Sequence Demarcation Tool - version 1 (SDTv1)

14 April 2014

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.

Download and installation

1. The program can be downloaded from http://web.cbio.uct.ac.za/SDT
2. Extract the **SDTv1.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"**.
5. The program folder **SDTv1** 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.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.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.

Program features

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.

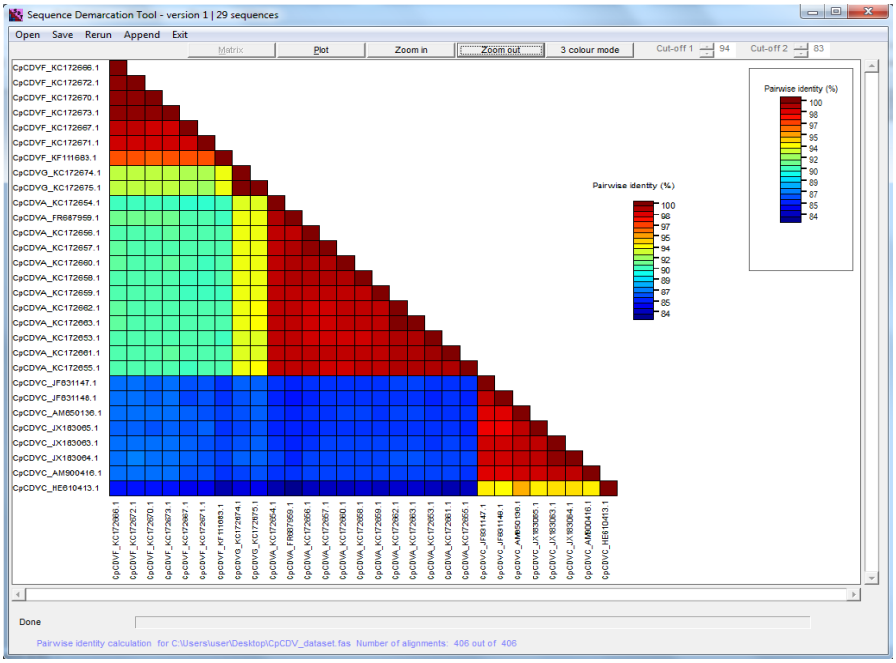


Fig 1. The SDT graphical user interface

Command buttons:

Pressing the **Matrix** command button will result in the colour coded matrix being displayed (Fig 2)

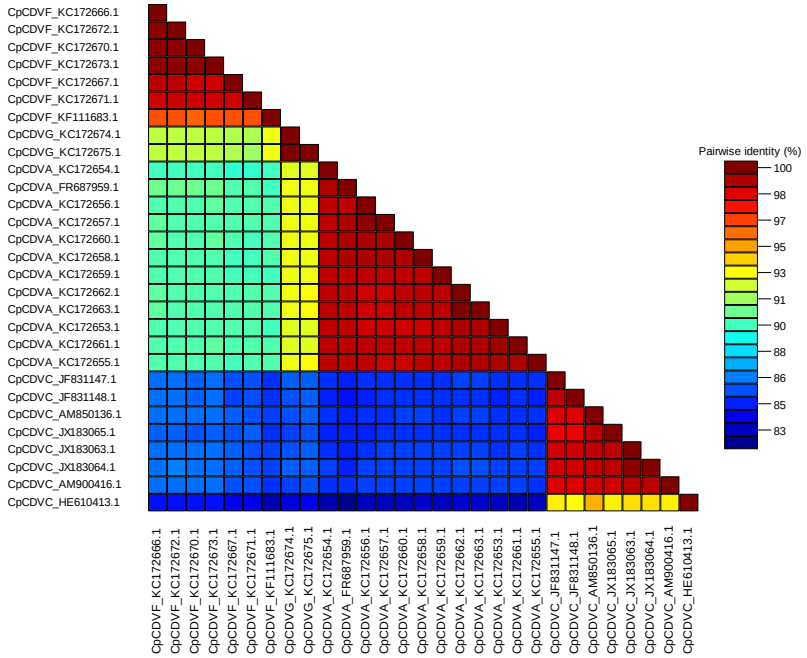


Fig 2. 2D Colour coded matrix: In this case with a **"**Full colour spectrum" colour scheme.

Pressing the **Plot** command button will result in the distribution plot of pairwise identity scores being displayed (Fig 3).

