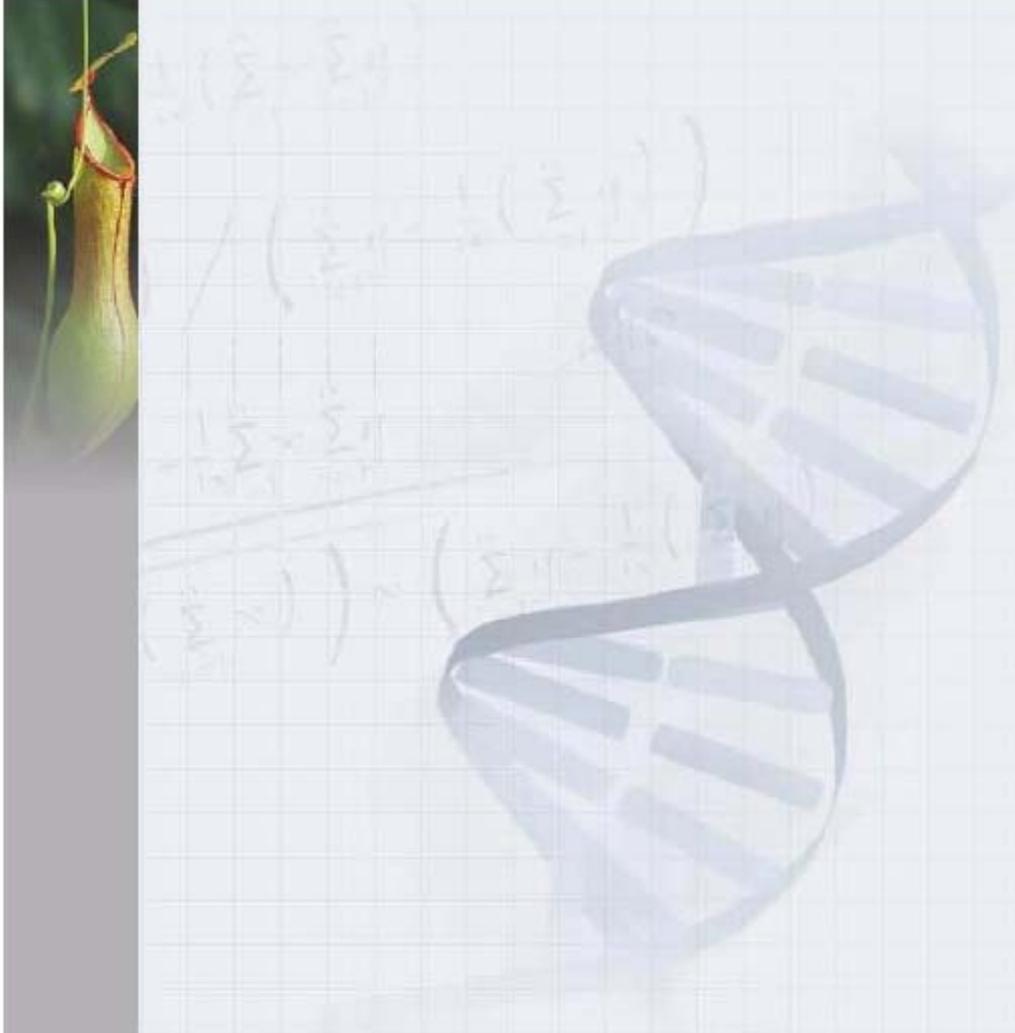


The Kodon quickguide



Version 3.5

Copyright © 2002-2007, Applied Maths NV. All rights reserved.

Kodon is a registered trademark of Applied Maths NV.
All other product names or trademarks are the property of their respective owners.



NOTES

SUPPORT BY APPLIED MATHS

While the best efforts have been made in preparing this manuscript, no liability is assumed by Applied Maths with respect to the use of the information provided.

No part of this guide may be reproduced by any means without prior written permission of the authors.

Kodon is a registered trademark of Applied Maths NV. All other product names or trademarks are the property of their respective owners.

Copyright © 2007 Applied Maths NV. All rights reserved.

Applied Maths NV
Keistraat 120
9830 Sint-Martens-Latem
Belgium

PHONE: +32 9 2222 100
FAX: +32 9 2222 102
E-MAIL: info@applied-maths.com

Applied Maths, Inc.
512 East 11th Street, Suite 207
Austin, Texas 78701
U.S.A.

PHONE: +1 512-482-9700
FAX: +1 512-482-9708
E-MAIL: info-US@applied-maths.com

URL: <http://www.applied-maths.com>

Table of contents

1. Kodon Basic Software	5	2.3 Clustering and phylogeny	16
1.1 Introduction	5	3. Molecular Analysis	19
1.2 Select functions	5	3.1 Frame analysis	19
1.3 Database querying	6	3.2 Primer design	20
1.4 Downloading sequences directly from the internet	7	3.3 Multiplex PCR design	21
1.5 Importing sequences from downloaded files ..	8	3.4 Pairwise matching and repeat analysis	22
1.6 Assembling sequences with Assembler	9	3.5 Motif search	24
1.7 Sequence editing	10	3.6 RNA secondary structure	24
1.8 Restriction enzyme analysis	12	3.7 Protein properties	25
1.9 Homology search	13	3.8 Vector construction	25
2. Multiple Alignment, Clustering & Phylog- eny	15	4. Chromosome Mapping	29
2.1 Sequence selection from feature matrix	15	4.1 Comparative chromosome mapping	29
2.2 Multiple alignment	16	4.2 Chromosome alignment	31
		4.3 Genome annotation	34

1. Kodon Basic Software

1.1 Introduction

This guide provides a general introduction to Kodon. Since Kodon is a powerful application, many features are not covered in this quickguide. Please refer to the manual for more detailed information on these features.

Two databases will be used in this quickguide: **Demobase** and **Bacterial chromosomes**. To install both databases on your computer, enable '*Install Demonstration Database*' and '*Install Bacterial Chromosome Database*' in the installation wizard.

1.1.1 Start Kodon by double-clicking on the  icon on your desktop or select *Start > Programs > Kodon*.

1.1.2 Make sure that **Demobase** is selected from the list and press *<Open database>*.

Kodon opens the database **Demobase**. This database contains bacterial genome sequences, eukaryote genes of plants, chromosomes of yeasts and a number of 5S, 16S, and 23S ribosomal RNA gene sequences of various bacteria (see Figure 1-1).

1.1.3 The *Main* window consists of a menu, a toolbar, a status bar and three panels (see Figure 1-1).

1.2 Select functions

1.2.1 Hold down the CTRL key and click on a sequence in the database panel (see Figure 1-1). The sequence is marked by a blue arrow.

1.2.2 Hold down the SHIFT key and click on another sequence. A range of sequences is now selected.

1.2.3 Press F5 to select all entries.

1.2.4 Clear the selection with F4.

Within each database, it is possible to create subsets, which can be saved and viewed. These subsets are called lists. In the **Demobase**, seven lists are already present in the Lists panel (see Figure 1-1).

1.2.5 Double-click on the fourth list ('Ribosomal RNAs'). A new tab appears in the left panel. Scroll down the page. The list contains 62 sequences.

1.2.6 Click on the  button in the list tab to close the list.

1.2.7 Click on the database field 'Organism' and then select *Edit > Sort by selected field* or click on .

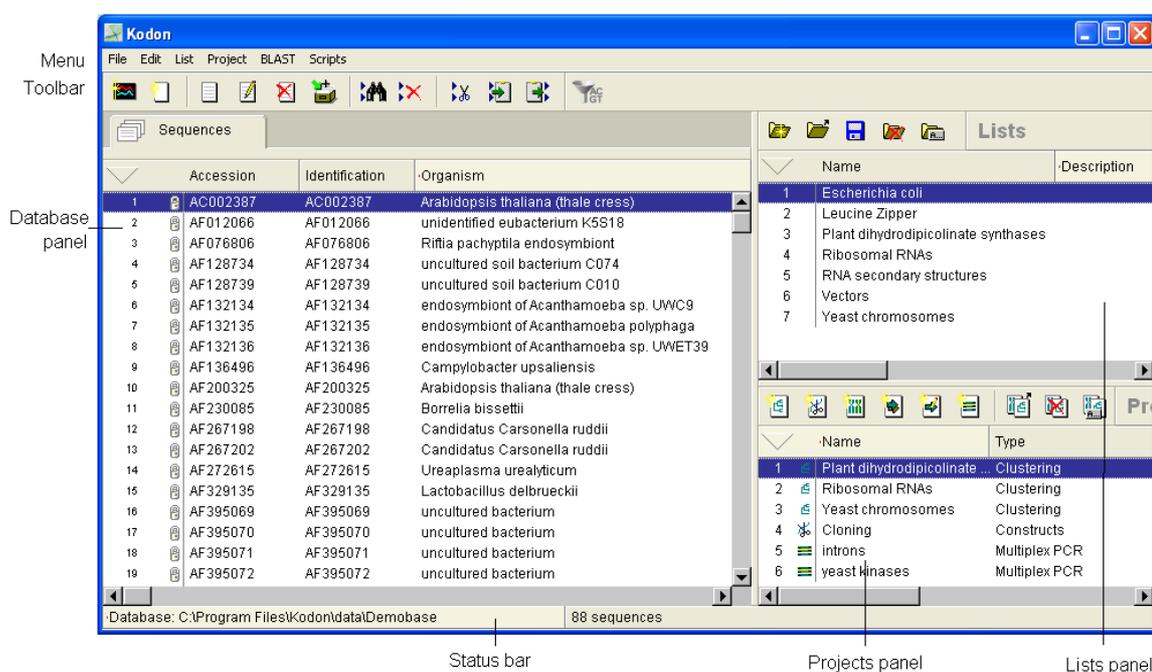


Figure 1-1. The *Main* window of the Kodon program.

The list is now sorted alphanumerically according to the field 'Organism'.

1.2.8 Select the first *Arabidopsis* sequence (CTRL + left-click), press the SHIFT key, and select the last *Arabidopsis* sequence.

All *Arabidopsis* sequences should now be marked with a blue arrow.

1.2.9 Select *List > Create new* or press .

1.2.10 Name the list 'Arabidopsis' and click <OK>.

The newly created list is listed in the upper right-hand panel.

1.2.11 Press the  button to save the list and click on the  button in the caption of the list tab to close the list.

1.2.12 Unselect the current selection by pressing F4.

1.3 Database querying

1.3.1 Select *Edit > Search* or  to call the *Query tool* window.

Three different types of query components can be queried: database fields, sequence headers, and subsequences.

• Database field query

1.3.2 In the *Query tool* window, click on <Database field> to open the *Database field search* window.

1.3.3 Type 'Arabidopsis' as the query and select 'Organism' as field (see Figure 1-2). Click <OK>.

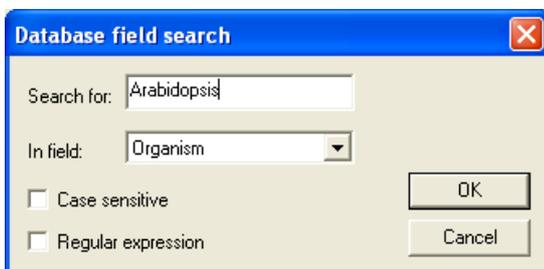


Figure 1-2. Database field search window.

A query component appears in the right panel, stating "Database field: Search 'Arabidopsis' in field 'Organism'".

1.3.4 Click <OK> to execute the query.

Verify that the *Arabidopsis* sequences in the database have been selected.

• Header query

1.3.5 Select *Edit > Search* or  to call the *Query tool* window again.

1.3.6 Clear the query panel by clicking <New> and press <OK> to confirm.

1.3.7 Click on <Sequence header> to open the *Sequence header search* window. This tool allows you to search for descriptive information that is contained in the GeneBank or EMBL sequence information header.

1.3.8 Type 'Vauterin' to search for sequences associated with this researcher and press <OK> (see Figure 1-3).



Figure 1-3. Sequence header search window.

1.3.9 Click <OK> to execute the query.

1.3.10 Verify that several sequences in the database have been selected.

• Subsequence query

1.3.11 Select *Edit > Search* or  to call the *Query tool* window again.

1.3.12 Clear the query panel by clicking <New> and press <OK> to confirm.

1.3.13 Click on <Subsequence> to open the *Subsequence search* window.

1.3.14 Type 'gcggtataac' in the input field, uncheck *Allow gaps* and set the mismatches to 0 (see Figure 1-4). Press <OK>.

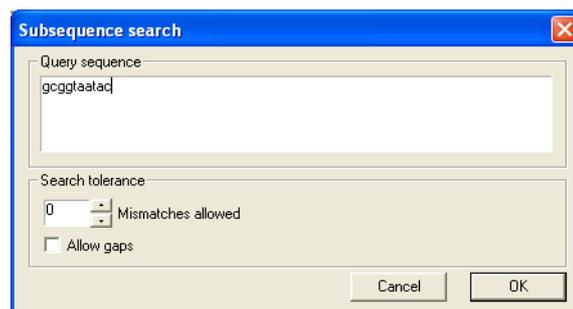


Figure 1-4. Subsequence search window.

1.3.15 Click **<OK>** to execute the query.

Verify that several sequences have been selected. Experiment with different mismatch tolerances and verify that a higher search tolerance results in more sequences being selected.

• **AND composite query**

The logical operator **AND** combines two components.

1.3.16 Open the *Query tool* window again with .

1.3.17 Clear the query panel by clicking **<New>** and press **<OK>** to confirm.

1.3.18 Create two new query components: a **Database field** containing 'Arabidopsis' in the field 'Organism' and a **Sequence header** containing 'Vauterin' (see Figure 1-5).

1.3.19 Select both components by click-dragging a box around them or CTRL-clicking on each of them.

1.3.20 Click on  to create a composite query (see Figure 1-5).

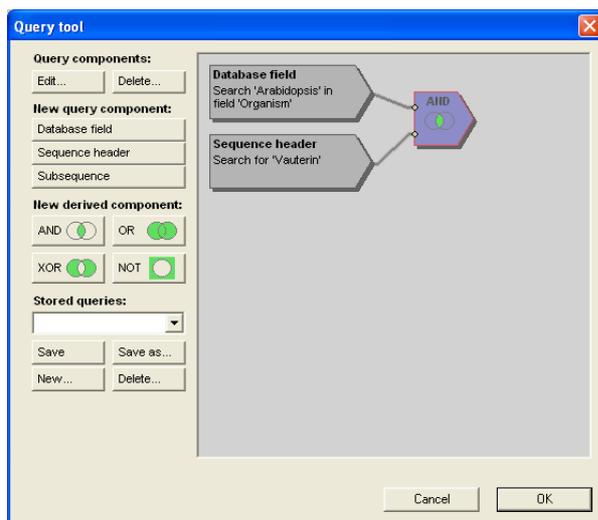


Figure 1-5. A composite query.

1.3.21 Click **<OK>** to execute the query.

Fewer sequences are selected than would have been selected if either component were queried separately.

• **OR composite query**

This operator is used to create a query that requires at least one component.

1.3.22 Open the *Query tool* window and delete the component **AND** by clicking on it and then clicking on **<Delete>** under 'Query components' (top of the

window). Press **<OK>** to confirm that you want to delete the selected component.

1.3.23 Select both components and click on the

 button.

1.3.24 Click **<OK>** to execute the query.

All sequences that would have been selected by querying the components separately are now selected.

• **NOT composite query**

This operator is used to select all sequences that do not match a component.

1.3.25 Return to the *Query tool* window and click on the **OR** component from the previous query.

1.3.26 Click on . The **NOT** component has been added after the **OR** component.

1.3.27 Click **<OK>** to execute the query.

The resulting selected sequences are all sequences that were not selected in the previous query.

• **XOR composite query**

This operator is used to select sequences that match *only one* of the components.

