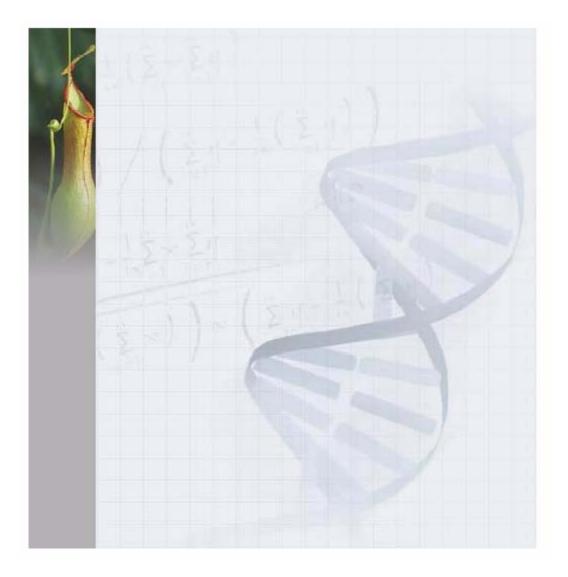
The Kodon quickguide



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 Software5
1.1 Introduction5
1.2 Select functions5
1.3 Database querying6
1.4 Downloading sequences directly from the
internet
1.5 Importing sequences from downloaded files8
1.6 Assembling sequences with Assembler9
1.7 Sequence editing10
1.8 Restriction enzyme analysis12
1.9 Homology search13
2. Multiple Alignment, Clustering & Phylog-
eny15
2.1 Sequence selection from feature matrix15

2.2 Multiple alignment16

2.3 Clustering and phylogeny16
3. Molecular Analysis19
3.1 Frame analysis
3.2 Primer design
3.3 Multiplex PCR design21
3.4 Pairwise matching and repeat analysis22
3.5 Motif search
3.6 RNA secondary structure24
3.7 Protein properties25
3.8 Vector construction
4. Chromosome Mapping
4.1 Comparative chromosome mapping
4.2 Chromosome alignment
4.3 Genome annotation

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 ricon

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 mark 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 \times 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

Menu	File Edi	t List Project	BLAST Scripts		
Toolbar	[[[]]] [[]] [[]] [[]] [[]] [[]] [[]] [🛛 📸 🖬 🕻	🗙 🔀 🕃 🌾	
	Þ	Sequences]		🕼 🖬 🗟 🞯 🔚 Lists
	$\overline{}$	Accession	Identification	Organism	Name ·Descriptio
	1	AC002387	AC002387	Arabidopsis thaliana (thale cress)	1 Escherichia coli
Database_	2	AF012066	AF012066	unidentified eubacterium K5S18	2 Leucine Zipper
panel	3	AF076806	AF076806	Riftia pachyptila endosymbiont	3 Plant dihydrodipicolinate synthases
	4	AF128734	AF128734	uncultured soil bacterium C074	4 Ribosomal RNAs 5 RNA secondary structures
	5	AF128739	AF128739	uncultured soil bacterium C010	6 Vectors
	6	AF132134	AF132134	endosymbiont of Acanthamoeba sp. UWC9	7 Yeast chromosomes
	7	AF132135	AF132135	endosymbiont of Acanthamoeba polyphaga	7 Teast circlinosomes
	8	AF132136	AF132136	endosymbiont of Acanthamoeba sp. UWET39	39
	9	AF136496	AF136496	Campylobacter upsaliensis	•
	10	AF200325	AF200325	Arabidopsis thaliana (thale cress)	
	11	AF230085	AF230085	Borrelia bissettii	12 🕺 12 💌 💌 🔁 🖬 🕅 12
	12	AF267198	AF267198	Candidatus Carsonella ruddii	Name Type
	13	AF267202	AF267202	Candidatus Carsonella ruddii	
	14	AF272615	AF272615	Ureaplasma urealyticum	1 C Plant dihydrodipicolinate Clustering
	15	AF329135	AF329135	Lactobacillus delbrueckii	2 🖆 Ribosomal RNAs Clustering 3 🖆 Yeast chromosomes Clustering
	16	AF395069 AF395069	AF395069	uncultured bacterium uncultured bacterium	4 ¥ Cloning Constructs
	17	AF395070 AF395071	AF395070 AF395071	uncultured bacterium	5 ≡ introns Multiplex PCR
	18 19	9	AF395071 AF395072	uncultured bacterium	_ 6 ≡ veast kinases Multiplex PCR
	19	AF395072	AF 39507 2	unculured baclenum	
	■	OAD	in all for the state of the second	h	
	·Databa	se: C:(Program F	iles\Kodon\data\Demo	base 88 sequences	

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 27

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 \times 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>*.

Database	field search	
Search for:	Arabidopsis	
In field:	Organism 💌	
🗖 Case se	ensitive	OK
🔲 Regula	r expression	Cancel

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 index 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).

Sequence header search	
Search for:	
Vauterin	-
	ОК
Case sensitive	Cancel

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 **math and a select** 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 'gcggtaatac' in the input field, uncheck *Allow gaps* and set the mismatches to 0 (see Figure 1-4). Press **<OK>**.

Subsequence search	
Query sequence gcggtaatad	
Search tolerance	
	Cancel 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 AND () to create a composite query (see Figure 1-5).

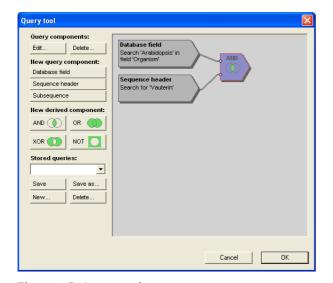


Figure 1-5. A composite query.

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



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 NOT . 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 XOR

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>*.

Search Nucleotide 💽 for AB022784 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).

Display GenE	3ank 📃 💌	Show	5 🔽	Text 💌
--------------	----------	------	-----	--------

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).

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



in the *Main* window of Kodon.

1.5.7 If the file does not appear in the file list, doubleclick 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).

Sequences Accessio 80 Y38080 81 Y13135 82 Y13138 83 Y15503 84 Y15507 85 Y11826	Identificatio Organism ATDHDPSG Arabidopsis thaliana (thale c Y13135 Saccharomyces cerevisiae Y13138 Saccharomyces cerevisiae	De ress) A.th Sa	scription naliana dhdps g ccharomyces cer	C-box; CAAT bo	Field 0		Name •Name Escheric Leucine) Li:
80 X98080 X 81 Y13135 X 82 Y13138 X 83 Y15503 Y 84 Y15507 Y 85 Y Z11826	ATDHDPSG Arabidopsis thaliana (thale c Y13135 Saccharomyces cerevisiae	ress) A.tr Sa	naliana dhdps g	C-box; CAAT bo		1	Escheric		
80 X98080 X 81 Y13135 Y 82 Y13138 Y 83 B Y15503 Y 84 B Y15507 Y 85 B Z11826 I	ATDHDPSG Arabidopsis thaliana (thale c Y13135 Saccharomyces cerevisiae	ress) A.th Sai	naliana dhdps g	C-box; CAAT bo		1			
87 👩 Z50067 5 88 🗿 Z97193 1	Streptomyces cerevisits SY15503 Streptomyces by SY15507 Streptomyces lydicus BMPB1R5S Bacillus methanolicus SC16SR117 Saccharomonospora azurea SH16SR120 Saccharopolyspora hordei MTCY180 Mycobacterium tuberculosis AB022784 Arabidopsis thaliana (thale c	Stri Stri B.n S.c S.h H37Rv Myo	eptomyces lydic nethanolicus 5S aesia ribosomal	 16S ribosomal 16S ribosomal 5S ribosomal R 5S rRNA; ribos 5S rRNA; ribos			Plant dih Riboson RNA sec Name Plant dih	nydrodipico nal RNAs condany str 	uctures

Figure 1-8. Demobase with imported sequence AB022784.

Query Result set					
1 sequences (644 b	ase pairs)				
File 🗙 Siz	e (bp) 🗙 Accession 🗴	Organism	× Description	× Keywords	×
AY429384.gb	644 AY429384	Arabidopsis th	nalia Arabidopsis the	alia	

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

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 2.

