



CotQuest User's Guide

For CotQuestG and CotQuestU
Program Versions 1.0, August 2008

Philippe Chouvarine¹, Daniel G. Peterson¹, and John Bunge²

¹Mississippi Genome Exploration Laboratory (www.mgel.msstate.edu), Department of Plant & Soil Sciences, Life Sciences and Biotechnology Institute, and Institute of Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA

²Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA

Table of Contents

Overview of CotQuest	1
The Nature of Cot Data	1
CotQuestG	
Download and Setup	3
User Interface	3
Creating a Data File	4
Analyzing a Dataset	5
Outlier Detection	8
CotQuestU	
Download and Setup	10
Directory and File List	11
Data Input Format	12
Required User Specifications	12
Running the Cot Analysis Programs	13
Outlier Detection	13

Overview of CotQuest

CotQuest is a suite of programs designed for automated nonlinear regression analysis of DNA reassociation kinetics (i.e., Cot) data. It consists of several scripts that, when used in association with the statistical software package SAS® (www.sas.com), provide users with a powerful, efficient, and statistically robust means of analyzing their Cot data. The purpose, utility, and underlying logic of CotQuest are detailed in the article “*CotQuest: Improved algorithm and software for nonlinear regression analysis of DNA reassociation kinetics data*” by Bunge, Chouvarine, and Peterson (manuscript submitted). This manual does not duplicate the content of that paper but rather provides information to assist readers in installation and operation of CotQuest.

CotQuest is packaged in two formats:

(1) **CotQuestG** comes with a graphical user interface (GUI) that *guides* users through the Cot analysis process. In short, the GUI *asks* users to answer a few relevant questions about their data and their research goals. This information is used by CotQuestG to produce customized SAS scripts which the GUI loads into the SAS program. With one click of the mouse, the customized CotQuestG SAS code will be used by SAS to automatically generate a summary results page through which one can access Cot curves, comparative model statistics/graphs, etc. CotQuestG requires Microsoft Windows with .NET Framework 2.0 (or higher), which is included in Windows XP (with current updates) and Windows Vista or can be downloaded from the Microsoft website (www.microsoft.com or more specifically, [click here](#)).

(2) **CotQuestU**, unlike CotQuestG, can be used with any SAS-compatible operating system (Windows, Macintosh, Linux, and UNIX). The operating system flexibility of CotQuestU comes at a slight cost with regard to automation as users must make changes in the CotQuestU SAS files without the aid of a GUI. However, this user’s manual explains in detail how to make these minor script changes. The output files produced through use of CotQuestU are identical to those generated using CotQuestG.

The Nature of Cot Data

A schematic diagram of a Cot experiment is shown in Figure 1. The data resulting from such an experiment consists of “Cot points.” Each Cot point consists of a pair of x and y values where $x = \text{Cot}$ and $y = \text{fraction single-stranded DNA}$. The required dataset consists of n Cot points $(x_1, y_1), \dots, (x_n, y_n)$, where typically $n \sim 10^1 - 10^2$, although denser experiments with larger sample sizes are possible. The main program fits a particular family of nonlinear curves to such a dataset. For each member of the family, the program returns various quantitative and graphical data analyses and diagnostics which help the user identify an optimal member of the family as the preferred model. If desired, a second program can be used to examine the data for outliers (relative to the given family of models). This program “nominates” (or identifies) outliers, which may then be deleted; the main program is then re-run on the reduced (outlier-deleted) dataset.