1.3.28 Return to the *Query tool* window and delete all but the first two query components.

1.3.29 Select the two components and click on

.

1.3.30 Click **<OK>** to execute the query.

The sequences that match both components are no longer selected as they would be if **AND** or **OR** were used.

New and derived components can be mixed together in an infinite variety of ways.

1.3.31 In the *Main* window, press F4 to unselect the current selection.

1.4 Downloading sequences directly from the internet

Kodon can import sequences directly from the Internet into the database.

1.4.1 Select *File > Import sequences via internet connection*. The Kodon HTML document viewer opens.

1.4.2 Navigate to the NCBI website: <http://www.ncbi.nlm.nih.gov>.

*NOTE: You can add this page to your list of favorites by selecting **Favorites > Add to list**.*

1.4.3 Search the Nucleotide database for AB022784 (see Figure 1-6) and press <Go>.

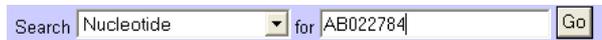


Figure 1-6. Search the nucleotide database.

1.4.4 Click on the AB002784 sequence in the next window.

1.4.5 In the window that pops up, set the visualization mode to 'Text' (see Figure 1-7).



Figure 1-7. Visualize the sequence as a text-formatted HTML document.

1.4.6 In the next window, select  or **File > Import sequence**.

A message box appears informing the user that one sequence has been detected in the HTML document.

1.4.7 Click <OK> to confirm the download action.

1.4.8 Close the *Kodon Browser* window.

1.4.9 In the **Demobase**, scroll down the list of entries. The imported sequence is listed at the bottom of the list and is marked with a blue arrow (see Figure 1-8).

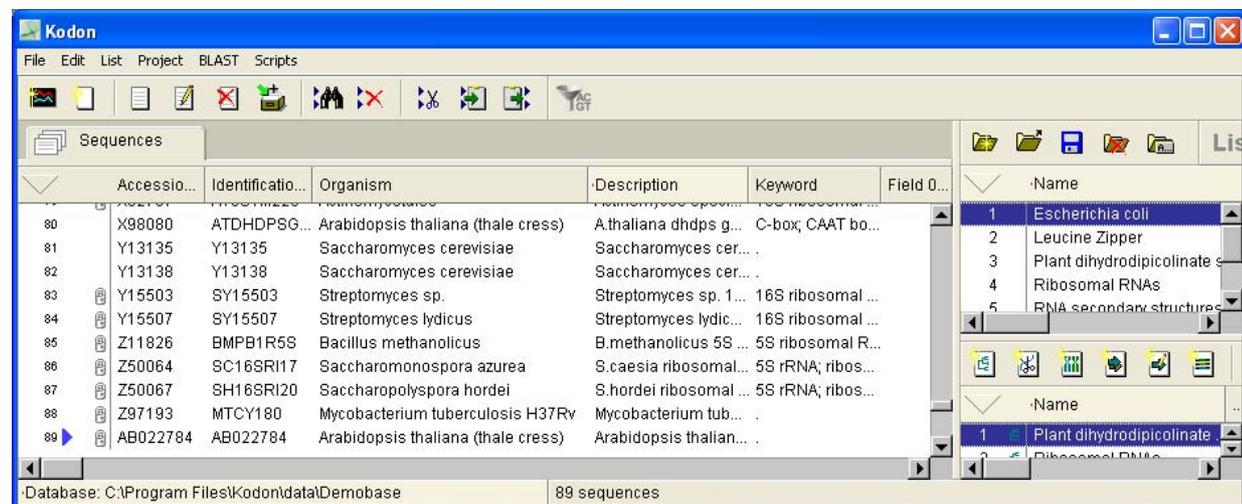


Figure 1-8. Demobase with imported sequence AB022784.

1.5 Importing sequences from downloaded files

Kodon can also import sequences from a folder containing downloaded sequences.

1.5.1 Using your regular web browser, navigate to the NCBI website (<http://www.ncbi.nlm.nih.gov>).

1.5.2 Search the Nucleotide database for AY429384 (see 1.4.3).

1.5.3 Select the AY429384 sequence in the next window.

1.5.4 In the window that pops up, select 'File' as destination (see Figure 1-9).

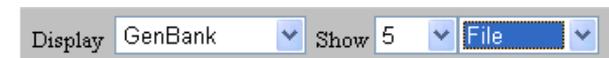


Figure 1-9. Select File as destination.

1.5.5 Save the file in the import folder of the Kodon program (default: `C:\Program Files\Kodon\import`). Rename the file 'AY429384.gb'.

1.5.6 Once the file has been saved in the import folder, select **File > Import sequences from downloaded files** or

press  in the *Main* window of Kodon.

1.5.7 If the file does not appear in the file list, double-click on '..' in the directory list, then double-click on 'import'.

1.5.8 A confirmation dialog box appears asking if you want to create updated map files of the files that are present in the import folder. Map files allow fast searching. Answer <Yes to all>.

1.5.9 The *External sequence file query* window opens (see Figure 1-10).



Figure 1-11. Only one sequence is present in our file.

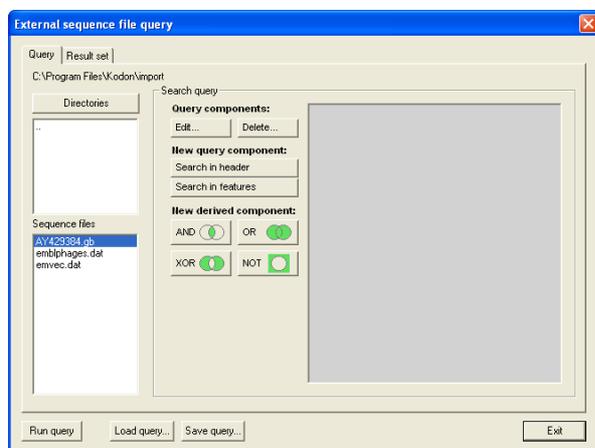


Figure 1-10. The *External sequence file query* window.

1.5.10 Select the file 'AY429384.gb' and press **<Run query>** to create a list of all sequences in the file. In our case, only one sequence is listed (see Figure 1-11).

NOTE: If more than one sequence is listed, a query can be created to narrow down the search.

1.5.11 Click on **<Import>** and confirm the import. Press **<Exit>**.

In the **Demobase**, scroll down the list of entries. The sequence is listed at the bottom of the list and is marked by a blue arrow, indicating that it is currently selected.

1.6 Assembling sequences with Assembler

Assembler is a plugin tool to assemble contig sequences from partial sequences resulting from sequencing experiments. The program accepts flat text files as well as binary chromatogram files (ABI, Beckman, and Amersham).

• Import sequences files

1.6.1 In the *Main* window, click the  button or select **File > Create new contig assembly**.

1.6.2 Name the sequence 'ICMP' and press **<OK>**.

1.6.3 In the *Assembler* window, select **File > Import sequence files** or click on .

1.6.4 Select all files in the folder 'seqassem' (default path: **C:\Program Files\Kodon\seqassem**). To select all files, select the first file, press **SHIFT** and select the last file.

1.6.5 Press **<Open>**.

The six sequences are shown in the *Assembler* window (see Figure 1-12). The top panel shows a graphical view of the sequences. Bad parts that are automatically trimmed from the sequences are underlined with a black bar. Gray underlined regions are ignored during assembly. Unknown bases (ambiguous positions) are indicated with a dark red flag on top of the sequence. Regions are color-coded based upon the quality of the raw curves, with red indicating poor quality and orange, yellow and green indicating increasing quality. The lower panel shows the raw chromatogram and the individual bases for the selected sequence.

• Trimming and quality assignment

1.6.6 The quality assignment and trimming parameters can be changed by selecting **File > Quality assignment** or by selecting .

1.6.7 In this example, change the settings slightly: under **Curve quality parameters**: *Sliding window size* 5; *Minimum good/bad peak ratio* 1.30; *Minimum short/long distance ratio* 0.60; under **Base calling quality parameters**: *Sliding window size* 51; *Minimum resolved* 30. The other settings can remain unchanged.

NOTE: Detailed information about the functionality of each of these parameters can be found in chapter 6 of the manual.

1.6.8 Automatic cleanup happens after pressing the **<OK>** button.

NOTE: You can manually activate () or inactivate () selected regions in the lower panel. To trim the sequences manually use the  and  buttons. If there are vectors in the sequences, click on the .

• Assembly

Once the sequences are trimmed and their quality is assigned, they can be assembled.

1.6.9 Select **File > Assemble sequences** or click on .



Figure 1-12. The *Assembler* main window.

1.6.10 In this exercise, keep the default settings and press **<OK>** to calculate the assembly.

NOTE: Detailed information about the functionality of each of these parameters can be found in chapter 6 of the manual.

After assembling the sequences, the second view is shown. The upper panel view displays the aligned trace sequences. The bottom panel displays the chromatogram file of the selected trace sequence and central panel shows the consensus sequence and the individual trace sequences.

1.6.11 Press the  button or select **View > Show aligned sequences**.

In this view, multiple trace chromatograms are shown and are aligned to each other and to the consensus.

1.6.12 Move the scroll bar to obtain a good view of the chromatograms.

1.6.13 Unresolved positions on the consensus are indicated in pink. Problem positions on individual trace sequences, which have been solved under the current settings are indicated in orange.

*NOTE: Select **Assembly > Consensus determination** to change the current settings.*

1.6.14 Select **File > Save** or press  and close *Assembler* and the *Sequence editor* window of the ICMP sequence.

1.7 Sequence editing

1.7.1 In the *Main* window of **Demobase**, double-click on the U72354 sequence to open its *Sequence editor* window.

The *Sequence editor* window is divided in three panels: **Sequence Editing**, **Graphical Presentation** and **Text Presentation**. You can navigate from one view to another by clicking on the tabs at the top of the panel.

1.7.2 If the padlock symbol (top left of the **Sequence editing** view) is green-checked, the sequence is locked and cannot be changed.

1.7.3 Unlock the sequence by pressing the padlock symbol. Confirm that you want to unlock the sequence.

• **Sequence editing view**

The long panel in the middle displays a graphical view of the sequence, with coding sequence (CDS) genes indicated as arrows.

1.7.4 Click and hold down the left mouse button on a CDS. A menu pops up listing its name and some descriptive information (see Figure 1-13).

1.7.5 Select **Edit > Find** or press .

1.7.6 In the next window, enter the sequence "ggatct" in the right input field and press  **Find**. Three exact matches are found (see Figure 1-14).

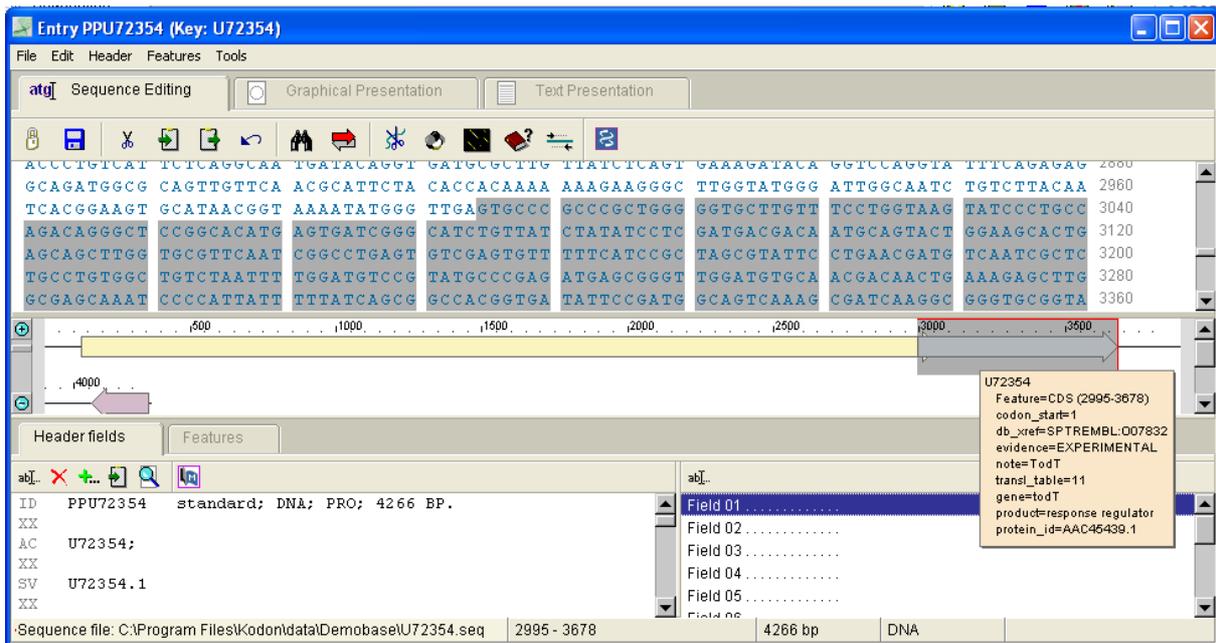


Figure 1-13. The *Sequence editor* window.

1.7.7 Raise the number of *Mismatches allowed* to 1 and press the **Find** button again. More matches are found.

1.7.8 Double-click on any matching region. The corresponding subsequence is selected in the top panel of the **Sequence Editing** view.

1.7.9 To close the *Subsequence search* window press **Close**.

1.7.10 Click on the **Features** tab in the lower left panel to view a list of the sequence's annotated features (see Figure 1-15).

1.7.11 Select one of the CDS features from the list. The qualifiers associated with this feature are listed in the right panel. The subsequence associated with this feature is selected in the upper panel.

1.7.12 Click on the button on the right side of the **Features** tab and select '/product=' from the list.

A new column appears, showing the product names for the three CDS.

• **Graphical presentation view**

1.7.13 Select the **Graphical Presentation** panel.

1.7.14 The sequence is shown schematically, with features indicated as blocks on a line. By default only CDS features are shown.

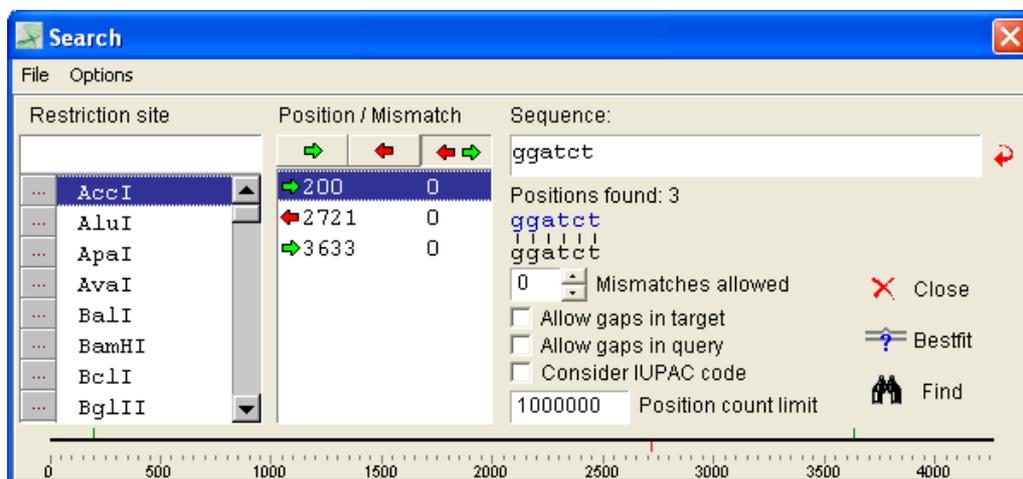


Figure 1-14. The *Subsequence search* window.