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

sequences manually use the 🔬 and 🖽 buttons. If

there are vectors in the sequences, click on the **Q** button to define and remove them.

Assembly

button or

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

1.6.9 Select *File > Assemble sequences* or click on

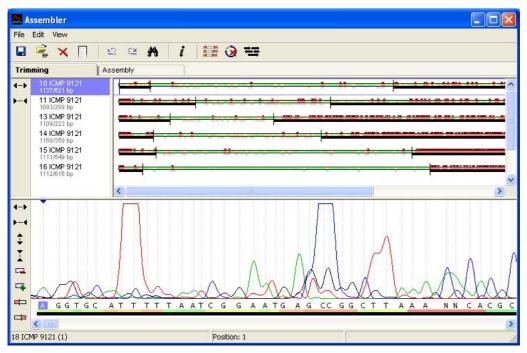


Figure 1-12. The Assembler main window.

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).

Entry PPU/235	54 (Key: U72354)							
File Edit Header F	Features Tools							
atg[Sequence E	Editing	Graphical Presentati	ion 🔲 Te	xt Presentation				
8 🔒 🐰	된 📑 🗠		o 🔝 🥩 🗧	• —				
ACCUTGTUAT	TUTUAGGUAA	TGATACAGGT	GATGUGUTTG	TTATUTUAGT	GAAAGATACA	GGTUUAGGTA	TTTUAGAGAG 2000 TGTCTTACAA 2960	
							rororimonni	
		AAAATATGGG		GCCCGCTGGG	GGTGCTTGTT		TATCCCTGCC 3040	
AGACAGGGCT			CATCIGITAT		GATGACGACA		GGAAGCACTG 3120	
AGCAGCTTGG			GTCGAGTGTT	TTTCATCCGC	TAGCGTATTC		TCAATCGCTC 3200	1
Teccteteec							AAAGAGCTTG 3280	
GCGAGCAAAT	CCCCATTATT	TTTATCAGCG	GCCACGGTGA	TATTCCGATG	GCAGTCAAAG	CGATCAAGGC	GGGTGCGGTA 3360	
Image: state to the s								
Header fields	Features						db_xref=SPTREMBL:007832 evidence=EXPERIMENTAL note=TodT	
ab] 🗙 🛻 된 🧧					ab][transl_table=11	
∎ <mark>. × + </mark> ID PPU72354 XX		NA; PRO; 4266	BP.		 Field 01		gene=todT product=response regulator	Ŀ
ID PPU72354		NA; PRO; 4266	BP.		Field 01 Field 02		gene=todT	
ID PPU72354 XX LC U72354; XX		NA; PRO; 4266	BP.		Field 01 Field 02 Field 03		gene=todT product=response regulator	
ID PPU72354 XX AC U72354; XX XV U72354.1		NA; PRO; 4266	BP.		Field 01 Field 02 Field 03 Field 03	· · · · · · · · · · · · · · · · · · ·	gene=todT product=response regulator	
ID PPU72354 XX AC U72354; XX BV U72354.1 XX	standard; D	NA; PRO; 4266 data\Demobase\U7;			Field 01 Field 02 Field 03	· · · · · · · · · · · · · · · · · · ·	gene=todT product=response regulator	

Figure 1-13. The Sequence editor window.

1.7.7 Raise the number of *Mismatches allowed* to 1 and

press the **Pind** 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 X 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.

Search		×
File Options		
Restriction site	Position / Mismatch Sequence:	
	🔹 🗢 🖛 ggatct	₽
AccI AluI ApaI AvaI BalI BamHI BclI BglII		
0 500	1000 1500 2000 2500 3000 3500 4000	

Figure 1-14. The Subsequence search window.

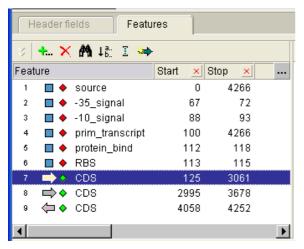


Figure 1-15. The *Features* panel in the sequence editor.

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 **T** 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 \blacklozenge 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. Alul). The enzyme is listed below and its restriction sites are mapped on the sequence.

1.7.24 Click on ¹²³¹ to hide the site positions on the map.

1.7.25 Click on the button <u>enzyme</u> 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 **F** to save the map with the changed layout.

• Text presentation view

The **Text Presentation view** provides a publicationready 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 $\langle OK \rangle$.

1.7.30 Select *File > Save* or click on **F** 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 BcII.

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

 \downarrow 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

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).

🔄 Homology search window File View -→X98080 0.00 3.15 → AL162295 →L77616 58.39 61.45 → AF230085 → AF136496 61.56 63.79 → AF395074 64.78 →X57791 65.17 → AF395073 → AF267202 65.28 65.86 → L36436 😂 🛃 4-> >-4 CDS CDS

Figure 1-16. *Homology search* window.

1.9.4 Press the **to** 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:

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 **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

Đ.

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.

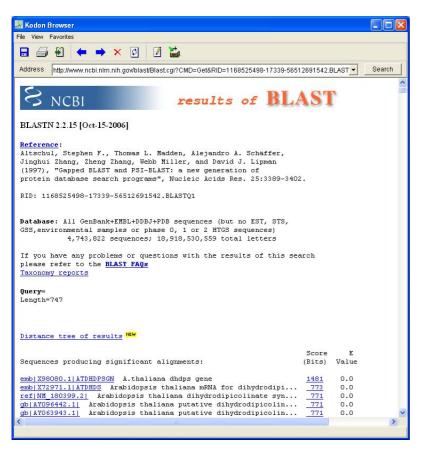


Figure 1-17. Results of a Web-based BLAST.

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.



2.1.5 Choose *View* > *Search subset* or press

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

Feature subset	
Show only features containing the str	ing
dihydrodipicolinate	
in the qualifier < All >	•
- Method	
 Replace current subset 	
Add to current subset	
C Search in current subset	ОК
Negative search	Cancel

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 **+ 10** 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.

Project 'Plant dihy	drodipicolinate synth	ases'															
File Edit View																	
Feature matrix	Selected features	Subset .	Alignme	ent													
🗖 🗛 🥵 🕻	* 된 📑 \star	3+ 🖹 🔍															
All	CDS	Repeat regions															
Source	Proteins	Binding sites		87	52		32										
Genes	Immunoglobulin	DNA structures		AC002387	AF200325 X98080	971	AL162295	U61730	L77616	090	X79675 V72742	598	M60599	850			
Feature 🗙 produ	ct 🗙			ACC	AF2 X98	X72	AL1	190 1 26	L79	X79	X7967	MBC	MBC	X52			
	lpyruvylsh		+ 1	1													~
	like protein ubulin cof		+ 1				1										
	rodipicoli		+ 1				-							1			
CDS dihydr	rodipicoli		+ 10		1	1		1 1	1 1	1	1 1	1	1				
	rodipicoli		+ 1		1												
	rodipicoli sperm sp		+ 1+ 1		_		1										_
	speini sp						'										<u> </u>
< > □ 🛛 🚽																	_
Feature 1/1																	
CDS 55460-57811 (23	352 bp)														^		
Forward /codon start="1"																ID: AC002387 OS: Arabidopsis thaliana (thale cre	iss)
/db xref="SWISS-PRO	T: P05466"															DE: Arabidopsis thaliana chromos	
/gene="At2g45300"																section 242 of 255 of the o	
	uvylshikimate-3-pho	sphate (EPSP)	syntha	se"												sequence. Sequence from T14P1, F4L23.	n clones
/protein_id="AAB826							-								- 1	KW: HTG.	
	SRICNGVQNPSLISNLSKSS SNRILLLAALSEGTTVVDN																
	LDGVPRMRERPIGDLVVGL																
	ERFGVSVEHSDSWDRFFVK																
	FGMRHLRAIDVNMNKMPDV. SLAACADVPITINDPGCTRI			ASWF	RVKET	FER	MIAI	CTE	LRK	LGA	TVEF	GSI	OYCI	/ITPPKKVKTAE	I		
DITUDIRTANAP	SERVEDVETTEDFUCTR	ATT DIE QUBREIT													×		

m

Figure 2-1. The Project window.

