Dover Analyzer
User's Manual

By developer team

Table of Contents

GENERAL INFORMATION

System Overview

Dover Analyzer is a wizard like application that takes collections of databases annotated in FASTA format and guides the user through a few steps to compute the overlap,

diversity and non-redundant sets of peptide sequences. It is implemented in Java to achieve platform independence and has been designed initially to analyze the publicly available antimicrobial peptide databases. However, additional analysis can be done by simply adding new databases or replacing the existing ones.

System requirements

- Hardware
  - Memory (RAM): 4 GB or above.
  - Processors: We recommend a multi-core processor due to the fact that the software has been implemented to enable parallel processing of computationally intensive tasks.
  - Hard Disk: a minimum of 130 MB of free space is required for the output files with the computed results.
- Software
  - Java(TM) 7 Runtime Environment or superior version on the target system.

Downloading the program

Dover Analyzer is available for download at http://mobiosd-hub.com/doverAnalyzer. It is provided as one zip file and all that is to be done is to download and unzip this file into a directory (<destiny-directory>) located in the computer where the users intends to run the program.

Running Dover Analyzer

Dover Analyzer must be run from a command line to allocate more memory than the default. The amount of demanded memory depends on the size of loaded data to be processed. Once the command window has been opened, the command may be executed as follows:

```
cd <destiny-directory>
java –Xmx4G -jar DoverAnalyzer.jar
```

First, you have to switch to the directory <destiny-directory> where the program Dover Analyzer.jar is located, by using the change directory (cd) command. Once in the <destiny-directory>, the user may run the java program by passing some arguments, after a hyphen (-), to the java virtual machine. The first option Xmx4G allows the use of 4GB of RAM as the maximum allowable memory. This amount of memory should be increased when an out of memory error is thrown upon adding more databases for analysis. The second option *jar* is to specify the program to run. There must be a blank character between the option named *jar* and its value DoverAnalyzer.jar.

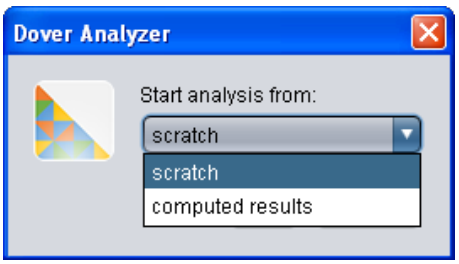
Windows users can start the program with 4 GB of main memory assigned by double clicking on the Run-4GB.bat file. In the same way GNU Linux users may run the Run-4GB.sh file. These files are inside the downloadable zip archive.

User Access Levels

Dover Analyzer program is free of charge and can be used by everyone. The source code is available upon request to the authors.

Getting started

This section explains how to use the Dover Analyzer program through the wizard and the results explorer window of the application. When the program is run, a loading splash screen is shown while the databases are being loaded. Once the application is ready; the user is prompted to start the analysis from scratch or using previously computed results. The analysis from scratch initializes the wizard while the other option for analysis displays a file chooser to select the directory whose content are the computed results. If the selected directory contains some computed results then the result explorer window will be displayed.

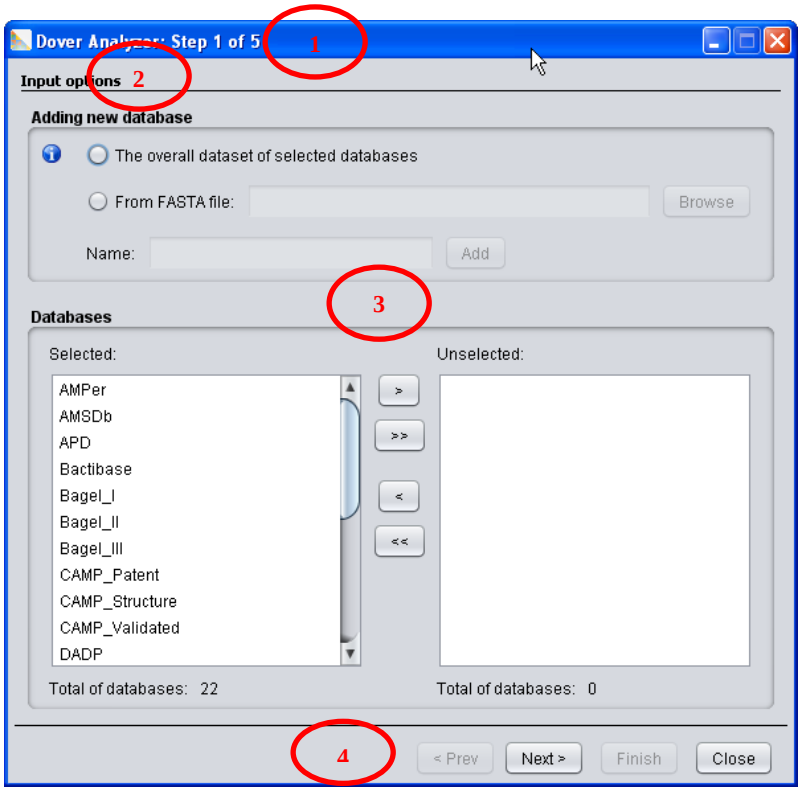


Dover Analyzer wizard

The Dover Analyzer wizard is visible if the “Start analysis from scratch” option is selected. A wizard page is comprised of 1) the title “Dover Analyzer: Step <no. of step> of <total of steps>” in the windows title bar, 2) a description text of the step in bold font, 3) the content area of the page between two separator lines to capture the user requirements and 4) the footer navigation. The footer navigation is static while the other parts of the page vary depending on the step. The “<Prev” button goes back and the “Next>” button will take the user to the next page if the content items on the current page are valid. The “Finish” button is available in the last step after the computations are made and the “Close” button is always available to close the wizard application. The following subsections show the wizard pages and what may be done, step-by-step.

