



Sequence Demarcation Tool - version 1.2 (SDTv1.2)

27 August 2014

I. Description

SDT is a free Windows based program that allows classification of virus sequences based on pairwise sequence identity calculations. Given a FASTA file containing DNA sequences, the program aligns all possible pairs of sequences using MUSCLE [1], ClustalW2 [2] or MAFFT [3], calculates a sequence identity score for each pair and uses a rooted neighbour joining phylogenetic tree to cluster closely related sequences based on these identity scores.

The identity scores are calculated as $1-(M/N)$ where **M** is the number of mismatching nucleotides and **N** the total number of positions along the alignment where neither sequence has a gap character. The program outputs a 2D graphical representation of the identity scores in a colour coded matrix and a plot of the distribution of the scores which can then be used to either classify viruses within an existing taxonomic framework, or determine the percentage identity cut-offs that should be used to formulate a novel taxonomic framework [4]. After computation SDT allows pairwise identity scores to be saved along with the sequence alignments into a file with an “.sdt” extension. The program can read such a file in cases where one or a few new sequences have been obtained and one would like to add these to a previous SDT analysed dataset. In such cases SDT will use the previously computed pairwise identity and will only recalculate the pairwise alignments and identity scores involving the newly added sequences. SDT also allows the creation of datasets containing user-defined degrees of genetic diversity by partitioning the input sequences into groups of sequences sharing a given range of pairwise identities.

II. Download and installation

1. The program can be downloaded from <http://web.cbio.uct.ac.za/SDT>
2. Extract the **SDTv1.2.zip** file into a temporary folder.
3. Windows XP users can simply double click on the file “SETUP.EXE”. Windows VISTA and Windows 7 users should right click on the file “setup.exe” and select the “run as administrator” option on the popup menu that appears.
4. Follow the instructions of the setup program and **use the default installation directory “C:\SDTv1.2”.**
5. The program folder **SDTv1.2** contains the following:
 - A **bin** directory containing the following programs: MUSCLE [1], Clustalw2 [2], MAFFT [3] and, Neighbor (<http://evolution.genetics.washington.edu/phylip.html>).

- A **tmp** directory that is used to store temporary files.
 - The executable **SDTv1.2.exe** which is the main program.
 - A **README** file which contains these instructions.
 - A FASTA file, **test.fas**, and a .SDT file, **test.sdt**, to test the program.
 - Two SDT files: **mastrevirus_references.sdt** and **mastrevirus_references.fas** used previously for classification of new mastreviruses.
 - Program configuration files: **sdt.in**, **ST5UNST.txt** and **SDTv1.exe.manifest**.
6. To start an analysis double-click the **SDTv1.2.exe** file and, when the program has launched (Fig 1), drag and drop on the main screen the file to analysed or click on the “Open” menu on the top left side of the program window. Within the open dialog box select the FASTA file you want to analyse and click the open command button. On the small window that appears select the alignment program you want to use and then click the Run button. Similarly, you can click on the “Open” menu item to load a .SDT file and then click on the “Append” menu item to load a FASTA file containing one or more new sequences and run an analysis including these. This uses pre-calculated values from the .SDT file and only determines pairwise identity scores for the new sequences. When the analysis completes you can use the “Save” menu option to save all the output files. Also you can right-click on the pictures to copy them onto the clipboard so that you can paste them to other programs such as PowerPoint or Word.

III. **Program features**

1. **Main menus**

The **Open** menu (Fig 1) enables you to load an input file, which the program will detect either as a FASTA file (in which case the user is prompted to choose an alignment program to use for pairwise identity score calculation), or a SDT file (in which case the program will give an option to append one or more sequences in FASTA format) choose the alignment program to use and run the analysis using pre-computed pairwise identity scores.

The **Save** menu enables you to save various analysis results in a variety of formats after an analysis has completed.

The **Rerun** menu is used to run the analysis on the loaded alignment using a different alignment program.

The **Append** menu is used to add sequences to a loaded SDT file.

The **Exit** menu allows you to stop the analysis and exit the program.