All		CDS	Repeat regions														
Source		Proteins	Binding sites			87	25		95						_		
Genes		Immunoglobulin	DNA structures			0023	2003	പത	622	P~-	616 1616	090	9675	2743	នាន	0599 0560	5
Feature	× product		×			\odot	ir Lē	5 6-	AL1	ω	L36	i Ž	ĽX	X7		M 61 2 2 2 2 2 2	Ż
CDS	dihydrodipicolinate			+	1											1	Ť
CDS	dihydrodipicolinate s	ynthase		+	10			1 1		1	1 1	1	1	1	1	1	
CDS	dihydrodipicolinate s	ynthase 2		+	1		1										
CDS	dihydrodipicolinate s	ynthase precursor		+	1				1								
CDS	putative dihydrodipico	olinate synthase		+	1	1											

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 崫 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).

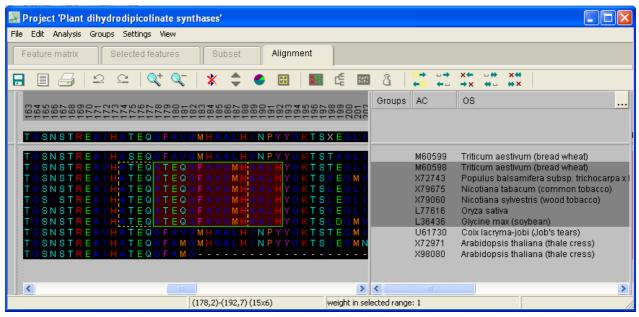


Figure 2-4. The drag-and-drop alignment tool.

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

Groups	os
	Triticum aestivum (bread wheat) Triticum aestivum (bread wheat) Populus balsamifera subsp. trichocarpa x Populus deltoides Nicotiana tabacum (common tobacco) Nicotiana sylvestris (wood tobacco) Oryza sativa Glycine max (soybean) Coix lacryma-jobi (Job's tears) Arabidopsis thaliana (thale cress) Arabidopsis thaliana (thale cress)

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).

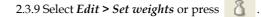
2.3.5 Select *View* > *Show distances as numbers* to see the pairwise distance values in the distance matrix (see Figure 2-6).

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.



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

Set weight in range		
Method C Automatic using entropy (C Manual (base positions) C Manual (reference numbers)	Parameters Begin nr: 145 * End nr: 174 * Set weights to: 3 *	OK Cancel

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.

Project 'Plant dihyd	rodipicolinate syntha	ses'									
<u>File E</u> dit <u>A</u> nalysis <u>G</u> roup:	s <u>S</u> ettings ⊻iew										
Feature matrix	Selected features	Subset Ali	gnment								
🔒 🗏 🎒 🗅	으 🍳 🍕	* 🗘 🔹 🛙	E 🛛	t 🖾 🖏		x← □₩ →x ↔□	x↔ ⇔x				
UPGMA Nr of bootstraps: 0				Groups AC	08		Multiple	e aligned	distance	es - No ci	orrectio
NI 01 0001511 ap 5. 0											
	XKRXEFVXIVK	XIGRENFVO	XXXV				Ó	20 4	0 60	0 80	ı 10
	EKRTEFVRIVE	AIGRENFVO	QKEV		Triticum aestiv			15.38	14.71	22.1	32.44
	EKRVEFVRIVE	ALGRENEVO	QKES		Triticum aestiv	um (bread wl		13.54	13.54	21.22	34.56 29.48
	EKRVEFVRIVE	ALGRENEVO	OVET		Oryza sativa Coix lacryma-jo	hi / loh'e taa			14.74	14.74	33,95
	AQRVEFVNLVK	KIGRENFVO	EEDV		Populus balsa		32.44	34.56	29.48	33.95	0
	DKRIEFANLVK	EIGREHFVO	NKVV		Glycine max (s		26.81	26.51	28.31	27.71	19.88
	AKREEFVKIVK	EIGRENFIC	ERDV	X796	Nicotiana taba	cum (commo	27.48	29.25	28.12	29.69	23.96
	A K R E E F V K I V K		BERDV		Nicotiana sylve		21.75	22.73	23.7	23.7	
	S K R I E F V K L V K	EIGREHFVO	DRDV		Arabidopsis the		28.49	29.95	28.41	30.94	23.35
				- X980	Arabidopsis the	aliana (thale	33.78	35.44	28.57	34.42	24.68
	<			<		>	<				>
	×),9)-(373,9) (1×1)		weight in selected r	ange: 1						

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.

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

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

19

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

3.1.6 Press the **to** 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.

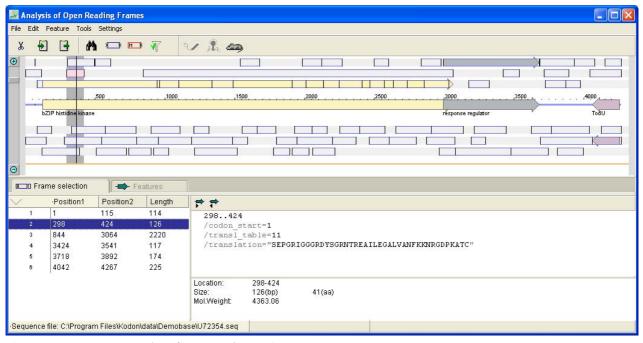


Figure 3-1. The Open reading frame analysis window.

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 +1).

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 16^{-1} 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 **Q** 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.

a Prime	r Design			
File Prime	ers PCR Sele	ctions		
⇒ 1	ar 1 8 1 1	-		
conser	nsus aa	acgcatta :gcgtaat	tccttaact aggaattga	1040 1050 1060 1070 1080 1080 1100 1110 tgetttecaatgegtttaaattea <mark>ceetgaeggggggtttat</mark> eegttgetgtettagtttgagtegaecaaattatgeettggt acgaaaggttaggeaatttaagt <mark>gggggetgeeegaata</mark> ggeaacgaegaatcaaatteagetggtttaataeggaacea
U7235	4 PPLat	acgcatta	teettaaet	tgetttecaatgegtttaaatteacccetgaeggggggettateegttgetgtettagtttgagtegaecaaattatgeettggtt
	100	(
Primer	s PI	ots	PCR pro	ducts Print-out Listing
La. Pos	. Len.	Deg.	Temp.	
1059	3 19	1	67.5	1 010 1020 1030 1040 1050 1060 1070 1080 1090 111
1957	7 20	1	67.5	
558	21	1	67.5	
1587	21	1	67.5	agaaatacgaacgcattatccttaacttgctttccaatgcgtttaaattcacccctgacggggggcttatccgttgctgtcttagtttgagtgag
2231	21	1	67.5	ctttatgcttgcgtaataggaattgaacgaaaggttacgcaaatttaagtggggactgcccccgaataggcaacgacagaatcaaactca
4218	3 19	1		
4210			67.5	
833	21	1	67.5 67.5	
	21	1 1		Primer1: (P) 5-cccctgacggggggcttat 3(0H)
833	21 3 21	' 1 1 1	67.5	Primer1: (P) 60 cccctgacggggggcttat-3(0H) Position: 1059
833 1958	21 3 21 21	' 1 1 1	67.5 67.5	Primer1: (P)6':cccctgacggggggcttat3(0H) Position: 1059 Primer length: 19
833 1958 1431	21 3 21 1 21 5 21	1 1 1 1 1	67.5 67.5 67.5	Primer1: (P)6'cccctgacggggggcttat3(0H) Position: 1059 Primer length: 19 Primer degeneration: 1
833 1958 1431 4066	21 3 21 1 21 5 21 7 21	1 1 1 1 1 1	67.5 67.5 67.5 67.5	Primer1: (P)6':cccctgacggggggcttat3(0H) Position: 1059 Primer length: 19
833 1958 1431 4066 4067	21 3 21 1 21 5 21 7 21 5 20	' 1 1 1 1 1 1	67.5 67.5 67.5 67.5 67.5	Primer1: (P) 6-cccctgacggggggcttat3(0H) Position: 1059 Primer length: 19 Primer degeneration: 1 Melting temperature: 67.5°C

Figure 3-2. The Primer design window.

