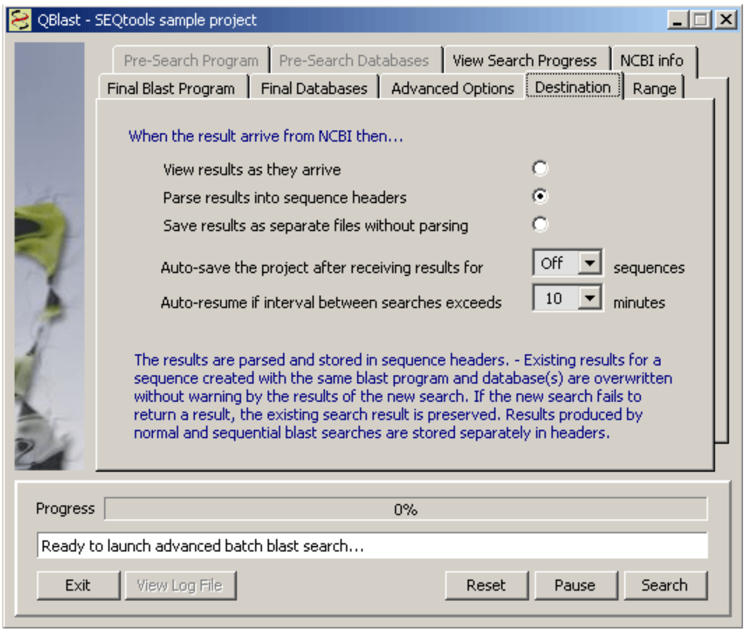
Activate - Sequential Batch Blast - search □

Progress 0%

Ready to launch advanced batch blast search...

The *Destination* tab contains an option to save the project for the specified number of completed searches. As all data in SEQtools are stores in PAM until the project is saved this setting should be active and be set to for example save/100 searches. If the blast engine at NCBI is very busy it may be an advantage to set the auto-resume value to 10 - 30 min to re-launch the sequence if more than the set amount of time has elapsed without a result has been received. Under normal circumstances it takes about 20-40 sec to search a 500 bp sequence with blastn.

You can choose to have the results *displayed as they arrive* in which case they are not stored in sequence headers. It is also possible to have the results *saved as separate files* (for example in html format). The default is *Parse results* into sequence headers.



When the result arrive from NCBI then...

View results as they arrive ○

Parse results into sequence headers ●

Save results as separate files without parsing ○

Auto-save the project after receiving results for Off sequences

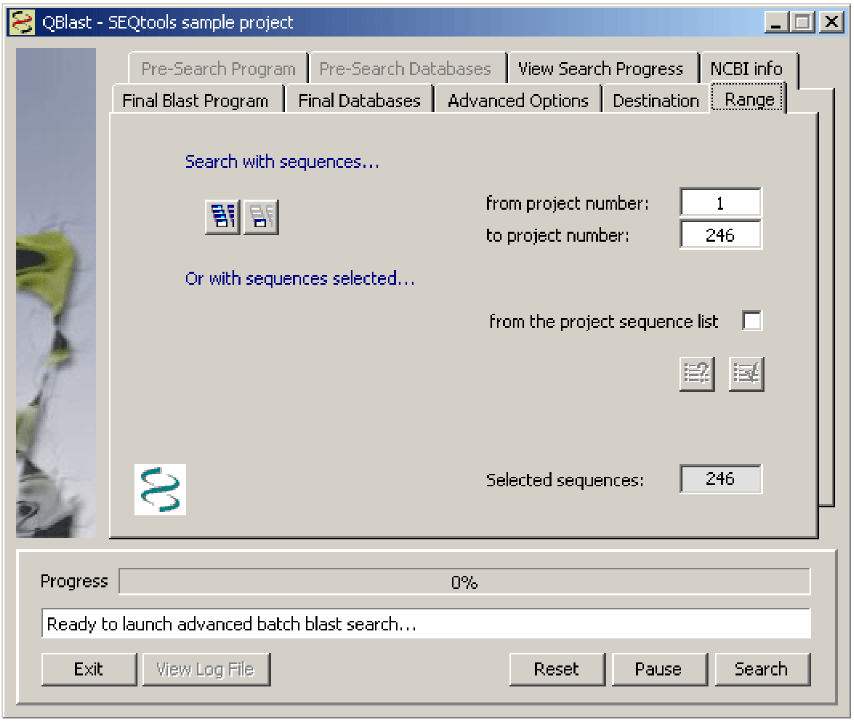
Auto-resume if interval between searches exceeds 10 minutes

The results are parsed and stored in sequence headers. - Existing results for a sequence created with the same blast program and database(s) are overwritten without warning by the results of the new search. If the new search fails to return a result, the existing search result is preserved. Results produced by normal and sequential blast searches are stored separately in headers.

Progress 0%

Ready to launch advanced batch blast search...

In this tab you can set the *range* of sequences you wish to search. This can be the *entire project* or the *currently displayed* sequence - or a *discontinuous series* of sequences selected from the project sequence list (described in detail on a separate page of the manual).



Search with sequences...

from project number: 1

to project number: 246

Or with sequences selected...

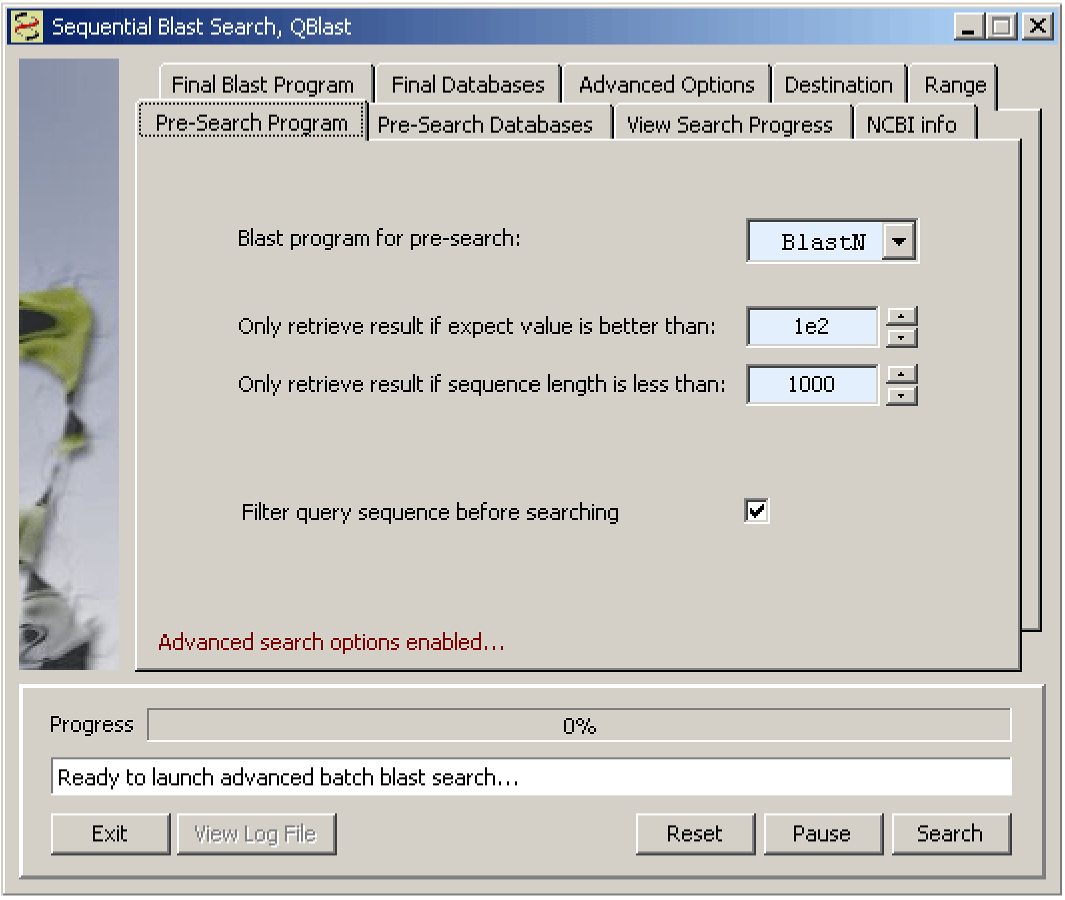
from the project sequence list □

Selected sequences: 246

Progress 0%

Ready to launch advanced batch blast search...

The tab for setting program options for the *first* search when *Advanced options* are enabled.