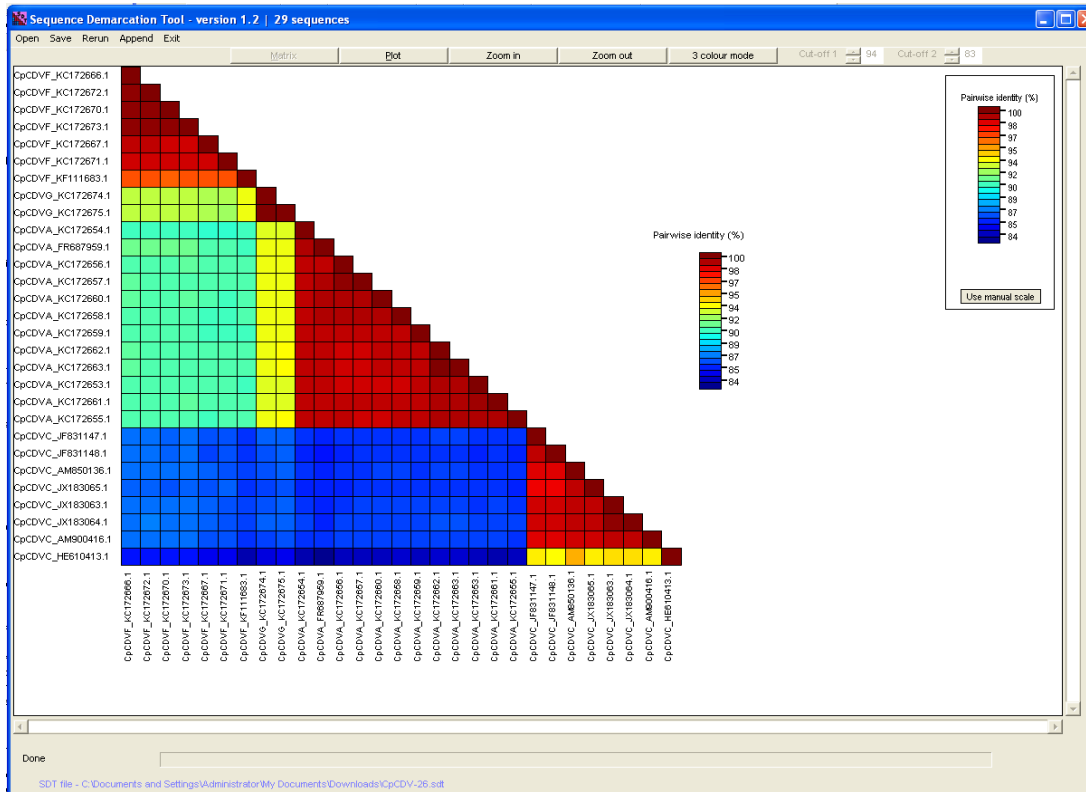


Fig 1. The SDT graphical user interface

2. Main command buttons

Pressing the **Matrix** command button (Fig 1) will result in the colour coded matrix being

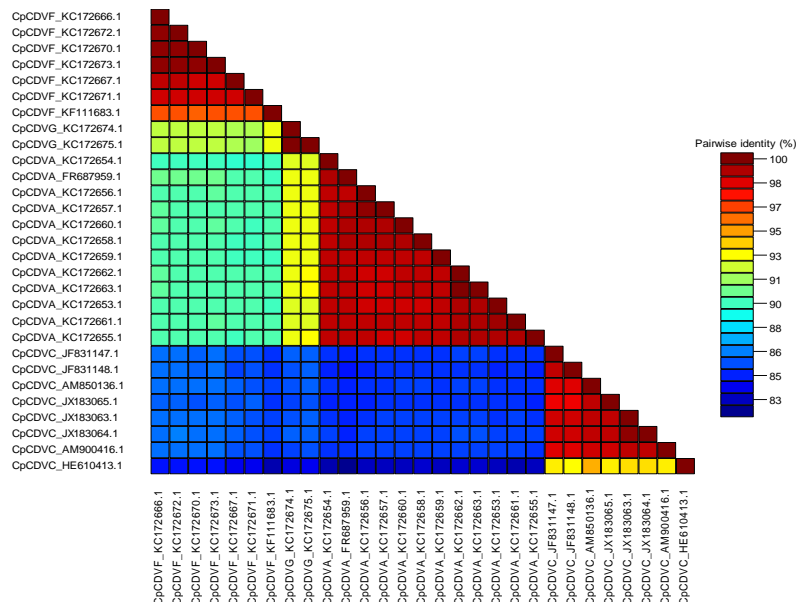


Fig 2. 2D Colour coded matrix

displayed (Fig 2), in this case with a “Full colour spectrum” colour scheme. Pressing the **Plot** command button will result in the distribution plot of pairwise identity scores being displayed (Fig 3).