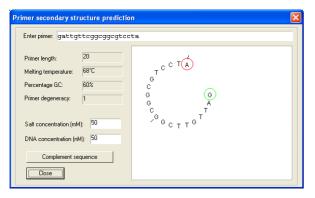


Figure 3-3. Primer properties.

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 even 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 f 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***>**.

Select features to display	×
Select features: source gene COS LTR tRNA misc_feature misc_RNA repeat_region	
Within selected features, search for text: kinase	_
Cancel	

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).

	955000	960000 9	70000 . !
•			

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



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

Add PCR locus from feature
Locus type: CDS
Locus code: Kinase_1
Comment:
PCR locus falls within the target sequence
C Extend PCR locus with
0 bases upstream
0 bases downstream
Exclude target sequence as priming site
 Fuse joined sequences and intervening sequences Fuse joined sequences only, removing intervening sequences (DNA to RNA)
C Consider joined sequences independently
Defaults Cancel OK

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 \times 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.

ile Locus Experiment											
। 🗯 ९१ ९२ 🖻 🖻	🗙 🐴 🛛 at]	at∐									
45000 50000	55000	6000		65000	70000	7500	0	80000	85000		90000 95000
			_	hh /		61					
CDS CDS CDS CDS	CDS C CDS	CDS			CDS CDS	CDS	C			s co	S CDS C CDS CDS
	נטס נ נטס	CDS	· ···	CDS	Kina		U			5 00	··· ···· ···· ···· ·····
					INITA	24-1					
					11	-					
					Kinase 1						
45000 50000	55000	6000	0	65000	70000	7500	0	80000	85000		90000 95000
concentra concentra co	n an tara an	and the		and the t		the second		at company			elecceretere.
	4										
	~ ~								_		
											-
]			Locus	Evne	eriments						
.tggcgetta actetaatae cagee	taree 400	1000	Loode	, II cybe	Sinnonits						
			1	Target	Code	Entry	Comme	Pos1	Pos2	Size	Status
			Y	ranger			0000000	1.001	1.001		
				000	LOnene T	14 24 25		0004.00	00000	4740	A
patgatgatg acgatgatga cgatg	acqat 540	-		CDS	Kinase_7	Y13135	-	668183	669902	1719	Accepted: experiment 1
gatgatgatg acgatgatga cgatg gacgacgacg acgatgaatc cgatc	acqat 540 tggaa 570			CDS CDS	Kinase_7 Kinase_6	Y13135 Y13135	-	668183 575392	669902 576499	1719 1107	Accepted: experiment 1 Accepted: experiment 1
gatgatgatg acgatgatga cgatg qacqacqacq acqatqaatc cgatc agcttagaag gggaagtgga tactg	acqat 540 tggaa 570 acacg 600		2 (
gatgatgatg acgatgatga cgatg qacqacqacq acqatqaatc cgatc agcttagaag gggaagtgga tactg gacgataaca acgaaggaga tggct	acqat 540 tggaa 570 acacg 600		2 0	CDS	_ Kinase_6 Kinase_0	Y13135 Y13135	-	575392 23935	576499 25396	1107 1461	Accepted: experiment 1 Accepted: experiment 1
gatgatgatg acgatgatga cgatg gacgacgacg acgatgatc cgatc agcttagaag gggaagtggg tactg gacgataaca acgaaggaga tggct aatcacgagg aaggtggaga agaag	acqat 540 tggaa 570 acacg 600 ctgac 630 gaagt 660		2 3 4	CDS CDS CDS	_ Kinase_6 Kinase_0 Kinase_4	Y13135 Y13135 Y13135	-	575392 23935 353056	576499 25396 355471	1107 1461 2415	Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1
agettagaag gggaagtgga taetg gaegataaca aegaaggaga tgget aateaegagg aaggtggaga agaag agaggagetg atgetgatgt gteea	acqat 540 tggaa 570 acacg 600 ctgac 630 gaagt 660		2 3 4 5	CDS CDS CDS CDS CDS	_ Kinase_6 Kinase_0 Kinase_4 Kinase_9	Y13135 Y13135 Y13135 Y13135 Y13135	20 20 20	575392 23935 353056 886067	576499 25396 355471 887870	1107 1461 2415 1803	Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1
gatğatgatg acğatgatga cyatg <u>gacqacqacq</u> acqatqatc cyatg agcttagaag gggaagtgga tactg agcgataaca acgaaggaga tggct aatcacgagg aagtggaga agaag agaggagctg atgctgatgt gtcca agcagcgtg ccgaaagggt tgccg	acqat 540 tggaa 570 acacg 600 ctgac 630 gaagt 660 gtgca 690		2 3 4 5	CDS CDS CDS CDS CDS CDS	_ Kinase_6 Kinase_0 Kinase_4	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135	-	575392 23935 353056 886067 160069	576499 25396 355471 887870 162763	1107 1461 2415 1803 2694	Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1
yatgatgatg acgatgatga cgatg acgacgacg acgatgatb gacgatgaag gggaagtgga tactg gacgataaca acgaaggaga tggct gacgataaca acgaaggaga gagggggtg aggtggag agaggagctg atgctgatgt gtcca agcagcgtp cogaagggt tgcgg sggatatatc aagatctag atcag	acqat 540 tggaa 570 acacg 600 ctgac 630 gaagt 660 gtgca 690 accca 720 ctatt 750		2 (3 (4 (5 (6 (CDS CDS CDS CDS CDS	_ Kinase_6 Kinase_0 Kinase_4 Kinase_9	Y13135 Y13135 Y13135 Y13135 Y13135	20 20 20	575392 23935 353056 886067	576499 25396 355471 887870	1107 1461 2415 1803	Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1
jatóstejátg acgatójatója cijátój gacgacgacg acgatógató gactagaag gygaagtoga tacto gacgataaca acgaaggaga tyget gacgataaca acgaaggaga gaggagetg atgetgatga gaggagetg atgetgatój tycca sagcagegtg cegaaagggt tyccg sigdatatat aaggatetag atcag aatatagaaa cagaatcaag gaatt gatacateg acagaacaag tygte	acqat 540 tggaa 570 ctgac 600 ctgac 630 gtgca 660 gtgca 690 accca 720 ctatt 750 tgtgg 780 tgtag 810		2 (3 (4 (5 (7 (CDS CDS CDS CDS CDS CDS CDS CDS	– Kinase_6 Kinase_0 Kinase_4 Kinase_9 Kinase_2 Kinase_5	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135	5 5 2 5	575392 23935 353056 886067 160069	576499 25396 355471 887870 162763	1107 1461 2415 1803 2694	Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1
gatgatgatg acgatgatga cgatg acgacqacg acgatgatga cgatg agcttagaag gggaagtgatga tactg gacgataaca acgaaggaga tggct aatcacgagg aaggtggaga agaag agaggagtg atgctgatgt gtcca aagcaggtg cggaaaggt tgcgg aatatagaaa cagaacgag gaat gatatagaaa cagaatcag gaat gatatactg acagaacag tggtc actacccc ctgatcaqt tggtc	accat 540 tggaa 570 acacg 600 ctgga 630 ggagt 660 gtgca 690 accca 720 ctatt 750 tgtgg 780 tgtaga 810 gttct 840		2 3 4 5 6 7 8	CDS CDS CDS CDS CDS CDS CDS CDS CDS	Kinase_6 Kinase_0 Kinase_4 Kinase_9 Kinase_2 Kinase_5 Kinase_3	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135		575392 23935 353056 886067 160069 458152 207036	576499 25396 355471 887870 162763 458989 208575	1107 1461 2415 1803 2694 837 1539	Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1
jatóstojátój acijatója cijátój jacgacgacg acgatogato agottagaag gggaagtoga tacto gacgataaca acgaaggaga tggot aatcacgagg aaggtogga tggot agaggagtog atgctogatog sagcagotg cogaagggot tgogg toggatatato aaggatotag atcag aataatagaaa cagaatcag gaatt aataatagaaa cagaatcag tggot cectttott tototagat atcag	accat 540 tggaa 570 acacg 600 ctgac 630 gaagt 660 gtgca 690 accca 720 ctatt 750 tgtgg 780 tacaa 810 gttct 840 ctgt 870		2 3 4 5 6 7 8 9	CDS CDS CDS CDS CDS CDS CDS CDS CDS CDS	Kinase_6 Kinase_0 Kinase_4 Kinase_9 Kinase_2 Kinase_5 Kinase_3 Kinase_1	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135		575392 23935 353056 886067 160069 458152 207036 69670	576499 25396 355471 887870 162763 458989 208575 72547	1107 1461 2415 1803 2694 837 1539 2877	Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1