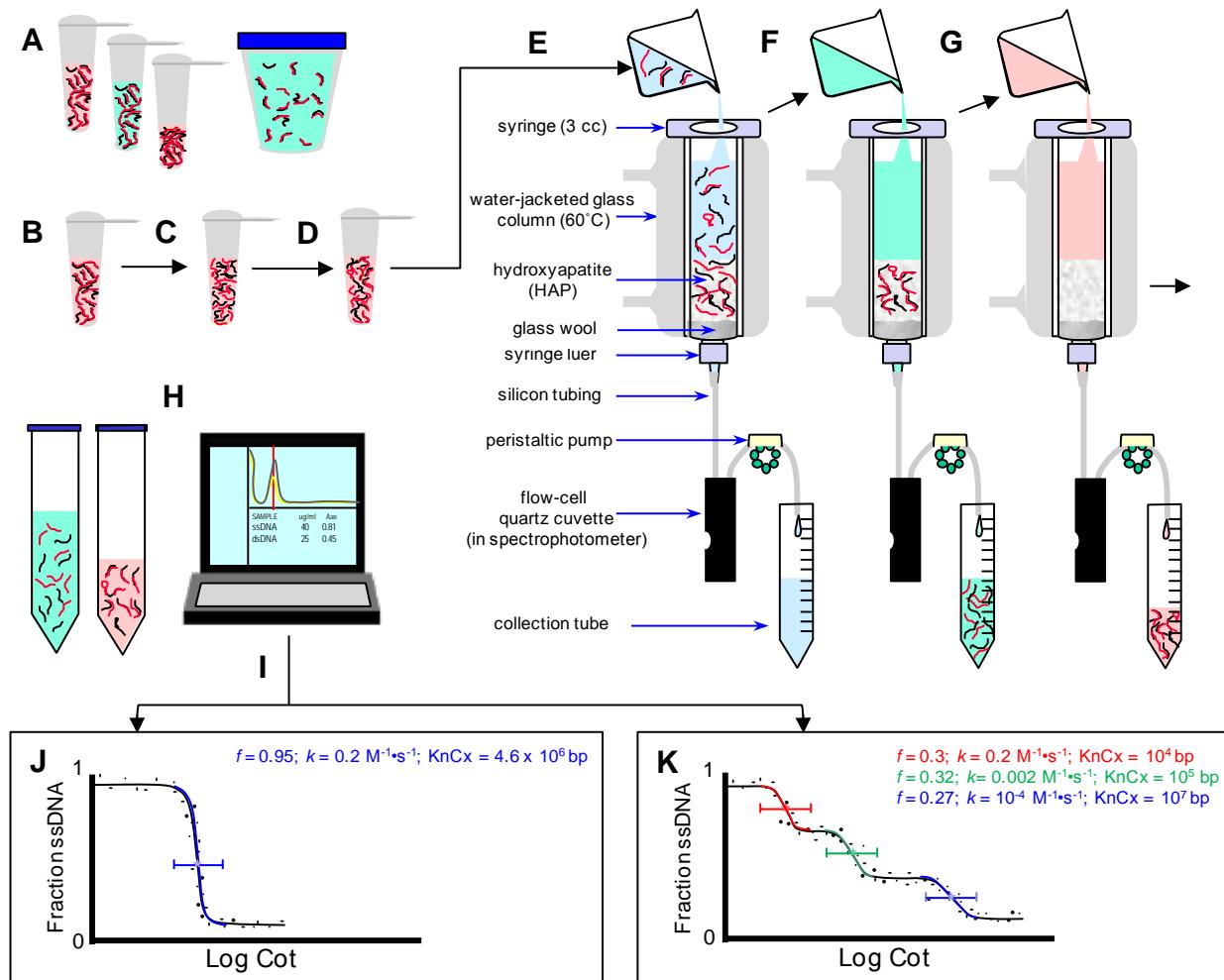


Fig. 1. Overview of Steps Involved in Cot Analysis Using Hydroxyapatite (HAP) Chromatography. (A) DNA from a particular source is sheared into 450 bp fragments and aliquots are dissolved in 0.03, 0.12, or 0.5 M sodium phosphate buffer (SPB) and sealed into glass or plastic tubes. (B) One tube is (C) placed in boiling water to denature DNA duplexes, and then transferred (D) into a water bath set at a temperature 25°C below the melting temperature for genomic DNA in that particular buffer (determined previously). Renaturation is allowed to proceed to a specific Cot value (see text for definition). (E) Once the sample has reached the desired Cot, it is quickly diluted in a 100-fold excess of 0.03 M SPB and loaded onto a HAP column equilibrated with 0.03 M SPB. At this buffer concentration all DNA binds to the HAP. (F) 0.12 M SPB is added causing single-stranded DNA (ssDNA) to elute. (G) 0.50 M SPB is added to the column to elute double-stranded DNA (dsDNA). (H) The volumes and absorbance values of the ssDNA eluant and the dsDNA eluant are used to determine the fraction of ssDNA for the Cot value (Peterson et al., 1998). (I) Steps B-H are repeated for all the DNA samples with each sample being renatured to a different Cot value. The fraction of ssDNA for each sample is plotted against the logarithm of its Cot value to yield a Cot point, and ultimately a graph of Cot points ranging from essentially no renaturation to nearly complete renaturation is prepared. Non-linear regression analysis, typically performed using a computer program, is used to generate a best-fit *Cot curve* for the data. (J) If the DNA source used in step A is a prokaryotic, viral, or organellar genome, the resulting Cot curve will exhibit a shape approximating a single second-order kinetics reaction (see text for details). In fitting the curve, the Cot analysis program will generate the reassociation constant k (and/or the $\text{Cot}_{\frac{1}{2}}$), the relative fraction f of the genome encompassed in the curve, and the genome's kinetic complexity (KnCx). (K) If the starting DNA is from a eukaryotic genome or an environmental sample, the resulting Cot curve will be an amalgam of several second-order subcurves or components. Each component represents DNA sequences of similar iteration within the genome (in the figure, the curve is composed of three components indicated by red, green, and blue, respectively). For each component, the program identifies its $\text{Cot}_{\frac{1}{2}}$ and/or k , f , and KnCx .