Blast program for pre-search: BlastN

Only retrieve result if expect value is better than: 1e2

Only retrieve result if sequence length is less than: 1000

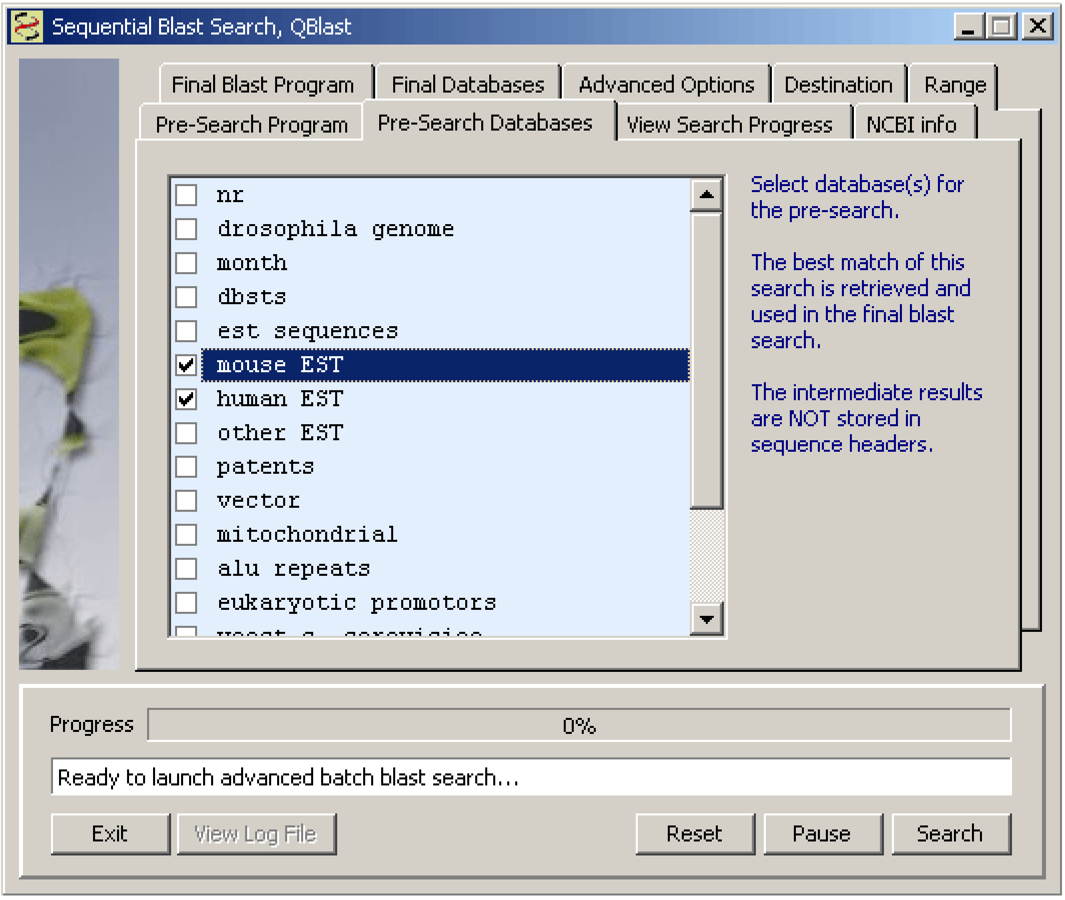
Filter query sequence before searching ☑

Advanced search options enabled...

Progress 0%

Ready to launch advanced batch blast search...

Available databases for the *first* search when *Advanced options* are enabled.



☐ nr
☐ drosophila genome
☐ month
☐ dbsts
☐ est sequences
☑ mouse EST
☑ human EST
☐ other EST
☐ patents
☐ vector
☐ mitochondrial
☐ alu repeats
☐ eukaryotic promotors
☐ yeast s. cerevisiae

Select database(s) for the pre-search.

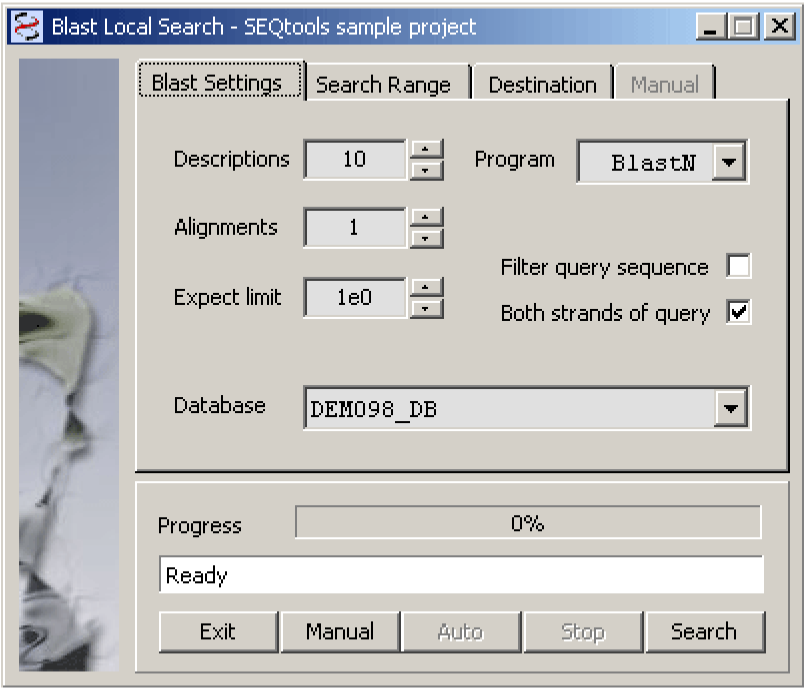
The best match of this search is retrieved and used in the final blast search.

The intermediate results are NOT stored in sequence headers.

Progress 0%

Ready to launch advanced batch blast search...

4.4.4.2 Local single database search - This function is for searching the project sequences with a local database, created by the user. Local databases can be created in different ways as described under the *Tools* menu. When you launch the local database search form, a message box (see below) informs you about available databases and displays a link to the function used to create local databases. The results are stored in sequence headers in exactly the same way as results from searches at NCBI.



Descriptions 10

Program BlastN

Alignments 1

Expect limit 1e0

Filter query sequence □

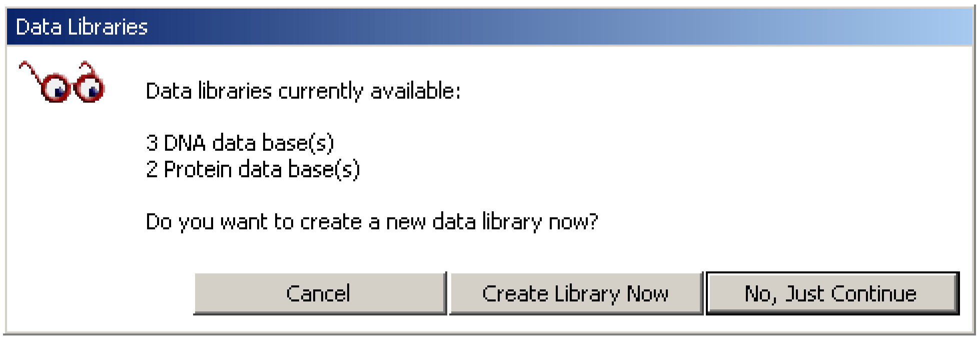
Both strands of query ☑

Database DEM098_DB

Progress 0%

Ready

The message box informing you about available local databases.

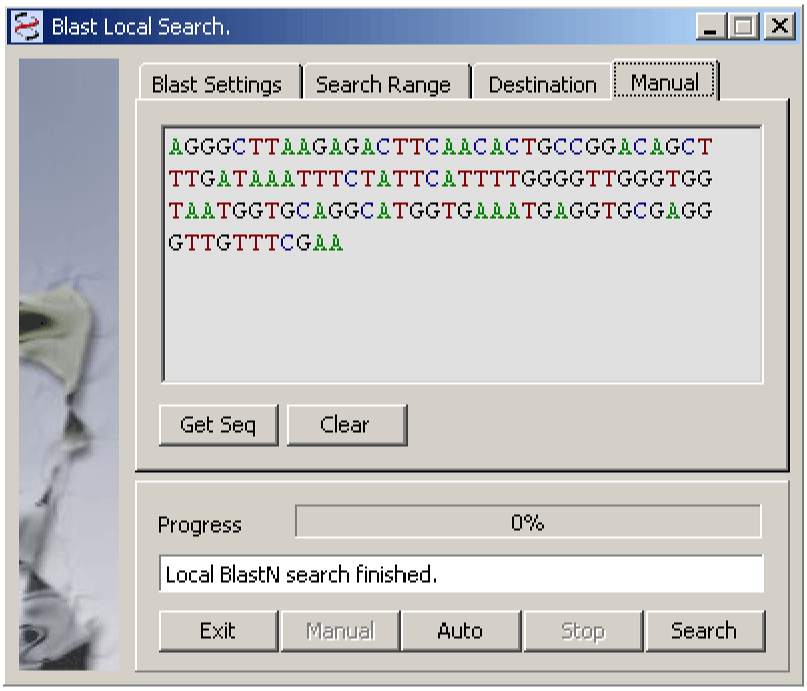


Data libraries currently available:

3 DNA data base(s)
2 Protein data base(s)

Do you want to create a new data library now?

To perform a *Manual search*, just highlight a region in the displayed sequence and click *Get Seq* to import the query into the local blast search form. Click *Search* to run the search.



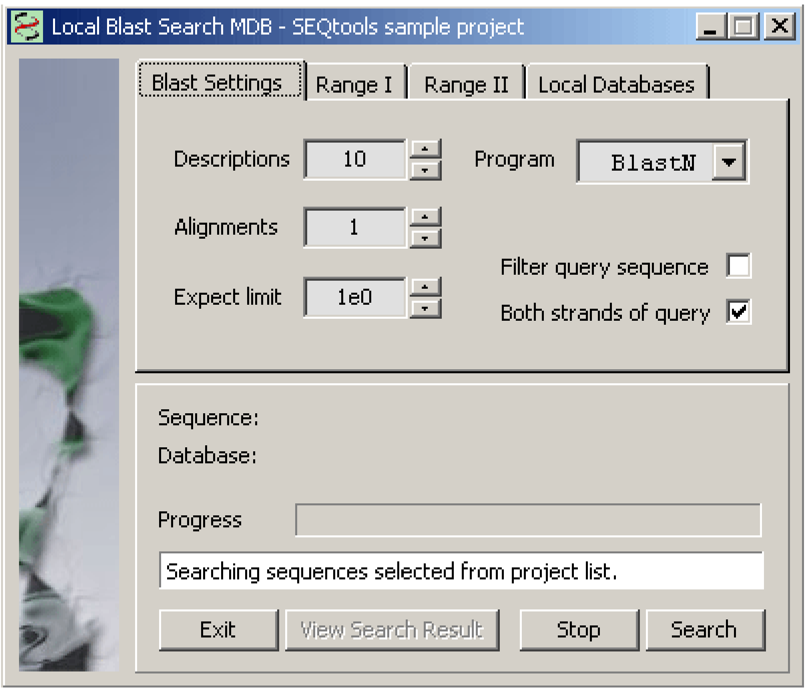
```
AGGGCTTAAGAGACTTCAACACTGCCGGACAGCT
TTGATAAATTTCTATTCATTTTGGGGTTGGGTGG
TAATGGTGCAGGCATGGTGAAATGAGGTGCGAGG
GTTGTTTCGAA
```

Local BlastN search finished.