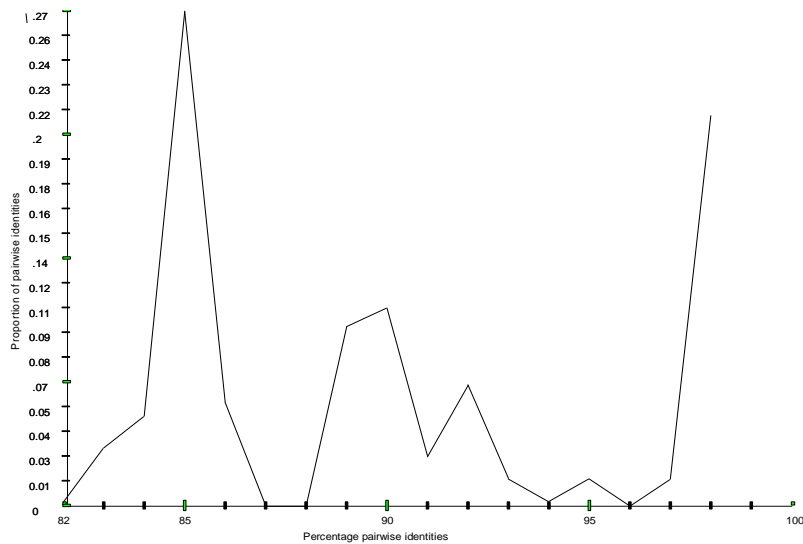


Fig 3. A plot displaying the distribution of pairwise identity scores

The **Zoom in / Zoom Out** command buttons enable you to enlarge or reduce the plot and matrix images.

The **3 colour mode** command button allows switching between “full colour spectrum” mode and the “three colours” mode (Fig 4) for the colour coded matrix representation of identity scores.

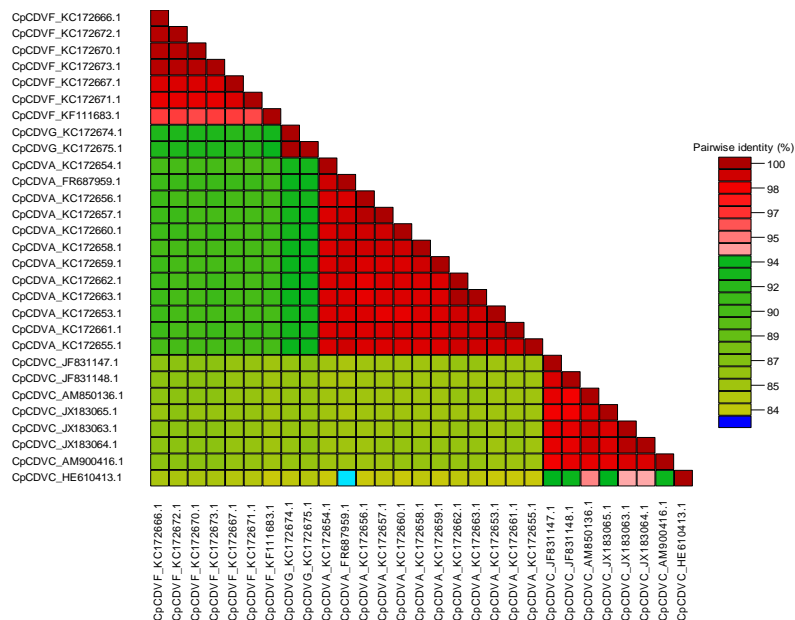


Fig 4. 2D Colour coded matrix

In this case with the “three colours mode” that uses a discontinuous range of red, green and blue shades differentiating the scores above and below two user defined cut-off values. In this case one cut-off value is at 95% and the other is at 78%. Such cut-offs might represent species or strain demarcation thresholds.