Input options

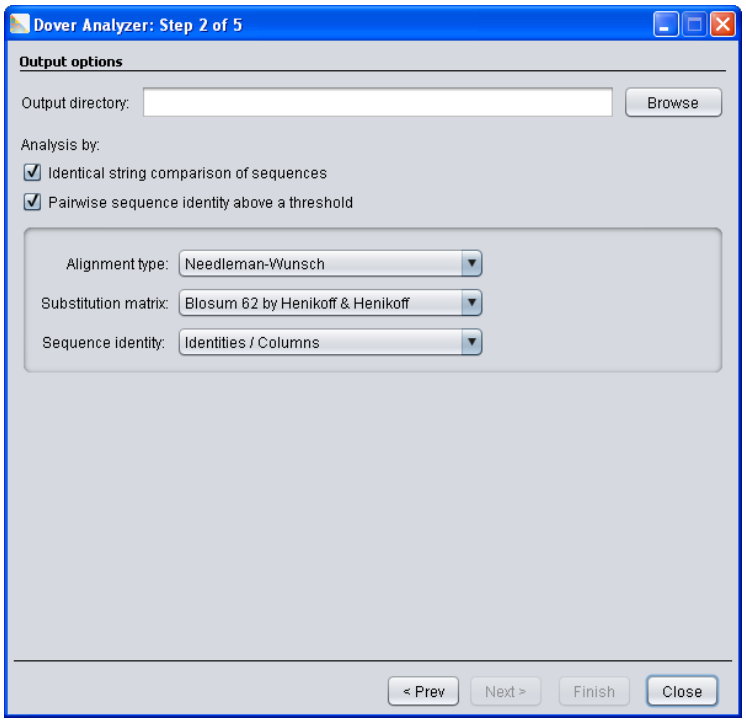
In the first step the user may select or deselect the databases to analyze. These databases are embedded into the application but in addition one may add the entire set, which is comprised of all entries of the selected databases, or a new database from a FASTA file. The additional databases are used only for the current analysis. If one wishes to include a new database permanently, this has to be done manually by adding the FASTA file into the “db” folder located in the same <destiny-directory> where the DoverAnalyzer.jar file is. Note that the name of the FASTA file corresponds with the name of the database in the “db” directory.



1. Windows Title: “Dover Analyzer: Step <no. of step> of <total of steps>”.
2. A description text of the step.
3. The content area of the page between two separator lines.
4. The footer navigation.

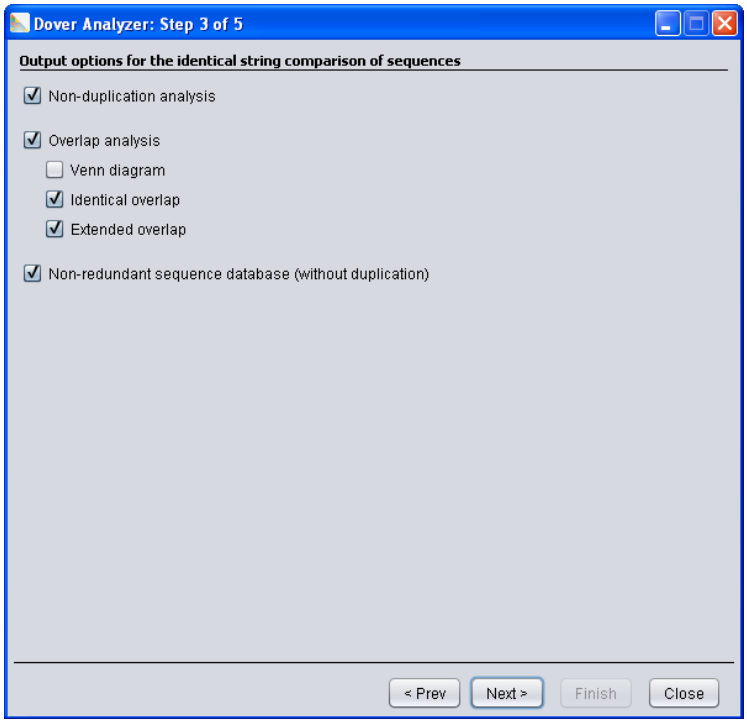
Output options

In the second step the user specifies an output directory and a criterion of comparison. There are two criteria to compare the entries in databases: 1) Identical string comparison of sequences or 2) Pairwise sequence identity above a given threshold. If the second option is chosen, then configurations may be carried out to set the sequence alignment type, the scoring matrix and the pairwise sequence identity definition.