4.4.4.3 Local multi-database search - Occasionally you may want to perform a local search on more than one local database. This can be accomplished with the *Search Multi-Database* function. Settings are the same as for other blast functions except it is possible to select more than one local databases.

Note that it is not possible to store more than one multi-database blast search in the sequence header. Running a second search overwrites the first one without warning. You should consider this function primarily as a help assisting you in getting an overview of the project rather than a proper analysis of individual sequences.

Clicking a line retrieves the selected sequence/sequence header (if displayed). You can the use the facilities (*Compose New Project*) in the sequence *Header* form to isolate interest ring sequences into a separate project - simply by selecting and clicking.



Descriptions 10

Alignments 1

Expect limit 1e0

Program BlastN

Filter query sequence □

Both strands of query ☑

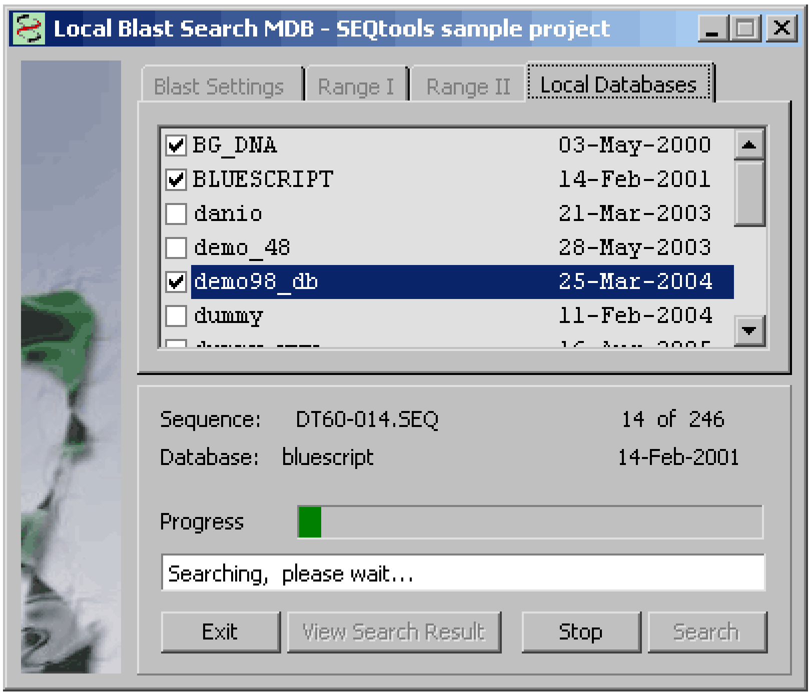
Sequence:

Database:

Progress

Searching sequences selected from project list.

Selection of local databases for a local multi-database blast search.



☑	BG_DNA	03-May-2000
☑	BLUESCRIPT	14-Feb-2001
□	danio	21-Mar-2003
□	demo_48	28-May-2003
☑	demo98_db	25-Mar-2004
□	dummy	11-Feb-2004

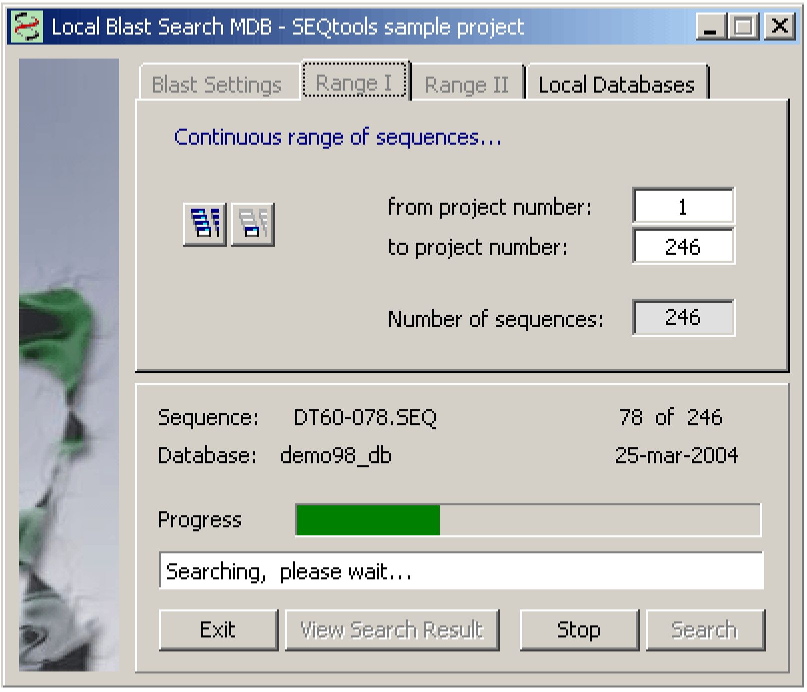
Sequence: DT60-014.SEQ 14 of 246

Database: bluescript 14-Feb-2001

Progress

Searching, please wait...

Setting the search range. The range can either be the displayed sequence or all project sequences (Range I) or a discontinuous series of sequences selected from the project sequence list (Range II).



Continuous range of sequences...

from project number: 1

to project number: 246

Number of sequences: 246

Sequence: DT60-078.SEQ 78 of 246

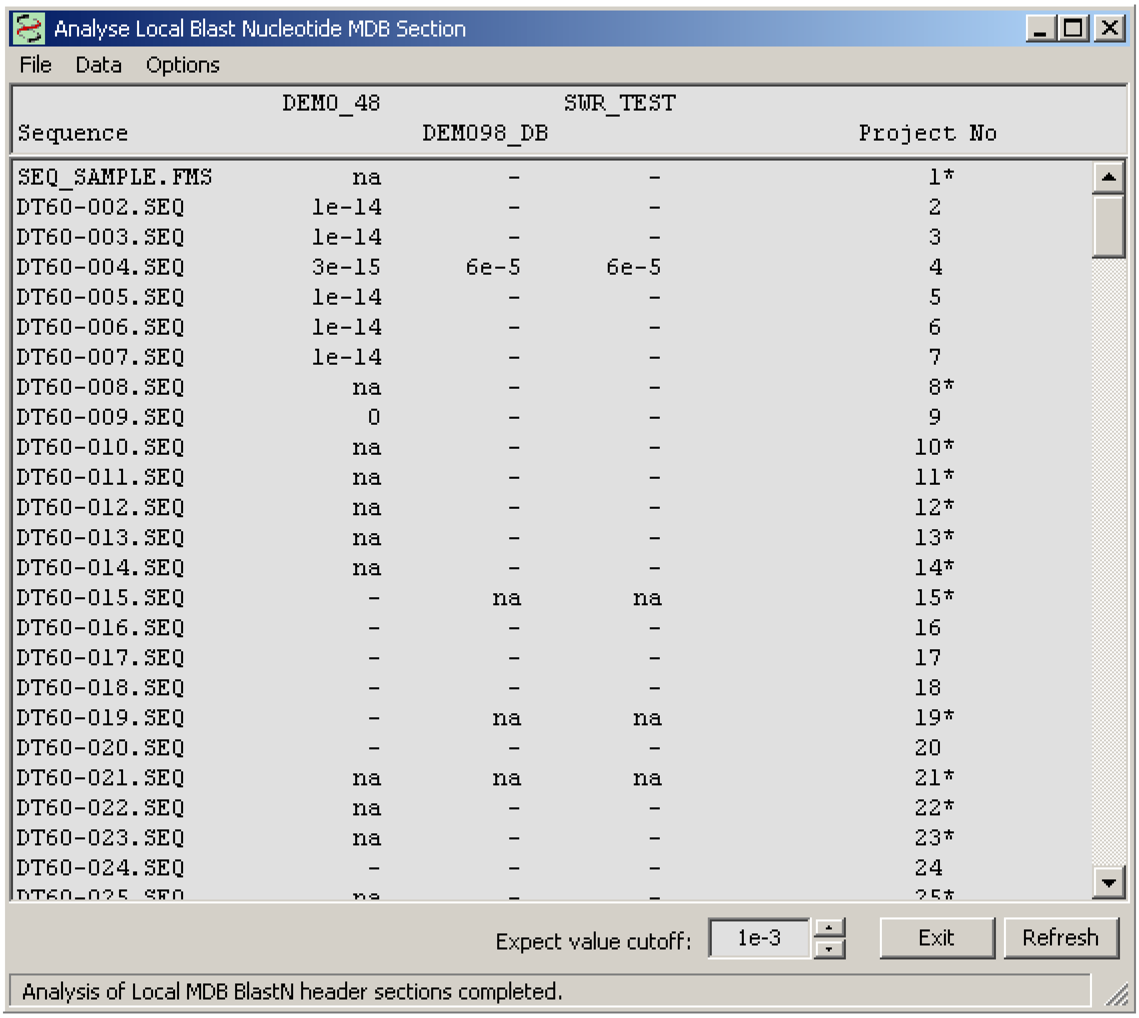
Database: demo98_db 25-mar-2004

Progress

Searching, please wait...

The results of a multi-database blast search is arranged somewhat differently in the sequence headers. All search results for a given sequence with the selected databases are contained in a single section of the header.

To create an overview of the multi-database search results the form shown below retrieves and displays the best multi-database search results for *all* sequences of the project.