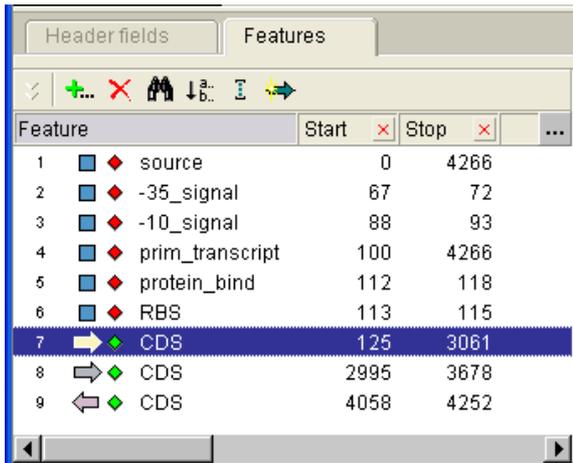


Figure 1-15. The Features panel in the sequence editor.

1.7.15 Click and hold down the left mouse button on a CDS to cause a pop-up box to appear, listing the name and all qualifiers for the selected gene.

1.7.16 Click on **FTS** to open the *Feature toolbox* window.

1.7.17 Select a CDS feature from the list. The feature becomes selected in the **Graphical Presentation** panel (red circle in the center).

1.7.18 Click on **?** to select all features equivalent to the selected feature.

1.7.19 Select the '-35_signal' from the list and press the **◆** button (double clicking on this feature does the same). The feature is shown on the image.

1.7.20 To hide the feature again, press the **◆** button.

NOTES:

1) To change size, position, or appearance of a feature, select the feature and click on the icon representing the desired effect:



2) To change the appearance of the entire figure play around with these buttons:



1.7.21 Close the *Feature toolbox* window.

1.7.22 Click on **ENZ** to open the *Restriction enzyme toolbox* window.

1.7.23 Double-click on an enzyme (e.g. AluI). The enzyme is listed below and its restriction sites are mapped on the sequence.

1.7.24 Click on **1834** to hide the site positions on the map.

1.7.25 Click on the button **...** next to the AluI enzyme. The *Restriction enzyme identification card* window is shown.

1.7.26 Close the *Restriction enzyme identification card* and *toolbox* window.

1.7.27 Select *File > Save* or click on **Save** to save the map with the changed layout.

• Text presentation view

The **Text Presentation view** provides a publication-ready image of the sequence in text form. The same features that are visible in the graphical view are also visible in the text view.

1.7.28 Select the **Text Presentation** panel.

1.7.29 Select *Text > Font size* or click on **a**. Change the size of the font to e.g. 10 and press <OK>.

1.7.30 Select *File > Save* or click on **Save** to save the image with the changed layout.

1.7.31 Close the *Sequence editor* window with *File > Exit*.

1.8 Restriction enzyme analysis

1.8.1 In the *Main* window of Kodon, open the sequence X98080 by double-clicking on it.

1.8.2 Select *Tools > Restriction enzyme analysis* or press



1.8.3 Double-click on the enzyme AluI. The cleavage sites of this enzyme are displayed in the upper right panel.

1.8.4 Double-click on the enzyme AluI in the upper right panel. The bottom panel now lists the restriction fragments with their positions and number of basepairs.

1.8.5 Repeat the two previous steps for the enzyme BclI.

1.8.6 The bottom panel now lists the fragments generated by the combination of the two enzymes AluI and BclI.

1.8.7 Select a fragment in the bottom panel. The fragment will appear in blue in the schematic.

Next, we are going to select enzymes having specific properties and a specific number of cleavage sites in the sequence.

1.8.8 Select *Settings > Filter* or press .

1.8.9 We are going to filter out all enzymes that generate blunt ends, that cut minimally 1 time and maximally 4 times: under *Cleavage type*, check *blunt end* only; under *Cleavage sites allowed*, set *Minimal* to 1 and *Maximal* to 4; leave the other settings unchanged.

1.8.10 Press <OK> to apply the settings and click  or *Enzyme > Add all*.

1.8.11 Select *List > Arrange by number of sites* or press . The list is sorted by the number of sites.

1.8.12 Select *List > Arrange by site positions* or press  to sort the list by site position.

1.8.13 Select one of the enzymes from the list, e.g. HaeIII and select *List > Add item to gel image* or press .

The predicted gel pattern of the enzyme is displayed in the *Gel image* window.

NOTE: You can add more enzymes to construct a virtual gel by repeating the previous actions for all the other enzymes you want to include in your gel.

1.8.14 Close the *Gel image* window and the *Restriction Enzyme Mapping* window.

1.9 Homology search

• Homology search (FASTA-based)

1.9.1 In the *Sequence editor* window of sequence X98080 select *Tools > Nucleic Acid Homology search* or press



1.9.2 Click on <Advanced settings>. In the window that pops up, you can adjust the parameters of the different steps of the homology search tool. Leave the default settings and press <OK>.

1.9.3 In the *Homology search* window, leave the settings unaltered and press <OK>.

NOTE: More information on these parameters can be found in the manual.

The homology search algorithm in Kodon is based on FASTA. After calculation, the *Homology search* window appears. The top panel shows schematic diagrams of significant regions of homology, which are shown in orange. The results are ranked by a *Distance calculation*. The lower the value the better the match, with zero being a perfect match (see Figure 1-16).

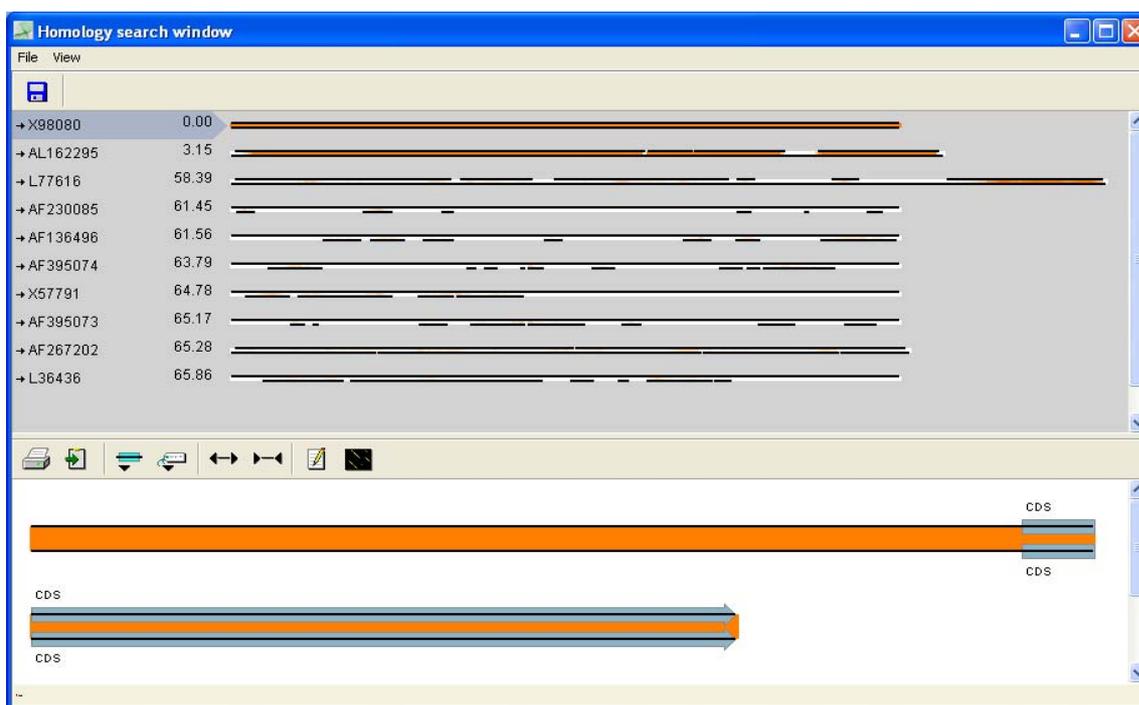


Figure 1-16. Homology search window.

1.9.4 Press the  button and select 'CAAT_signal' from the list.

1.9.5 Press the  button and select the 'product' qualifier.

1.9.6 Select *View* > *Show sequence* to view the text sequence.

1.9.7 Close the *Homology search* window.

NOTES:

1) To perform a homology search at protein level, select **Tools** > **Amino acid homology search**.

2) For a detailed study of the matches between two sequences, see paragraph 3.4 in this guide.

• **Web-based search (BLAST)**

1.9.8 In the *Sequence editor* window of the sequence X98080, select the **Features** tab and select the CDS feature from the list.

1.9.9 Select *Scripts* > *NCBIBlastn* in the *Main* database window, or click on  in the upper toolbar of the *Sequence editing* view.

1.9.10 The *Kodon Browser* pops up with the message 'Waiting for response...'.

1.9.11 Once the calculation is finished, the results are displayed in the *Kodon Browser*.

1.9.12 The results can be copied to the clipboard with .

1.9.13 Close the *Kodon Browser* and the *Sequence editor* window.

• **BLAST against Kodon databases**

In Kodon, you can make your own BLAST database from a selection of database entries. Once a nucleic acid and/or protein BLAST database is created, you can select sequences to match against the created BLAST database.

Information on how to create a BLAST database and performing a BLAST against this database can be found in the manual in chapter 7.



BLASTN 2.2.15 [Oct-15-2006]

Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, 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.

RID: 1168525498-17339-56512691542.BLASTQ1

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, environmental samples or phase 0, 1 or 2 HTGS sequences)
4,743,822 sequences; 18,918,530,559 total letters

If you have any problems or questions with the results of this search please refer to the [BLAST FAQs](#)
[Taxonomy reports](#)

Query=
Length=747

[Distance tree of results](#) NEW

Sequences producing significant alignments:	Score (Bits)	E Value
emb X98080.1 ATDHDPSGN A.thaliana dhdps gene	1481	0.0
emb X72971.1 ATDHDS Arabidopsis thaliana mRNA for dihydrodipi...	773	0.0
ref NM_180399.2 Arabidopsis thaliana dihydrodipicolinate syn...	771	0.0
gb AY096442.1 Arabidopsis thaliana putative dihydrodipicolin...	771	0.0
gb AY063943.1 Arabidopsis thaliana putative dihydrodipicolin...	771	0.0

Figure 1-17. Results of a Web-based BLAST.

2. Multiple Alignment, Clustering & Phylogeny

2.1 Sequence selection from feature matrix

2.1.1 In the *Main Demobase* window, double-click on the clustering project 'Plant dihydrodipicolinate synthases' (bottom right panel, see Figure 1-1).

NOTE: A detailed description on how to create a project can be found in the manual.

In the *Project* window, the sequences are listed vertically as columns, whereas the features are listed horizontally as rows. The number of copies of a feature is indicated in the feature matrix.

2.1.2 Click on a particular *feature instance* in the feature matrix. Its row and column are highlighted by a darker blue. Information about the selected feature is listed in the lower left panel (see Figure 2-1).

2.1.3 Double-click on a *feature instance* in the feature matrix. The instance becomes bordered by a red square.

2.1.4 Press  to unselect all.

2.1.5 Choose *View > Search subset* or press .

2.1.6 Enter 'dihydrodipicolinate' in the input field and press <OK> (see Figure 2-2).

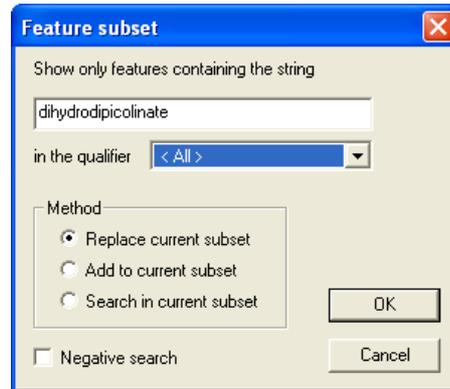


Figure 2-2. The *Feature subset* window.

2.1.7 The results of the search are listed as a separate subset (the **Subset** tab is now selected).

2.1.8 Click on the  button to select all 'dihydrodipicolinate synthase' features in the matrix (see Figure 2-3).

In a next step, we are going to create a multiple alignment of the 10 selected sequences.

Figure 2-1. The *Project* window.

All	CDS	Repeat regions																
Source	Proteins		Binding sites															
Genes	Immunoglobulin		DNA structures															
Feature	product																	
CDS	dihydrodipicolinate																	
CDS	dihydrodipicolinate synthase																	
CDS	dihydrodipicolinate synthase 2																	
CDS	dihydrodipicolinate synthase precursor																	
CDS	putative dihydrodipicolinate synthase																	

Figure 2-3. All 10 'dihydrodipicolinate synthases' features are selected in the feature matrix.

2.2 Multiple alignment

2.2.1 Click on the **Alignment** tab in the *Project* window.

2.2.2 A window appears asking if you want to update the alignment. Click **<No>**.

In the **Alignment** tab, the unaligned sequences and the information fields shown.

2.2.3 Select **Analysis > Align all** or press  to start the multiple alignment. Leave the default settings in the next window and press **<OK>**.

NOTE: A detailed description of the parameters can be found in the manual.

When the calculations are done, the sequences are aligned.

2.2.4 Use the scroll bar at the bottom of the window to move back and forth along the aligned sequences.

2.2.5 Select **Edit > Use consensus as reference**. The base positions are shown along the top of the alignment.

2.2.6 Select **View > Show consensus blocks**. Consensus regions are highlighted (gray).

2.2.7 Select **View > Show consensus as dots**. Consensus regions are shown as dots. The bases deviating from the consensus are written in full.

2.2.8 Select **View > No consensus blocks** to restore the original view.

2.2.9 Click and drag the cursor to draw a yellow box around a block of bases. Click and hold the left mouse button down within the selection and drag the selection to the left (or right) side (see Figure 2-4).

2.2.10 Click on  to save the alignment.

2.3 Clustering and phylogeny

2.3.1 Click on the 'OS field' (see **Alignment** tab in Figure 2-4), and select **Groups > Create from selected field**.

2.3.2 Organisms with the same name, are assigned to the same color (see Figure 2-5).

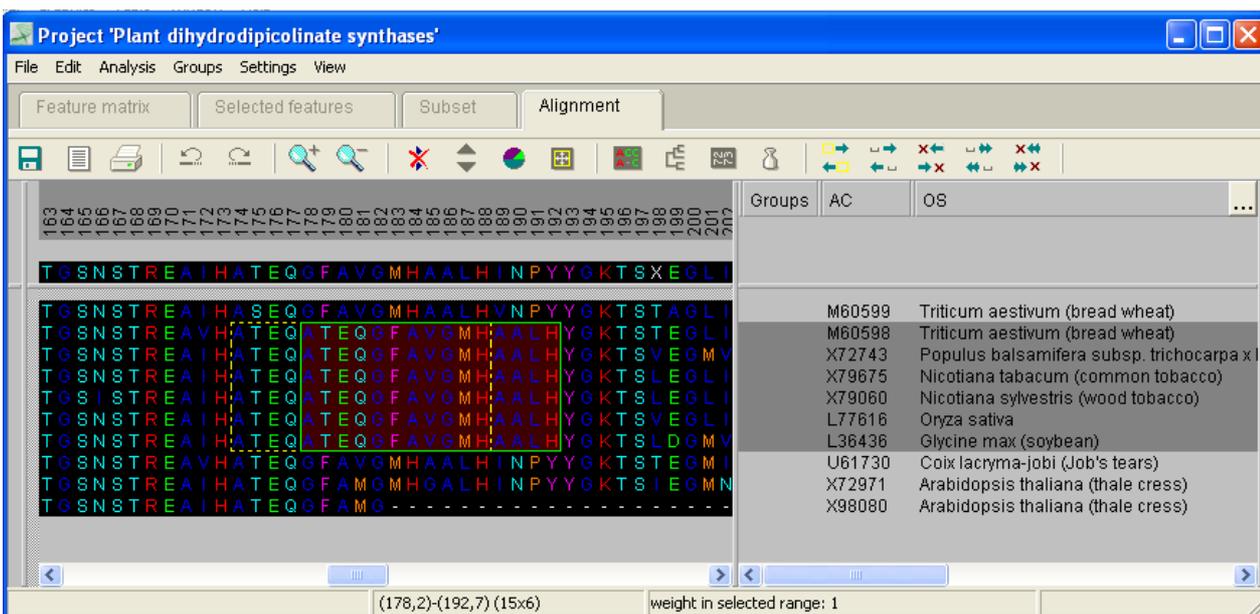


Figure 2-4. The drag-and-drop alignment tool.