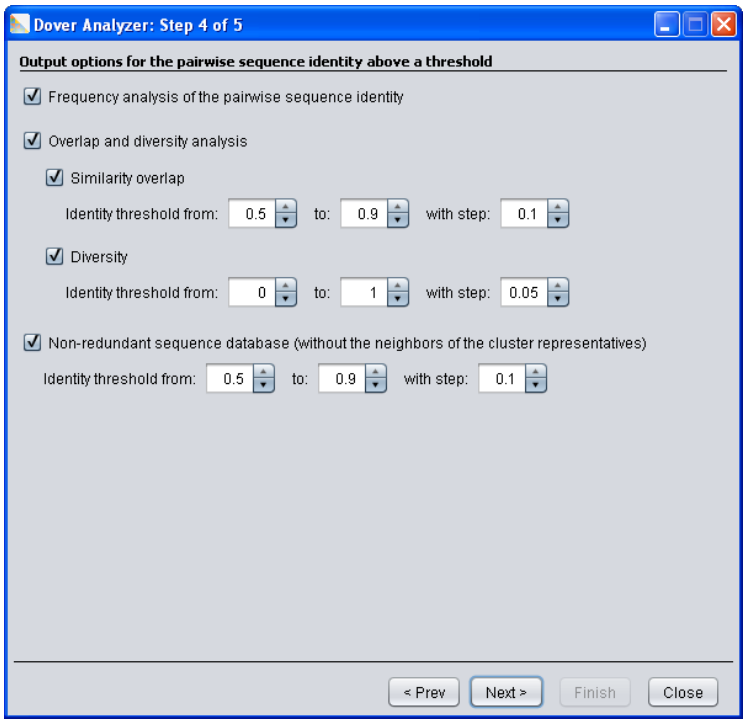
Output options for the identical string comparison of sequences

If the “Identical string comparison of sequences” was selected in the step 2, then the wizard page in the step 3 displays the output options for this criterion. With this type of comparison, the analysis for non-duplicate sequences, Venn diagram, identical or extended overlap and data sets without duplication of sequences could be computed. The identical overlap illustrates the sequences shared between two data sets while the extended overlap shows the sequences stored exclusively in a given number of databases.



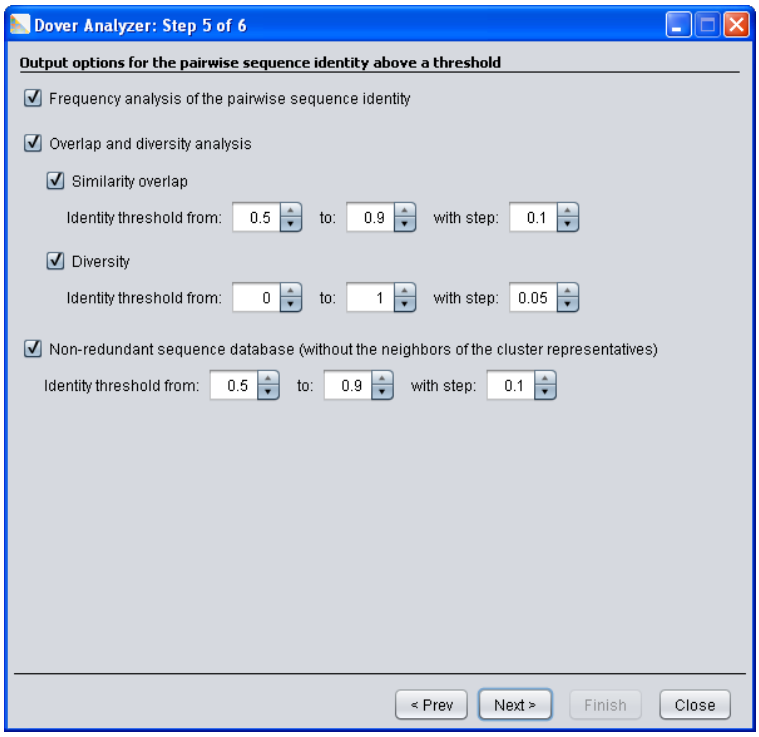
Customize Venn diagram

If the “Venn diagram” option was selected in the previous step, then this diagram may be customized in the following wizard page. Clicking the Add or Del button allows the user to specify the databases for the different sets. The label for each set may be modified as well.



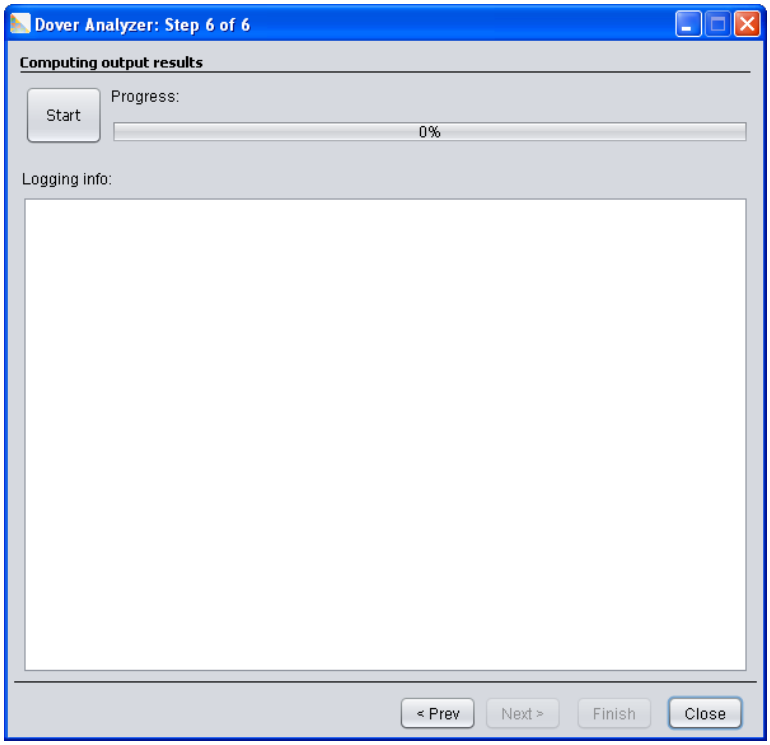
Output options for the pairwise sequence identity above a threshold