CotQuestG

Download and Setup

CotQuestG requires Windows with .NET Framework 2.0 or higher and SAS® statistical software installed. Chances are you already have one of the required versions of .NET Framework installed if you use Windows XP with current updates or Windows Vista. Otherwise you can download it from Microsoft ([click here](#)). At many research institutions SAS is available for a small fee under an institution-wide site license. CotQuestG with all required SAS scripts can be downloaded from <http://www.mgel.msstate.edu/tools.htm> as a ZIP archive. Create a directory C:\CotQuestG and extract the archive to that directory. No additional installation is needed. Double click on the CotQuestG.exe icon (Figure 2) to start the program.

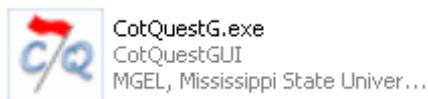


Fig. 2. The CotQuestG.exe icon

User Interface

Starting the program will bring up the Welcome window (Figure 3). From this window you can either choose to create an input file with Cot data or analyze your existing input file.

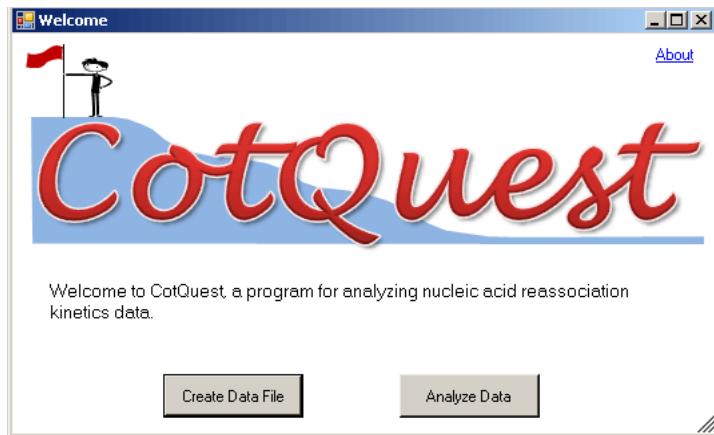


Fig. 3. The Welcome window

Creating a Data File

If you choose to create a new data file you will see the following window (Figure 4):