NOTE: You might have to scroll to the left to see the group colors.

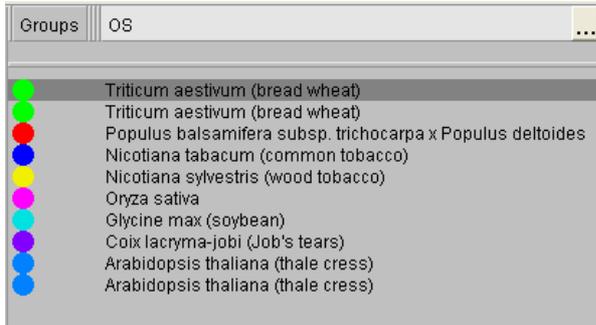


Figure 2-5. Groups based on the organisms.

2.3.3 Select *Analysis > Clustering* or press  to calculate the distance matrix and a dendrogram.

2.3.4 Select *Multiple aligned distances, UPGMA, and No correction* and click on **<OK>**.

NOTE: More information about the parameters can be found in the manual.

After calculation, the dendrogram and the matrix are shown (see Figure 2-6).

2.3.5 Select *View > Show distances as numbers* to see the pairwise distance values in the distance matrix (see Figure 2-6).

2.3.6 Select a cluster of sequences by clicking on the node where the branch of the sequences is connected to the other branches. A small white circle is placed on the node.

2.3.7 Select *Edit > Swap branches*. The branches connected at the selected node are swapped.

In phylogenetic study, it can be useful to assign different weights to different regions in a multiple alignment.

2.3.8 Select an area (preferably an area which is conserved along the sequences) using the selection rectangle (click and drag). The left and right sides of the rectangle will determine the region for which you will change the weight.

2.3.9 Select *Edit > Set weights* or press .

2.3.10 Select *Manual (base positions)* and set the weights to e.g. 3. Press **<OK>**.

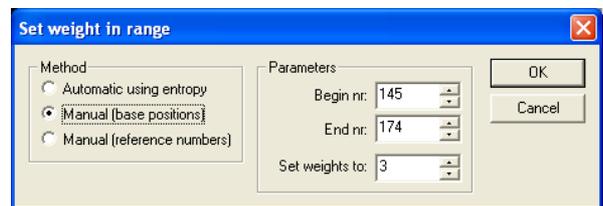


Figure 2-7. Settings the weights manually for a selected region.

Kodon assigns the weight to the selected region. You can see the weight of the currently selected position in the status bar.

2.3.11 Select *Analysis > Bootstrap analysis* to test the significance the dendrogram.

2.3.12 Set the number of bootstraps to 100 and press **<OK>** to start the bootstrap analysis.

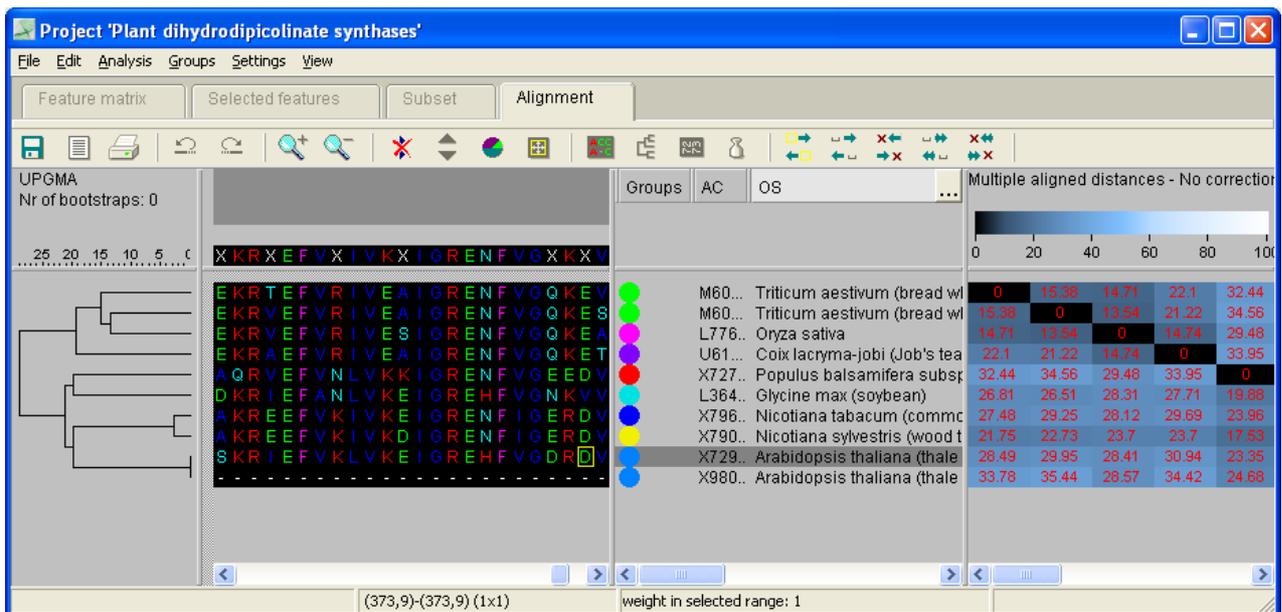


Figure 2-6. Dendrogram, multiple alignment and distance matrix.

When finished, each branch is documented with a bootstrap value, i.e. a percentage measure of its significance.

NOTE: For details on sequence alignment, please see the manual.

2.3.13 Save the *Clustering Project* window () and close it.

3. Molecular Analysis

3.1 Frame analysis

In Kodon it is possible to analyze a sequence to find open reading frames (ORF).

3.1.1 Double-click on the entry U72354 in the **Demobase**.

3.1.2 Select **Tools > Open Reading Frame analysis** or press .

The computer will notify that the 'Bacterial and Plant Plastid Code Translation Table' has been detected.

3.1.3 Choose **<OK>** to start the frame determination analysis.

The sequence plot panel (upper panel) shows a graphical view of the sequence. The three reading frames of the forward strand are mapped above the sequence plot; those of the reverse strand are mapped below (see Figure 3-1).

3.1.4 Use the zoom scroller at the left side of the plot to zoom in (up to base level), and zoom out (full-length).

In the sequence plot, open reading frame (ORF) stretches are drawn as blue boxes with black lines marking the upstream and downstream stop codons of the ORF.

3.1.5 To plot only the protein coding sequences (PCS)

press the  button. Red boxes are now shown with a black line indicating the stop codon.

3.1.6 Press the  button to remap the ORFs on the sequences.

3.1.7 Call the *Frame settings* window with **Settings > Frame settings** and change the minimal sizes for the open reading frames to 150. Press **<OK>**.

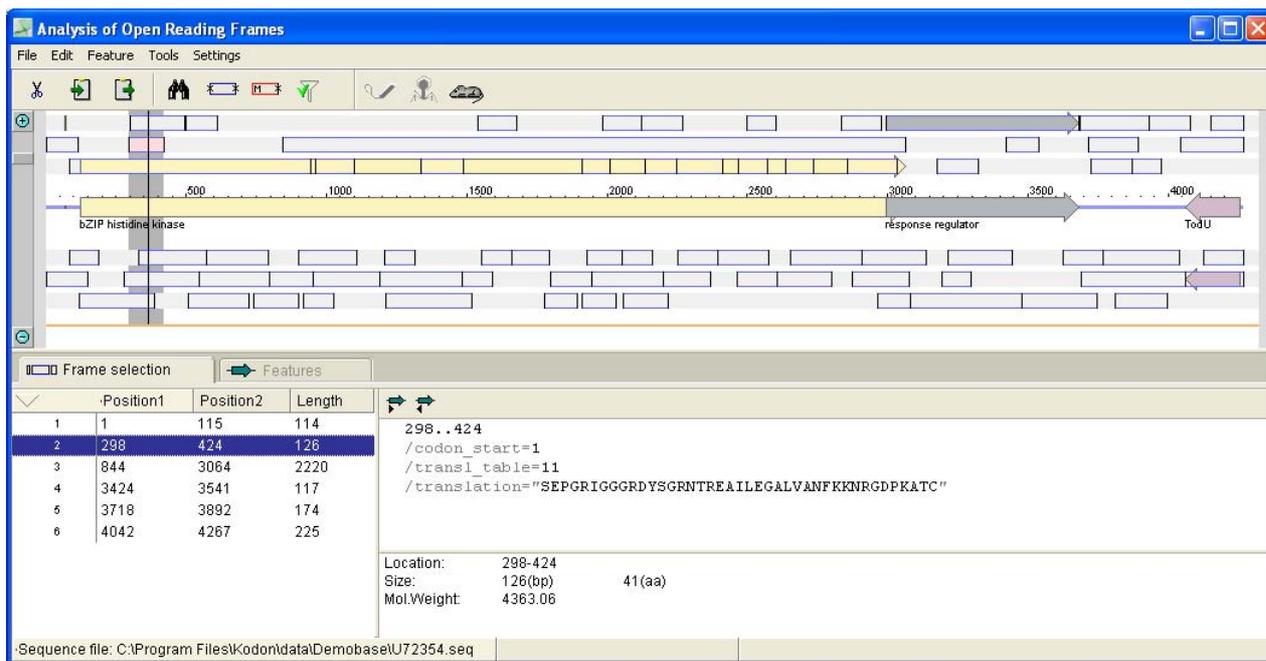
Open reading frames with a minimal number of 150 bp are mapped on the figure.

The lower panel displays the frame selection panel (**Frame selection**) and the feature listing panel (**Features**).

3.1.8 In the **Frame selection** panel, the ORFs (or PCSs) of the currently selected reading frame are listed. In the right panel, details of the currently selected ORF (or PCS) are shown.

3.1.9 Select an ORF in the sequence plot. The ORF is highlighted in pink (stop-to-stop) and the information is automatically updated in the lower panel.

3.1.10 Select the **Features** tab.



The screenshot displays the 'Analysis of Open Reading Frames' window. The top panel shows a sequence plot with ORFs mapped as blue boxes. The 'Frame selection' panel lists ORFs with their positions and lengths. The 'Features' panel shows details for the selected ORF, including its location, size, molecular weight, and translation.

	Position1	Position2	Length
1	1	115	114
2	298	424	126
3	844	3064	2220
4	3424	3541	117
5	3718	3892	174
6	4042	4267	225

298..424
 /codon_start=1
 /transl_table=11
 /translation="SEPGRIGGGRDYSGRNTREAILLEGALVANFKKINRGDPKATC"

Location: 298-424
 Size: 126(bp) 41(aa)
 Mol.Weight: 4363.06

Sequence file: C:\Program Files\Kodon\data\Demobase\U72354.seq

Figure 3-1. The *Open reading frame analysis* window.

3.1.11 Select a feature from the list (e.g. a CDS). The region is automatically selected in the sequence plot.

Further functionality and layout is exactly the same as in the *Sequence editor* window.

3.1.12 Select the **Frame selection** tab and select an ORF.

3.1.13 Select *Tools > Map selected sequence as CDS*, give the product a name, press <Add> and press <OK>.

The CDS is mapped on the sequence.

3.1.14 Close the *Open reading frame analysis* window.

3.2 Primer design

3.2.1 In the *Sequence editor* window of sequence U72354 select *Tools > Primer design*.

3.2.2 If part of the sequence is selected in the *Sequence editor* window, an *Information* window pops up. In that case, select *Use full entry sequence as template* and press <OK>.

3.2.3 In the *Primer design* window, click on  to begin the calculations of the primer locations and PCR products.

NOTE: Settings for the primer design can be changed in the *Primer design window* (*File > Settings* or press ).

The results of the primer calculations are displayed in the lower panel (see Figure 3-2).

3.2.4 In the left lower panel, select 'Pos.' in the toolbar and press the  button.

The primers are now arranged according to their positions.

3.2.5 Select a primer from the list in the left lower panel.

The description is shown in the lower right panel. The primers are displayed in blue and red in the upper panel.

3.2.6 Zoom in with the  button until the nucleotides are visible in the right panel.

The nucleotides of the primer are indicated in blue on the figure (see Figure 3-2).

3.2.7 Click on the **PCR products** tab.

The PCR products are listed in the lower left panel and the primers in the lower right panel.

3.2.8 Select a PCR product from the list (lower left panel).

The primers are displayed in blue and red in the upper panel and the PCR product is displayed in dark green.

3.2.9 Select a primer in the lower right panel and select *Primers > Secondary structure prediction* or click



The melting temperature and the secondary structure of the primer are given as a function of salt and DNA concentration, which can be optimized.

The screenshot shows the 'Primer Design' window with the following components:

- Sequence Plot:** Shows the consensus sequence for U72354. A region from position 1059 to 1100 is highlighted in red. The sequence is:


```

      consensus  aacgcattatccttaactgtctttccaatgctgtttaaatcaaccctgacggggggcttatccggttgctgtcttagtttgagtcgaccaaattatgccttggtt
      ctgcgtaataggaattgaacgaaagttacgcaaatttaagtggggactgccccccaataaggcaacgacagaatcaaacctcagctggtttaacggaaccac
      U72354      PPLaacgcattatccttaactgtctttccaatgctgtttaaatcaaccctgacggggggcttatccggttgctgtcttagtttgagtcgaccaaattatgccttggtt
      
```
- Primer Table:**

↓ Pos.	Len.	Deg.	Temp.
1059	19	1	67.5
1957	20	1	67.5
558	21	1	67.5
1587	21	1	67.5
2231	21	1	67.5
4218	19	1	67.5
833	21	1	67.5
1958	21	1	67.5
1431	21	1	67.5
4066	21	1	67.5
4067	21	1	67.5
1326	20	1	67.5
4147	21	1	67.5
4148	21	1	67.5
- Primer Details Panel:**

Primer1: (P) 5'-ccctgacggggggcttat-3'(OH)
 Position: 1059
 Primer length: 19
 Primer degeneration: 1
 Melting temperature: 67.5°C
 Minimal difference: -

Primer2: (P) 5'-ataagcccccgctcagggg-3'(OH)
 Position: 1059

Figure 3-2. The *Primer design* window.

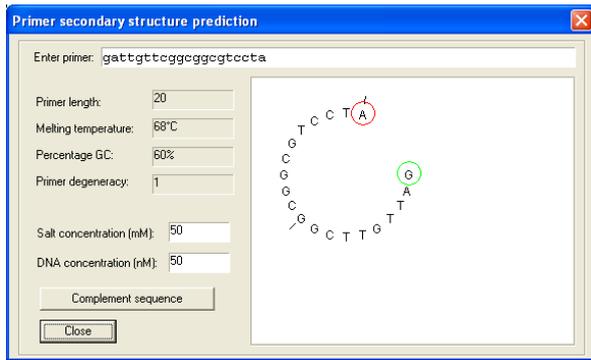


Figure 3-3. Primer properties.

3.2.10 Press <Close> to close the window.

3.2.11 Select the **Primers** tab again.

3.2.12 Click-drag along the sequence shown in the middle right panel to create a gray region. The gray region is highlighted in pink in the upper panel.

3.2.13 Select **File > Settings** again or press .

3.2.14 Check **Search primers inside selection** and press <OK>.

3.2.15 Click on  to calculate the primer locations and PCR products.

Primers (**Primers** tab) and PCR products (**PCR products**) that fall within the selection are listed.

3.2.16 Close the *Primer design* window and the *Sequence editor* window (without saving the sequence).

3.3 Multiplex PCR design

The multiplex PCR application has been designed for the creation of primer sets which will amplify multiple target regions with a minimum of reactions.