If the “Pairwise sequence identity above a threshold” was selected in the step 2, then the wizard page in this step displays the output options for this criterion. Within this criterion, the analysis for the frequency of pairwise sequence identity, similarity overlap, diversity ratio and non-redundant sequence database could be computed. Note that the identity threshold may be configured to change from a low to a high value, using a given step size.



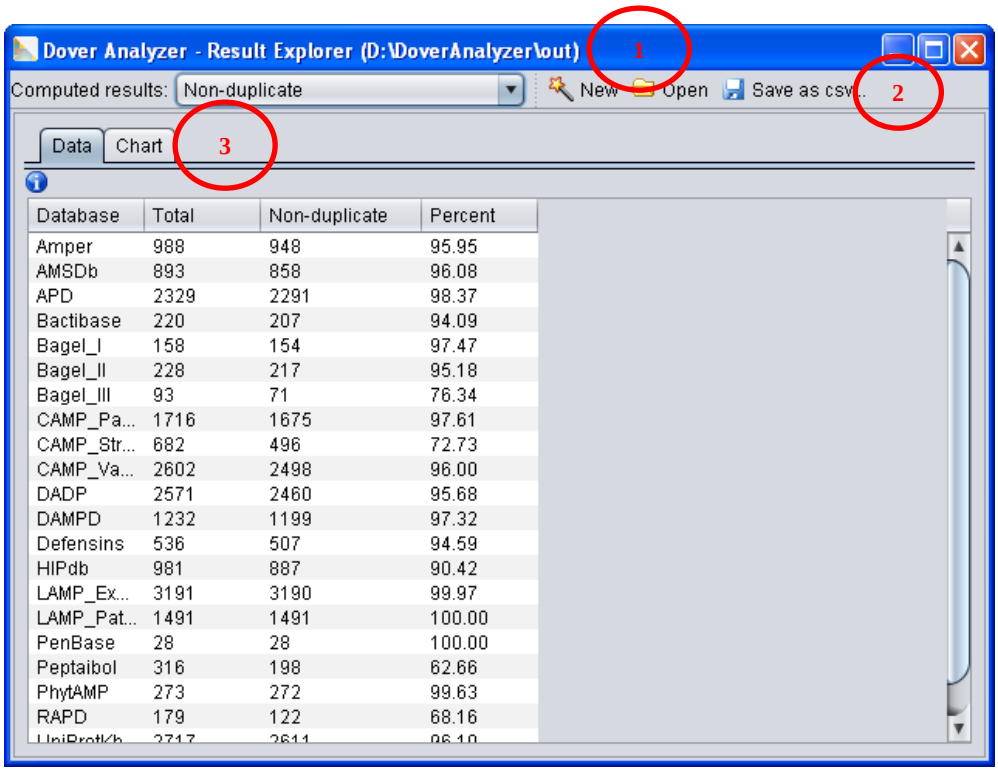
Computing the output results

Finally, the “Start” button will compute the results according to the specifications in previous steps. The logging info area will display messages about the running process and the progress bar indicates the percentage of task completion. When the task is completed the Finish button is activated and the user may click on it to open the result explorer window or close the application.



Dover Analyzer results explorer window

The result explorer window may be opened either from the initial dialog box by selecting “Start analysis from computed results” or via the “Finish” button of the last step of the wizard. This window is comprised of 1) the title, which is of the form “Dover Analyzer – Result Explorer (<path_of_the_folder>)” to visualize the path of the folder that contains the computed results, 2) the tool bar and 3) the tab panel where the data and chart are shown.