The **Up and down** buttons allow you to adjust two demarcation cut-offs when displaying the matrix in the three colour mode.

The **Run** command button: After choosing an alignment program (Fig. 5) this initiates pairwise sequence alignment, the calculation of pairwise identity scores and the display of these scores in the form of a colour coded matrix. The checkbox at the bottom is by default checked; if unchecked the program does not rearrange the scores.



Fig 5. Small window for selection of sequence alignment program

The **Use manual scale** command button (Fig 6) allows changing the colour scale used in pairwise identities matrix.

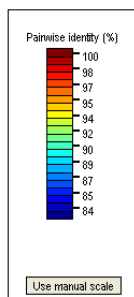


Fig 6. Command button to adjust colour scale

3. Submenus

The **Save SDT file** submenu (Fig 7) allows you to save the sequence analysis results and the pairwise identity scores into a .SDT file for future analyses. This saves time in that the program will then not need to recalculate pairwise identity scores that have already been calculated. New sequences or even whole alignments can be appended to old .SDT dataset and analysed much more quickly than if the entire dataset was analysed from scratch.

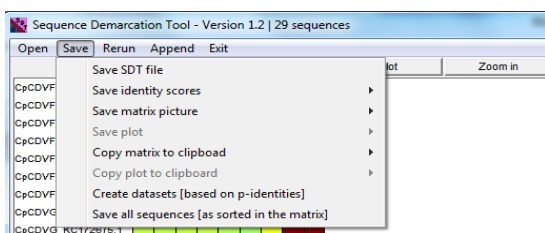


Fig 7. Save SDT file submenu

The **Save identity scores** submenu (Fig 8) allows you to save sequence identity scores in a spreadsheet-readable text file in either a single column or a matrix (i.e. multiple columns; Fig 8).