jatgatgatg acgatgatga cgatg jatgatgatg acgatgatga cgatg jacgatgatga gggatggat gggaagtgga tactg jacgataaca acgaaggaga tgget jacgatgatga aggtggatg aggaggotg atgetgatg tgecg jaggatgatg aggtggatg jatacatcag acgatgatgatgatga jatacatcag acagaacaag ggtt jatacatcag acagaacaag tggtc jatacatca acgatcagt tccaa ccttttett etcetagat atgg acgatasa atgacaatgg ggcaa	acqat 540 tggaa 570 acacg 600 ctgac 630 gtgca 690 accca 720 ctatt 750 tgtgg 780 tgtgg 780 tgtgg 810 gtctct 840 ctggt 870 ccaat 900		2 4 4 6 6 6 6 7 4 6 6 6 6 6 6 6 6 6 6 6 6	CDS CDS CDS CDS CDS CDS CDS CDS CDS CDS	Kinase_6 Kinase_0 Kinase_4 Kinase_9 Kinase_2 Kinase_5 Kinase_3 Kinase_1 Kinase_11	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135		575392 23935 353056 886067 160069 458152 207036 69670 970766	576499 25396 355471 887870 162763 458989 208575 72547 973730	1107 1461 2415 1803 2694 837 1539 2877 2964	Accepted: experiment 1 Accepted: experiment 1
jatjatgitg acgatgatga jacqacgacg acgatgatc cgatg jacqacgacg gggaagtgga tactg jacqataaca acgaaggaga tggt jacgataaca acgaaggaga tggt gagagctg atgctgatgt gtcca aggagggt ccga aggtcgggt tgcgg gggatatatc aaggatctag atcag jatatagaa cagaatcag gatt jatacactg acagaacag tggt ccaacccc ctgatcagt tggt acctaccca acagtagta tggg acgacaaa atgacaatga ggat	actat 540 tggaa 570 acacg 600 ctgac 630 ggagt 660 ggtgca 690 accca 720 ctatt 750 tgtgg 780 tgtgg 780 tgtgg 840 ctagt 840 ctggt 870 caaaa 930		2 4 4 6 6 6 6 7 4 6 6 6 6 6 6 6 6 6 6 6 6	CDS CDS CDS CDS CDS CDS CDS CDS CDS CDS	Kinase_6 Kinase_0 Kinase_4 Kinase_9 Kinase_2 Kinase_5 Kinase_3 Kinase_1	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135		575392 23935 353056 886067 160069 458152 207036 69670	576499 25396 355471 887870 162763 458989 208575 72547	1107 1461 2415 1803 2694 837 1539 2877	Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1 Accepted: experiment 1
jatgatgatg acgatgatga cigatg gacgacgacg acgatgagtga gacgatgaag gggaagtgga tactg gacgataaca acgaaggaga tggat gacgataaca acgaaggaga gaag gaggagtg acgtgatga gaag gaggagtg acgtgatctag tcca sigcagtgt ccgaaagggt tgcg ggatatata aaggatctag atcag gataatata gaag acgataagaa cagaatcaag gaat gatacatcta caagaacaag tggtc actaccee ctgatcatte tecea actgacaaa atgacaatg ggcaa attfgtcca acagtggtag caaaa actcaatgg tecetgat ttaca	acag 540 tggaa 570 acacg 600 ctgac 630 ggagt 660 ggagt 660 ggagt 640 accca 720 ctatt 760 tgtgg 780 tacag 810 gttet 840 ctggt 870 ccagt 970 agaaa 930 agaaa 960		2 4 4 5 6 6 7 6 9 0 10 0	CDS CDS CDS CDS CDS CDS CDS CDS CDS CDS	Kinase_6 Kinase_0 Kinase_4 Kinase_9 Kinase_2 Kinase_5 Kinase_3 Kinase_1 Kinase_11	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135		575392 23935 353056 886067 160069 458152 207036 69670 970766	576499 25396 355471 887870 162763 458989 208575 72547 973730	1107 1461 2415 1803 2694 837 1539 2877 2964	Accepted: experiment 1 Accepted: experiment 1
jatjatyäty acjatjatja cijatj jagettagaag gogatjatja cijatj jagettagaag gogaagtaga taetg jacgataaca acgaaggaga tyget jacgataaca acgaaggaga tyget jageagetg atgetgatgt geegg toggatatate aaggatetag ateag jatatagaa cagaateag ateag jatatagaa cagaateag ateg jatatatga cagaateag ateg jatatatga aagtagat gege jatataga atag jatataga atag jatataga atag jatataga atag jatataga atag jatataga atag jatataga ateg jatataga atge jatataga atge jatataga atge jatataga atge jatataga atge jatataga atge jatataga atge jataga atge atge atge atge atge atge atge atge atge atge at	accat 640 tggaa 570 acaca 600 ctgac 630 ggaat 660 gtgca 690 accca 720 ctatt 750 tgtg 780 actat 840 ctggt 870 agtaa 900 agaaa 930 aaata 960 ggcaa 990		2 4 4 5 6 6 7 6 9 0 10 0	CDS CDS CDS CDS CDS CDS CDS CDS CDS CDS	Kinase_6 Kinase_0 Kinase_4 Kinase_2 Kinase_2 Kinase_3 Kinase_11 Kinase_10	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135		575392 23935 353056 886067 160069 458152 207036 69670 970766 954666	576499 25396 355471 887870 162763 458989 208575 72547 973730 958203	1107 1461 2415 1803 2694 837 1539 2877 2964 3537	Accepted: experiment 1 Accepted: experiment 1
jatjatgitg acgatgatja cijatg gacgacgacg acgatgagtga tactg gacgatgag gggaagtgga tagtg gacgataaca acgaaggaga tggct gacgataaca acgaaggaga tggct gacgatgatg aggtgggt gggg gggatatac aaggatctag atcag gatacactag acaggagt gggg gggatatac acagaacaag tggtc actacacco cigatatte tccaa contectitiett eteetagat acgg ggttggacaaa atgacatga ggcaa ggttggcaa tectgata ttag gatcactag tecctgata ttacac gtttggcaa atacaggt tggga ggtatatga acaggaagt acgg aacgacaaa atgacatga ggcaa gtttggcaa acaggat ttag gtttggteet tittacette tagat	abgat 540 tggaa 570 acacg 600 ctgac 630 ggagt 660 gggag 660 gggag 680 tgtgg 780 tgtgg 780 tgtgg 780 tgtgg 780 tgtgg 780 tgtgg 870 caaat 910 ggcaa 930 agaat 960 ggcaa 990 tatet 1020		2 4 4 5 6 6 7 6 9 0 10 0	CDS CDS CDS CDS CDS CDS CDS CDS CDS CDS	Kinase_6 Kinase_0 Kinase_4 Kinase_2 Kinase_2 Kinase_3 Kinase_11 Kinase_10	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135		575392 23935 353056 886067 160069 458152 207036 69670 970766 954666	576499 25396 355471 887870 162763 458989 208575 72547 973730 958203	1107 1461 2415 1803 2694 837 1539 2877 2964 3537	Accepted: experiment 1 Accepted: experiment 1
jatjatgistg acgatgatga cigatg jacgatgistg acgatgatga gggaagtgg acgatgaatc cigatc jacgataaca acgaaggaga tactg jacgataaca acgaaggaga tigog aggatgatg cigaaggagg tigog gggatatatc aaggatctag atcag jatatagaa cigaatcag atcag jatatagaa cigaatcag atcag jatatagaa cigaatcag tacg jatatagaa cigaatcag tacg jatatagaa cigaatcag tacg jatatagaa cigaatcag atcag jatatagaa cigaatcag atcag jatatagaa atgacatga ggta jatacactu cicatagag atcgt acgacaaa atgacatga ggca atctgtbut tittacatti taca	acag 640 tggaa 570 acacg 600 ctgac 630 ggagt 660 gggagt 660 gtgca 720 ctatt 750 tgtgg 780 tgtgg 780 tgtgt 870 ccggt 870 ccgat 900 agaaa 930 agata 960 ggca 990 tatet 1020 ctat 1050		2 4 4 5 6 6 7 6 9 0 10 0	CDS CDS CDS CDS CDS CDS CDS CDS CDS CDS	Kinase_6 Kinase_0 Kinase_4 Kinase_2 Kinase_2 Kinase_3 Kinase_11 Kinase_10	Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135 Y13135		575392 23935 353056 886067 160069 458152 207036 69670 970766 954666	576499 25396 355471 887870 162763 458989 208575 72547 973730 958203	1107 1461 2415 1803 2694 837 1539 2877 2964 3537	Accepted: experiment 1 Accepted: experiment 1

Figure 3-7. The Multiplex PCR analysis window.