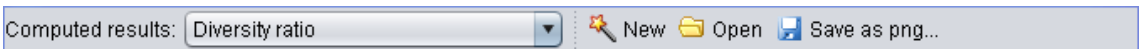


Database	Total	Non-duplicate	Percent
Amper	988	948	95.95
AMSDb	893	858	96.08
APD	2329	2291	98.37
Bactibase	220	207	94.09
Bagel_I	158	154	97.47
Bagel_II	228	217	95.18
Bagel_III	93	71	76.34
CAMP_Pa...	1716	1675	97.61
CAMP_Str...	682	496	72.73
CAMP_Va...	2602	2498	96.00
DADP	2571	2460	95.68
DAMPD	1232	1199	97.32
Defensins	536	507	94.59
HIPdb	981	887	90.42
LAMP_Ex...	3191	3190	99.97
LAMP_Pat...	1491	1491	100.00
PenBase	28	28	100.00
Peptaibol	316	198	62.66
PhytAMP	273	272	99.63
RAPD	179	122	68.16

1. Title: “Dover Analyzer – Result Explorer (<path_of_the_folder>)”
2. Tool bar
3. Tab panel

Tool bar

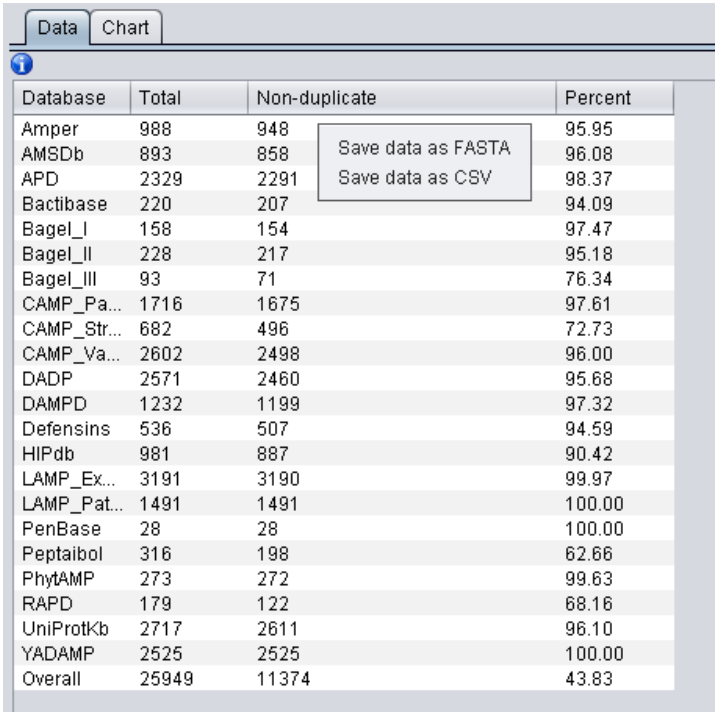
A tool bar shows from left to right the options to explore the computed results and the buttons to display the wizard for new computations, open another folder with computed results and save data as comma separated values (CSV) file or chart as PNG file.



For each computed result selected in the toolbar, the main content area of the window will change to present, in most cases, a tab-panel with two tabs: Data and Chart.

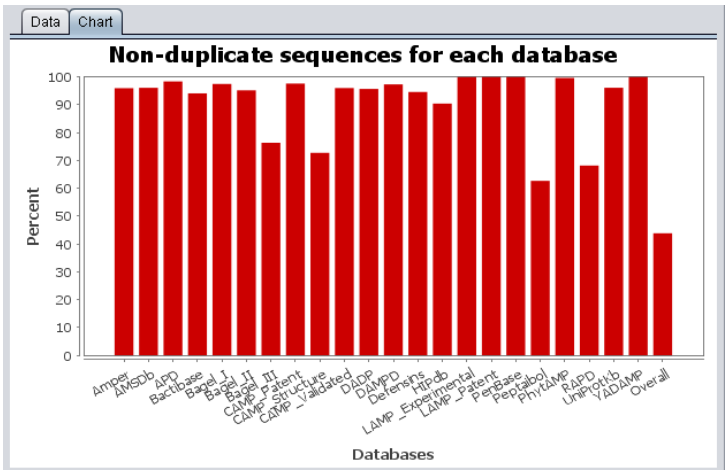
Tab panel: Non-duplicate

The Data tab displays the number and percentage of non-duplicate sequences for each database. Also the non-duplicate sequences for a particular database can be exported by right clicking on a cell of the Non-duplicate column.



Database	Total	Non-duplicate	Percent
Amper	988	948	95.95
AMSDb	893	858	96.08
APD	2329	2291	98.37
Bactibase	220	207	94.09
Bagel_I	158	154	97.47
Bagel_II	228	217	95.18
Bagel_III	93	71	76.34
CAMP_Pa...	1716	1675	97.61
CAMP_Str...	682	496	72.73
CAMP_Va...	2602	2498	96.00
DADP	2571	2460	95.68
DAMPD	1232	1199	97.32
Defensins	536	507	94.59
HIPdb	981	887	90.42
LAMP_Ex...	3191	3190	99.97
LAMP_Pat...	1491	1491	100.00
PenBase	28	28	100.00
Peptaibol	316	198	62.66
PhytAMP	273	272	99.63
RAPD	179	122	68.16
UniProtKb	2717	2611	96.10
YADAMP	2525	2525	100.00
Overall	25949	11374	43.83

The Chart tab shows in a bar graph the percentage of non-duplicate sequences for each database.

