User's Guide to GCPAT (Greedy Clique Partition pAckage Tool) – August 31, 2005

There are three major forms of output:

- a fingerprint UPGMA tree, where clones with similar fingerprints will tend to cluster together (although it should be noted that not all clusters with similar fingerprints will be located close together in the tree)

- a taxonomic table, where detailed data describing the clones belonging to each fingerprint group is generated, and

- files describing which clone clusters have a fingerprint within a given distance from a specified clone cluster fingerprint. Distance is measured in terms of the sum of the differences between fingerprints, where each mismatch between a 0 and a 1 = 1, and each match between an N and a 0, 1, or N = 0.5.

Files containing (1) clone names, (2) fingerprints, (3) sequences predicted to have identical fingerprints, or (4) "consensus" sequences (which have a sequence closest to the actual consensus of the sequences in 3) for a specified range of fingerprint groups can also be generated.

Fingerprint Trees

In order to generate a fingerprint tree you will need a fingerprint file:

- Copy the the contents of the "Reduced_Unequal_Variance" worksheet page from the **Main Processing Tool**, or another file or page containing your fingerprint information, into Word using the "Paste Special" function: highlight the information in Excel, press CTRL + C, go to Word, Edit-> Paste Special, and choose "Unformatted Text". The following assumes you have two blank columns between the column containing clone name information and the fingerprint information for the first probe.
- Remove the lines containing header information (probe names, "Gene ID")
- Use the "Replace" function to remove tabs: Edit -> Replace -> More -> Special -> Tab Character. Place two Tab Characters (^t) in the "Find what" box, and type a space in the "Replace with" box. Click Replace All.
- Now repeat only with a single Tab Character in the "Find" box and no space in the "Replace" box. Click Replace All.
- You should have a list of clone names followed by a space and the binary fingerprint after.
- Enter the first line of the fingerprint file, which should contain the number of clones, a space, then the number of probes, and no other characters. Eg. for 96 clones and 30 probes, the first line looks like this:
  96 30
  [Clone name1] [fingerprint1]

[Clone name2] [fingerprint2] …
- Save as a plain text file. The name of the file should contain no blank spaces.

To generate a fingerprint tree using GCPAT:
- File -> Open fingerprint file
- Browse until you find the fingerpint file. Select the file and click “Open”.
- You will see a message:

Open: "[filename].txt".
Please wait...
[number of clones] fingerprints were read

- Clustering -> GCP
- You will see a message in a popup window confirming that you want to compute a tree for the selected file.
- Click “Yes”.
- You will see a message:

GCP running on “[filename].txt”

- The program will automatically cluster the fingerprints, tell you how many clusters there are and save the output in a cluster file (same name as your fingerprint file, only an extension “.cluster” is added). The more fingerprints there are and the more “N”s you have in your fingerprints, the longer the clustering step will take.
- The computation process may take up to several hours for large (7000 clones or more) trees. The program will not be responsive in this time and may look as if it has stopped running. Be sure to wait an appropriate length of time before stopping the program in this situation.
- When the program is finished, a second window in which the tree is displayed, labeled “Tree”, will pop up.
- The tree information (which you can use to open the tree with Tree Explorer also) is saved in the file indicated in the program window (same name as your fingerprint file, only an extension “.tree” is added)
- The program will automatically also generate a cluster file (same name as your fingerprint file, only an extension “.cluster” is added)

The numbers to the right of the clone names in the tree are index numbers for each cluster. These numbers are found in the cluster file that was used to generate that tree. These numbers will also appear in the taxonomic table, and are used to identify clusters.

<u>To add or remove treatment colours on the tree:</u>

- Add a colour: Tools -> Highlight treatment -> One
- type the name of the treatment in the box at the bottom (Between Add and Remove)

- Click on the colour you want
- Drag the colour from the highlighted text in the centre of the dialog box down to the text you have just typed
- Click “Add”

- To change a colour, go through the above process, and select a different colour.

- To remove all colours, Tools -> Highlight treatment -> None

- To change colour of a single clone, type in the entire clone name and do the rest as if it were a treatment.

- To change the colour of multiple clones simultaneously (all clones will become the same colour), Tools -> Highlight treatment -> Multiple
- Click on “Browse” and select the text file containing a list of clone names. The format of the file will consist of one clone name per line:

  CloneName1
  CloneName2
  CloneName3 (etc.)

- Then select the colour you want the clones to become, drag it down to the name of the file which will be displayed in the dialog box at the bottom, and click “Add”.
- Be patient, if you are colouring a lot of clones it might take some time before GCPAT is finished with this step. It may look as if it is no longer responding.

<u>To save or print a tree in various formats:</u>

- To save colour tree: File -> Save. The file that is generated can only be read by GCPAT. Note that in order to display it in the future, you will also need the cluster file that was used to generate that tree.