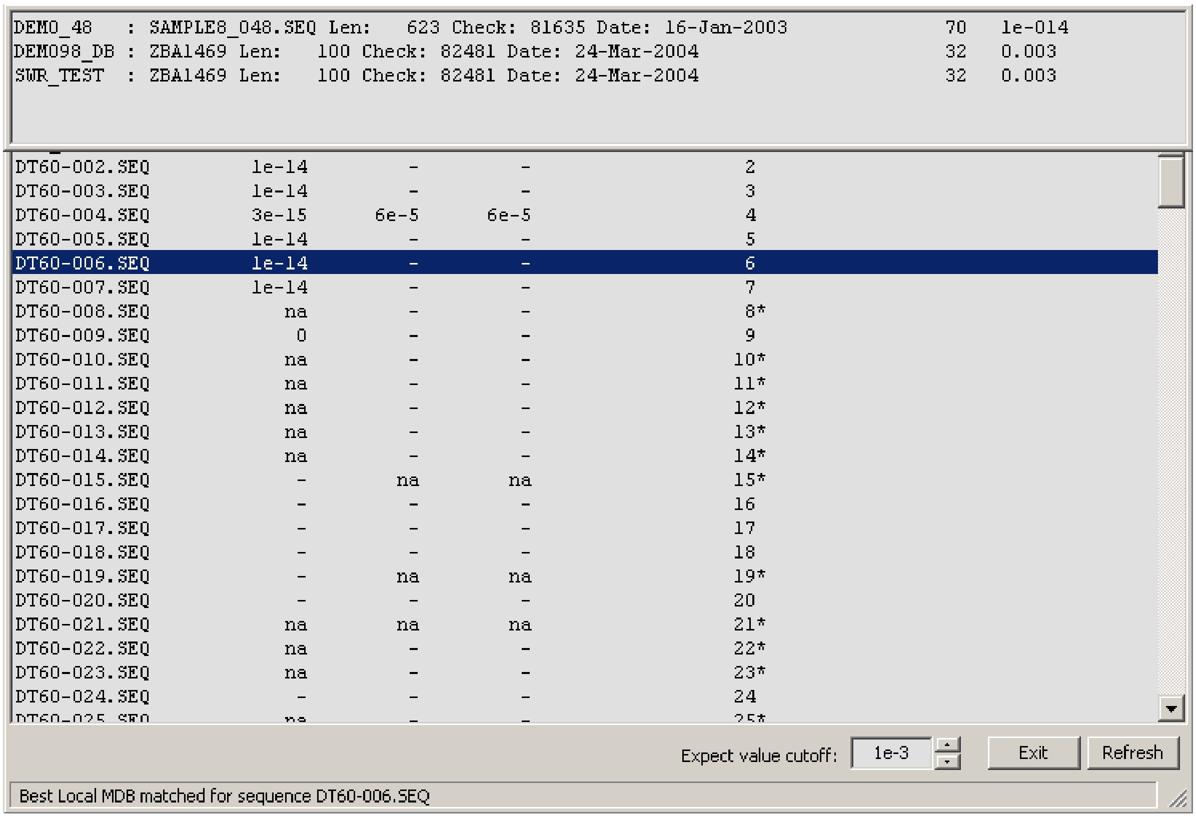


Sequence	DEMO_48	DEMO98_DB	SWR_TEST	Project No
SEQ_SAMPLE.FMS	na	-	-	1*
DT60-002.SEQ	1e-14	-	-	2
DT60-003.SEQ	1e-14	-	-	3
DT60-004.SEQ	3e-15	6e-5	6e-5	4
DT60-005.SEQ	1e-14	-	-	5
DT60-006.SEQ	1e-14	-	-	6
DT60-007.SEQ	1e-14	-	-	7
DT60-008.SEQ	na	-	-	8*
DT60-009.SEQ	0	-	-	9
DT60-010.SEQ	na	-	-	10*
DT60-011.SEQ	na	-	-	11*
DT60-012.SEQ	na	-	-	12*
DT60-013.SEQ	na	-	-	13*
DT60-014.SEQ	na	-	-	14*
DT60-015.SEQ	-	na	na	15*
DT60-016.SEQ	-	-	-	16
DT60-017.SEQ	-	-	-	17
DT60-018.SEQ	-	-	-	18
DT60-019.SEQ	-	na	na	19*
DT60-020.SEQ	-	-	-	20
DT60-021.SEQ	na	na	na	21*
DT60-022.SEQ	na	-	-	22*
DT60-023.SEQ	na	-	-	23*
DT60-024.SEQ	-	-	-	24
DT60-025.SEQ	na	-	-	25*

Expect value cutoff: 1e-3

Analysis of Local MDB BlastN header sections completed.

Highlighting a line in the form and then holding down the right mouse button retrieves the best match for *all* databases for the selected sequence.



```
DEMO_48    : SAMPLE8_048.SEQ Len:    623 Check: 81635 Date: 16-Jan-2003          70    1e-014
DEMO98_DB  : ZBA1469 Len:    100 Check: 82481 Date: 24-Mar-2004                  32    0.003
SWR_TEST   : ZBA1469 Len:    100 Check: 82481 Date: 24-Mar-2004                  32    0.003
```

```
DT60-002.SEQ        1e-14       -        -            2
DT60-003.SEQ        1e-14       -        -            3
DT60-004.SEQ        3e-15    6e-5     6e-5            4
DT60-005.SEQ        1e-14       -        -            5
DT60-006.SEQ        1e-14       -        -            6
DT60-007.SEQ        1e-14       -        -            7
DT60-008.SEQ           na       -        -            8*
DT60-009.SEQ            0       -        -            9
DT60-010.SEQ           na       -        -           10*
DT60-011.SEQ           na       -        -           11*
DT60-012.SEQ           na       -        -           12*
DT60-013.SEQ           na       -        -           13*
DT60-014.SEQ           na       -        -           14*
DT60-015.SEQ            -      na       na           15*
DT60-016.SEQ            -       -        -           16
DT60-017.SEQ            -       -        -           17
DT60-018.SEQ            -       -        -           18
DT60-019.SEQ            -      na       na           19*
DT60-020.SEQ            -       -        -           20
DT60-021.SEQ           na      na       na           21*
DT60-022.SEQ           na       -        -           22*
DT60-023.SEQ           na       -        -           23*
DT60-024.SEQ            -       -        -           24
DT60-025.SEQ           na       -        -           25*
```

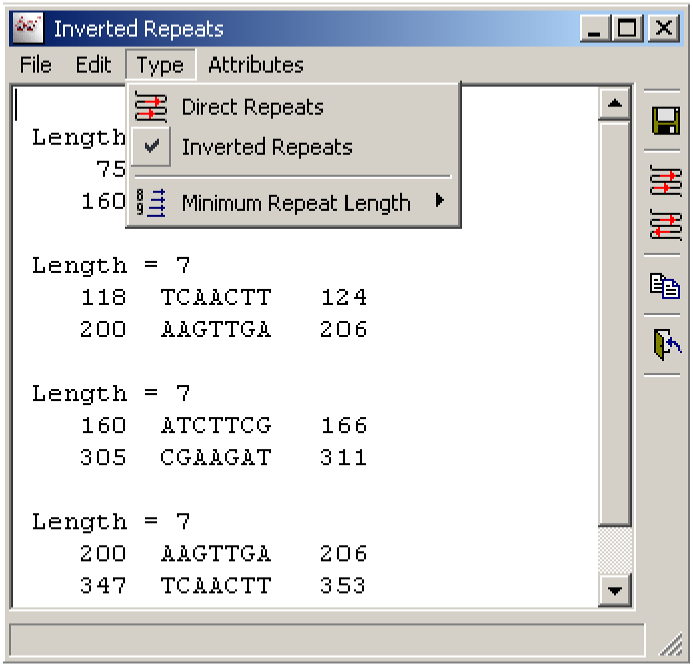
Expect value cutoff: 1e-3

Best Local MDB matched for sequence DT60-006.SEQ

4.4.5 find in sequence

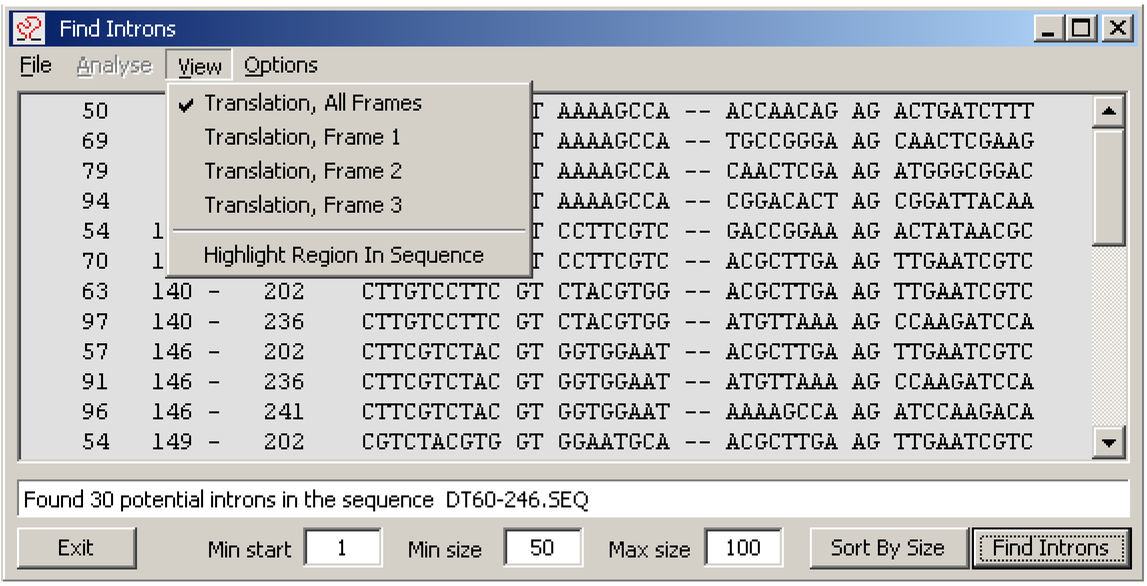
Two are included for revealing the existence of repeats in the sequence and for detecting/indicating the presence of introns (primarily in yeast). Neither function should be considered as perfect. Much more sophisticated functions are required to identify introns in mammalian genes and the user is strongly advised to visit websites specifically directed towards this analysis.