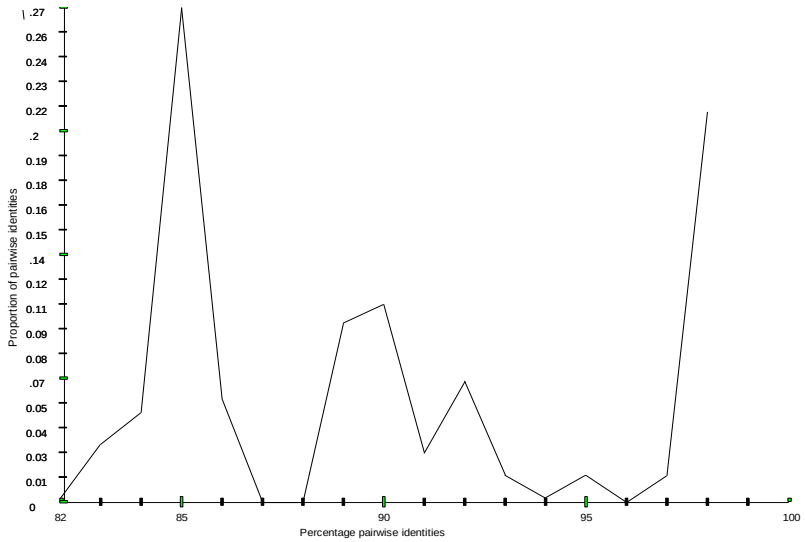


Fig 3. A plot displaying the distribution of pairwise identity scores

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.

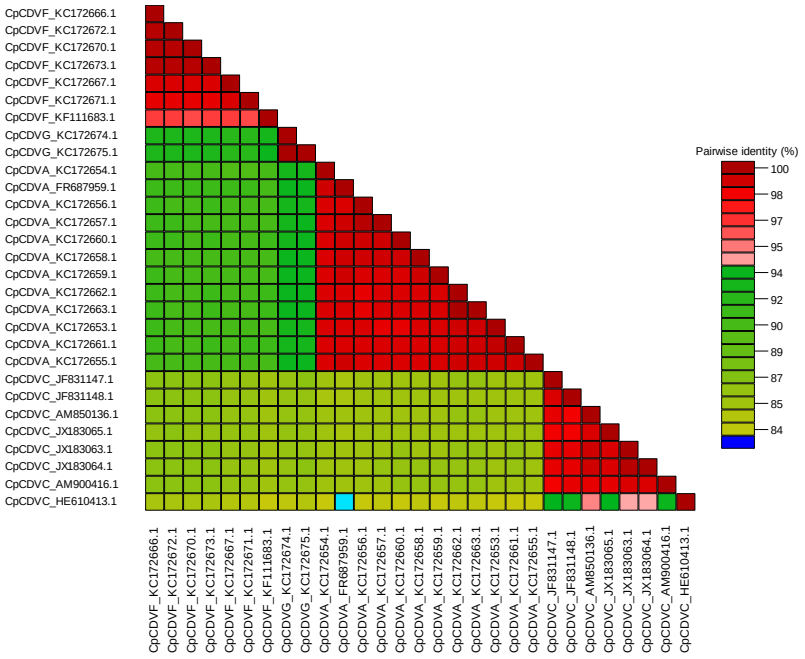


Fig 4. 2D Colour coded matrix: In this case with the **“**three colours mode” that uses a discontinuous range of red, green and blue shades differentiating the scores above and below two user defined cut-off values. In this

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.

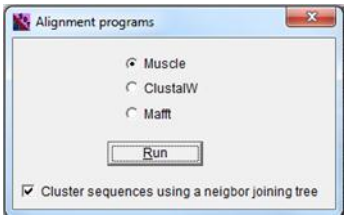


Fig 5. Small window for selection of sequence alignment program

The **Save SDT file** submenu (Fig 6) 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.

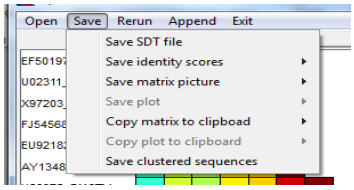


Fig 6. Save SDT file submenu

The **Save identity scores** submenu (Fig 7) 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).