3.3.1 In the *Main* window of the **Demobase**, make sure that no entries are selected by pressing F4.

3.3.2 Open the list 'Yeast chromosomes' (upper right panel) by double-clicking on it.

3.3.3 Select the three *Saccharomyces* entries in the list (CTRL + left-click) and press  or select **Project > New multiplex primer design**.

3.3.4 Name the new project e.g. 'Saccharomyces' and press <OK>.

The *Multiplex PCR analysis* window opens. The upper panel plots the three *Saccharomyces* chromosomes contained within this project.

In a next step, we are going to search for possible amplification strategies for all mapped features encoding a kinase function located on the three *Saccharomyces* chromosomes.

NOTE: Within the multiplex application, a target for PCR amplification is called a '**locus**'.

In a first step we are going to select all features that encode a kinase.

3.3.5 Left-click in the first sequence plot. The plot is now drawn on a pink-colored background.

3.3.6 Select **Locus > Search for features** or .

3.3.7 In the *Feature toolbox* window, press the  button.

3.3.8 In the next window, select **CDS** from the list and search for the text 'kinase' (see Figure 3-4). Press <OK>.

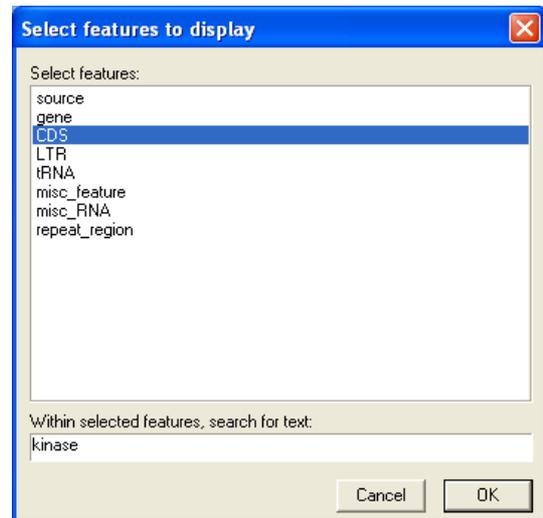


Figure 3-4. Select features to display.

In the *Feature toolbox* window, only the CDS encoding for kinase are listed.

3.3.9 In the *Feature toolbox* window, select all CDS listed (select the first CDS, hold the SHIFT-button and select the last CDS). Once all the CDS are selected, close the window.

3.3.10 The CDS that were selected in the *Feature toolbox* window, are marked with a red circle in the sequence plot (see Figure 3-5).

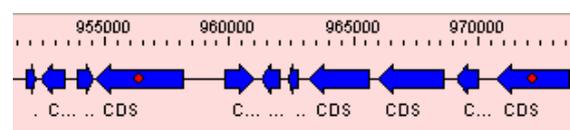


Figure 3-5. Selected CDS (red circle).

3.3.11 Select **File > Define locus prefix**, enter 'Kinase' and press **<OK>**.

Next, we are going to add a PCR locus for the selected features.

3.3.12 Select the menu item **Locus > Add** or press the



button.

3.3.13 In the next window, enable **PCR locus falls within the target sequence** and press **<OK>** (see Figure 3-6).

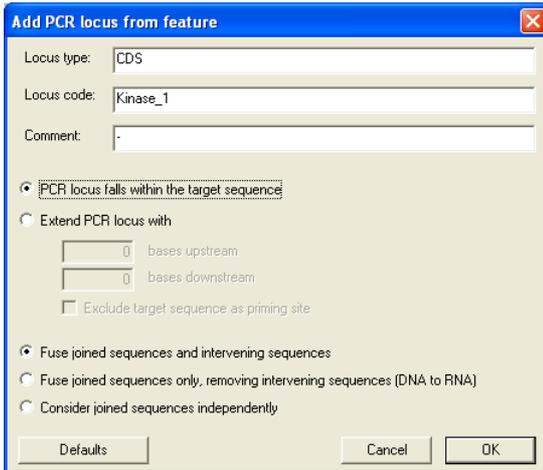


Figure 3-6. The PCR locus description dialog box.

3.3.14 Start calculating the multiplex PCR by pressing the



button.

Once the calculations are finished, the lower right panel displays the targets mapped on the sequence (**Locus** tab).

3.3.15 Select a target from the list.

The lower left panel shows a sequence detail of the currently selected target. The primers are indicated on the sequence (blue and red underlined, see Figure 3-7).

3.3.16 Click on the **Experiments** tab in the lower right panel.

The primer pairs are listed by locus and experiment.

3.3.17 Save the project and close the **Multiplex PCR** window. In the **Main** window, close the list 'Yeast chromosomes' (press the **X** button).

3.3.18 Clear the selection by pressing **F4**.

3.4 Pairwise matching and repeat analysis

• Repeat analysis

3.4.1 Open the entry AL162295 by double-clicking on it.

3.4.2 Select **Tools > Repeat analysis** or press



3.4.3 Leave the default settings in the next window and press **<OK>** to perform the calculations.

NOTE: More information about the settings can be found in the manual.

Target	Code	Entry	Comme...	Pos1	Pos2	Size	Status	
1	CDS	Kinase_7	Y13135	-	668183	669902	1719	Accepted: experiment 1
2	CDS	Kinase_6	Y13135	-	575392	576499	1107	Accepted: experiment 1
3	CDS	Kinase_0	Y13135	-	23935	25396	1461	Accepted: experiment 1
4	CDS	Kinase_4	Y13135	-	353056	355471	2415	Accepted: experiment 1
5	CDS	Kinase_9	Y13135	-	886067	887870	1803	Accepted: experiment 1
6	CDS	Kinase_2	Y13135	-	160069	162763	2694	Accepted: experiment 1
7	CDS	Kinase_5	Y13135	-	458152	458989	837	Accepted: experiment 1
8	CDS	Kinase_3	Y13135	-	207036	208575	1539	Accepted: experiment 1
9	CDS	Kinase_1	Y13135	-	69670	72547	2877	Accepted: experiment 1
10	CDS	Kinase_11	Y13135	-	970766	973730	2964	Accepted: experiment 1
11	CDS	Kinase_10	Y13135	-	954666	958203	3537	Accepted: experiment 1
12	CDS	Kinase_8	Y13135	-	872038	875104	3066	Accepted: experiment 1

Figure 3-7. The Multiplex PCR analysis window.

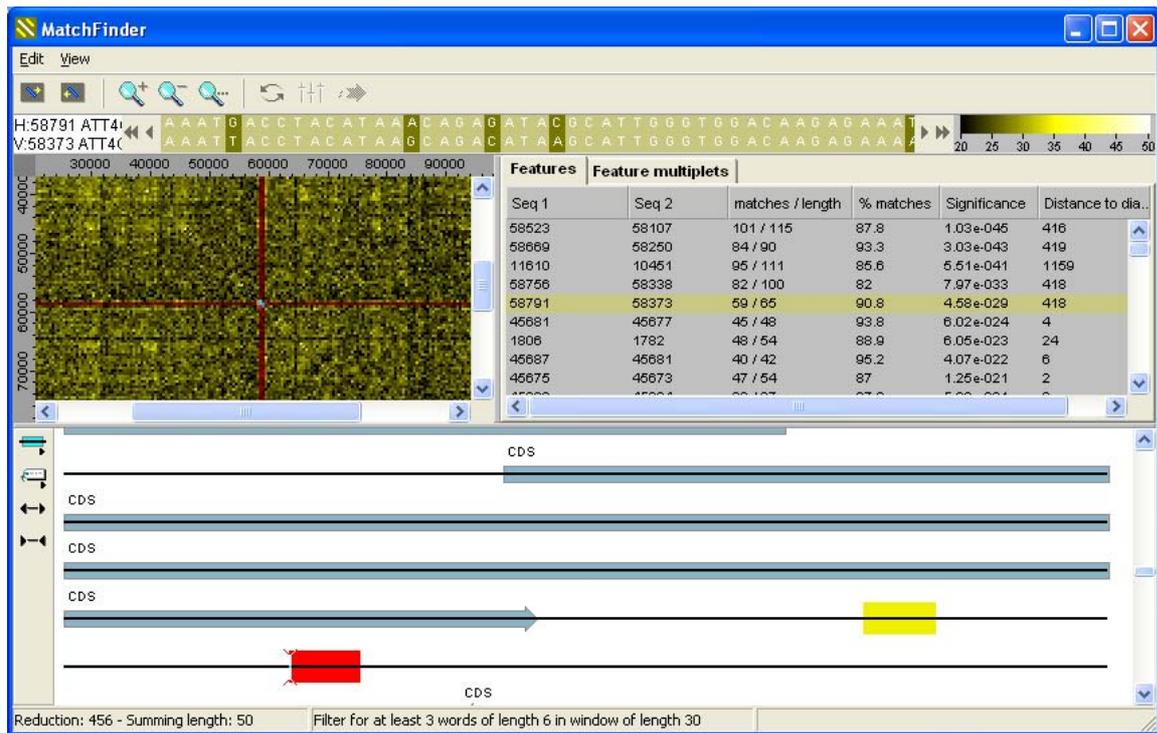


Figure 3-8. The *MatchFinder* window.

When the calculations are completed, the *MatchFinder* window pops up (see Figure 3-8). The top left panel contains a dot plot of similarity values, with higher similarity shown in lighter shades of yellow. Each block represents two positions on the sequence, one on the X-axis and one on the Y-axis. The top right panel contains a list of *Features* and *Feature multiplets*.

3.4.4 Click on a feature from the list *Features* to show its position in the dot plot on the left and the graphical map in the lower panel.

3.4.5 Select *Edit > Invert complement horizontal sequence* or select  to search for inverted repeats.

3.4.6 Click on the second view tab: *Feature multiplets*. In this view, only stretches that are repeated multiple times are listed. Non-exact matches are also included.

3.4.7 Close the *MatchFinder* window.

3.4.8 In the *Sequence editor* window select *Tools > Multiple repeat search*.

3.4.9 Leave the settings unaltered and press <OK>.

Unlike the repeat search, only exact matches are considered, thereby simplifying the results.

3.4.10 Close *Multiple exact repeats* window and the *Sequence editor* window.

• Matching analysis

Matching analysis can be launched from the result of a homology search (see chapter 1.9).

3.4.11 In the *Main* window of the database **Demobase**, double-click on sequence X98080 to open the *Sequence editor* window.

3.4.12 In the *Sequence editor* window of sequence X98080 select *Tools > Nucleic Acid Homology search* or press



3.4.13 In the *Homology search* window, leave the settings unaltered and press <OK>.

In the results window, the best matching sequences are shown together with their identity score (in %).

3.4.14 Select the best matching sequence, in this example entry AL162295.

3.4.15 To view the matches of the two sequences (X98080 and AL162295) into more detail, select .

3.4.16 Leave the settings unaltered and press <OK>.

3.4.17 The *MatchFinder* window pops up.

NOTE: Some features of the MatchFinder window and the associated settings are explained under 'Repeat analysis'. Extra features can be found in the manual.

3.4.18 Close the *MatchFinder* window, the *Homology search* window, and the *Sequence editor* window.

3.5 Motif search

3.5.1 Double-click on the *Arabidopsis* entry AL162295 in the **Demobase**.

3.5.2 Click on the **Features** tab in the lower panel and select a CDS feature from the list.

3.5.3 Press the  button or select *Tools > Motif search*.

The upper panel displays at its left side a list box listing all known patterns and conserved profiles (matrices) present in the Prosite database.

3.5.4 Select the motif 'N-glycosylation site' from the list and press the  button or press *Motif > Search item*.

If the pattern is found in the protein sequence, the pattern is listed with a short description in the result list box in the upper panel.

3.5.5 Press the  button or select the menu item *Motif > Search all*.

All patterns found in the sequence are now listed in the upper right panel.

3.5.6 Close the *Motif search* window and the *Sequence editor* window.

3.6 RNA secondary structure

3.6.1 In the **Demobase Main** window, open the list 'RNA secondary structures' in the Lists panel (upper right panel).

3.6.2 Double-click on 'tRNAala' to open the *Sequence editor* window for this entry.

3.6.3 Select *Tools > RNA folding*.

The *Secondary structure prediction* window opens. Initially, no secondary structure is calculated.

3.6.4 Press the  button or select the menu item *Edit > Set parameters*.

In the dialog box that pops up, parameters can be changed to influence base pair formation.

3.6.5 Uncheck *Allow GU pairs*, uncheck *Allow GU pairs at the end of helices* and uncheck *Allow lonely pairs*. Press <OK>.

NOTE: More information about RNA folding and the parameters can be found in the manual.

3.6.6 Select  or press *Edit > Calculate secondary structure* to start the calculation.

When the calculation is finished, a plot of the predicted secondary structure appears. The total free energy of the predicted secondary structure is written at the bottom of the window. The predicted folding is the result of minimizing the total free energy (see Figure 3-9).

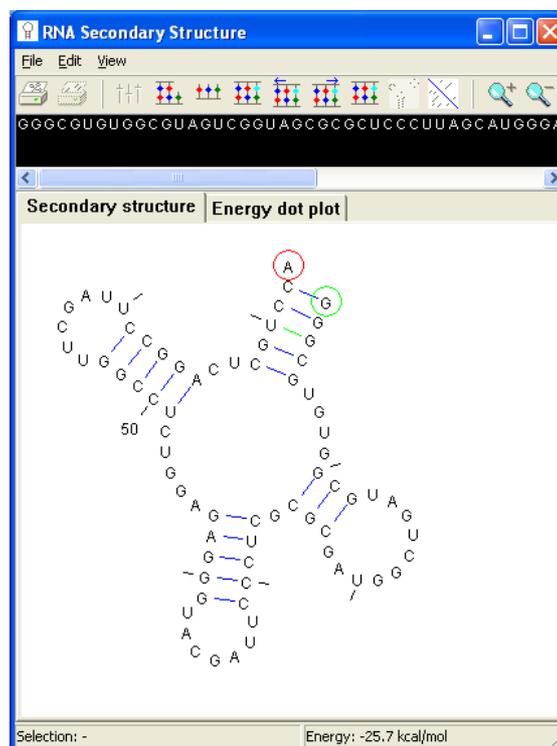


Figure 3-9. The yeast alanyl-tRNA secondary structure prediction.

3.6.7 Click on a base in the plot. Hold the SHIFT key and select another base in the plot. A stretch of bases is now selected in the sequence panel.

3.6.8 To make changes to the selected stretch, select one of the items  and

recalculate the secondary structure by pressing .

Besides the predicted secondary structure of the molecule, the program can also show an energy dot plot. More information about this feature can be found in the manual.

3.6.9 Close the *RNA secondary structure* window and the *Sequence editor* window.

3.6.10 In the *Main* window, close the 'RNA secondary structures' list (press the  button).

3.7 Protein properties

3.7.1 Double-click on the AL162295 sequence in the *Main* window of the **Demobase**.

3.7.2 Click on the **Features** tab in the lower panel and select the first CDS feature from the list.

3.7.3 Select **Tools > Protein properties** or select  to open the *Protein Structure Viewer* window.

The window contains three views: **Composition, Alpha helices, L zippers**.

3.7.4 In the **Composition** panel select **File > Add curve** or press .

3.7.5 Click on the **<Hydrophobic>** button and press **<OK>**.

The red curve in the lower right panel shows the relative contribution, as a percentage of the surrounding window, of hydrophobic amino acids (see Figure 3-10).

3.7.6 Select **File > Add curve** or press  again, select **<Hydrophilic>** and press **<OK>**.

The green curve in the lower right panel shows the relative contribution of the hydrophilic amino acids (see Figure 3-10).