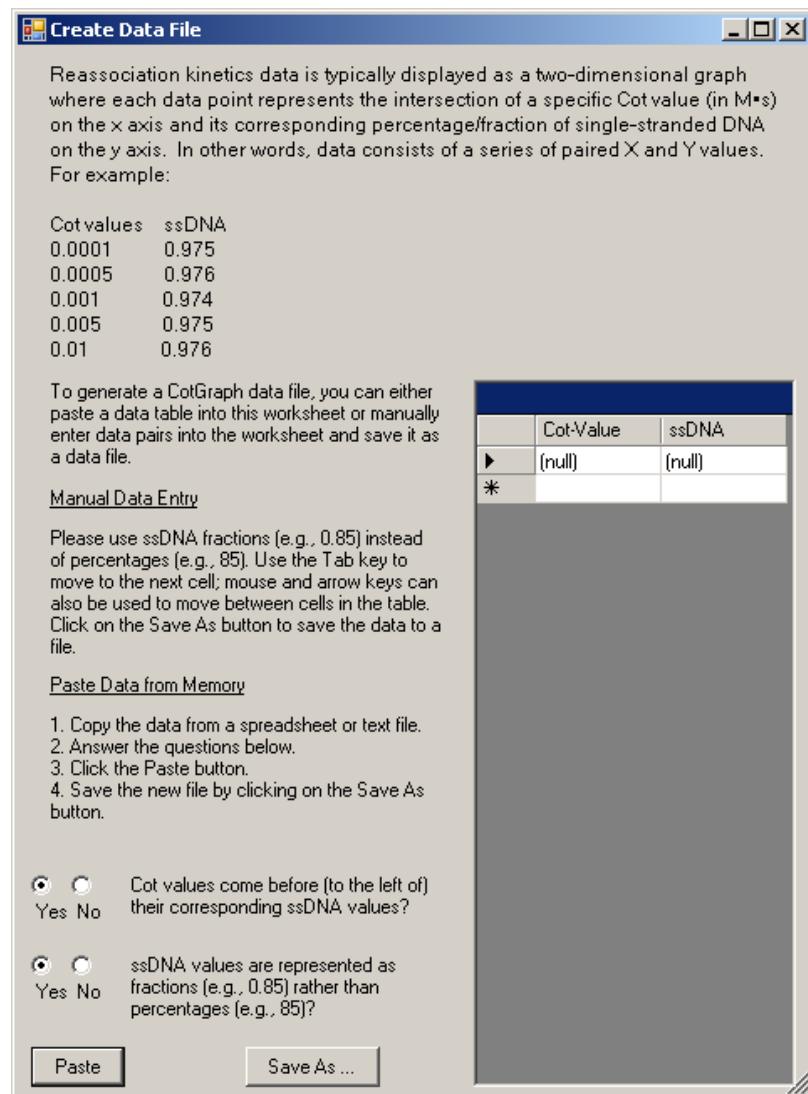


Fig. 4. Create Data File dialog window

The generation of a Cot data file (*.cot) is explained in the text of the "Create Data File" window. Once you have created your data file, close the "Create Data File" window. The Welcome window should still be open.

Analyzing a Dataset

To analyze a dataset press the Analyze Data button on the Welcome window. The following dialog window will be displayed (Figure 5):

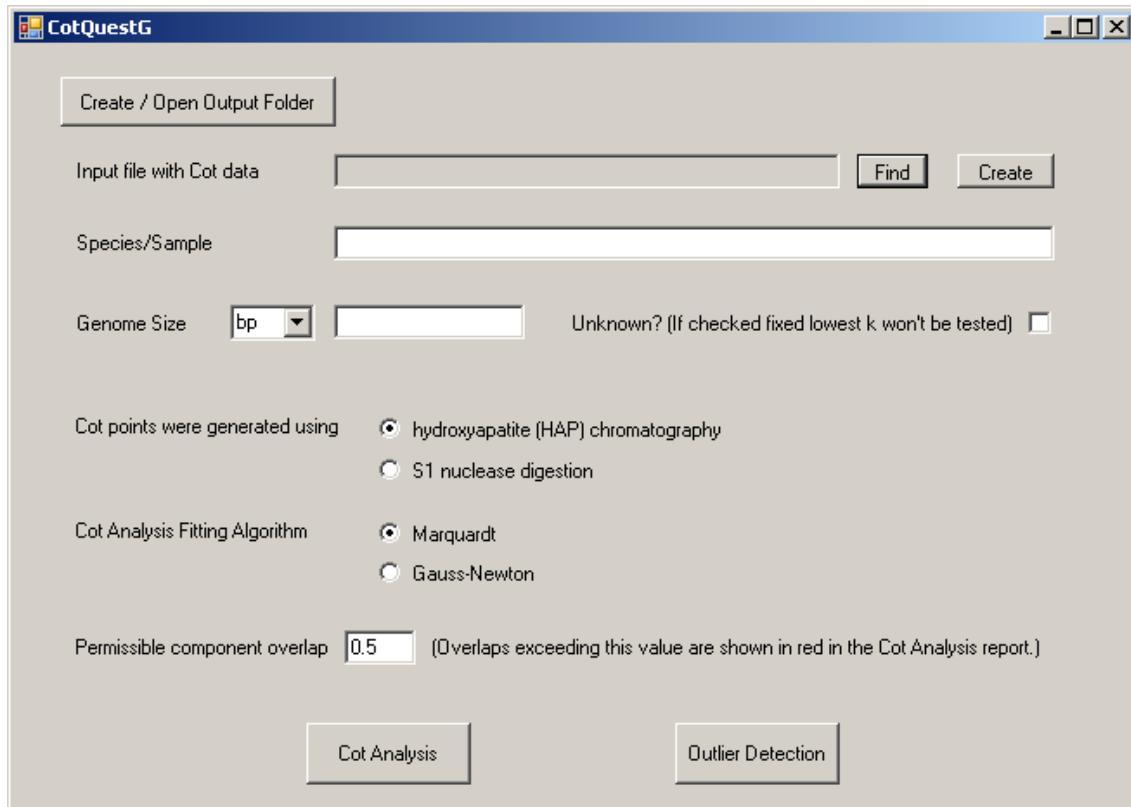


Fig. 5. Dialog window for collection of initial settings and algorithm selection.

Create/Open Output Folder: This button opens a folder browsing dialog window that lets you select an existing folder or create a new folder in which your results will be placed (Figure 6). To create a new folder, for example in C:\CotQuestG\Output\, press the “+” button by the Output folder (Figure 6, left frame), then press the ‘Make New Folder’ button and type in folder name. Select your new folder and press OK. Once a folder has been selected/created, the folder’s path will be displayed to the right of the button (Figure 6, right frame).