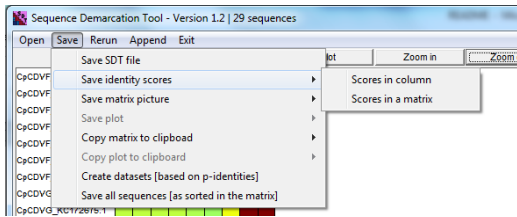


Fig 8. Save identity scores submenu

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE				
1	>CpCDVF_KC172666.1	1.000																																	
2	>CpCDVF_KC172672.1	0.998	1.000																																
3	>CpCDVF_KC172670.1	0.997	0.996	1.000																															
4	>CpCDVF_KC172673.1	0.998	0.997	0.997	1.000																														
5	>CpCDVF_KC172667.1	0.989	0.988	0.988	0.988	1.000																													
6	>CpCDVF_KC172671.1	0.987	0.986	0.986	0.986	0.987	1.000																												
7	>CpCDVF_KF11683.1	0.965	0.964	0.963	0.965	0.965	0.964	1.000																											
8	>CpCDVG_KC172674.1	0.925	0.924	0.924	0.924	0.922	0.920	0.932	1.000																										
9	>CpCDVG_KC172675.1	0.925	0.924	0.923	0.924	0.920	0.919	0.932	0.939	1.000																									
10	>CpCDVA_KC172654.1	0.903	0.902	0.902	0.903	0.901	0.901	0.901	0.929	0.930	1.000																								
11	>CpCDVA_FR687959.1	0.911	0.910	0.909	0.911	0.906	0.906	0.903	0.932	0.933	0.991	1.000																							
12	>CpCDVA_KC172656.1	0.906	0.905	0.906	0.906	0.905	0.904	0.904	0.931	0.931	0.989	0.990	1.000																						
13	>CpCDVA_KC172657.1	0.907	0.906	0.905	0.906	0.905	0.904	0.905	0.932	0.932	0.989	0.991	0.998	1.000																					
14	>CpCDVA_KC172660.1	0.907	0.906	0.905	0.906	0.905	0.903	0.906	0.931	0.932	0.989	0.991	0.993	0.993	1.000																				
15	>CpCDVA_KC172658.1	0.905	0.904	0.904	0.905	0.903	0.903	0.905	0.932	0.933	0.989	0.991	0.992	0.991	0.992	1.000																			
16	>CpCDVA_KC172659.1	0.906	0.905	0.906	0.906	0.903	0.903	0.903	0.933	0.933	0.989	0.990	0.989	0.989	0.989	0.990	1.000																		
17	>CpCDVA_KC172662.1	0.906	0.906	0.904	0.906	0.904	0.904	0.906	0.933	0.934	0.986	0.990	0.987	0.988	0.988	0.988	0.988	1.000																	
18	>CpCDVA_KC172663.1	0.906	0.906	0.904	0.906	0.904	0.904	0.906	0.933	0.934	0.986	0.990	0.987	0.988	0.988	0.988	0.988	1.000	1.000																
19	>CpCDVA_KC172653.1	0.906	0.905	0.906	0.906	0.904	0.905	0.903	0.931	0.930	0.989	0.989	0.987	0.988	0.987	0.988	0.988	0.991	0.991	1.000															
20	>CpCDVA_KC172661.1	0.906	0.905	0.906	0.906	0.906	0.906	0.906	0.930	0.931	0.989	0.989	0.989	0.988	0.988	0.988	0.989	0.991	0.991	0.994	1.000														
21	>CpCDVA_KC172655.1	0.905	0.904	0.904	0.905	0.903	0.904	0.904	0.933	0.934	0.989	0.990	0.988	0.987	0.988	0.989	0.990	0.991	0.991	0.991	0.991	1.000													
22	>CpCDVC_IF631147.1	0.867	0.867	0.865	0.867	0.863	0.861	0.855	0.863	0.864	0.855	0.852	0.856	0.857	0.857	0.857	0.857	0.860	0.860	0.857	0.857	0.856	1.000												
23	>CpCDVC_IF631148.1	0.866	0.867	0.864	0.866	0.859	0.861	0.855	0.862	0.863	0.853	0.852	0.854	0.855	0.856	0.855	0.855	0.857	0.857	0.855	0.856	0.855	0.990	1.000											
24	>CpCDVC_AM950136.1	0.868	0.868	0.866	0.868	0.863	0.862	0.858	0.862	0.863	0.856	0.854	0.857	0.857	0.858	0.858	0.857	0.859	0.859	0.857	0.857	0.856	0.985	0.984	1.000										
25	>CpCDVC_XJ183085.1	0.864	0.864	0.862	0.864	0.861	0.863	0.857	0.860	0.861	0.854	0.853	0.855	0.856	0.856	0.856	0.856	0.857	0.857	0.855	0.856	0.854	0.982	0.982	0.988	1.000									
26	>CpCDVC_XJ183083.1	0.867	0.868	0.866	0.867	0.863	0.862	0.858	0.862	0.863	0.857	0.854	0.857	0.858	0.859	0.859	0.859	0.860	0.860	0.857	0.859	0.857	0.986	0.986	0.990	0.988	1.000								
27	>CpCDVC_XJ183084.1	0.868	0.869	0.867	0.868	0.864	0.863	0.858	0.863	0.865	0.858	0.855	0.858	0.859	0.860	0.859	0.859	0.861	0.861	0.858	0.860	0.858	0.986	0.986	0.990	0.988	0.998	1.000							
28	>CpCDVC_AM900416.1	0.868	0.868	0.866	0.868	0.863	0.862	0.856	0.860	0.861	0.857	0.856	0.858	0.858	0.859	0.859	0.859	0.860	0.860	0.858	0.858	0.857	0.983	0.985	0.987	0.986	0.989	0.989	1.000						
29	>CpCDVC_HE610413.1	0.851	0.852	0.850	0.852	0.846	0.845	0.836	0.843	0.844	0.835	0.829	0.838	0.838	0.836	0.839	0.837	0.839	0.839	0.837	0.835	0.836	0.938	0.936	0.949	0.938	0.940	0.940	0.936	1.000					