3.7.7 Click on the **Alpha helices** tab.

3.7.8 Click on an amino acid sphere and drag the hand icon to flip and rotate the alpha-helical structure.

3.7.9 Select an amino acid in the top panel, and then click on  to center the selection.

3.7.10 Press the  and  buttons to shift the helix up or down.

3.7.11 Next, select the **L zippers** tab.

The middle view shows the protein sequence, on which coiled coils (green color), possible leucine repeat regions and leucine zippers are indicated (not present in this protein sequence).

3.7.12 Close the *Protein Structure Viewer* window and the *Sequence editor* window.

3.8 Vector construction

In Kodon it is possible to create a cloning project to model the insertion of a gene into a vector.

3.8.1 In the *Main Demobase* window, make sure that no entries are selected by pressing F4.

3.8.2 Select the sequences 'U25267' (Cloning vector pBluescript II KS) and 'X98080' (*Arabidopsis thaliana* DHDPS gene). To select both entries hold the CTRL-button and left-click. Only these two entries should now be marked with a blue arrow in the *Main* window.

3.8.3 Select **Project > New cloning** or press  to open a new cloning project.

3.8.4 Name the project, for example **DHDPS**, and press **<OK>**.

The upper panel shows a graphical representation of the cloning strategy (see Figure 3-11).

3.8.5 To identify useful restriction enzymes, click on the linear sequence (X98080) in the top panel and select **Cloning > Restriction digestion**.

3.8.6 In the next window (see "Restriction enzyme analysis" on page 12), add all enzymes to the list in

order to see their cut sites by pressing the  button (or select **Enzyme > Add all**).



Figure 3-10. Curves representing the relative contribution of selected amino acids in a moving window.

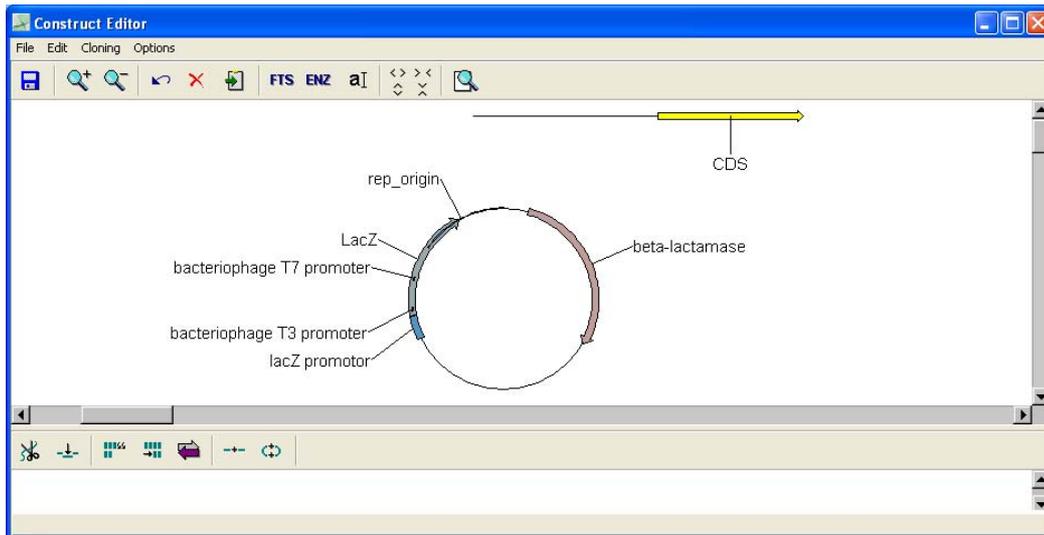


Figure 3-11. The *Construct editor* window.

3.8.7 Double-click on the yellow CDS feature in the middle panel. You may have to use the scroll bar and zoom out () to see the CDS feature.

The CDS feature is now displayed with a gray background color in the middle panel (see Figure 3-12).

3.8.8 Select *Map > Search sites close to selection (outside)* or press .

The *Combinations* window shows various combinations of restriction enzymes which are ranked by how closely they cut out the selected region.

3.8.9 Double-click on the first pair of enzymes, **NcoI** and **SacI** and close the *Combinations* window.

The fragments created by the NcoI and SacI enzymes are shown in the bottom panel (see Figure 3-12).

3.8.10 Close the *Restriction Enzyme Mapping* window.

3.8.11 In the *Construct editor* window, select the third fragment from the list (length 765 bp) in the bottom panel and click on  to release the fragment.

3.8.12 Click on the circular sequence (U25267) and select *Cloning > Restriction digestion*.

3.8.13 Add all enzymes to the list in order to see their cut sites by pressing the  button (or select *Enzyme > Add all*).

3.8.14 Scroll down the list and double-click on SacI to add it to the fragment list in the bottom panel.

3.8.15 Close the *Restriction Enzyme Mapping* window.

3.8.16 In the *Construct editor* window, select the fragment from the list (length 2979 bp) in the bottom panel and click on  to release the fragment.

3.8.17 First select the longer fragment (SacI-SacI-fragment) and then select the shorter fragment (SacI-NcoI-fragment) using SHIFT-click.

3.8.18 Click on  to ligate the fragments together at one end.

NOTE: If the fragments were selected in the opposite order, the ligation would fail.

3.8.19 Select the ligation product in the top panel and click four times on the  button to remove the sticky ends.

3.8.20 Click on  to perform a circular ligation.

3.8.21 Resize the image using the , , and  buttons.

3.8.22 Click on  to save the project, and then close the *Construct editor* window.

3.8.23 Close the **Demobase**.

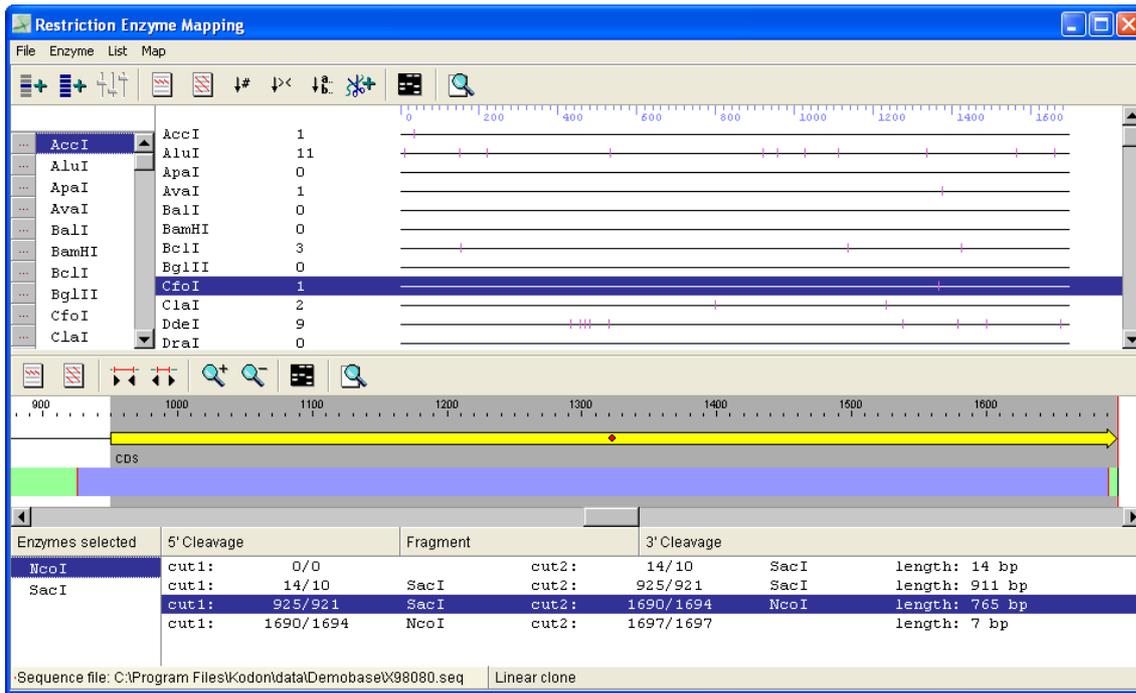


Figure 3-12. Mapping restrictions enzymes on the sequence.

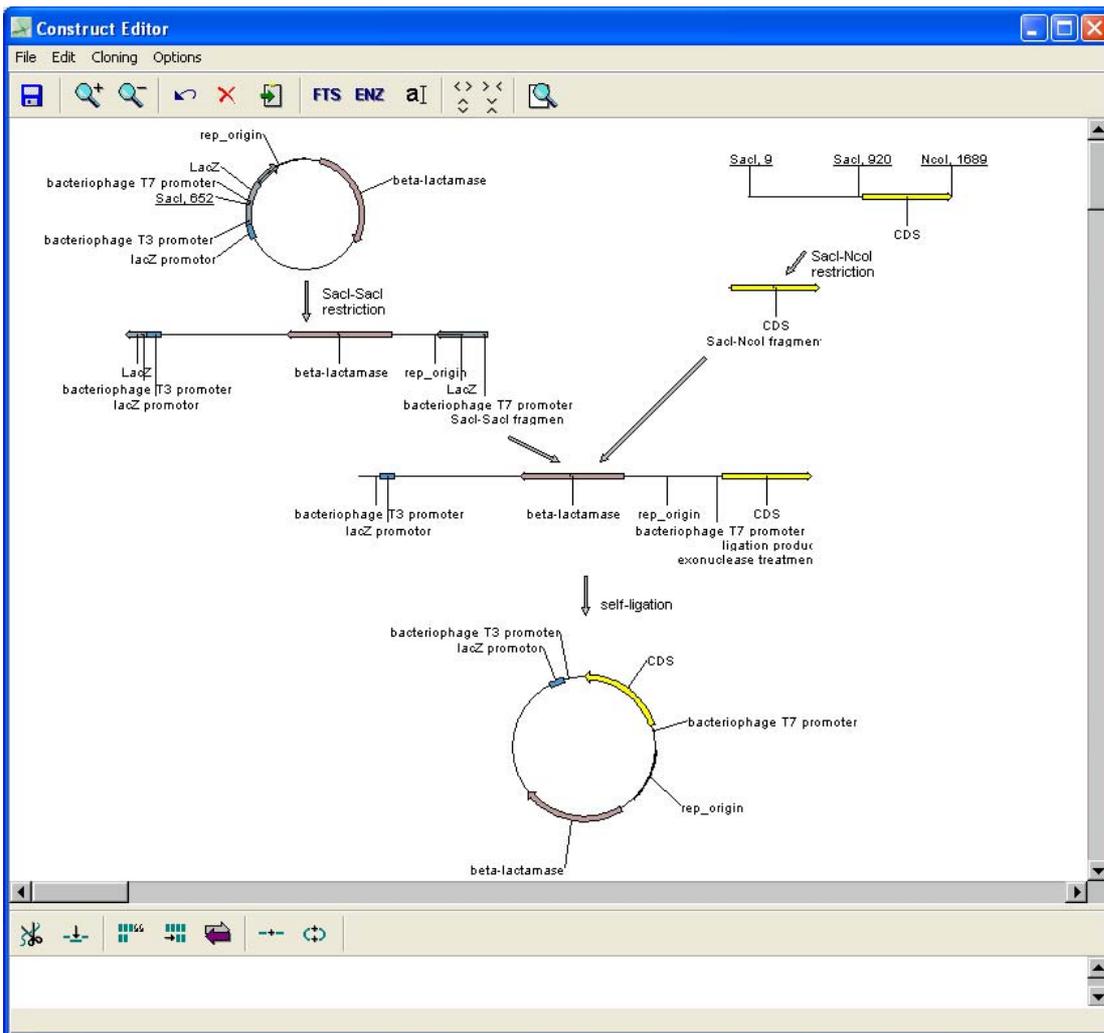


Figure 3-13. A vector construction project.

4. Chromosome Mapping

4.1 Comparative chromosome mapping

The comparative chromosome mapping functionality in Kodon allows users to cluster whole chromosomes and map homologous regions between pairs of chromosomes.

Two types of comparative chromosome mapping can be created: a first type based on the DNA sequences and a second type which only uses the information present in annotated CDS features mapped on the sequences.

4.1.1 In the Kodon startup screen, select **Bacterial chromosomes** and select **<Open database>**.

*NOTE: The **Bacterial chromosomes** database is listed in the startup screen only if you have enabled 'Install Bacterial Chromosome Database' in the installation wizard.*

4.1.2 In the *Main* window, select the following sequences (CTRL + left-click): the 2 "Salmonella" sequences (AE006468, AL513382), the "Shigella" sequence (AE005674) and 3 "Escherichia" sequences (AE005174, BA000007, U00096).

4.1.3 Once these six sequences are selected in the database, select **Project > New DNA comparative**

chromosome mapping or press .

4.1.4 Name the new project e.g. 'Chromosome clustering' and press **<OK>**.

NOTE: The creation of a CDS-based comparative chromosome mapping project can be found in the manual.

• The Matrix view

The *Comparative chromosome mapping* window displays the selected sequences as rows and columns of a matrix representing all possible combinations of the pairwise comparisons.

4.1.5 Call the search settings by pressing **File > Matrix and stretch calculation settings**.

These settings will direct the calculation of the pairwise comparisons.

4.1.6 For the current comparative chromosome mapping project, the default values can be accepted. Press **<OK>**.

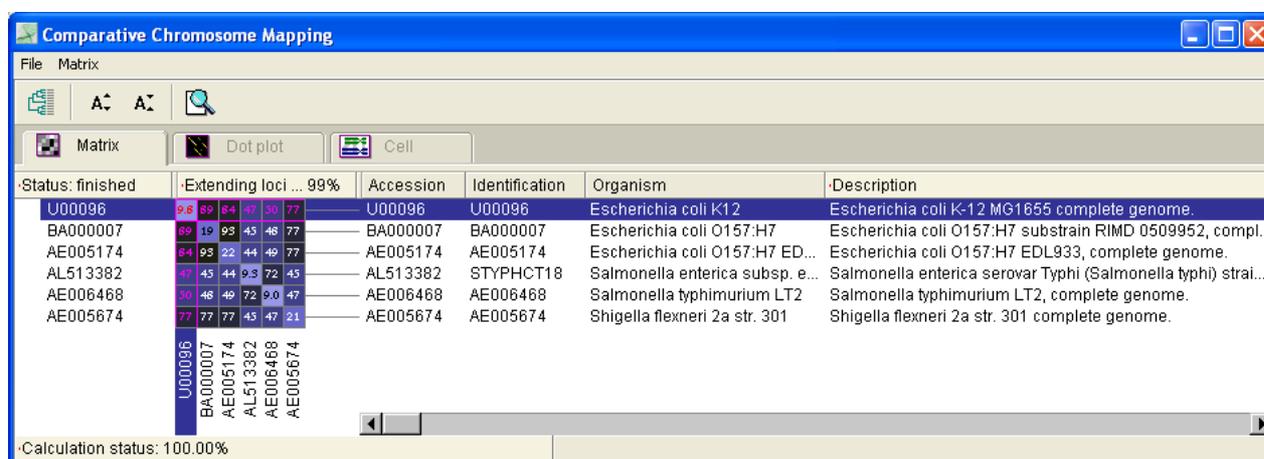
NOTE: More information about these parameters can be found in the manual.

4.1.7 To start the calculation, press the  button or select **File > Run calculation**.

4.1.8 The button will change in a  button, and the status bar indicates in red "**Status: running**".

4.1.9 If a cell is fully calculated, it will be depicted in green.

4.1.10 After calculation (this may take a couple of minutes), the status bar indicates: "**Status: finished**", two new tabs appear and two new buttons. The cells in the matrix display an identity score ranging from 0-100% (see Figure 4-1).