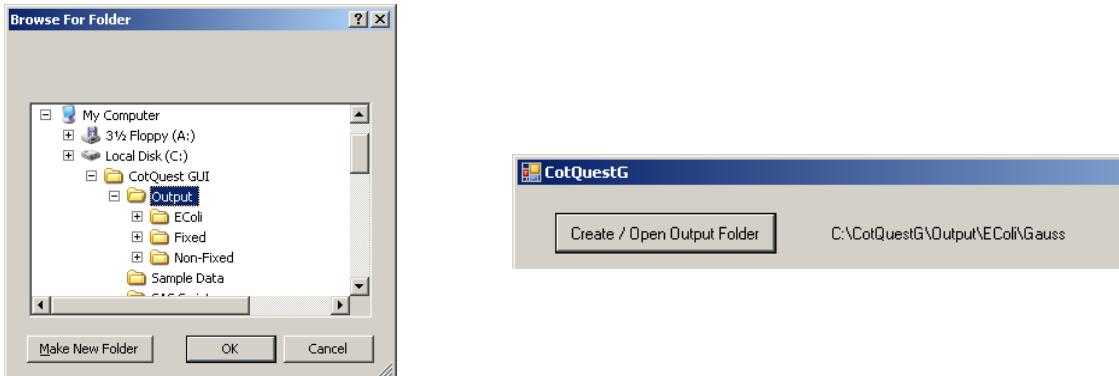


Fig. 6. Creating a new folder.

Input file with Cot data: Use the “Find” button to select an input Cot data file.

Species/Sample: Enter a descriptive name into this box. This text will be displayed in Cot analysis report.

Genome Size: Genome size (1C) can be specified by entering a value in this box. The genome size is used to create additional models by fixing the lowest k (see the original article for details). If the genome size is unknown or you do not want to generate fixed k models, check the ‘Unknown?’ checkbox.

Note: Genome size values in picograms (pg) can be converted into base pairs (bp) using the following formula: $bp = pg \cdot (0.978 \times 10^9 \text{ bp/pg})$.

Cot points were generated using: Select the appropriate option.

Cot Analysis Fitting Algorithm: Choose either Marquardt or Gauss-Newton (see research paper for details).

Permissible Component Overlap: The default value is 0.5. Values exceeding this constant will be shown in red (as warnings) in the resulting report. Fraction overlaps > 0.5 typically indicates that the data could be explained by a simpler model.

Cot Analysis: To start a Cot analysis, press the Cot Analysis button. The following window will appear (Figure 7):

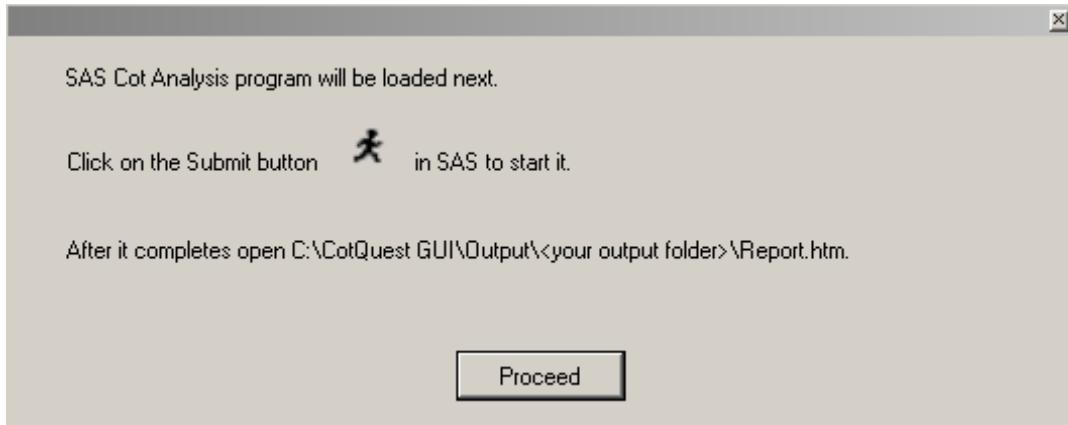


Fig. 7. Window that opens after the Cot Analysis button is pressed

Once the Proceed button has been pressed, SAS will automatically open (assuming that it has been properly installed on your computer). A CotQuestG script customized based on your entries in the main dialog window (Figure 5) will be visible in the lower half of the screen (Figure 8).