Edit										
*	🌆 🔍 - 🖓 -	G thit /#								
:587 :583	91 ATT4 A A A T G A A A T G A A A T G A A A T G A A A T T A A A A		T A A A C A G A G T A A G C A G A G	BATACGI ATAAGI			BAAAT BAAAA	20 25 30	35 40	45 8
	30000 40000 50000 1	60000 70000	80000 90000	Features	Feature multipl	lets				
E.	a de la deserve	a la se la	in addate	Seq 1	Seq 2	matches / length	% matches	Significance	Distance f	to dia.
100	这个问题,我们就是我们的 是	100000	计算机 医子宫上的	58523	58107	101 / 115	87.8	1.03e-045	416	^
15	and the second	1 - 2 - 2 - 2	CONTRACTOR OF	58669	58250	84/90	93.3	3.03e-043	419	
	14日1日年1月1日日1月1日日	1.10 10.00	an and	11610	10451	95 / 111	85.6	5.51e-041	1159	
1.	and the second second second	1 2 2 2	10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	58756	58338	82 / 100	82	7.97e-033	418	
1.0	and the second second second		Contraction of the local sectors in the	58791	58373	59/65	90.8	4.58e-029	418	
	中国建物的中国管理和任何	1	Are, When the	45681	45677	45 / 48	93.8	6.02e-024	4	
1.00	and the second second	1000	治疗法治 中的	1806	1782	48 / 54	88.9	6.05e-023	24	
100	教育の大学の教育			45687	45681	40 / 42	95.2	4.07 e-022	6	
1	the second second second	A REAL PROPERTY.	10 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	45675	45673	47 / 54	87	1.25e-021	2	~
-		CONTRACTOR OF STREET		<	4500.4		~~~		-	>
. <		1	>						_	
₹.				CDS						1
_	<i>1</i>			808						2 I I
-				19-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-1-						
->	CDS									
	6) 6)									
-4	CDS									
										£
	CDS						_			
	4									
	<u>7</u>	_								

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 **S** 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>*.

Unlik 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 **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).

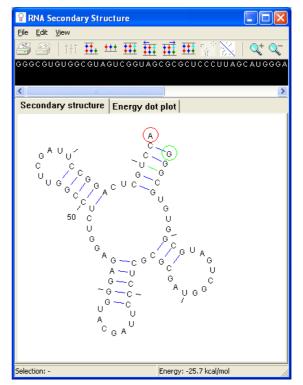


Figure 3-9. The yeast alanyl-tRNA secondary structure prediction.

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	‡ ‡₊	+++	‡‡ ‡	Ī	Ħ	‡‡ ‡	and
recalculate the se	econda	ary st	ructur	re by p	oressii	ng	2.

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 \times 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 **b** 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

20 40 60 80 100 120 140

on 📒 to center the selection.

Percentage

Curve name

3.7.10 Press the $\stackrel{\leftarrow}{=}$ and $\stackrel{\rightarrow}{=}$ 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 open a new cloning project.

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

The upper panel shows a graphical respresentation 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 **t** button (or select *Enzyme* > *Add all*).

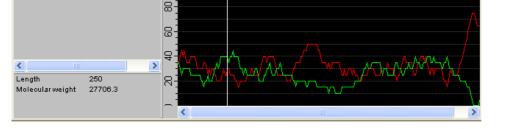


Figure 3-10. Curves representing the relative contribution of selected amino acids in a moving window.

160 180

200 220

to

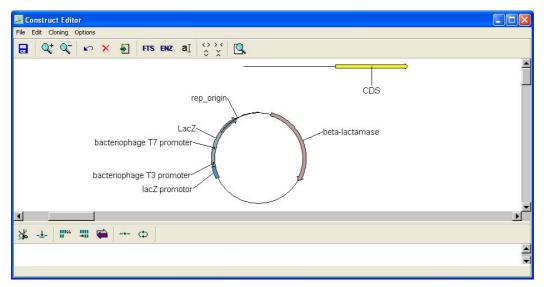


Figure 3-11. The Construct editor window.

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 (\bigcirc) 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 ezymes 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

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-SacIfragment) and then select the shorter fragemnt (SacI-NcoI-fragment) using SHIFT-click.

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

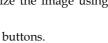
NOTE: If the fragments were selected in the opposite order, the ligation would fail.

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 \mathbf{Q}^{-} , \mathbf{Q}^{+} , and



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

3.8.23 Close the **Demobase**.

> <

╸┋╸┽╡╡╎	🔤 🖹 I#	-t><_t₽∷ ⅔+							
	AccI	1	'0 '20 	0 400	500 80	0 1000	1200 14	400 150	0
AccI 🔺	AluI	11	++		+		+		_
AluI	ApaI	0					-	-	_
ApaI	AvaI	1							_
AvaI	Ball	0							_
BalI	BamHI BclI	0					1		_
BamHI	BglII	3				i			_
BelI	CfoI	1							_
BglII	ClaI	2							_
CfoI	DdeI	9			+				+-
ClaI 🗨	DraI	0							_
		🕻 📰 🛛 🕵							
				1300			ро Г	1600	
00 1			1200	1300	1400 • • • • • • • • • • • • •	15	00	1600	
							00	1600	· · · · ·
			1200				00	1600	· · · · ·
			Fragment				0	1600	· · · · ·
cos ymes selected	. 1000				 3' Cleavage 14/10 	5ac I	length:	14 bp	· · · · ·
	5'Cleavage cut1: cut1:	0/0 14/10	Fragment Sac I	cut2: cut2:	• 3'Cleavage 14/10 925/921	Sac I Sac I	length: length:	14 bp 911 bp	· · · · · ·
ymes selected	5'Cleavage	0/0	Fragment	cut2:	 3' Cleavage 14/10 	SacI	length:	14 bp 911 bp 765 bp	· · · · · ·

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

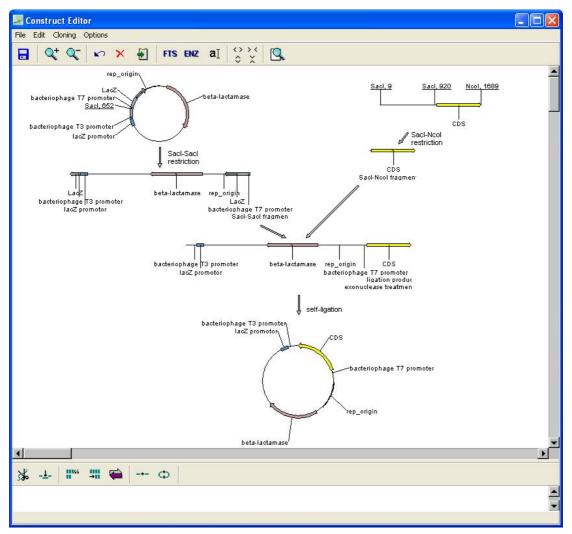


Figure 3-13. A vector construction project.

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 \longrightarrow 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).