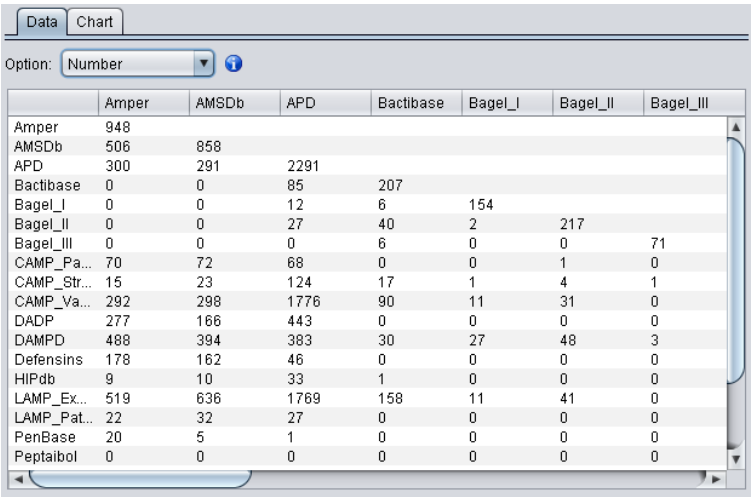


Tab panel: Venn diagram

The Chart tab displays the Venn diagram and the user may save the intersected sequences by right clicking on the numbers.

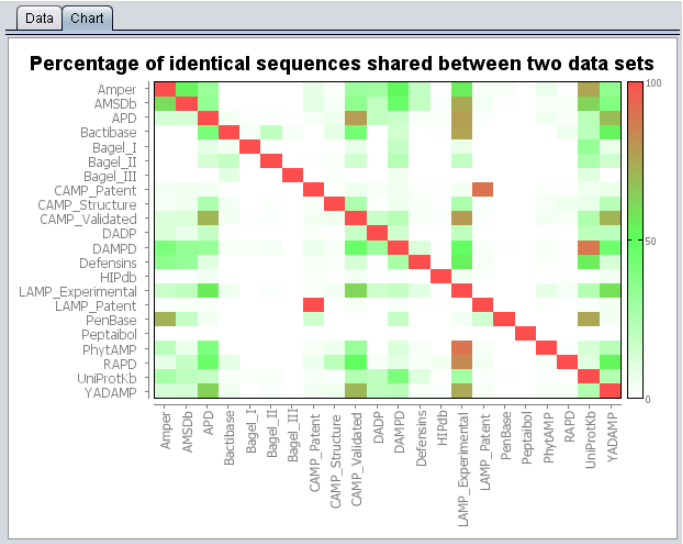
Tab panel: Identical Overlap

The Data tab displays the number and percentage of sequences in a row database for which there is an identical sequence in a column database. The overlap sequences can be exported by right clicking on a non-empty cell. For the number option this table is symmetric and only the lower triangular part is used.



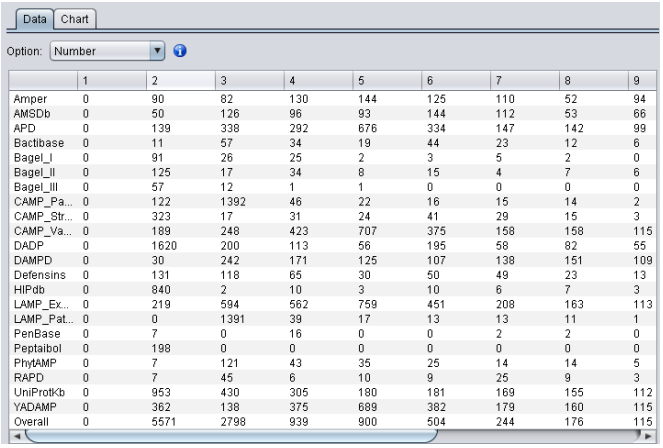
	Amper	AMSDb	APD	Bactibase	Bagel_I	Bagel_II	Bagel_III
Amper	948						
AMSDb	506	858					
APD	300	291	2291				
Bactibase	0	0	85	207			
Bagel_I	0	0	12	6	154		
Bagel_II	0	0	27	40	2	217	
Bagel_III	0	0	0	6	0	0	71
CAMP_Pa...	70	72	68	0	0	1	0
CAMP_Str...	15	23	124	17	1	4	1
CAMP_Va...	292	298	1776	90	11	31	0
DADP	277	166	443	0	0	0	0
DAMPD	488	394	383	30	27	48	3
Defensins	178	162	46	0	0	0	0
HIPdb	9	10	33	1	0	0	0
LAMP_Ex...	519	636	1769	158	11	41	0
LAMP_Pat...	22	32	27	0	0	0	0
PenBase	20	5	1	0	0	0	0
Peptaibol	0	0	0	0	0	0	0

The Chart tab shows the percentages of the row database included in the column database.



