Cot Analysis using Models 1, 1F, 2, 2F, 3, 3F, 4, and 4F – Pseudo-code

Assign values to variables of input file, genome size, nonlinear regression algorithm, and permissible fraction overlap

If Cot generation method is HAP chromatography

Assign 1 to variable *cgc*

If Cot generation method is S1 nuclease digestion

Assign 0.44 to variable *cgc*

Calculate fixed k as 1020628.74 divided by genome size

Read the input file and count the number of observations

Perform nonlinear regression of Cot data using model m=1: ssDNA = f0+f1*(1+k1*cot)**(-1**cgc*) with the following bounds: f0, f1, k1 >= 0 and f0, f1 < 1; and the following parameter search grids: f0=.01 to .26 by .05, f1=.01 to .96 by .05, and k1=1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2.

Save the output parameter estimates, convergence status, ANOVA, and predicted ssDNA values for model m=1.

Perform nonlinear regression of Cot data using model m=1F with fixed k1: ssDNA = f0+f1*(1+kfixed*cot)**(-1**cgc*) with the following bounds: f0, f1 >= 0 and f0, f1 < 1; and the following parameter search grids: f0=.01 to .26 by .05 and f1=.01 to .96 by .05.

Save the output parameter estimates, convergence status, ANOVA, and predicted ssDNA values *ssDANp* for model m=1F.

Obtain upper bound *maxf0* for f0 for models m=2 and m=2F within the 95% confidence interval for f0 given by model m=1 (as 1 – lower bound of f1)

Calculate grid search step for f0 of models m=2 and m=2F as (*maxf0* - .01)/5

Create bounds and step sizes using model m=1 parameter output file for the remaining parameters for models m=2 and m=2F as follows: lowercl = max(0,lowercl); uppercl = min(uppercl,1); step = (uppercl-lowercl)/5.

Calculate residuals for models m=1 and m=1F as *ssDNA – ssDNAp*

Using the output data for models m=1 and m=1F create the following graphs:

Normal Probability Plot of the Residuals

Histogram of the Residuals

Residuals versus the Fitted Values

Residuals versus the Order of the Data

Perform nonlinear regression of Cot data using model m=2: ssDNA = f0+f1*(1+k1*cot)**(-1**cgc*)+f2*(1+k2*cot)**(-1**cgc*) with the following bounds: f0, f1, f2, k1, k2 >= 0 and f0, f1, f2 < 1; and the following parameter search grids: for f0, f1, and k1 use lower bounds, upper bounds and steps calculated above; f2 = 0.01 to *maxf0* by step (calculated above), and k2=1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2.

Save the output parameter estimates, convergence status, ANOVA, and predicted ssDNA values for model m=2.

Find the value of the lowest k of model m=2 and assign it to *lk2.*

Perform nonlinear regression of Cot data using model m=2F. Use the same ssDNA equation as in model m=2, but substitute *lk2* for the lowest k parameter. The bounds and parameter search grids are the same as in model m=2, except the lowest k parameter is omitted from them.

Save the output parameter estimates, convergence status, ANOVA, and predicted ssDNA values for model m=2F.

Obtain upper bound *maxf0* for f0 for models m=3 and m=3F within the 95% confidence interval for f0 given by model m=2 (as 1 – sum of lower bounds of f1 and f2)

Calculate grid search step for f0 of models m=3 and m=3F as (*maxf0* - .001)/5

Create bounds and step sizes using model m=2 parameter output file for the remaining parameters for models m=3 and m=3F as follows: lowercl = max(0,lowercl); uppercl = min(uppercl,1); step = (uppercl-lowercl)/5.

Calculate residuals for models m=2 and m=2F as *ssDNA – ssDNAp*

Using the output data for models m=2 and m=2F create the following graphs:

Normal Probability Plot of the Residuals

Histogram of the Residuals

Residuals versus the Fitted Values

Residuals versus the Order of the Data

Perform nonlinear regression of Cot data using model m=3: ssDNA = f0+f1*(1+k1*cot)**(-1**cgc*)+f2*(1+k2*cot)**(-1**cgc*)+f3*(1+k3*cot)**(-1**cgc*) with the following bounds: f0, f1, f2, f3, k1, k2, k3 >= 0 and f0, f1, f2, f3 < 1; and the following parameter search grids: for f0, f1, k1, f2, k2 use lower bounds, upper bounds and steps calculated above; f3 = 0.001 to *maxf0* by step (calculated above), and k3=1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2.

Save the output parameter estimates, convergence status, ANOVA, and predicted ssDNA values for model m=3.