Comparative Ch	romosome Mapping				
File Matrix					
貸 A: AI	9				
💽 Matrix	Dot plot	Cell			
·Status: finished	•Extending loci 99%	Accession	Identification	Organism	•Description
U00096	9.8 89 84 47 50 77	- U00096	U00096	Escherichia coli K12	Escherichia coli K-12 MG1655 complete genome.
BA000007	89 19 93 45 46 77	- BA000007	BA000007	Escherichia coli 0157:H7	Escherichia coli 0157:H7 substrain RIMD 0509952, compl
AE005174	84 93 22 44 49 77	- AE005174	AE005174	Escherichia coli 0157:H7 ED	Escherichia coli O157:H7 EDL933, complete genome.
AL513382	47 45 44 9.3 72 45	- AL513382	STYPHCT18	Salmonella enterica subsp. e	Salmonella enterica serovar Typhi (Salmonella typhi) strai
AE006468	30 48 49 72 9.0 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.
	U00096 BA00007 AE005174 AL513382 AL513382 AE006468 AE005674				
					<u> </u>
•Calculation status: 10	00.00%				

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

File Matrix						
	2					
🔝 Matrix	Dot plot	Cell				
		•	Accession	Identification	Organism	Description
	AL513382 AE006468 BA000007 AE005174 U00096 AE005674	A 10000001 A 10000000 A 100000000 A 100000000	AL513382 AE006468 BA000007 AE005174 U00096 AE005674	STYPHCT18 AE006468 BA000007 AE005174 U00096 AE005674	Salmonella enterica subsp. e Salmonella typhimurium LT2 Escherichia coli 0157:H7 Escherichia coli 0157:H7 ED Escherichia coli K12 Shigella flexneri 2a str. 301	Salmonella enterica serovar Typhi (Salmonel Salmonella typhimurium LT2, complete geno Escherichia coli 0157:H7 substrain RIMD 05 Escherichia coli 0157:H7 EDL933, complete Escherichia coli K-12 M01655 complete geno Shigella flexneri 2a str. 301 complete genom

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.

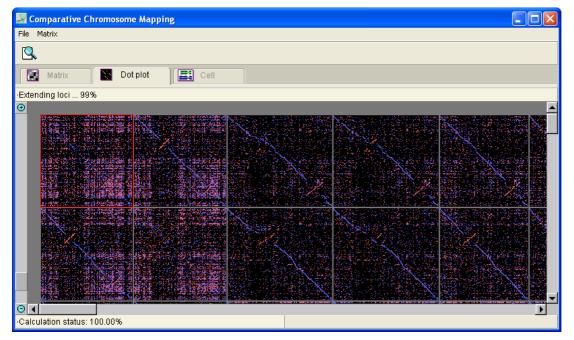


Figure 4-3. The Dot plot view in a comparative chromosome mapping project.

liot 🔛 Cell							
		Pos	1	Pos2	Lenath	Identity	
		903	86	554716			
		904		555388	155		İ
		905		555517	260	75	1
		906	53	555763	136	75	
		907	114	556022	153	74	
		908	:50	556258	115	70	
	100 C	909	12	556720	128	73	
		910		557929	669	80	
		911					
		912					
	14 C		1 · · ·				
<u>X</u>		917	114	2841040	100	12	
							<u> </u>
rotein (553500 / 554000 / 55	4500	hypotheti	cal protein putative lig 0	ase hyp ,556500,55	othetičal protCu 7000	(I)-translocation P-type A 00	,558500
	000 .570500 .571000 .57 rotein .553500 .554000 .55	000 .571000 .571000 .572000	Pos 903 904 905 906 907 908 909 910 911 912 913 914 905 907 908 910 911 912 913 914 915 916 917 ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●	Pos1 903 86 904 78 905 07 906 53 907 14 908 50 909 12 910 48 911 153 912 154 913 145 914 40 915 56 918 75 917 114 114 114 115 56 917 114 114 114 114 114 114 114 115 56 917 114 114 114 114 114 114 114 114 114 114 114 114 114 114 114 115 56 115 56 114 114	Pos1 Pos2 903 86 554716 904 78 555388 905 107 565617 906 53 555763 907 14 565022 908 50 565258 909 12 556720 910 48 567929 911 53 770726 912 554 400529 913 45 556826 914 40 558821 915 56 560134 916 75 560483 917 14 2841040 14 0 573000 573500 57 000 ,570500 ,573000 ,573000 57 000 ,570500 ,573000 ,573000 ,573500 ,57	Pos1 Pos2 Length 903 86 554716 540 904 78 555388 155 905 107 556517 260 906 53 555763 136 907 14 556022 153 908 50 556258 115 909 12 56720 128 910 48 557929 669 911 53 770726 91 912 54 400529 82 913 45 558626 203 914 40 558821 1350 915 56 560134 327 916 75 560483 77 917 14 2841040 100 000 570500 57200 57300 57400 57400 000 570500 555000 555000 555000 555000 557000 57200	Pos1 Pos2 Length Identity 903 86 554716 540 77 904 78 555388 155 72 905 107 655517 260 75 906 53 555763 136 75 907 14 556022 153 74 908 50 556258 115 70 909 12 56670 128 73 910 48 557929 669 80 911 53 770726 91 71 912 54 400529 82 70 913 45 558626 203 74 914 40 558821 1350 84 915 56 560134 327 79 918 75 560483 77 70 917 14 2841040 100 72 916 55

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 parellelism 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 **unit** 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.

6	Alignments Mutations	Options										
	A tgac N	T	7									
	Template: U00096		.500000	,10000	00 .150	0000	.2000000	,2500000	.3000000	,3500000	.4000000 .4	45000
E	9A000007 (75.0)	1000	081 T	n		X		1 P - 12	1111 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1	10 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
1	AE005174 (74.0)	100 2 3	141 T 1 146	an an er eg	Successive free	VALUE AND A	allout a Second of	1.19.11.1.1.1	real real of the second	11.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1	an yan an and an an	-
7	AL513382 (53.D)	party sections of	and I happene service and it and	manantalia deles auda			WHAT IS ANY IN THE SAME	autorial laness and slope .	antitations, ships, countration in	W. In pre-states we design to defend	Talifetania provide and an addition of	Kaday
7	AE006468 (54.0)	where the articles is	A Property of the second second second	entine-trajes via an and	Number of the state	ale maintain on the	NAMES OF TAXABLE PARTY OF	a state latter and shipe	water and the same of the second	Will be seen to a south without	Tablecteride/Antralizitaneni unattifalcan (Apr hart
7	AE005674 (67.0)	1.72	10 T 10 T 10 T 10 T	11 Acres 10 - 5		A THE OWNER OF	when it was a second	× 1.765	A. S. M. S. Market Street, Name	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Sector of the sector	7.177
-	10.0											
	,2000		,4000	,6000		,8000	, 10000	a general a serie	,12000	,14000	,16000	_
	aspartokinase I, homoseri	bomose	threepine sup	r off b	inner membrane	transaldo	requirer off	posi orf	chanempe Hsp70:	chanerone in L	IS186 Gef n N	
	,2000										,160	
	-1-12				-							
	aspartokinase I-homoseri.										Gef p N. ,160	
		N	. 14000	10000		loning .	· · · · · · · · · · · · · · · · · · ·	Carlos And	12000	14000		00
	aspartokinase I, homoseri	homose.	. threonine syn o	r orf, h	inner membrane	transaldo	. requ orf,	posi orf,	. chaperone Hsp70;	chaperone	Gef p N.	
	,2000		,4000		Acres and	,8000			12000	,14000	100	
	aspartokinase I/hornoseri.								Pro Konstala And	- Deside		
	,2000				putative ami	transald ,8000	mo co	con	DnaK protein (heat 12000	. DnaJ p 14000	N.	
		No						- Care - Ca -				in l
	aspartokinase I		. threonine s	putati	putative alan			posit	chaperone Hsp70	. heat s		
	,2000		,4000	,6000		,8000	,10000)	,12000	,14000		
			. threonine syn o		van L	Chain	Mag of	autor of	. chaperone Hsp70;	abanamana ui	N	- 8

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). 4.2.9 Press the

m

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>*.