Matrix format

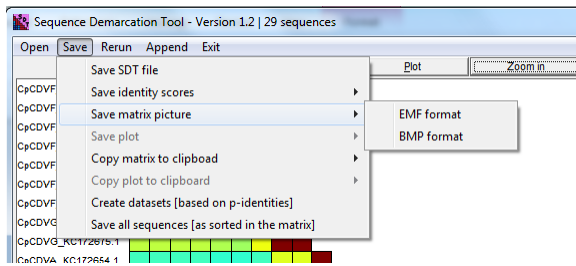
	A	B	C
1	First Sequence	Second Sequence	Identity Score
2	>CpCDVF_KC172666.1	>CpCDVF_KC172672.1	0.9984429
3	>CpCDVF_KC172666.1	>CpCDVF_KC172670.1	0.9972752
4	>CpCDVF_KC172666.1	>CpCDVF_KC172673.1	0.9980537
5	>CpCDVF_KC172666.1	>CpCDVF_KC172667.1	0.9887072
6	>CpCDVF_KC172666.1	>CpCDVF_KC172671.1	0.9867653
7	>CpCDVF_KC172666.1	>CpCDVF_KF11683.1	0.9646962
8	>CpCDVF_KC172666.1	>CpCDVG_KC172674.1	0.925
9	>CpCDVF_KC172666.1	>CpCDVG_KC172675.1	0.9246094
10	>CpCDVF_KC172666.1	>CpDVA_KC172654.1	0.9028861
11	>CpCDVF_KC172666.1	>CpDVA_FR687959.1	0.9109926
12	>CpCDVF_KC172666.1	>CpDVA_KC172656.1	0.9060062
13	>CpCDVF_KC172666.1	>CpDVA_KC172657.1	0.9067863
14	>CpCDVF_KC172666.1	>CpDVA_KC172660.1	0.9067863
15	>CpCDVF_KC172666.1	>CpDVA_KC172658.1	0.9052262
16	>CpCDVF_KC172666.1	>CpDVA_KC172659.1	0.9060062

Column format

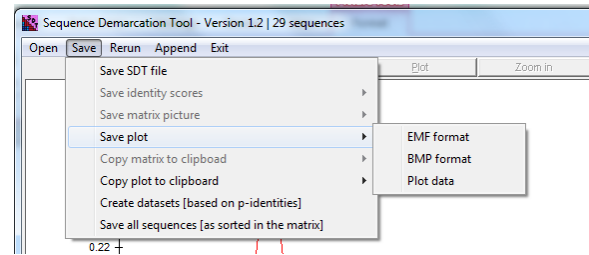
Fig 9. Output formats of spreadsheet readable text files.

The **Save matrix picture** submenu (Fig 10) allows you to save a 2D colour coded graphical image of the matrix in either .EMF (vector graphic good for rescaling) or .BMP (bitmap graphic that is not good for rescaling) formats.

The **Save plot** submenu (Fig 10) allows you to save the pairwise identity distribution plot in either .EMF or .BMP formats, or as a spreadsheet-readable list of the data used to produce the plot.



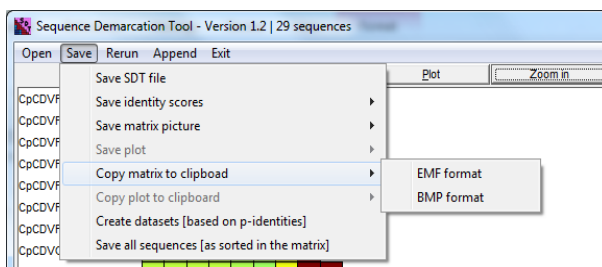
Save matrix picture submenu



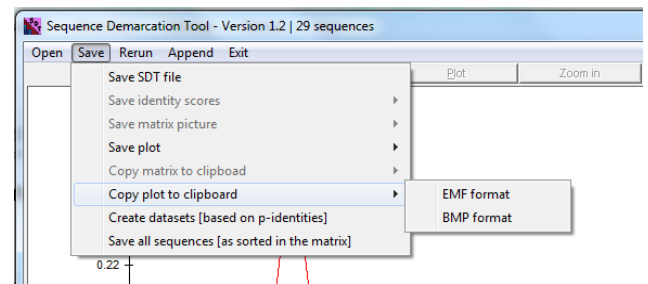
Save plot submenu

Fig 10. The save matrix picture and save plot menu items

Copy matrix to clipboard submenus (Fig 11) allows you to copy the matrix to the clipboard in either .EMF or .BMP formats so that it can be pasted into another program such as Microsoft Word or Powepoint.