Find the value of the lowest k of model m=3 and assign it to *lk3.*

Perform nonlinear regression of Cot data using model m=3F. Use the same ssDNA equation as in model m=3, but substitute *lk3* for the lowest k parameter. The bounds and parameter search grids are the same as in model m=3, except the lowest k parameter is omitted from them.

Save the output parameter estimates, convergence status, ANOVA, and predicted ssDNA values for model m=3F.

Obtain upper bound *maxf0* for f0 for models m=4 and m=4F within the 95% confidence interval for f0 given by model m=3 (as 1 – sum of lower bounds of f1, f2, and f3)

Calculate grid search step for f0 of models m=4 and m=4F as (*maxf0* - .0001)/5

Create bounds and step sizes using model m=3 parameter output file for the remaining parameters for models m=4 and m=4F as follows: lowercl = max(0,lowercl); uppercl = min(uppercl,1); step = (uppercl-lowercl)/5.

Calculate residuals for models m=3 and m=3F as *ssDNA – ssDNAp*

Using the output data for models m=3 and m=3F create the following graphs:

Normal Probability Plot of the Residuals

Histogram of the Residuals

Residuals versus the Fitted Values

Residuals versus the Order of the Data

Perform nonlinear regression of Cot data using model m=4: ssDNA = f0+f1*(1+k1*cot)**(-1**cgc*)+f2*(1+k2*cot)**(-1**cgc*)+f3*(1+k3*cot)**(-1**cgc*)+f4*(1+k4*cot)**(-1**cgc*) with the following bounds: f0, f1, f2, f3, f4, k1, k2, k3, k4 >= 0 and f0, f1, f2, f3, f4 < 1; and the following parameter search grids: for f0, f1, k1, f2, k2, f3, k3 use lower bounds, upper bounds and steps calculated above; f4 = 0.0001 to *maxf0* by step (calculated above), and k4=1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1, 1e2.

Save the output parameter estimates, convergence status, ANOVA, and predicted ssDNA values for model m=4.

Find the value of the lowest k of model m=4 and assign it to *lk4.*

Perform nonlinear regression of Cot data using model m=4F. Use the same ssDNA equation as in model m=4, but substitute *lk4* for the lowest k parameter. The bounds and parameter search grids are the same as in model m=4, except the lowest k parameter is omitted from them.

Save the output parameter estimates, convergence status, ANOVA, and predicted ssDNA values for model m=4F.

Calculate residuals for models m=4 and m=4F as *ssDNA – ssDNAp*

Using the output data for models m=4 and m=4F create the following graphs:

Normal Probability Plot of the Residuals

Histogram of the Residuals

Residuals versus the Fitted Values

Residuals versus the Order of the Data

Using ANOVA output and the number of observations obtained earlier calculate AICc of each model.

Using the obtained results plot Cot curves and arrange the output in a user-friendly HTML output.

NOTE: The other scripts that perform Cot analysis without fixing the lowest k or one-component Cot analysis are subsets of this script.

Outlier Detection – Pseudo-code

Set FDR threshold *Q*

Depending on the model selected for outlier detection perform the steps described in the Cot Analysis pseudo-code listing up to calculation of residuals for this model. For this example we use Model 3F. Lower-order fixed models (1F and 2F) can be omitted.

Calculate mean absolute deviation *mad*.

Fit nonlinear regression model with Cauchy (Lorentzian) distributed error terms with Cauchy scale parameter *sc* using the following model: ssdna = f0+f1*(1+k1*cot)**(-1**cgc*)+f2*(1+k2*cot)**(-1**cgc*)+f3*(1+k3*cot)**(-1**cgc*) , where the lowest k is replaced with *lk3*; set bounds as f0, f1, k1, f2, k2, f3, k3, sc >= 0 (omitting the lowest k) and f0, f1, f2, f3 < 1; set starting parameter values of f0, f1, f2, f3, k1, k2, k3 (omitting the lowest k) to those obtained from model m=3F and *sc* to *mad*/2, *mad*, 2**mad*.

Save output to a dataset.

Calculate the 68.27% quantile *qntl* of the absolute values of the residuals.

For each observation

Calculate Cauchy normalized residuals as *cnormres* = *ctrueres*/(*qntl***N*/(*N*-*P*)), where *ctrueres* is Cauchy residual, *N* is the number of observations, and *P* is the number of parameters in the operative model (6 for model 3F).

For each observation

Calculate p-values as *raw_p* = 2*(1-CDF('T',abs(*cnormres*),(*N*-*P*)))

For each observation

Calculate FDR-corrected p-values

For each observation

If FDR-corrected p-value is less than the FDR threshold *Q*

Mark the observation as outlier.

Save the resulting dataset to HTML file.