.Status: finished	.Extending loci ... 99%	Accession	Identification	Organism	.Description
U00096	95 89 64 47 30 77	U00096	U00096	Escherichia coli K12	Escherichia coli K-12 MG1655 complete genome.
BA000007	89 19 93 45 48 77	BA000007	BA000007	Escherichia coli O157:H7	Escherichia coli O157:H7 substrain RIMD 0509952, compl...
AE005174	64 93 22 44 49 77	AE005174	AE005174	Escherichia coli O157:H7 ED...	Escherichia coli O157:H7 EDL933, complete genome.
AL513382	47 45 44 93 72 45	AL513382	STYPHCT18	Salmonella enterica subsp. e...	Salmonella enterica serovar Typhi (Salmonella typhi) strai...
AE006468	30 48 49 72 90 47	AE006468	AE006468	Salmonella typhimurium LT2	Salmonella typhimurium LT2, complete genome.
AE005674	77 77 77 45 47 21	AE005674	AE005674	Shigella flexneri 2a str. 301	Shigella flexneri 2a str. 301 complete genome.

Figure 4-1. The *Comparative Chromosome Mapping* window: Matrix view.

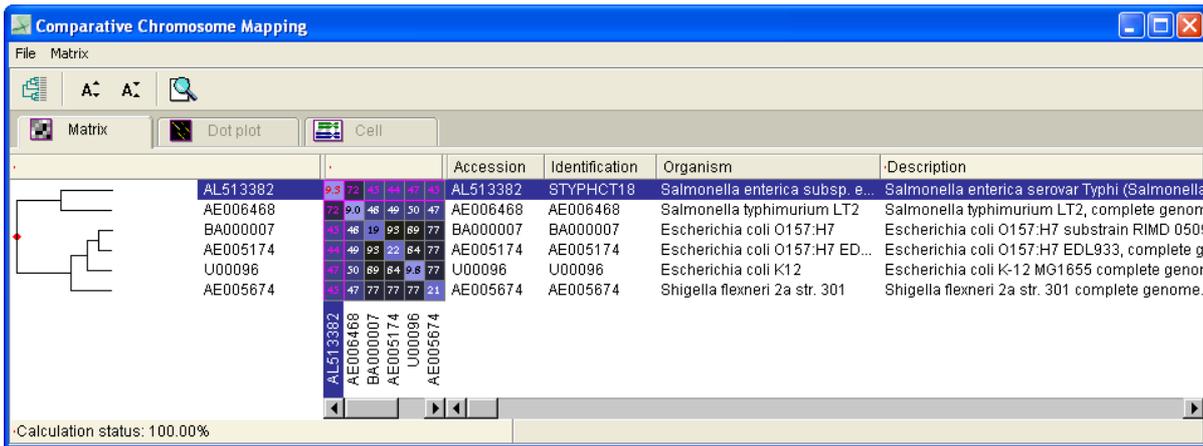


Figure 4-2. The Matrix view with a clustering of the sequences.

4.1.11 Select *Matrix* > *Clustering settings*. You can cluster the entries based on Neighbor joining or UPGMA.

4.1.12 Leave UPGMA enabled and press <OK>.

4.1.13 Select the  button or select *Matrix* > *Calculate clustering* to cluster the sequences based on their pairwise identity scores (see Figure 4-2).

• The Dot plot view

4.1.14 Select the menu item *Matrix* > *Show dot plots* or press the **Dot plot** tab.

Each cell is shown as a dot plot. Blue dots represent stretches of homology between both sequences in the

direct orientation, whereas red dots represent stretches of homology with one sequence inverted.

4.1.15 Select a dot plot. The organism information and the keys are indicated on the status bar (see Figure 4-3).

4.1.16 Use the zoom scroller at the left side to zoom in or out.

4.1.17 Select *File* > *Stretch import settings* and set the '*Minimal stretch identity*' to 70% and press <OK>.

Only stretches with a minimal identity score of 70% are now plotted.

• The Cell view

4.1.18 Select a cell in the **Dot plot** view (e.g. Escherichia vs. Salmonella) and press the **Cell** tab.

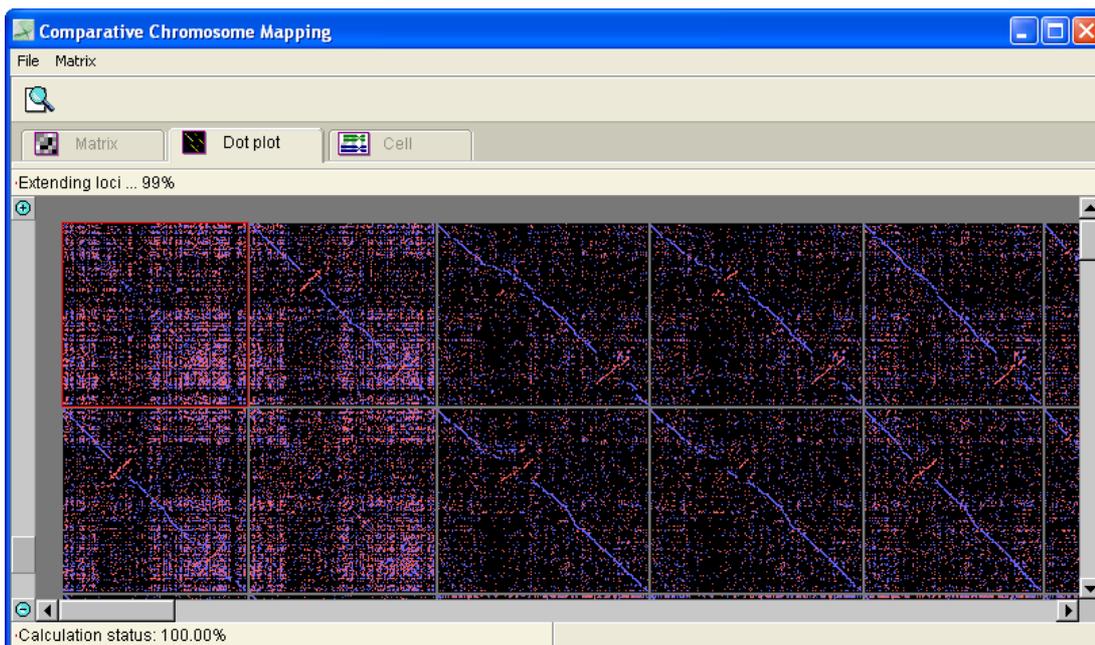


Figure 4-3. The Dot plot view in a comparative chromosome mapping project.

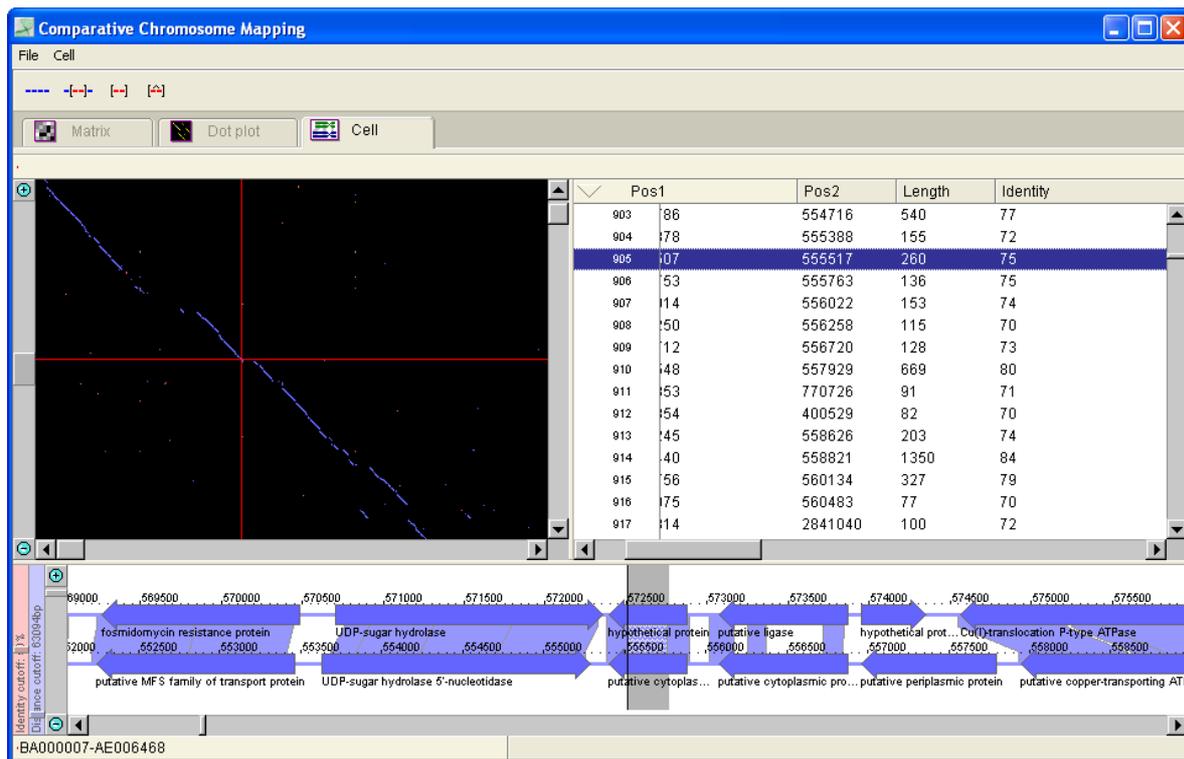


Figure 4-4. The Cell view: detailed dot plot view.

The upper left panel of the **Cell** view displays the dot plot of the two sequence entries and the upper right panel lists all stretches of homology found between the two sequences. The lower panel presents the sequence alignment view (see Figure 4-4).

4.1.19 Click with the mouse pointer on a spot within the dot plot. A red X-Y cross indicates the current position selected. The sequence corresponding to the spot is selected (gray) in the lower panel. The list scrolls to the corresponding item within the list (see Figure 4-4).

4.1.20 Use the zoom scroller to zoom in (and out).

If regions of discontinuous parallelism occur, one can try to link the individual stretches into one block. Such blocks are called '*superstretches*'.

4.1.21 Superstretches are mapped by pressing the  button. Superstretches are mapped in green and the list is updated.

4.1.22 Press the  button to map the stretches on the dot plot and the sequence.

NOTE: More settings are available and can be found in the manual.

4.1.23 Select the **Matrix** tab again. In the next section, we are going to align the chromosomes.

4.2 Chromosome alignment

4.2.1 After having selected the **Matrix** view, select the entry U00096 (E.coli genome).

4.2.2 Select **File > Superstretch calculation settings** and press the **<Defaults>** button. Check **Overwrite gaps within superstretches** and set the number to 500 bp. Press **<OK>**.

4.2.3 Select **File > Multiple alignment project > Run calculation**.

Once the calculations are finished, the *Chromosome alignment* window appears (see Figure 4-5).

4.2.4 Select **File > Save as** or press , name the alignment e.g. 'Multiple alignment' and press **<OK>**.

The upper panel, the **Alignment overview** panel, shows the guiding sequence ('template') and the other sequences ('queries') that are aligned against the guiding sequence.

The left column of the alignment contains the accessions and the overall identity score (between brackets) of the query sequences against the template. Blue blocks represent direct homology, whereas red blocks indicate inverted homology.

4.2.5 Use the zoom scroller at the left side to zoom in and out.

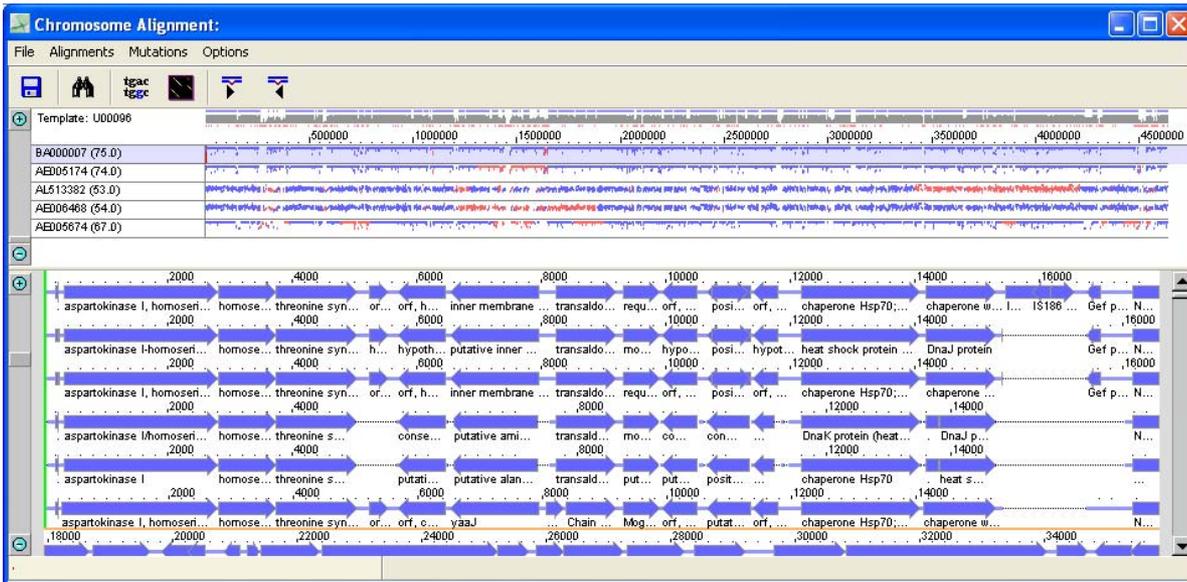


Figure 4-5. The Chromosome Alignment window.

In the **Alignment detail** panel (lower window panel), the multiple alignment is shown into more detail.

4.2.6 Make a selection (click and drag), or change cursor position in the overview panel. The selection and cursor position in the detail panel is updated.

4.2.7 Press the  button or select *Alignments* > *Find*.

4.2.8 Select the **Subsequence** tab, leave all entries selected, enter the sequence 'gatgaatgatgg' and press <Find>.

The **Sequence** tab with the results of the search is shown in the *Chromosome Alignment* window (see Figure 4-6).

The orientation of the match is indicated with an arrow (red: inverted; green: original). The last column indicates if the match found on the target sequence is present on the alignment (green diamond) or not (red diamond).

	Position on entry	Entry	Position on alignment	Mismatches	Match	Presence on alignment
1	614312	AL513382		0	gatgaatgatgg gatgaatgatgg	◆
2	4004476	AL513382		0	gatgaatgatgg gatgaatgatgg	◆
3	622684	AE006468		0	gatgaatgatgg gatgaatgatgg	◆
4	3844381	AE006468		0	gatgaatgatgg gatgaatgatgg	◆
5	4058531	AL513382	3674379	0	gatgaatgatgg gatgaatgatgg	◆
6	3784873	AE006468	3674379	0	gatgaatgatgg gatgaatgatgg	◆
7	3730079	U00096	3730079	0	gatgaatgatgg gatgaatgatgg	◆

Figure 4-6. The results of the sequence search function.

4.2.9 Press the  button again or select *Alignments* > *Find*.

4.2.10 Select the **Features** tab. Leave all entries selected, select CDS, and search for the text 'kinase' (see Figure 4-7). Press <Find>.