- To export tree to PNG format: File -> Export as PNG.
- Browse to the directory you want your tree to be saved in, and type in a name. The program will cut up the tree image into manageable sizes, and generate multiple PNG files, each with the name you entered plus a number indicating the order of the files, with 1 = the top of the tree. These files will be saved in a folder that has the name that you entered.

- To print a tree:

(1) using GCPAT and Adobe Acrobat (you must have Acrobat installed for this to work):

- File -> Print, then select “Acrobat Distiller” as the printer.
- Click “Properties”.

- Choose “Landscape” as the orientation of the paper.
- Click “Advanced”.
- Next to “Paper Size” there should be a pulldown menu. Select “A3” from this menu.
- Click “OK”, then print.
- Open the file you generate with this procedure with Adobe Acrobat, and select “Shrink oversized pages to paper size”. The images that have been generated should fit on the paper size you have selected for your printer.

(2) using GCPAT alone: File-> Print, then select the printer you wish to use. Set the printout orientation to “landscape” and the paper size to “A3”. If you can’t select these options in your printer’s “Properties” box, it is probable that you won’t be able to get all the information that is on the tree to fit onto the pages.

If you can’t get Adobe Acrobat, and your printer doesn’t give you the above options, you can also try the following:

(3) print using Tree Explorer, a free downloadable program that we obtained via the Internet. It’s hard to find due to similarly-named programs that are designed for other purposes- search for “Tree Explorer Tamura” on the web if the following link does not work: http://evolgen.biol.metro-u.ac.jp/TE/TE_man.html.

After you install Tree Explorer, you can use it to print black-and white trees that don’t have the cluster index information printed to the side (just clone names):

- File -> Import Tree Data -> Newick Standard File
- Change the file type from Newick Standard File to All Files (*.*)
- Browse until you find the saved tree file and click “Open”. The tree will be displayed.
- If it is a large tree, you can print out the entire tree if you DO NOT change any settings and simply select File -> Print. For some reason, it also helps to scroll through the tree before printing.

To find a particular clone or cluster in the tree:

- Tools->Search
- A dialog box will pop up asking what to search for. Enter the name of the clone you wish to find in the tree.
- The display will change so that clone is present on the screen. A green box will be centered over the clone name to help you locate it. If you click on the screen, the box will disappear.
- To find a cluster, simply enter the name of one of the clones from that cluster.

To redraw the UPGMA tree:

This is an option which allows you to take a UPGMA tree, then redraw it so that clones with a certain distance between their fingerprints are placed in the same cluster. Note that this option uses the UPGMA tree as the basis for the re-drawn tree,

so that clusters that have similar fingerprints within the distance specified, but are placed far apart in the tree, won't be clustered together but will remain in their separate areas. Distance is measured in terms of the sum of the differences between fingerprints, where each mismatch between a 0 and a 1 = 1, and each match between an N and a 0, 1, or N = 0.5.

- Tools -> Re-draw the tree
- A dialog box will pop up asking for the distance threshold. If you enter "2", then clones located in adjacent clusters in the UPGMA tree which have a distance equal to or less than two between them will be placed in the same cluster.

To determine which clusters are within a specified distance of a given cluster:

- You need to have the tree displayed in order to do this. To display the tree, open a cluster file, then select Tools->Display Tree, and select the tree file. The cluster file must be the one used to generate the tree.
- In the tree display window: Tools -> Search Nearby Clusters
- A dialog box will ask you to enter the index (number) of the cluster you are interested in.
- A second dialog box will then ask you to specify the distance threshold. The clusters that will be displayed in the resulting file will all have a distance equal to or less than this threshold value. Distance is measured in terms of the sum of the differences between fingerprints, where each mismatch between a 0 and a 1 = 1, and each match between an N and a 0, 1, or N = 0.5.
- If your tree has a lot of clusters in it, it may take a long time for the program to finish the calculations involved, so be patient. You may even want to run this calculation overnight.
- When the program is finished, it will display a "Done!" message and tell you the name of the file that the information is saved in.
- The name of the new file will consist of the name of your fingerprint tree, plus an extension "XaroundY.nearby", where X = the distance threshold and Y=the index number of the cluster you entered.
- The file will be in tab-delimited format, so you can open it in Excel. If you do this, be sure to specify that the "Fingerprint" column should be in text format rather than general format, or the program will attempt to display the fingerprints in scientific notation.
- There will be four columns of information:

Cluster Index	Distance	Clone Name	Fingerprint
92	1		
		S10_Plate_1-F6	00111101010110011001011010001011
		S10_Plate_1-K5	00111101010110011001011010001011
		S5_Plate_1-C3	00111101010110011001011010001011

- Data for each cluster that falls within the distance threshold specified will be listed in the file.

Taxonomic Tables

These tables contain detailed information on each of the fingerprint clusters identified in the process of making the fingerprint tree, including:

(1) the total number of clones in each cluster
(2) how many clones there are per treatment in each cluster
(3) the fingerprint that describes each cluster
(4) the minimum pairwise identity between sequences in your "taxon sequences" file which are predicted to have the same fingerprint as the cluster.
(5) the sequence from the "taxon sequences" file that most closely resembles the sequences in 4, above. This file is called a "consensus sequence" but it is not in fact a consensus sequence of the sequences in 4.