4.4.5.1 Repeats - Identifies direct and inverted repeated regions in sequences.



```
 Length
    75
   160

 Length = 7
   118   TCAACTT   124
   200   AAGTTGA   206

 Length = 7
   160   ATCTTCG   166
   305   CGAAGAT   311

 Length = 7
   200   AAGTTGA   206
   347   TCAACTT   353
```

4.4.6.2 Introns - Primitive function for identification of introns in yeast.



```
  50                                     T AAAAGCCA -- ACCAACAG AG ACTGATCTTT
  69                                     T AAAAGCCA -- TGCCGGGA AG CAACTCGAAG
  79                                     T AAAAGCCA -- CAACTCGA AG ATGGGCGGAC
  94                                     T AAAAGCCA -- CGGACACT AG CGGATTACAA
  54   1                                 T CCTTCGTC -- GACCGGAA AG ACTATAACGC
  70   1                                 T CCTTCGTC -- ACGCTTGA AG TTGAATCGTC
  63  140 -  202   CTTGTCCTTC GT CTACGTGG -- ACGCTTGA AG TTGAATCGTC
  97  140 -  236   CTTGTCCTTC GT CTACGTGG -- ATGTTAAA AG CCAAGATCCA
  57  146 -  202   CTTCGTCTAC GT GGTGGAAT -- ACGCTTGA AG TTGAATCGTC
  91  146 -  236   CTTCGTCTAC GT GGTGGAAT -- ATGTTAAA AG CCAAGATCCA
  96  146 -  241   CTTCGTCTAC GT GGTGGAAT -- AAAAGCCA AG ATCCAAGACA
  54  149 -  202   CGTCTACGTG GT GGAATGCA -- ACGCTTGA AG TTGAATCGTC
```

Found 30 potential introns in the sequence DT60-246.SEQ

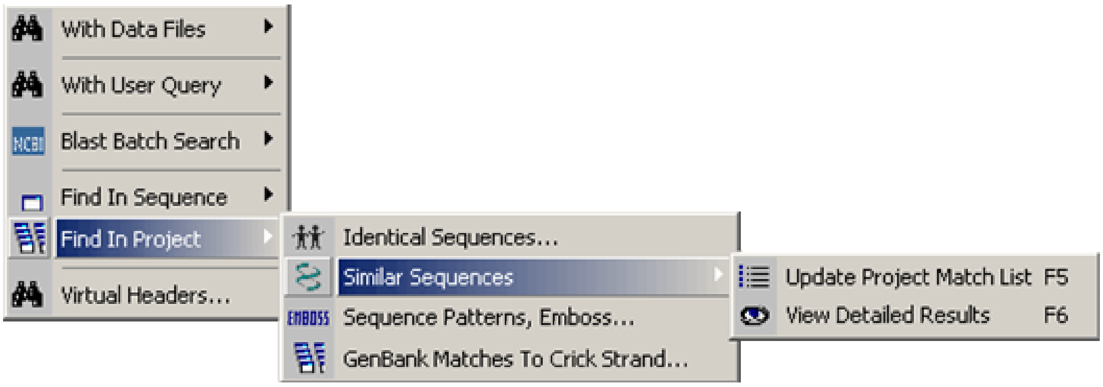
Min start: 1

Min size: 50

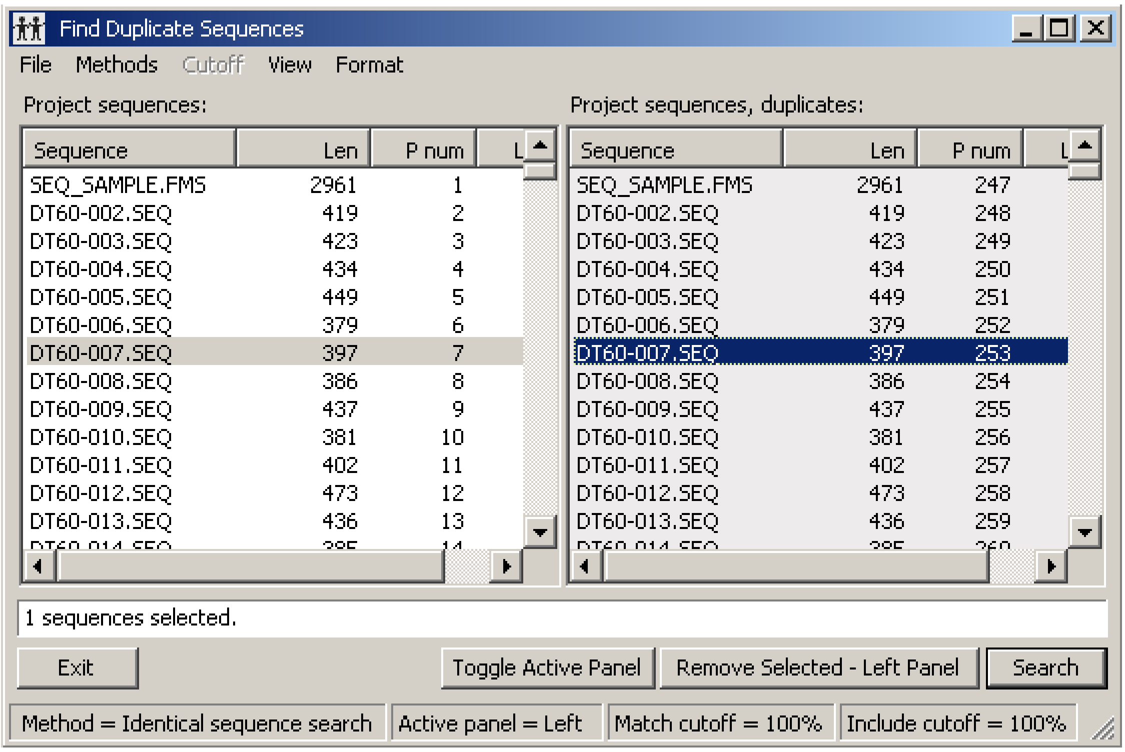
Max size: 100

4.4.6 find in project

The four functions described below are all designed to perform analyses on the entire project. This includes finding duplicate sequences, performing a quick project blast search to reveal internal similarity among project sequences, a emboss based pattern search and finally a function to identify project sequences with antisense blast matches to sequences in sequence header.



4.4.6.1 Duplicates - Scans project and lists duplicate sequences. Duplicate sequences can then be selected and removed from the project.



Project sequences:

Sequence	Len	P num
SEQ_SAMPLE.FMS	2961	1
DT60-002.SEQ	419	2
DT60-003.SEQ	423	3
DT60-004.SEQ	434	4
DT60-005.SEQ	449	5
DT60-006.SEQ	379	6
DT60-007.SEQ	397	7
DT60-008.SEQ	386	8
DT60-009.SEQ	437	9
DT60-010.SEQ	381	10
DT60-011.SEQ	402	11
DT60-012.SEQ	473	12
DT60-013.SEQ	436	13

Project sequences, duplicates:

Sequence	Len	P num
SEQ_SAMPLE.FMS	2961	247
DT60-002.SEQ	419	248
DT60-003.SEQ	423	249
DT60-004.SEQ	434	250
DT60-005.SEQ	449	251
DT60-006.SEQ	379	252
DT60-007.SEQ	397	253
DT60-008.SEQ	386	254
DT60-009.SEQ	437	255
DT60-010.SEQ	381	256
DT60-011.SEQ	402	257
DT60-012.SEQ	473	258
DT60-013.SEQ	436	259

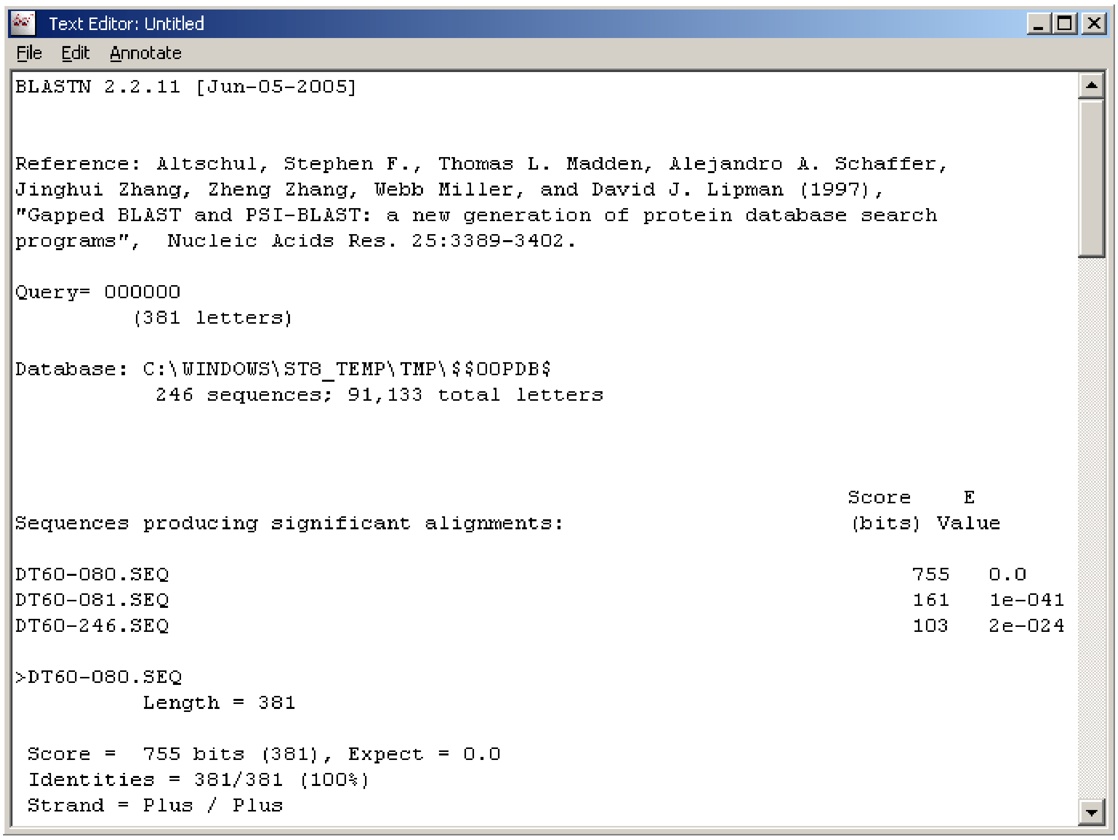
1 sequences selected.