Entries to be searched U00096	Features to be searcher	1
BA000007	-35_signal	
AE005174 AL513382	CDS misc_feature	
AE006468	misc_reactive	
AE005674	promoter	
	protein_bind RBS	
	rep_origin	
1	repeat_region	~
Text to be searched kinase		

Figure 4-7. The Find dialog box.

7	Po	sition on entry	Entry	·Position on alignment	Mismatches	Match	Presence on alignment
1	=>	614312	AL513382		0	gatgaatgatgg gatgaatgatgg	<u> </u>
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	٠
6	+	3784873	AE006468	3674379	0	gatgaatgatgg gatgaatgatgg	•
7	-	3730079	U00096	3730079	0	gatgaatgatgg gatgaatgatgg	•

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

A	Featu	res	×					
	+ X	↓a	1 🖻	7				+ × ab <u>[</u>
Feat 1 2 3 4 5 6 7 8 9 10 11		CDS CDS CDS CDS CDS CDS CDS CDS CDS CDS	Entry × U00096 U00096 U00096 U00096 U00096 U00096 U00096 U00096 U00096 U00096	Start × 337 2801 68348 157253 191855 259612 339389 405629 434858 496399 499349 349349	Stop 2799 3733 70048 157732 192580 260715 340339 406153 435835 497043 50653	1 1 1 1 1 1 1 1 1 1	/product= aspartokinase I, homoserine dehydrogenase I homoserine kinase L-ribulokinase 7,8-dihydro-6-hydroxymethylpterin- pyrophosphol- uridylate kinase gamma-glutamate kinase putative kinase (EC 2.7.2.2). shikimate kinase II thiamin-monophosphate kinase adenylate kinase activity; pleiotropic effects on gly inosine-guanosine kinase	<pre>28013733 / codon_start=1 / db_xref="SWISS-PROT:P00547" / note="0310; 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="ALC73114.1" / translation="HVKVYAPASSANMSVGFDVL(TPVDGALLGDVVTVEAAETFSLNNLGFFACKLPS]</pre>

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 **tgac** 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 **b**utton to find the next mutation downstream the current cursor position.

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

\sim	 Position 	Mutation	Entry	DNA cł	nange	Am	nino :	acid cha	Gap
49	152	silent	AE006468	g →	a				
50	153	silent	AL513382	c →	g				
51	153	silent	AE006468	c →	g				
52	161	silent	AL513382	g →	a				
53	161	silent	AE006468	g →	a				
54	170	silent	AL513382	t \rightarrow	g				
55	170	silent	AE006468	t \rightarrow	g				
56	194	missense	AL513382	a →	с	К	\rightarrow	N	
57	194	missense	AE006468	a →	с	К	\rightarrow	N	
58	200	silent	AL513382	t \rightarrow	с				
•	1								_

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

Chrome	osome Alignme	nt:																		
File Alignm	ients Mutations	Options																		
	h 🔶 📓	7 7																		
CCACCAC TCACCAC TCACCAC CCACCAC CCACCAC CCACCAC CCAACC CCAAACC CCAAACC	ູ່ເຈເບງ	Techecta Techecta Techecta Techecta Teatecta Teatecta Teatecta Teatecta	CLOPECTORIA	Fccfccf Fccfccf Fccfccf Fccfccf Fccfccf F	стяссія собсейс собсе собсейс собсейс собсейс собсе собсейс собсе собсейс собсе собс собс	C TEEEC C TEEEC C TEEEC C TEEEC C TEEEC 23210 TF C TEEEC 23210 TF C TEEEC 23210 TF C TEEEC C 23210 TF C 23210 TEEEC TEEC TEEEC TEEEC TEEEC TEEEC TEEEC TEEEC TEEEC T	AGCCAG AGCAAA AGCAAA AGCCAA AGCCAA CGCTTT CGCTTT CGCTTT CGCTTT	ce a ce	Teečee Sievitvi Sievi	CGAATCC TGAAGCC TGAAGCC TGAAGCC CGAATCC ACGCCT ACGCCT ACGCCT ACGCCT CACGCCT CACGCCT CACGCCT CACGCCT CACGCCT	CAÇCAN CAÇCAN CAÇCAN CAÇCAN CAÇCAN CAÇCAN GOŢTAŢ GOŢTAŢ GOŢTAŢ	сабалте сабалес сабалес сабалте Сесетес сесстес сесстес	CACCAC CCCCCC CCCCCC CCCCCCC CCCAC CCCAC CCCAC CCCAC	ардаран арадас арадас арадас арадас арадар ара ара	Stattg Stattg Stattg Stattg Stattg Cattg Icattag Icattag	eticei aticei aticei eticei eciecei	La Titic: La Titic: Le El Céc Ce El Céc Céc Céc Céc Céc Céc Céc Céc Céc Céc	Sceiece Sceiece Sceiece Sceiece Sceiece Sceitati Scitati Sceitati Sceitati	cchtvicci ccétvévci ccéceptci ccéceptci ccéceptci ccéteptci	ICC CCC CCC CCC CCC CCC CCC CCC CCC CCC
	itions ×																			
\sim	Position	Mutation	Entry	DNA ch	ange	Amino	acid ch	na Gap												
123555		missense	AE006468	с →	a	$\Gamma \rightarrow$	I													
123556		missense	AL513382	g →	с	$r \rightarrow$	I													
123557	10000000	missense	AE006468		с	$r \rightarrow$	I													
123558		silent	AL513382	g≯	t															
123559	0.757.718	silent	AE006468	g ⇒	t															
123560 123561		silent silent	AL513382 AE006468		t															
123561		silent	AE000408 AL513382		t c															
123002	1423210	andit	ALUT3302	a≯	C															
121																				

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



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.

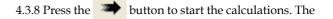
4.3.4 In the *Annotation* window, select *File* > *Open reading frame settings*.

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>*.



button will change in a 🛛 🗱 button.

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: $\sim 100\%$ identification score, green $\sim 50\%$). 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*



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

Annotation			
File Feature Options			
	* > 4 88		
	L - L 1914 LINDU U 110 - U 101- U 101- 1-13 - L - L - L - L - L - L - L - L - L -	\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\	
6 1 2	4 2 45 5 4 6		5 318 A412 A8 15
	1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 +	<mark>(11)()()()()()()()()()()()()()()()()()(</mark>	
Tined by Kodon	Ar orf, conserved hypothetical protein (99.2) putative tra putative DedA family membrane protein (proton e putative	ansport protein transport protein (98.0) I AED/0647 Hitw pump (00.5) I AED/0648 transport (90.1) AED/0648 d hypothetical protein (21.3) I AED/0 activity and transport (41.1) I AED/074 i I AED/07	transori
NA polymera. Letto. Larabin. Lettodokinase DNA polymera. Letto. Larabin. Lettodokinas DNA polymera. Letto. DNA polymera. Letto. DNA polymera. Letto. DNA polymera. Lettodokinas put	pu. putative tham. putative transport pr (0.03.) I.A putative than. putative transport pr (0.03.) I.A putative than putative transportsm tham tham that putative dispetition to universe Top 	i 3-isop 3-is 2-isopropyimalate synthase (99.4) AE0056	88 52 22
A polymera	orf pU thiami i	3-isopropy1 proba u b	ransorip
iApolymera (nb.). (arabino(nbulokinase	pu putative tthiami putative transport pros		anson, () (abC,) () testum for () UDP.N. () Datani, ()
⊖ ◀ J •Status: finished Que	rv sequence: U00096	Annotated sequence: no entry defined	<u>.</u>
Guido, Infondu Gue	y ooquonee000000	ramounes sequence, no entry defined	

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.

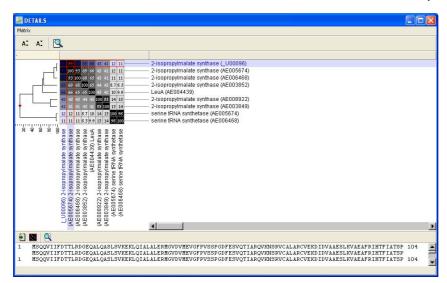


Figure 4-12. Detailed comparison of a single annotated CDS.