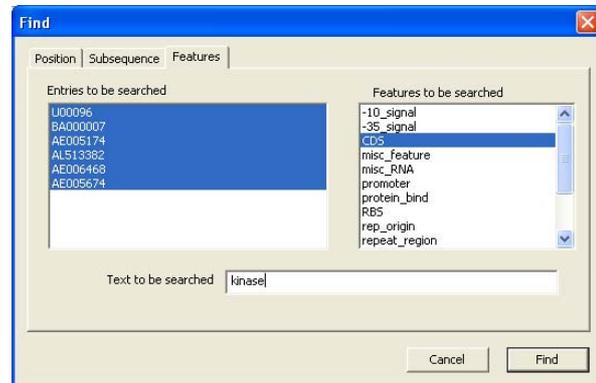
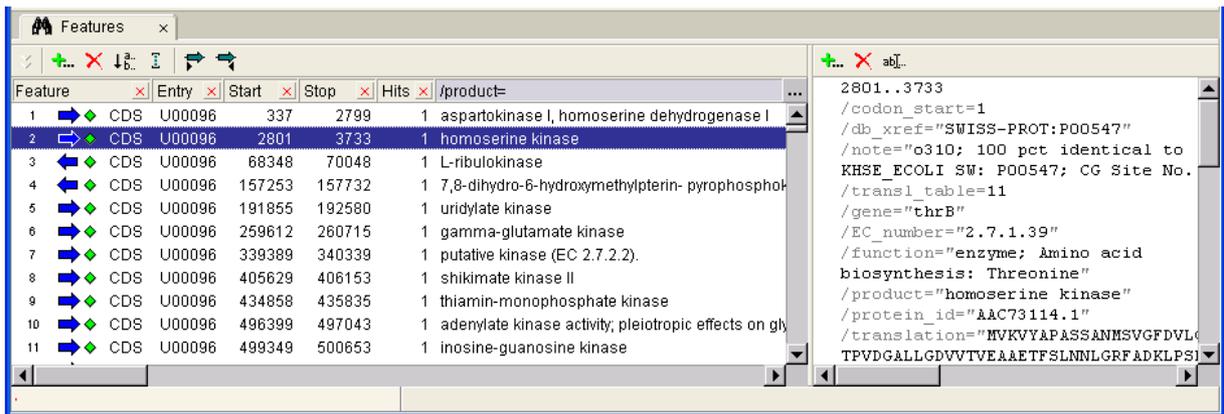


Figure 4-7. The Find dialog box.



Feature	Entry	Start	Stop	Hits	/product=
1	CDS U00096	337	2799	1	aspartokinase I, homoserine dehydrogenase I
2	CDS U00096	2801	3733	1	homoserine kinase
3	CDS U00096	68348	70048	1	L-ribulokinase
4	CDS U00096	157253	157732	1	7,8-dihydro-6-hydroxymethylpterin- pyrophosphok
5	CDS U00096	191855	192580	1	uridylate kinase
6	CDS U00096	259612	260715	1	gamma-glutamate kinase
7	CDS U00096	339389	340339	1	putative kinase (EC 2.7.2.2).
8	CDS U00096	405629	406153	1	shikimate kinase II
9	CDS U00096	434858	435835	1	thiamin-monophosphate kinase
10	CDS U00096	496399	497043	1	adenylate kinase activity; pleiotropic effects on gly
11	CDS U00096	499349	500653	1	inosine-guanosine kinase

```

2801..3733
/codon_start=1
/db_xref="SWISS-PROT:P00547"
/note="o310; 100 pct identical to
KHSE_ECOLI SW: P00547; CG Site No.
/transl_table=11
/gene="thrB"
/EC_number="2.7.1.39"
/function="enzyme; Amino acid
biosynthesis; Threonine"
/product="homoserine kinase"
/protein_id="AAC73114.1"
/translation="MVKVYAPASSANMSVGFVLC
TPVDGALLGDVVTEAAETFSLNNLGRFADKLPSI

```

Figure 4-8. The results of the features search function.

The **Features** tab with the results of the search is shown in the *Chromosome Alignment* window (see Figure 4-8).

4.2.11 Click on the  button next to the header fields in the **Features** tab and select '/product=' from the list.

A new column appears, showing the product names for the CDS. All the CDS listed are kinases (see Figure 4-8).

4.2.12 Call the dot plot view by pressing the  button or select *Options > Show dot plot*.

Details about the alignment of the selected query sequence (Y-axis) against the template sequence (X-axis) are shown.

In a next step, we are going to look for mutations in the sequences.

4.2.13 Select *Mutations > Calculations > Settings*.

4.2.14 Make sure that *Silent*, *Missense* and *Indel* mutations are selected and press the <OK> button.

4.2.15 Select *Mutations > Calculations > Run*.

The progress status is indicated in the statusbar at the bottom of the window. The results of the mutation

search function are listed in the **Mutations** panel (see Figure 4-9).

The *Position* gives the alignment position of the mutation, the *Mutation* states the mutation type, the *DNA change* gives the nucleotide change (template - query) and the *Amino acid change* gives the change at translation level, where applicable.

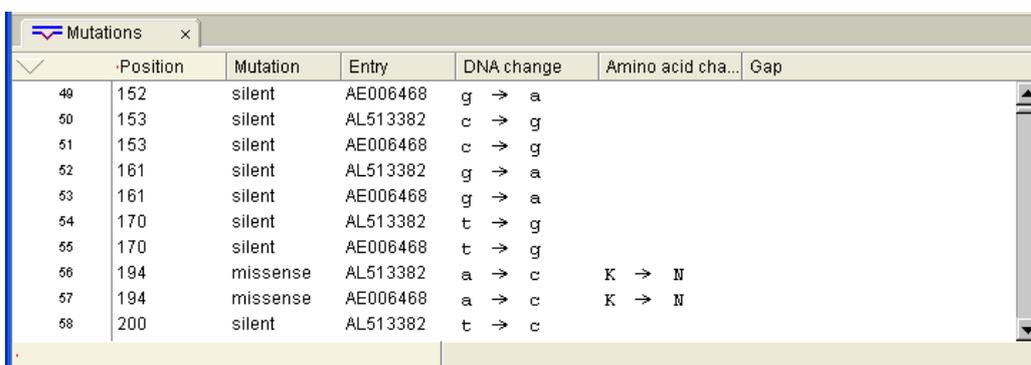
4.2.16 Select a mutation from the list. The alignment detail view scrolls to the mutation location.

4.2.17 Press the  button to switch to text mode (see Figure 4-10). All mutations are shown in color (red: silent mutations, blue: missense mutations, green: indel mutations).

4.2.18 Press the  button to switch back to the graphical view.

4.2.19 Select the  button to find the next mutation downstream the current cursor position.

4.2.20 Select the  button to jump to the next mutation upstream the current cursor position.



Position	Mutation	Entry	DNA change	Amino acid cha...	Gap
49	152	AE006468	g → a		
50	153	AL513382	c → g		
51	153	AE006468	c → g		
52	161	AL513382	g → a		
53	161	AE006468	g → a		
54	170	AL513382	t → g		
55	170	AE006468	t → g		
56	194	AL513382	a → c	K → N	
57	194	AE006468	a → c	K → N	
58	200	AL513382	t → c		

Figure 4-9. The results of the mutation search function.

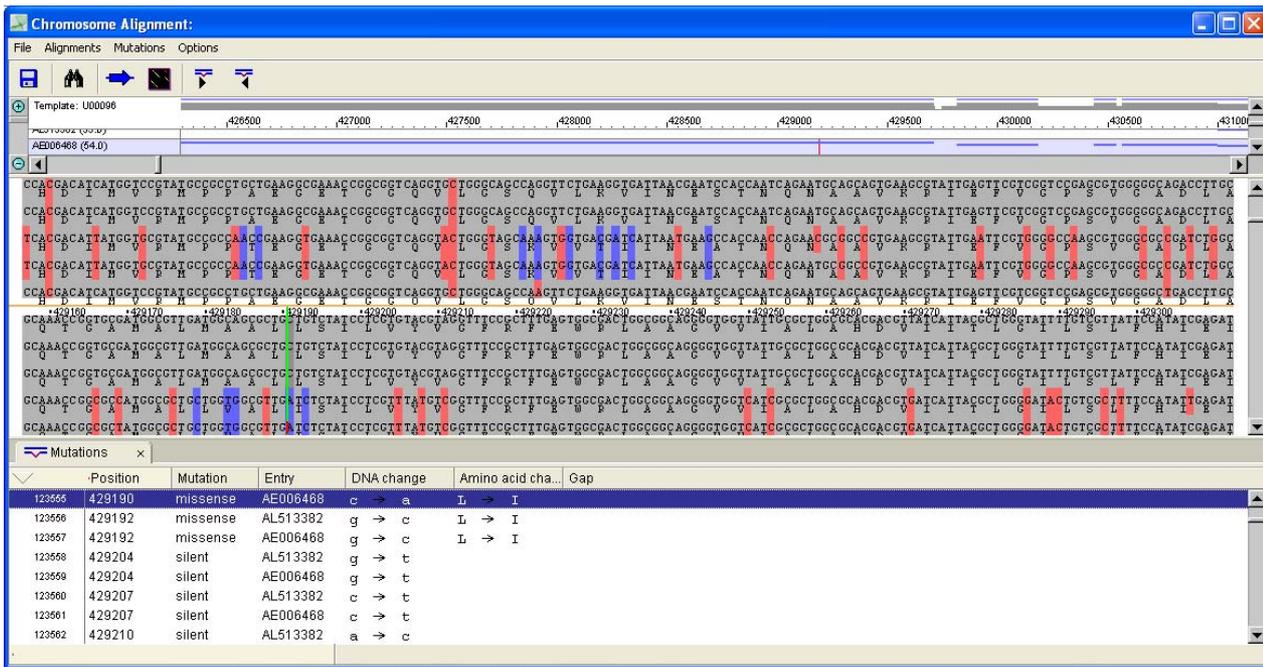


Figure 4-10. The Alignment detail view in text mode.

4.2.21 Save the *Chromosome Alignment* window and close it. Close the *Comparative Chromosome Mapping* window.

4.3 Genome annotation

In Kodon, one can annotate coding regions (CDS features) on sequences. The annotation is performed through comparison with existing annotated sequences and/or homology searches over the internet (EMBL/GENBANK).

4.3.1 In the *Main* window of the **Bacterial chromosomes** database, make sure that no entries are selected by pressing F4.

4.3.2 Select the entry 'U00096' (CTRL + left-click) and

press .

4.3.3 In the next window, select the list 'Annotation' to serve as template entries and press <OK>.

NOTE: The template entries can be a list of entries selected from the database or resulting from a query performed on external entry data files.

4.3.4 In the *Annotation* window, select *File > Open reading frame settings*.

4.3.5 Check the option '*Consider only open reading frames with initiation codon*' and uncheck the other two options. Press <OK> twice.

4.3.6 Select *File > Search settings*. Make sure that the word size is set to 4 amino acids. The default values can be accepted for the other options. Press <OK>.

The third set of options concern the evaluation and annotation of the query coding regions.

4.3.7 Select *File > Annotation settings*. Leave all settings unaltered and press <OK>.

4.3.8 Press the  button to start the calculations. The button will change in a .

4.3.9 After calculations, information fields appear for each protein coding sequence in the overview panel (see Figure 4-11).

4.3.10 Select a field. The field changes into a pop up chart (see Figure 4-11).

Hits are listed according to their identification score (red: ~100% identification score, green ~50%). Next to the color code, the product description of the template feature is shown, together with the identity score and the accession code of the template sequence.

4.3.11 Select one of the hits in the pop up chart. The query and template sequence are plotted at the bottom of the window, with the color of the blocks corresponding to the identification color of the hit (see Figure 4-11).

4.3.12 Select the menu item *File > Identification details* or press .

The identity score matrix and dendrogram of the template protein sequences that produced a hit with the

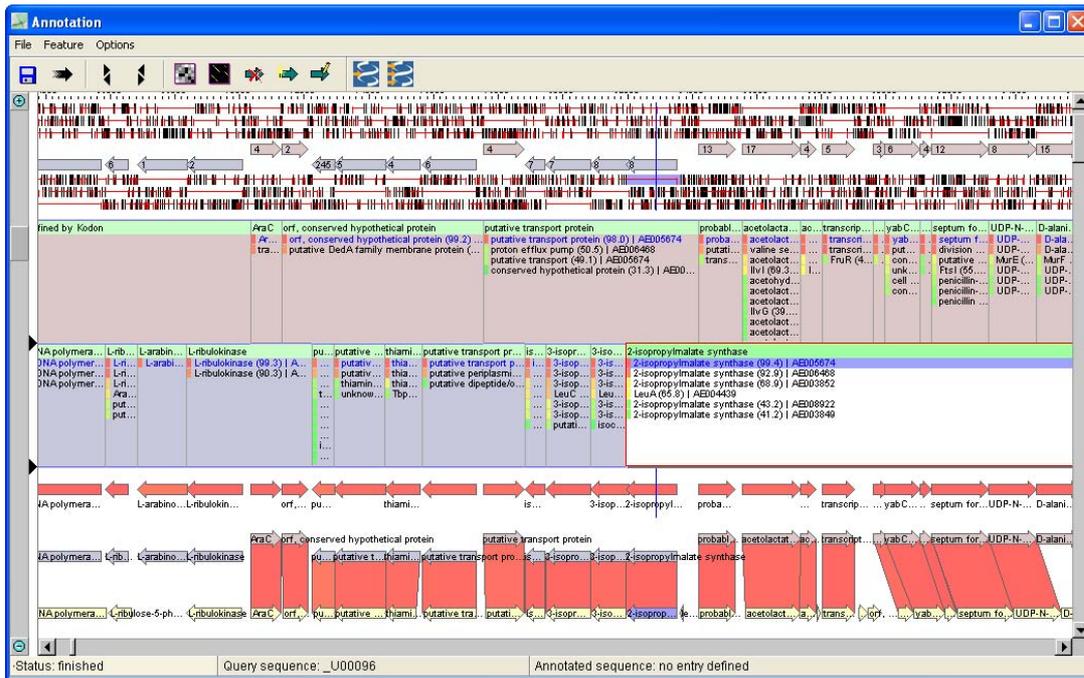


Figure 4-11. The *Annotation* window.

query coding sequence (in blue) are shown in the new window (see Figure 4-12).

NOTE: Hits with identity scores below the annotation threshold may be included in the Details window.

4.3.13 Close the *Details* window.

4.3.14 In the *Annotation* window, select  to view the dot plot of the selected hit.

The dot plot shows all stretches of homology between the query coding sequence and the currently selected template sequence. This allows a fast interpretation of

the significance of the homology stretches between the two protein sequences.

4.3.15 Click on a defined stretch within the dot plot. The corresponding sequence alignment is shown in the window below the dot plot.

4.3.16 Close the *Dot plot* window.

4.3.17 In the *Annotation* window, select *File > Save annotated sequence as*. Name the annotation project e.g. 'E.coli' and press <OK>.

4.3.18 Close the *Annotation* window.

NOTE: Screening of features against local databases or public databases can be found in the manual.

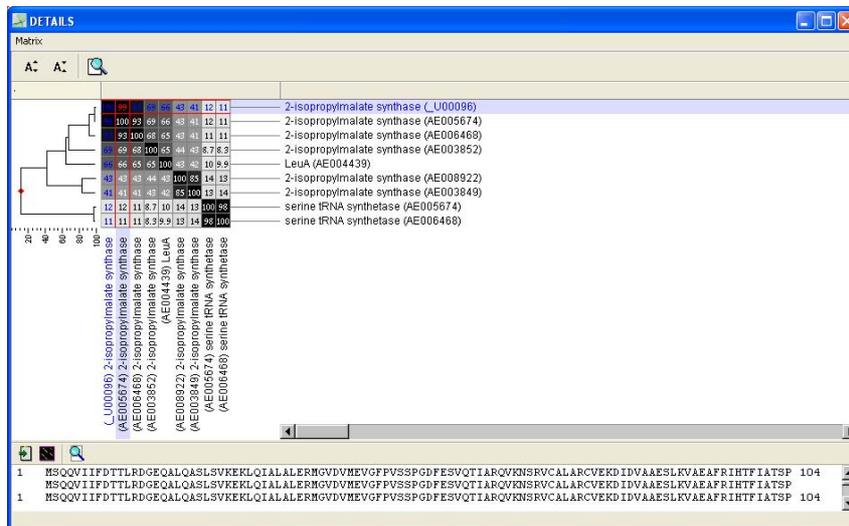


Figure 4-12. Detailed comparison of a single annotated CDS.