Method = Identical sequence search | Active panel = Left | Match cutoff = 100% | Include cutoff = 100%

4.4.6.2 Project blast - Project blast search builds a local database (if not already present) and performs a blast search of the currently displayed sequence against the project database.

The result is either displayed in a simple text form (below) or in the sequence list (with light blue background) of the main sequence editor of **SEQtools.**

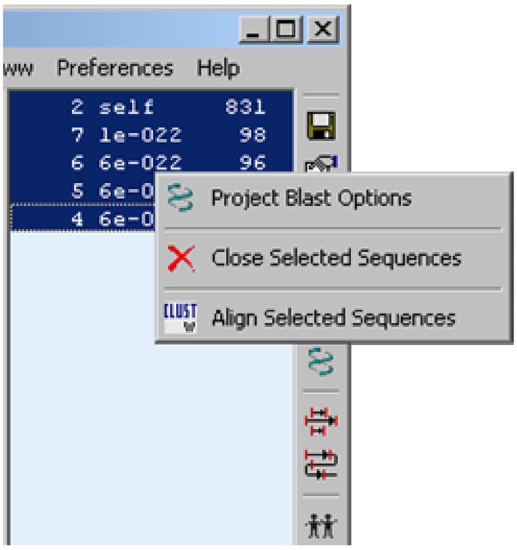
The latter display can be achieved by clicking <F5>.



```
BLASTN 2.2.11 [Jun-05-2005]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= 000000
         (381 letters)

Database: C:\WINDOWS\ST8_TEMP\TMP\$$OOPDB$
           246 sequences; 91,133 total letters



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

DT60-080.SEQ                                                          755   0.0
DT60-081.SEQ                                                          161   1e-041
DT60-246.SEQ                                                          103   2e-024

>DT60-080.SEQ
          Length = 381

 Score =  755 bits (381), Expect = 0.0
 Identities = 381/381 (100%)
 Strand = Plus / Plus
```

Project blast results for displayed sequence. Right-click the sequence list to return to the normal list (grey=load order or yellow=sorted).

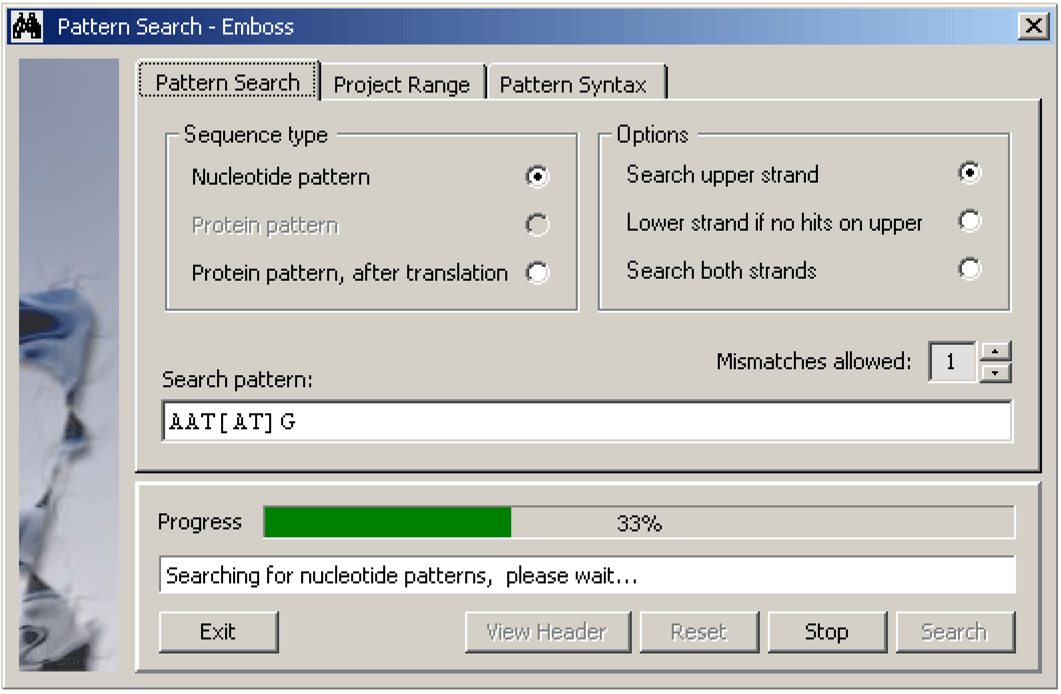
The list of project blast matches is linked to an alignment function (ClustalW): Select some or all matches (click while holding down <CTRL>) and Shift-Right click to open the popup menu offering to access to project blast preferences, *Close Selected Sequences* or *Align Selected Sequences* as illustrated by the screenshot below.



```
  2 self      831
  7 1e-022     98
  6 6e-022     96
  5 6e-0
  4 6e-0
```

4.4.6.3 Patterns - The pattern search function utilizes *emboss* functions to find patterns in project sequences. The Range settings are the same as for other search programs in SEQtools.

The syntax is briefly described in the form below (consult the emboss homepage for additional details). The result of the pattern search is stored in the sequence headers and are displayed by clicking the *View Header* command button.



Pattern Search - Emboss

Sequence type

● Nucleotide pattern

○ Protein pattern

○ Protein pattern, after translation

Options

● Search upper strand

○ Lower strand if no hits on upper

○ Search both strands

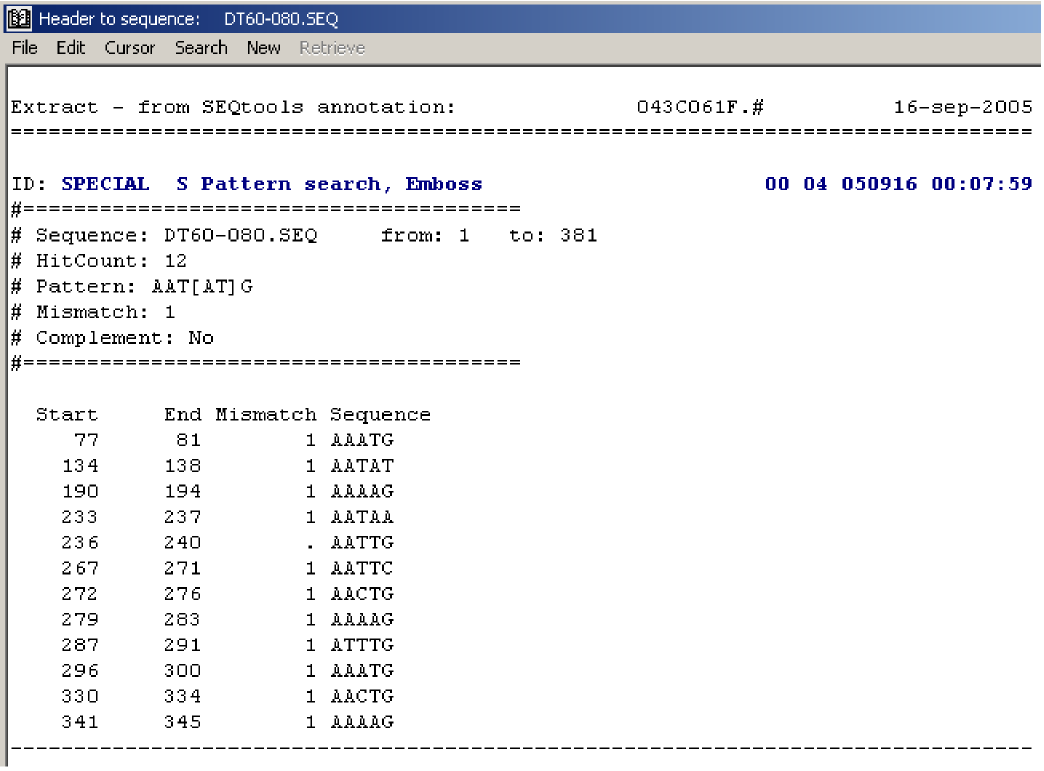
Mismatches allowed: 1

Search pattern: AAT[AT]G

Progress 33%

Searching for nucleotide patterns, please wait...