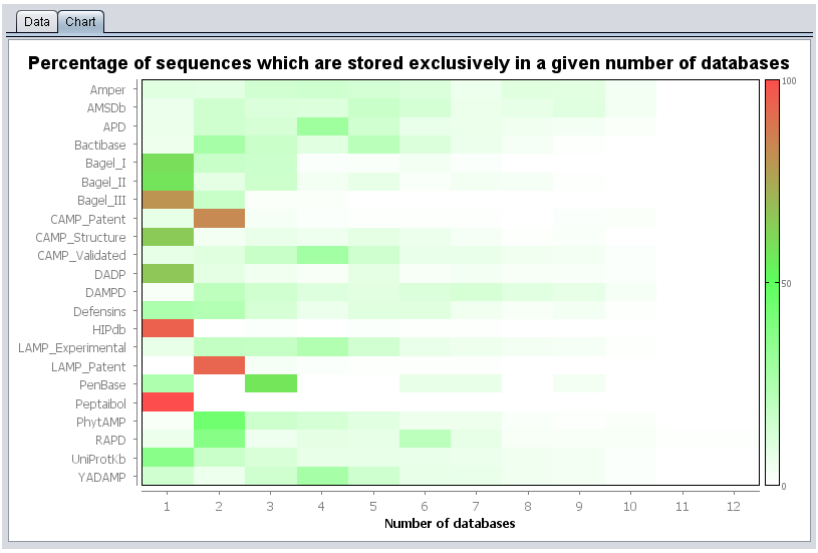
Tab panel: Extended overlap

The Data tab displays the number and percentage of sequences which are stored exclusively in a given number of databases.



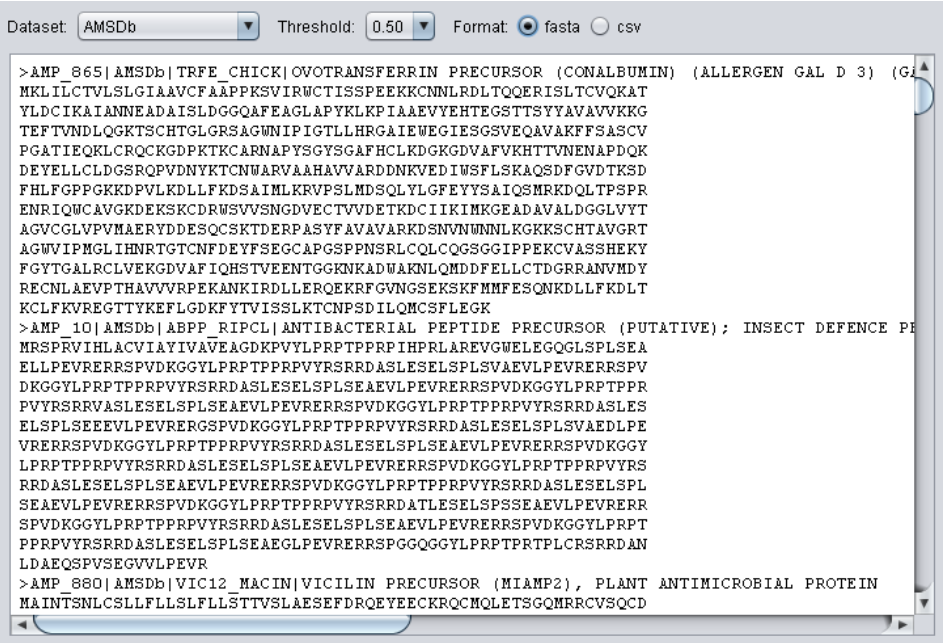
	1	2	3	4	5	6	7	8	9
Amper	0	90	82	130	144	125	110	52	94
AMSDb	0	50	126	96	93	144	112	53	66
APD	0	139	338	292	676	334	147	142	99
Bactibase	0	11	57	34	19	44	23	12	6
Bagel_I	0	91	26	25	2	3	5	2	0
Bagel_II	0	125	17	34	8	15	4	7	6
Bagel_III	0	57	12	1	1	0	0	0	0
CAMP_Pa...	0	122	1392	46	22	16	15	14	2
CAMP_Str...	0	323	17	31	24	41	29	15	3
CAMP_Va...	0	189	248	423	707	375	158	158	115
DADP	0	1620	200	113	56	195	58	82	55
DAMPD	0	30	242	171	125	107	138	151	109
Defensins	0	131	118	65	30	50	49	23	13
HIPdb	0	840	2	10	3	10	6	7	3
LAMP_Ex...	0	219	594	562	759	451	208	163	113
LAMP_Pat...	0	0	1391	39	17	13	13	11	1
PenBase	0	7	0	16	0	0	2	2	0
Peptaibol	0	198	0	0	0	0	0	0	0
PhytAMP	0	7	121	43	35	25	14	14	5
RAPD	0	7	45	6	10	9	25	9	3
UniProtKb	0	953	430	305	180	181	169	155	112
YADAMP	0	362	138	375	689	382	179	160	115
Overall	0	5571	2798	939	900	504	244	176	115

The Chart tab illustrates the percentage of the row database that is stored exclusively in a given number of databases.



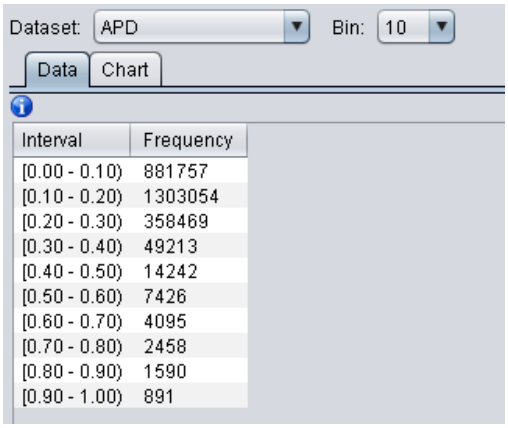
Tab panel: Non-redundant sequence database

For each specified database and threshold, a set of non-redundant sequences may be saved either in FASTA or CSV file formats.