Copy matrix to clipboard submenu



Copy plot to clipboard submenu

Fig 11. The copy matrix to clipboard and copy plot to clipboard menus

Copy plot to clipboard submenu (Fig 11) allows you to copy the plot image in either .EMF or .BMP formats to the clipboard so that it can be pasted into another program such as Word or Powerpoint.

Create datasets [based on p-identities] submenu (Fig 12) allows you to objectively create datasets of sequences sharing a desired level of diversity/similarity to suite further genome evolution analyses such as inference of patterns of natural selection or the identification of conserved genomic secondary structures [5,6]. After the sequence identity scores have been computed, you need to enter the maximum and minimum identity percentage in the create datasets window (Fig 12) , the program will then partition the input sequence dataset into sets of non-overlapping sequence files, with each file containing only sequence pairs with identities that are within the user specified range.

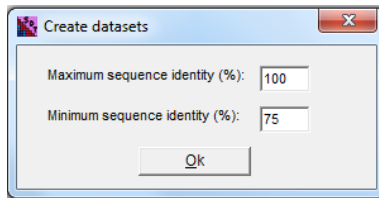


Fig 12. Small window for choosing the range of identities

Save all sequences [as sorted in the Matrix] submenu allows creating a dataset of all sequences arranged as they appear in the matrix after clustering using a neighbour joining tree.

Important notes:

- You must ensure that the new sequences start at the same position as the sequences already loaded.
- At the end of every run the program saves all the output files into the tmp folder which you can then manually retrieve either should you forget to save the analysis results before quitting, or if the program unexpectedly crashes at the end of a run (this will occasionally happen if the dataset being analysed is very large). The files stored in the tmp folder will all be overwritten during the next run of the program.
- SDT does not display a colour coded matrix for a dataset of more than 500 sequences. It does, however, allow you to save the matrix to a file in either .BMP or .EMF formats
- Large datasets containing about 1000 sequences that are each ~2,800nts long will take approximately 3 days to analyse. In such cases (and even for much larger datasets) SDTMPI, a parallelized version of SDT, could be used. SDTMPI only calculates a matrix of pairwise identity scores (i.e. it does not produce either the graphical identity score matrix or the pairwise identity score distribution plots) but can, given access to a computer cluster with several processors, produce a completed pairwise identity score matrix approximately 19 to 38 times quicker than SDT when running on 20 to 40 cores respectively (SDTMPI is available at <http://web.cbio.uct.ac.za/SDT>).

References

1. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797.
2. Larkin M a, Blackshields G, Brown NP, Chenna R, McGettigan P a, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.

3. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30: 772–780.
4. Muhire B, Martin DP, Brown JK, Navas-Castillo J, Moriones E, et al. (2013) A genome-wide pairwise-identity-based proposal for the classification of viruses in the genus Mastrevirus (family Geminiviridae). *Arch Virol* 158: 1411–1424.
5. Muhire BM, Golden M, Murrell B, Lefeuvre P, Lett J, et al. (2013) Evidence of pervasive biologically functional secondary-structures within the genomes of eukaryotic single-stranded DNA viruses. *J Virol*.
6. Stenzel T, Piasecki T, Chrzastek K, Julian L, Muhire BM, et al. (2014) Pigeon circoviruses display patterns of recombination, genomic secondary structure and selection similar to those of Beak and feather disease viruses. *J Gen Virol*.



Authors: Brejnev Muhire¹, Darren Martin¹ and Arvind Varsani²

¹Institute of Infectious Diseases and Molecular Medicine, Computational Biology Group, University of Cape Town, South Africa

²School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch, 8140, New Zealand

BM is funded by the University of Cape Town



And the Poliomyelitis Research Foundation (South Africa)



Web sites: <http://web.cbio.uct.ac.za/SDT>

Email: mhrbre001@myuct.ac.za

mubrejnev@gmail.com