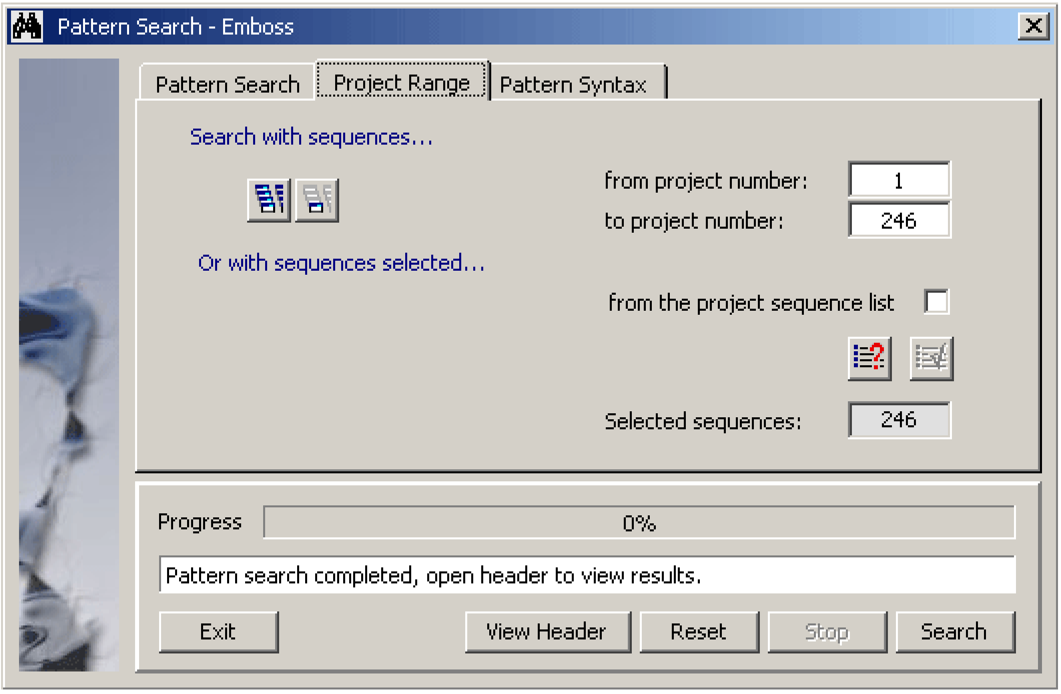
The sequence header displaying the results of a pattern search. As for local multi-database blast search it is only possible to store the results of a single pattern sear. The next search will overwrite the existing results without warning.



Header to sequence: DT60-080.SEQ
File Edit Cursor Search New Retrieve

```
Extract - from SEQtools annotation:              043C061F.#          16-sep-2005
================================================================================

ID: SPECIAL  S Pattern search, Emboss                      00 04 050916 00:07:59
#=======================================
# Sequence: DT60-080.SEQ     from: 1    to: 381
# HitCount: 12
# Pattern: AAT[AT]G
# Mismatch: 1
# Complement: No
#=======================================

  Start      End Mismatch Sequence
     77       81        1 AAATG
    134      138        1 AATAT
    190      194        1 AAAAG
    233      237        1 AATAA
    236      240        . AATTG
    267      271        1 AATTC
    272      276        1 AACTG
    279      283        1 AAAAG
    287      291        1 ATTTG
    296      300        1 AAATG
    330      334        1 AACTG
    341      345        1 AAAAG
--------------------------------------------------------------------------------
```

Setting the *Range* parameter for a pattern search.



Search with sequences...

from project number: 1

to project number: 246

Or with sequences selected...

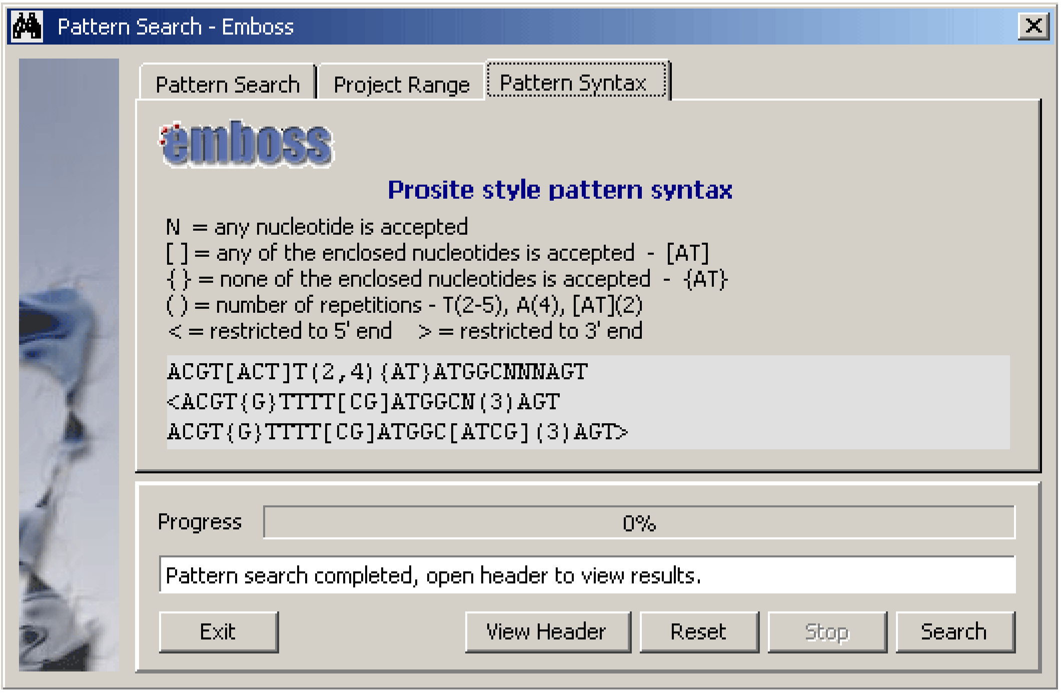
from the project sequence list □

Selected sequences: 246

Progress 0%

Pattern search completed, open header to view results.

The brief description of the pattern syntax. Consult the emboss homepage for details and additional examples.



Prosite style pattern syntax

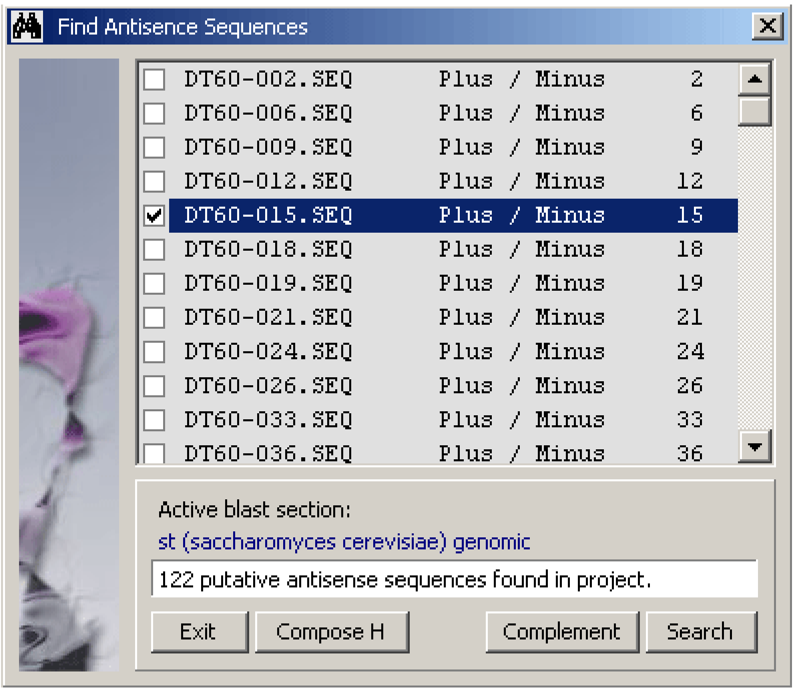
N = any nucleotide is accepted
[ ] = any of the enclosed nucleotides is accepted - [AT]
{ } = none of the enclosed nucleotides is accepted - {AT}
( ) = number of repetitions - T(2-5), A(4), [AT](2)
< = restricted to 5' end   > = restricted to 3' end

```
ACGT[ACT]T(2,4){AT}ATGGCNNNAGT
<ACGT{G}TTTT[CG]ATGGCN(3)AGT
ACGT{G}TTTT[CG]ATGGC[ATCG](3)AGT>
```

Progress 0%

Pattern search completed, open header to view results.

4.4.6.4 Antisense - This function searches all headers of the project and examines the alignment sections of blast results (if present) for the selected *Virtual Blast Search* and lists the strand orientation for the best match of each sequence in the project. It is possible to select project sequences to be complemented to make the orientation of the sequence and its database match the same.



```
□ DT60-002.SEQ    Plus / Minus    2
□ DT60-006.SEQ    Plus / Minus    6
□ DT60-009.SEQ    Plus / Minus    9
□ DT60-012.SEQ    Plus / Minus   12
☑ DT60-015.SEQ    Plus / Minus   15
□ DT60-018.SEQ    Plus / Minus   18
□ DT60-019.SEQ    Plus / Minus   19
□ DT60-021.SEQ    Plus / Minus   21
□ DT60-024.SEQ    Plus / Minus   24
□ DT60-026.SEQ    Plus / Minus   26
□ DT60-033.SEQ    Plus / Minus   33
□ DT60-036.SEQ    Plus / Minus   36
```

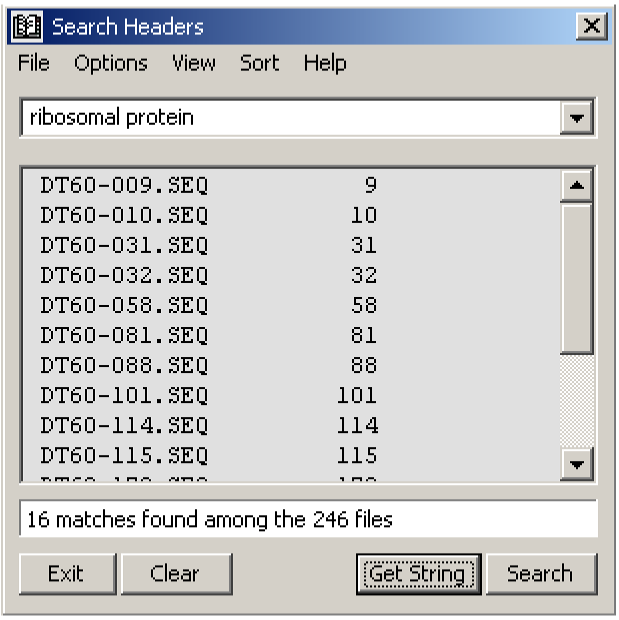
Active blast section:
st (saccharomyces cerevisiae) genomic

122 putative antisense sequences found in project.