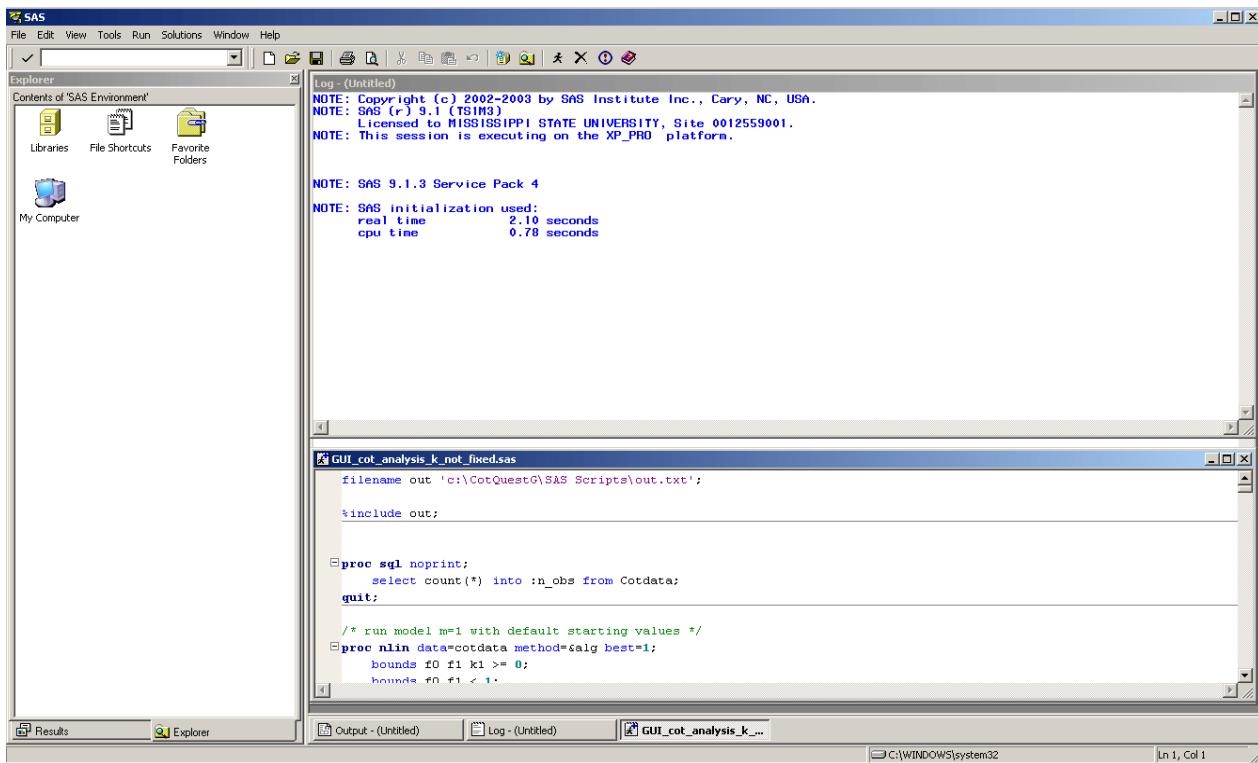


Fig. 8. SAS with the CotQuestG script loaded

Click on the Submit button (Figure 9) to start the analysis.



Fig. 9. The SAS Submit button

The Cot analysis process is computationally intensive and will take 10-30 minutes on an average single-processor desktop computer. The script generates a series of output reports which are saved to the ReportFiles subdirectory in the output folder selected/created as described above. The output folder will also contain a file called **Report.htm**. This page provides a summary of the analysis, biological values traditionally produced in Cot analyses, links to all generated HTML output files, and detailed consolidated statistics for model selection (Figure 10). You can compare AICc, A-sq, W-sq, convergence, and fraction overlap percentages to make an educated decision about the nonlinear regression model that is likely to be the best description of your data (see scientific paper for details).

Once SAS has finished running, the SAS program can be closed. The CotQuestG program will still be open.

Outlier Detection

From the dialog window shown in Figure 5 one can also conduct a SAS-based detection of potential outliers. In short, press the Outlier Detection button and proceed by following the instructions in the next two dialog windows. The Outlier Detection step may be repeated more than once until no more outliers are detected. If you save your dataset without outliers to a new file make sure you select this new file in the main window using the [Find] button before repeating the Outlier Detection step or proceeding with an additional Cot Analysis.



Version 1.0

Species/Sample: Sorghum bicolor BTx623
Algorithm: marquardt
Input file: C:\CotQuestG\Sample Data\sorghum.cot