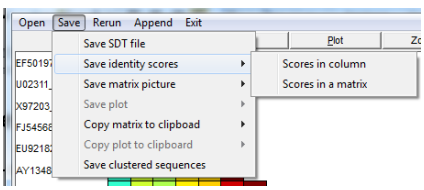
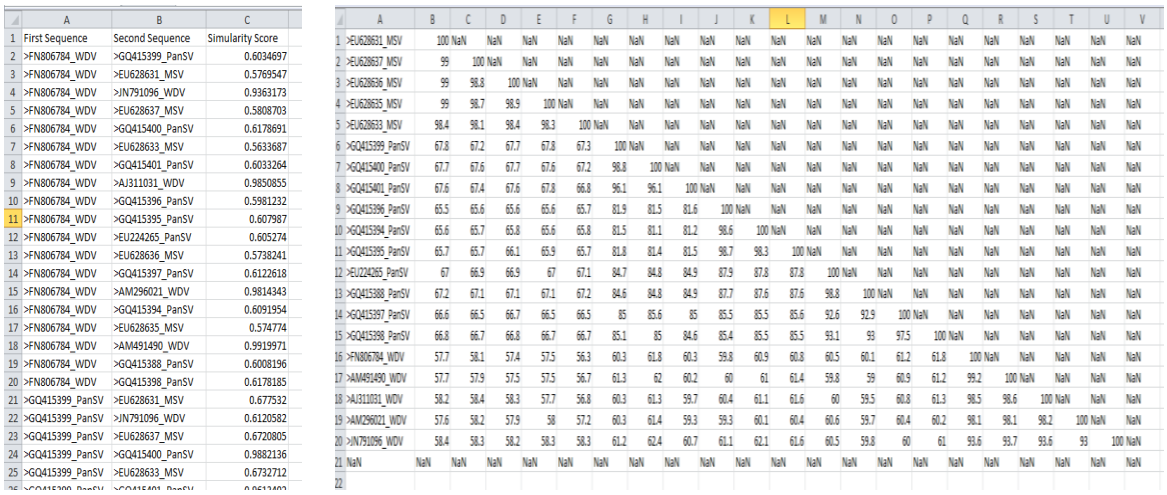


Fig 7. Save identity scores submenu



	A	B	C
1	First Sequence	Second Sequence	Simularity Score
2	>FN806784_WDV	>GQ415399_PanSV	0.6034697
3	>FN806784_WDV	>EU628631_MSV	0.5769547
4	>FN806784_WDV	>JN791096_WDV	0.9363173
5	>FN806784_WDV	>EU628637_MSV	0.5808703
6	>FN806784_WDV	>GQ415400_PanSV	0.6178691
7	>FN806784_WDV	>EU628633_MSV	0.5633687
8	>FN806784_WDV	>GQ415401_PanSV	0.6033264
9	>FN806784_WDV	>AJ311031_WDV	0.9850855
10	>FN806784_WDV	>GQ415396_PanSV	0.5981232
11	>FN806784_WDV	>GQ415395_PanSV	0.607987
12	>FN806784_WDV	>EU224265_PanSV	0.605274
13	>FN806784_WDV	>EU628636_MSV	0.5738241
14	>FN806784_WDV	>GQ415397_PanSV	0.6122618
15	>FN806784_WDV	>AM296021_WDV	0.9814343
16	>FN806784_WDV	>GQ415394_PanSV	0.6091954
17	>FN806784_WDV	>EU628635_MSV	0.574774
18	>FN806784_WDV	>AM491490_WDV	0.9919971
19	>FN806784_WDV	>GQ415388_PanSV	0.6008196
20	>FN806784_WDV	>GQ415398_PanSV	0.6178185
21	>GQ415399_PanSV	>EU628631_MSV	0.677532
22	>GQ415399_PanSV	>JN791096_WDV	0.6120582
23	>GQ415399_PanSV	>EU628637_MSV	0.6720805
24	>GQ415399_PanSV	>GQ415400_PanSV	0.9882136
25	>GQ415399_PanSV	>EU628633_MSV	0.6732712