To make a table after clustering fingerprints:

- File -> Open probe set file. This file will be a FASTA file containing all the probes used in the fingerprints, in the same order as they appear in the fingerprints. When the file has finished reading the probe set it will display a message saying so.
- File -> Open taxon sequences file. This file will be in FASTA format, and should contain complete and unambiguous rRNA gene sequences for as many different organisms within the set you are examining. We currently use the "training" data files used to generate our OFRG probe sets, which contain several thousand different, complete and unambiguous bacterial or fungal rRNA gene sequences. There will be a pause while the program opens and reads all the sequences. When the file has finished reading the taxon sequences file, it will display a message saying so.
- If you do not currently have a fingerprint cluster file opened: File -> Open cluster file. This file will be the one automatically saved when generating the fingerprint tree and will end with ".cluster".
- Tools ->Taxonomic tabulation. A dialog box will pop up telling you that "There is no taxon information in memory. If you want to find consensus names please open a taxon lineage file." Since there currently are no complete taxon lineage files in existence, you can't open one, so just click "No."
- A dialog box will pop up asking you to verify the files you want to run the tabulation on. Click "Yes".
- You will be told how many clusters are identified, or in other words, have a fingerprint that matches a predicted fingerprint of at least one gene sequence in the taxon sequences file. The number of identified clusters is partly a function of how many and how varied the sequences in your taxon sequence file are.
- The message "Proceed to compute identity value per group" will come up. The identity value is the minimum pairwise identity between sequences in an identified group where there is more than one sequence. It takes a long time to generate this information, so if you don't need it, click "No". If you want the minimum pairwise identity and "consensus" sequence information for each identified group, click "Yes".
- If you clicked "Yes", a window will pop up displaying a message indicating which of the identified groups it's currently processing.

- The program will take a while to run. On our machines it typically takes anywhere from 2 to 4 hours. The more data you have and the larger the clusters are, the longer it will take to run.
- The output will not be displayed or saved automatically. You must save the table file by going into File -> Save tabulation into file. The table will be saved in a tab-delimited text format.
- Open the table by using Excel. Be sure to specify during the opening process that the file is tab-delimited, and that the column containing fingerprints is in "text" format, not "general" format (or the fingerprint will be treated as a number and the output will be in scientific notation).

Output format of taxonomic table:
Column A: Group number. This is an arbitrarily assigned number given to each cluster so that each cluster can be identified.

Column B: No. clones in cluster. This is the number of clones in total in the cluster.

Columns C to ?: Each column will stand for a separate treatment, as indicated by a prefix in the clone name (see instructions for **Gene ID Generator** macro), and contain the number of clones within each treatment in the appropriate column.

Next column: Minimum pairwise identity. This contains the minimum pairwise identity between sequences in the "taxon sequence" file that are predicted to have the same fingerprint as the cluster. A taxon sequence will not always be found to match the fingerprint for a given cluster. In this situation, this column will be blank. Also, when there is only one sequence with a predicted fingerprint matching that of a cluster, there will be no value in this column, since there are no two sequences to compare and obtain a pairwise identity for. Generally high values for minimum pairwise identity (97% or higher) indicates that the probe set you are using can distinguish clones with different sequences to the species level.

Next column: Fingerprint. This contains the fingerprint that describes the cluster.

Next column: List of all clones in cluster. The first clone in this cluster is in the same row as all the other information for the cluster. Additional clones in the cluster are listed in subsequent rows.

Next column: Consensus sequence of cluster. This is not actually a true consensus sequence of the sequences used to generate the minimum pairwise identity value, but rather the sequence from the "taxon sequences" file that most closely resembles the consensus of these sequences. Where only one sequence from the "taxon sequences" file is predicted to have the same fingerprint as the cluster, this will be the "consensus" sequence.

The "consensus" sequences can also be used to generate a phylogenetic tree, but you'd need one sequence for every single cluster. If you don't have very many clusters you can

obtain these sequences for unidentified clusters by sequencing inserts from clones representative of that cluster. We've never bothered to do this.

The last three rows of output will contain calculated values for fingerprint richness (S), diversity (inverse of Shannon's index, H) and evenness (E) for each treatment.

Other output files

These can be generated by running through the same process as listed above for taxonomic tables. Once the program is finished, you indicate the desired output and save it. Be sure that if you want data concerning minimum pairwise identities or taxon sequences for identified groups, that you select "Yes" when asked to "proceed to compute the identity value per group".

-File -> Save. A popup window will appear in which you can select one of four types of data to save: clone names, sequences, fingerprints, or consensus sequences. Click to select the data type.
-Next select which fingerprint groups to save this data for. You can choose to save data for groups with a certain range of minimum pairwise identity values, or groups with a certain range of group numbers.
-Click "Save." A tab-delimited text file will be generated that you can open in Excel.