Date: 06/16/2008
Number of Cot points: 62

	AICc, A-sq, W-sq, Convergence	Component	Fraction	Kinetic Complexity	k	Cot½	Residual Analysis		
Model 1 Cot curve	-344.655, 0.97140860, 0.15991545, Converged NLIN output	Reassociated	0.6537	9757434	0.0680	14.7059	Plots Data		
		Unreassociated	0.1225						
Model 1F Cot curve	-224.184, 1.51459463, 0.25986591, Converged NLIN output	Reassociated	0.5737	416491324	0.001398	715.2454	Plots Data		
		Unreassociated	0						
Model 2 Cot curve	-415.658, 0.69943367, 0.10762848, Converged NLIN output	Highly Repetitive	0.4607	1643045	0.2846	3.5137	Plots Data		
		Single\Low	0.2963	113488491	0.00265	377.3585			
Model 2F Cot curve	-414.521, 0.55600124, 0.08845857, Converged NLIN output	Unreassociated	0.0493				Plots Data		
		Fraction overlap. HR-SL: 0%							
Model 3 Cot curve	-436.04, 0.48570032, 0.07194388, Converged NLIN output	Highly Repetitive	0.4998	2385035	0.2127	4.7015	Plots Data		
		Single\Low	0.2695	195667024	0.001398	715.3076			
Model 3F Cot curve	-438.65, 0.46189585, 0.06839740, Converged NLIN output	Unreassociated	0.0323				Plots Data		
		Fraction overlap. HR-SL: 0%							
Model 4 Cot curve	-436.04, 0.48570864, 0.07194535, Partially Converged NLIN output	Highly Repetitive	0.1472	19350	7.7214	0.1295	Plots Data		
		Moderately Repetitive	0.416	4075676	0.1036	9.6525			
Model 4F		Single\Low	0.2359	171272175	0.001398	715.3076	Plots Data		
		Unreassociated	0.0393						
Fraction overlaps. HR-MR: 6.45%, MR-SL: 7.21%									
The model is invalid									

Other Data/Graphs:

[Cot curves for Models 1-4](#)

[Cot Curves for Models 1F-4F](#)

[Cot Points](#)

[Parameter Estimates for Models 1-4](#)

[Parameter Estimates for Models 1F-4F](#)

[Stats for Models 1-4](#)

[Stats for Models 1F-4F](#)

Fig. 10. Report.htm provides a summary of the analysis, links to all program output files, and detailed model statistics

CotQuestU

Download and Setup

CotQuestU can be downloaded from <http://www.mgel.msstate.edu/tools.htm> as a ZIP archive. Create a directory C:\CotQuestU and extract the archive to that directory. No additional installation is needed.

The CotQuestU package consists of six SAS scripts and a few auxiliary files (see Table 1 for file descriptions). The scripts are platform independent and can run on any computer that has SAS statistical software installed. There are versions of SAS available for Windows, Macintosh, Linux, and UNIX. File paths in the provided SAS code and in this manual are in the Windows format. They should be modified to run the code on other operating systems. At many research institutions SAS is available for a small fee under an institution-wide site license.

To load a CotQuestU SAS script into SAS either double-click on a *.sas file which starts SAS and reads in the CotQuestU script or start SAS and open the appropriate *.sas program from the CotQuestU folder.

Directory and File List

Table 1. Essential directories and files included in the package and their functions.

Directory or File	Description
C:\CotQuestU\OutlierDetection\	Contains all outlier detection SAS scripts. Their HTML report (Outliers.htm) is generated in this directory.
C:\CotQuestU\OutlierDetection\OutlierDetection_m3f_ROUTf.sas	SAS script for outlier detection in Cot data with 3 or more components when genome size is known.
C:\CotQuestU\OutlierDetection\OutlierDetection_m3_ROUT.sas	SAS script for outlier detection in Cot data with 3 or more components when genome size is unknown or when Cot analysis will be performed with cot_analysis_k_not_fixed.sas and won't include additional models with the lowest k fixed.
C:\CotQuestU\OutlierDetection\OutlierDetection_m2f_ROUTf.sas	SAS script for outlier detection in Cot data with 2 or more components when genome size is known.
C:\CotQuestU\OutlierDetection\OutlierDetection_m2_ROUT.sas	SAS script for outlier detection in Cot data with 2 or more components, when genome size is unknown or when Cot analysis will be performed with cot_analysis_k_not_fixed.sas and won't include additional models with the lowest k fixed.
C:\CotQuestU\OutlierDetection\OneComponentOutlierDetection_ROUT.sas	SAS script for outlier detection in potential one-component Cot data.
cot_analysis_Fixed_Lowest_k.sas	SAS script that generates one-, two-, three-, and four-component models and the corresponding models with fixed lowest k .
cot_analysis_k_not_fixed.sas	SAS script that generates one-, two-, three-, and four-component models.
one_component_cot_analysis.sas	SAS script that generates a one-component model and has an option to keep the sum of reassociated and non-reassociated components within 100%. [Some one-component datasets have an optimal fit where f_1+f_0 is slightly greater than 1].
C:\CotQuestU\ReportFiles\	Default storage area for GIF and HTML report files generated by cot_analysis_Fixed_Lowest_k.sas and cot_analysis_k_not_fixed.sas.
C:\CotQuestU\OneComponentReportFiles\	Default storage area for GIF and HTML report files generated by one_component_cot_analysis.sas.
C:\CotQuestU\Report.htm	Generated HTML file to view Cot analysis results produced by cot_analysis_Fixed_Lowest_k.sas or cot_analysis_k_not_fixed.sas.
C:\CotQuestU\OneComponentReport.htm	Generated HTML file to view Cot analysis results produced by one_component_cot_analysis.sas.
C:\CotQuestU\out.txt	Auxiliary file used by SAS scripts.
C:\CotQuestU\Sample Data\	Contains sample files with Cot data in plain text format.
C:\CotQuestU\Output\	Contains sample Cot analysis reports.