4.4.7 search virtual headers

The last search function enables you to search sequence headers with plain text queries. The menu options include case sensitive/insensitive, whole word only, match listing as sequence data or descriptions.

Search headers - match listing by sequence data.

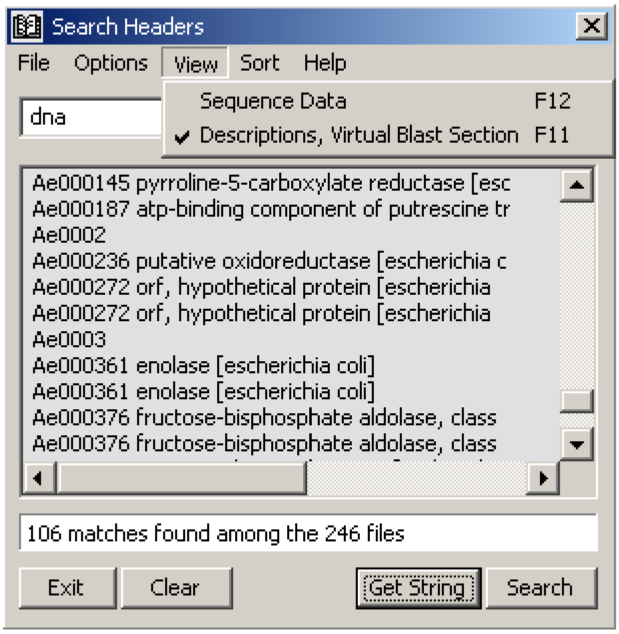


ribosomal protein

```
DT60-009.SEQ          9
DT60-010.SEQ         10
DT60-031.SEQ         31
DT60-032.SEQ         32
DT60-058.SEQ         58
DT60-081.SEQ         81
DT60-088.SEQ         88
DT60-101.SEQ        101
DT60-114.SEQ        114
DT60-115.SEQ        115
```

16 matches found among the 246 files

View setting set to *Descriptions, Virtual Blast Sections* causes match listing by description lines.

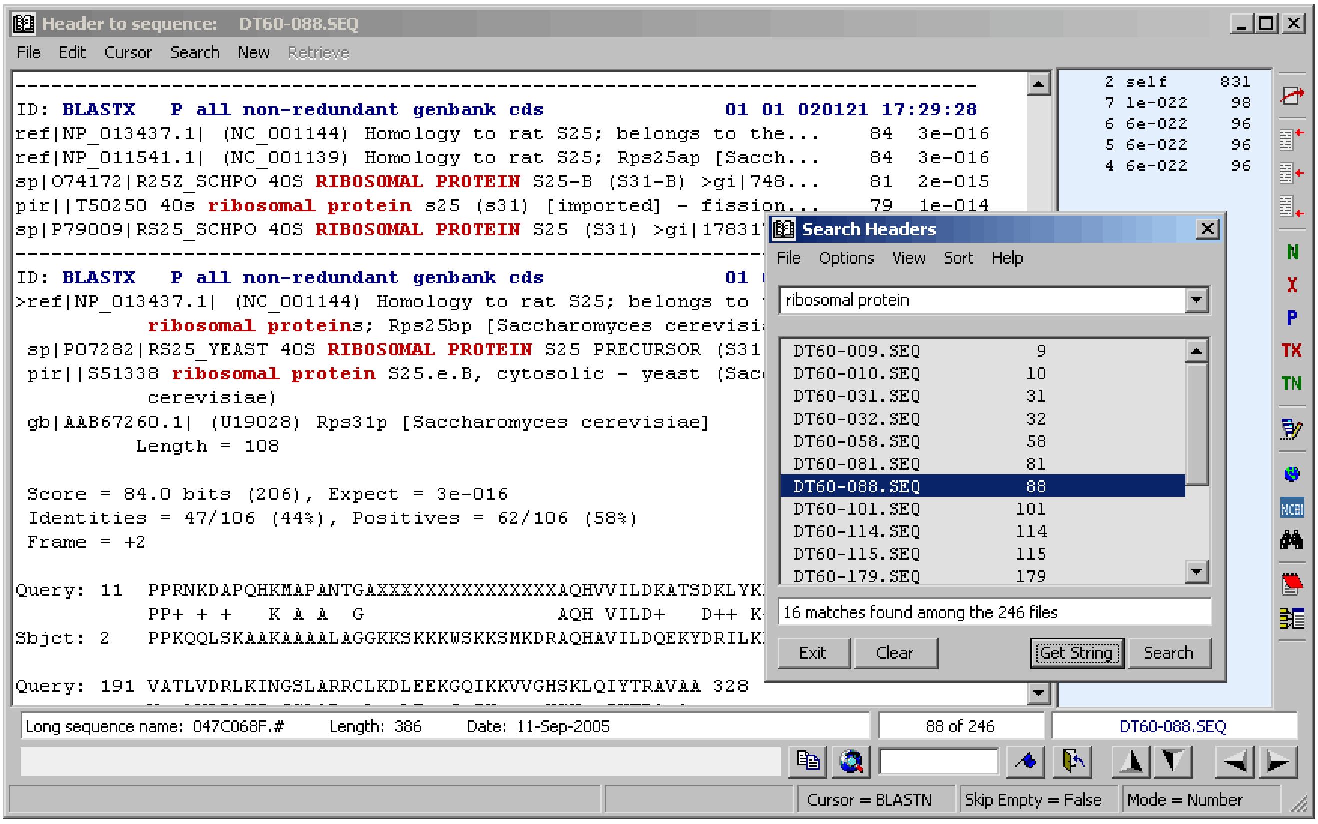


dna

```
Ae000145 pyrroline-5-carboxylate reductase [esc
Ae000187 atp-binding component of putrescine tr
Ae0002
Ae000236 putative oxidoreductase [escherichia c
Ae000272 orf, hypothetical protein [escherichia
Ae000272 orf, hypothetical protein [escherichia
Ae0003
Ae000361 enolase [escherichia coli]
Ae000361 enolase [escherichia coli]
Ae000376 fructose-bisphosphate aldolase, class
Ae000376 fructose-bisphosphate aldolase, class
```

106 matches found among the 246 files

Clicking a line in the header results form retrieves the relevant sequence header and paints matches to the query string red in the header text. Note that the header search is limited to the currently selected items (check marked in the *Compose Header* form).



Header to sequence: DT60-088.SEQ
File Edit Cursor Search New Retrieve

```
------------------------------------------------------------------------------
ID: BLASTX   P all non-redundant genbank cds              01 01 020121 17:29:28
ref|NP_013437.1| (NC_001144) Homology to rat S25; belongs to the...    84  3e-016
ref|NP_011541.1| (NC_001139) Homology to rat S25; Rps25ap [Sacch...    84  3e-016
sp|O74172|R25Z_SCHPO 40S RIBOSOMAL PROTEIN S25-B (S31-B) >gi|748...    81  2e-015
pir||T50250 40s ribosomal protein s25 (s31) [imported] - fission...    79  1e-014
sp|P79009|RS25_SCHPO 40S RIBOSOMAL PROTEIN S25 (S31) >gi|17831
--------------------------------------------------------------
ID: BLASTX   P all non-redundant genbank cds              01
>ref|NP_013437.1| (NC_001144) Homology to rat S25; belongs to
           ribosomal proteins; Rps25bp [Saccharomyces cerevisi
 sp|P07282|RS25_YEAST 40S RIBOSOMAL PROTEIN S25 PRECURSOR (S31
 pir||S51338 ribosomal protein S25.e.B, cytosolic - yeast (Sac
           cerevisiae)
 gb|AAB67260.1| (U19028) Rps31p [Saccharomyces cerevisiae]
          Length = 108

 Score = 84.0 bits (206), Expect = 3e-016
 Identities = 47/106 (44%), Positives = 62/106 (58%)
 Frame = +2

Query: 11   PPRNKDAPQHKMAPANTGAXXXXXXXXXXXXXXXAQHVVILDKATSDKLYK
             PP+ + +   K A A  G                AQH VILD+   D++ K
Sbjct: 2    PPKQQLSKAAKAAAALAGGKKSKKKWSKKSMKDRAQHAVILDQEKYDRILK

Query: 191 VATLVDRLKINGSLARRCLKDLEEKGQIKKVVGHSKLQIYTRAVAA 328
```

```
  2 self     831
  7 1e-022    98
  6 6e-022    96
  5 6e-022    96
  4 6e-022    96
```

ribosomal protein

```
DT60-009.SEQ       9
DT60-010.SEQ      10
DT60-031.SEQ      31
DT60-032.SEQ      32
DT60-058.SEQ      58
DT60-081.SEQ      81
DT60-088.SEQ      88
DT60-101.SEQ     101
DT60-114.SEQ     114
DT60-115.SEQ     115
DT60-179.SEQ     179
```

16 matches found among the 246 files

Long sequence name: 047C068F.# Length: 386 Date: 11-Sep-2005

88 of 246

DT60-088.SEQ

Cursor = BLASTN

Skip Empty = False

Mode = Number