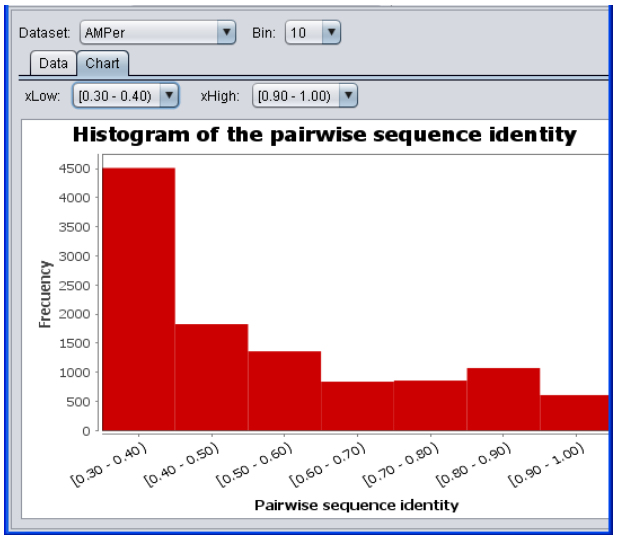
Tab panel: Pairwise sequence identity

The Data tab displays for each database and number of bins the intervals and frequency of the pairwise sequence identity values. The pair of sequences for which the identity value falls in an interval can be exported by right clicking on a cell of the “Frequency” column. In the FASTA format the pair of sequences are stored one after the other.



Interval	Frequency
[0.00 - 0.10)	881757
[0.10 - 0.20)	1303054
[0.20 - 0.30)	358469
[0.30 - 0.40)	49213
[0.40 - 0.50)	14242
[0.50 - 0.60)	7426
[0.60 - 0.70)	4095
[0.70 - 0.80)	2458
[0.80 - 0.90)	1590
[0.90 - 1.00)	891

The Chart tab shows the histogram of the pairwise sequence identities for each database and number of bins. The start (x-low) and end value (x-high) of the visible range of X axis may be modified to zoom in a portion of the figure.



Tab panel: Similarity Overlap

This type of analysis shows the number and percentage of sequences in a row database for which there is, at least, one similar sequence in the column database. The Data and Chart sections are analogous to the identical overlap. In the same way the user may save the sequences by right clicking on a cell.

Tab panel: Clusters and Diversity

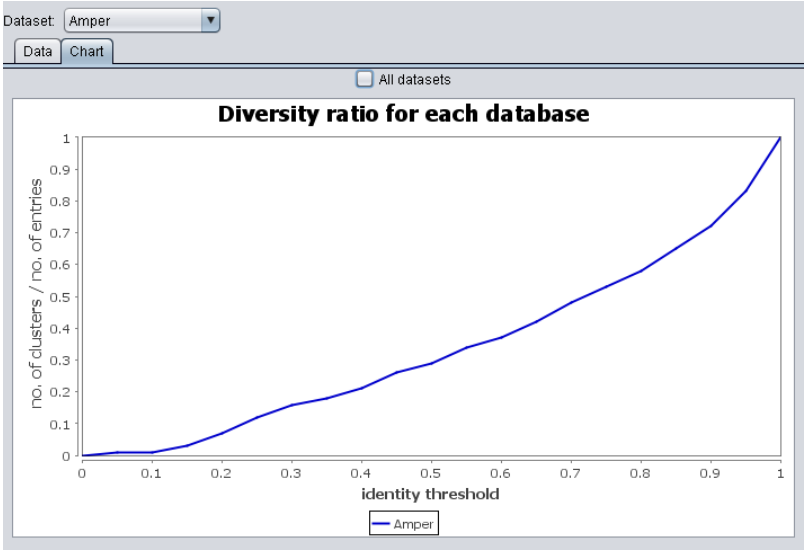
The Data tab displays the number of clusters and diversity ratio as a function of the threshold. The clusters can be exported by right clicking on any cell of the “No. of clusters” column.

Dataset: Amper

Data | Chart

Threshold	No. of clusters	Diversity
0.00	1	0.00
0.05	6	0.01
0.10	9	0.01
0.15	27	0.03
0.20	63	0.07
0.25	112	0.12
0.30	147	0.16
0.35	174	0.18
0.40	200	0.21
0.45	245	0.26
0.50	277	0.29
0.55	326	0.34
0.60	355	0.37
0.65	399	0.42
0.70	452	0.48
0.75	499	0.53
0.80	553	0.58
0.85	615	0.65
0.90	678	0.72
0.95	790	0.83
1.00	948	1.00

The Chart tab shows the diversity ratio as a function of the identity threshold. Clicking on the “All datasets” check box permits the inclusion of the rest of databases into this graph.



Troubleshooting

Out of memory error

The java virtual machine handles the program’s memory and when an object does not fit in the assigned memory space then the java.lang.OutOfMemoryError exception is thrown. This error can occur at the moment of creating the pairwise sequence identity matrix, if the user is working with a large amount of data. To solve this problem the user needs to allocate enough memory by setting the parameter -Xmx#g even higher if possible, where # is the amount of gigabytes of memory that one wishes to allocate. See the “Running Dover Analyzer” section to see how to run the program with the –Xmx option.