Data Input Format

Data must be in a space-delimited standard text (ASCII) file with record delimiters (carriage returns), of the form

y1[space] x1[return]
y2 [space]x2[return]
etc.

or as it would actually look in a text file

y1 x1
y2 x2
...
yn xn

Note that for each Cot point the y-value (ssDNA) is listed first and the x-value (Cot) is listed second. The entire file consists of data; there are NO variable names in the first row. The program automatically assigns variable names. Sample input files are included in the Sample Data folder.

Required User Specifications

Once a CotQuestU is loaded into SAS a few settings must be adjusted by changing the code in the lower SAS window (e.g., Figure 8). For the CotQuestU SAS scripts, the user MUST set the data input file pathname, the organism name, and the Cot generation constant (1 for hydroxyapatite; 0.44 for S1 nuclease digestion – see the scientific paper for details). If the organism's genome size is known, this can be entered as well to allow fixing of the lowest k (this applies only to those scripts designed for use when genome size is known – see Table 1).

Example:

To enter the settings for an analysis of the bald cypress (*Taxodium distichum*) dataset using the script *cot_analysis_Fixed_Lowest_k.sas*, open *cot_analysis_Fixed_Lowest_k.sas* and modify the following lines of code at the beginning of the SAS file (SAS code is shown in blue; explanatory notes not actually in the SAS script are shown in red italics):

```
infile 'C:\CotQuestU\Sample Data\cypress.cot';                                location of input file
%let G = 9751000000;                                                        genome size in base pairs
%let organism = Taxodium distichum;                                         name of species/sample
%let cgc = 1;                                                               Cot generation constant (1 for HAP, 0.44 for S1 nucl.)
The user MAY also change fitting algorithm and permissible overlap:
%let alg = gauss;                                                       gauss and marquardt are acceptable values
%let permis_fraction_overlap = 0.5;                                         0.5 is the default; can use any value between 0 and 1
```

For the outlier-detection programs, the same lines of code MUST be modified to specify the input file pathname and the organism genome size. The user may also change the FDR threshold (see Motulsky and Brown, 2006, BMC Bioinformatics, 7:123).

```
%let Q = 0.01;
```

Running the Cot Analysis Programs

We will work through a typical example using the sorghum (*Sorghum bicolor*) dataset included in the archive in the sorghum.cot file.

(1) Open the main program (C:\CotQuestU\cot_analysis_Fixed_Lowest_k.sas) in SAS. The easiest way to do this is to double click on the cot_analysis_Fixed_Lowest_k.sas file.

(2) Enter the input file pathname (e.g., `infile 'C:\CotQuestU\Sample Data\sorghum.cot'`) and the sorghum genome size (`%let G = 730000000;`). Change other settings if needed.

(3) Click the Submit button. It looks like this:  . The CotQuest Cot analysis program is computationally intensive and will run for 10-30 minutes on an average single-processor desktop computer. It generates a long list of output reports in the Results panel. To facilitate easy results navigation, the most vital results are also saved to the ReportFiles subdirectory. The program also creates HTML pages in this subdirectory for organization and viewing of the generated image files. To view this essential program output open C:\CotQuestU\Report.htm. This page provides links to all generated HTML output and detailed consolidated model statistics for model selection (Figure 10). You can compare AICc, A-sq, W-sq, convergence, and fraction overlap percentages to make an educated decision about the model you want to select. The models that failed to stay in bounds are marked as invalid and their statistics are not reported. This page also includes links to the detailed non-linear regression (NLIN PROC) output. If it is hard to select the model based on the statistical values you can also compare the residual plots and Cot graphs using links from Report.htm.

(4) Every time the program executes, it will overwrite the contents of the ReportFiles subdirectory and Report.htm. To prevent this, copy the ReportFiles folder and the Report.htm page to a new folder. A few sample reports are stored in organism-specific folders in C:\CotQuestU\Output\.

Outlier Detection

For the outlier-detection programs, the input file pathname and the organism genome size must be entered in the loaded SAS script. The user may also change the FDR threshold (see Motulsky and Brown, 2006, BMC Bioinformatics, 7:123). The default is

```
%let Q = 0.01;
```