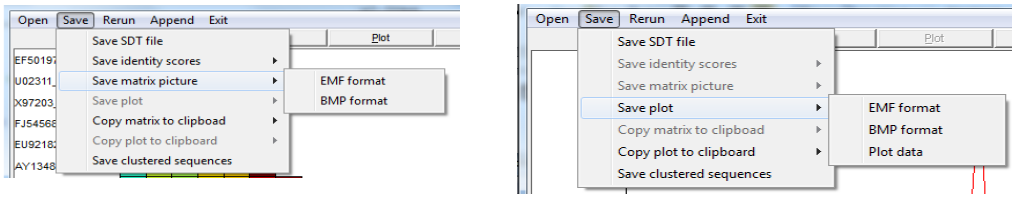
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	>EU628631_MSV	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	>EU628637_MSV	99	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	>EU628636_MSV	99	98.8	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	>EU628635_MSV	99	98.7	98.9	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	>EU628633_MSV	98.4	98.1	98.4	98.3	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	>GQ415399_PanSV	67.8	67.2	67.7	67.8	67.3	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	>GQ415400_PanSV	67.7	67.6	67.7	67.6	67.2	98.8	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	>GQ415401_PanSV	67.6	67.4	67.6	67.8	66.8	96.1	96.1	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	>GQ415396_PanSV	65.5	65.6	65.6	65.6	65.7	81.9	81.5	81.6	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10	>GQ415394_PanSV	65.6	65.7	65.8	65.6	65.8	81.5	81.1	81.2	98.6	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11	>GQ415395_PanSV	65.7	65.7	66.1	65.9	65.7	81.8	81.4	81.5	98.7	98.3	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
12	>EU224265_PanSV	67	66.9	66.9	67	67.1	84.7	84.8	84.9	87.9	87.8	87.8	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13	>GQ415388_PanSV	67.2	67.1	67.1	67.1	67.2	84.6	84.8	84.9	87.7	87.6	87.6	98.8	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
14	>GQ415397_PanSV	66.6	66.5	66.7	66.5	66.5	85	85.6	85	85.5	85.5	85.6	92.6	92.9	100	NaN	NaN	NaN	NaN	NaN	NaN	NaN
15	>GQ415398_PanSV	66.8	66.7	66.8	66.7	66.7	85.1	85	84.6	85.4	85.5	85.5	93.1	93	97.5	100	NaN	NaN	NaN	NaN	NaN	NaN
16	>FN806784_WDV	57.7	58.1	57.4	57.5	56.3	60.3	61.8	60.3	59.8	60.9	60.8	60.5	60.1	61.2	61.8	100	NaN	NaN	NaN	NaN	NaN
17	>AM491490_WDV	57.7	57.9	57.5	57.5	56.7	61.3	62	60.2	60	61	61.4	59.8	59	60.9	61.2	99.2	100	NaN	NaN	NaN	NaN
18	>AJ311031_WDV	58.2	58.4	58.3	57.7	56.8	60.3	61.3	59.7	60.4	61.1	61.6	60	59.5	60.8	61.3	98.5	98.6	100	NaN	NaN	NaN
19	>AM296021_WDV	57.6	58.2	57.9	58	57.2	60.3	61.4	59.3	59.3	60.1	60.4	60.6	59.7	60.4	60.2	98.1	98.1	98.2	100	NaN	NaN
20	>JN791096_WDV	58.4	58.3	58.2	58.3	58.3	61.2	62.4	60.7	61.1	62.1	61.6	60.5	59.8	60	61	93.6	93.7	93.6	93	100	NaN
21	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
22																						

Column format **Matrix format**

Fig 8. Output formats of spreadsheet readable text files.

The **Save matrix picture** submenu (Fig 9) 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.

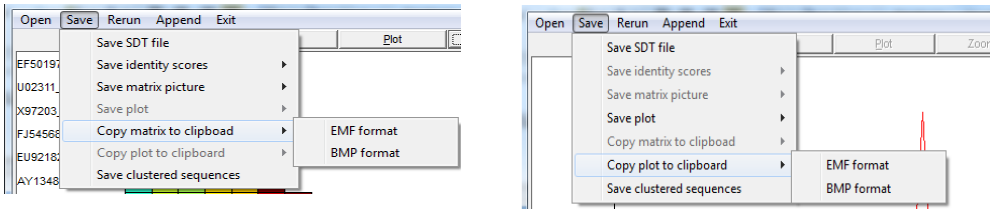
The **Save plot** submenu (Fig 9) allows you to save the pairwise identity distribution plot in either .EMF or .BMP formats, or as a spreadsheet-readable list of the data used to produce the plot.



Save matrix picture submenu **Save plot submenu**

Fig 9. The save matrix picture and save plot menu items

Copy matrix to clipboard submenus (Fig 10) 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 10. The copy matrix to clipboard and copy plot to clipboard menus

Copy plot to clipboard submenu (Fig 10) 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.

Save clustered sequences submenu (Fig 10) 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 11) , 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.

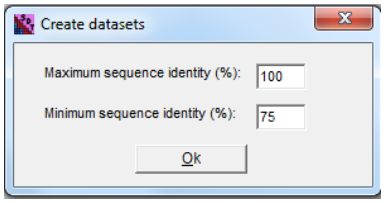


Fig 11. Small window for choosing the range of identities

Rerun menu which is enabled when the analysis is completed allows you to re-run an analysis using a different alignment program.

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, Chrząstek 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 Muhire[1], Darren Martin[1] and Arvind Varsani[2]

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

[2]School 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)

http://web.cbio.uct.ac.za/SDT mhrbre001@myuct.ac.za

mubrejnev@gmail.com