Analytical Biochemistry 388 (2009) 322–330

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

CotQuest: Improved algorithm and software for nonlinear regression analysis of DNA reassociation kinetics data

John Bunge [a], Philippe Chouvarine [b], Daniel G. Peterson [b,*]

[a] Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA
[b] Mississippi Genome Exploration Laboratory, Department of Plant and Soil Sciences, Life Sciences and Biotechnology Institute, and Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA

ARTICLE INFO

Article history:
Received 8 January 2009
Available online 12 March 2009

Keywords:
DNA reassociation kinetics
Cot analysis
Cot
Nonlinear regression
Algorithm
Software

ABSTRACT

Cot analysis (DNA reassociation kinetics) has long been used to explore genome structure in individual species, estimate genome similarity among organisms, and evaluate diversity in ecological samples, yet the algorithms and computational tools designed for analyzing Cot data are outdated, difficult to use, and prone to error. We report a new nonlinear regression procedure for analysis of Cot data and describe our algorithms in detail. Our procedure is implemented as CotQuest, a suite of scripts designed for use with the statistics package SAS. Unlike previous programs, CotQuest does not require users to input guesses as to the final values of parameters; rather, it employs a novel algorithm to step through a sequence of progressively more complex models, with the results from a given analysis being used to generate starting values for the next model. Moreover, CotQuest returns a statistical comparison of potential models and provides a variety of model assessment and selection diagnostics to help users in model selection. In situations where two models possess similar goodness-of-fit assessments, visual analysis of the Cot curves and comparison of CotQuest-generated graphs and statistics reflecting the normality and homoscedasticity of residuals can be employed to make educated choices between models.

The ability of DNA to denature and reassociate in a base-specific manner is central to cellular processes such as DNA replication and transcription. It also underlies many analytical molecular biology techniques, including the polymerase chain reaction, blot-based hybridization, microarray/gene chip technologies, and most DNA sequencing and resequencing methods.

Some of the earliest DNA reassociation research was conducted more than 40 years ago by Britten and his coworkers at the Carnegie Institution of Washington. Specifically, Roy Britten's group developed Cot analysis, a technique in which DNA reassociation kinetics is used to explore sequence composition/diversity [1,2]. Cot analysis has been used to characterize the genomes of individual species [1], estimate relatedness between species [3,4], and study diversity in complex environmental samples [5,6]. The work of Britten and coworkers paved the way for development of subsequent reassociation-based molecular techniques [7] and led to one of the most important discoveries in genome biology—specifically, the finding that eukaryotic genomes, on the whole, are dominated by repetitive nongenic sequences [1]. With the development of molecular cloning and DNA sequencing during the 1970s, Cot analysis was performed less frequently [8]. However, Cot research has experienced a resurgence in popularity due, in part, to the use of its underlying principles in complementary DNA (cDNA)[1] library "normalization" [9] and the development of Cot-based sequencing strategies that permit preferential isolation and sequencing of low- and/or high-copy sequences from a genome [8,10–13].

In Cot analysis, the product of DNA concentration (C_0), reassociation time (t), and a "buffer factor" accounting for cation concentration (δ) has a predictable effect on the amount of reassociation occurring in a denatured DNA sample [2]. The major unknown factor influencing reassociation is the underlying sequence composition of the DNA. Sequence composition is studied by exploring how changes in $C_0 t\delta$ (known by the colloquialism "Cot") influence reassociation. Typically, a graph is created where the fraction of reassociated DNA is plotted against the logarithm of Cot (from Cot $\sim$ 0 to Cot values at which reassociation is complete[2]) for DNA from a particular source. The resulting scatter plot is analyzed

* Corresponding author. Fax: +1 662 325 8742.
E-mail address: dpeterson@pss.msstate.edu (D.G. Peterson).

[1] *Abbreviations used:* cDNA, complementary DNA; HAP, hydroxyapatite; dsDNA, double-stranded DNA; ssDNA, single-stranded DNA; AICc, corrected Akaike's information criterion; AIC, Akaike's information criterion; FDR, false discovery rate; GUI, graphical user interface.

[2] Cot curves will rarely, if ever, start at complete denaturation (100% ssDNA) or end at complete reassociation (0% ssDNA). Intramolecular "foldback" will create some duplex regions at Cot values too small to allow actual intermolecular pairing. Moreover, a fraction of the DNA (usually < 5%) will never reassociate, possibly due to damage caused during DNA shearing [2].

0003-2697/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2009.03.007

using nonlinear regression analysis, and a least squares curve is fit through the data. This graph, known as a Cot curve, provides a visual representation of the genome.

There are two methods commonly used in generating Cot data. The older and more widely practiced of these approaches is the hydroxyapatite (HAP) chromatography method, which is based on the novel DNA binding properties of the calcium compound HAP. In short, HAP chromatography can be used to fractionate a partially reassociated sample (i.e., a sample reassociated to a specific Cot value) into double-stranded DNA (dsDNA) and single-stranded DNA (ssDNA). Quantification of the amount of DNA remaining single-stranded can be performed by comparing the product of the volume and spectrophotometric absorbance at 260 nm (A_{260}) of the ssDNA eluant with the corresponding product (volume · A_{260}) of the dsDNA eluant [14]. Alternatively, a small random portion of the DNA sample can be labeled with a radioisotope and reassociation can be accessed by measuring the radioactivity and volumes of the ssDNA and dsDNA fractions collected after HAP fractionation [15]. The general steps in a Cot analysis performed using HAP chromatography are shown in section A of the supplementary material.

Britten and his colleagues discovered that HAP-based Cot analyses of prokaryotic, viral, and organellar genomes (i.e., largely nonrepetitive genomes) produced curves with shapes approximating ideal second-order kinetic reactions [1]. As shown in Fig. 1, the point along the x axis at which reassociation is 50% complete is the curve's Cot½. The reassociation rate, k, for the curve is the inverse of its Cot½ and is proportional to the slope of the linear region. Roughly 80% of reassociation occurs in the "two Cot decade region" flanking the Cot½ (i.e., the region between 0.1·Cot½ and 10·Cot½). When Britten and Kohne compared Cot curves of different nonrepetitive genome species, they discovered that genome size is directly correlated with Cot½ [1]. The functional form of a second-order kinetic reaction for a nonrepetitive genome is $f(1 + k\ \mathrm{Cot})^{-1}$ [16], where f is a constant between 0 and 1 (see below).

In HAP-based exploration of the genomes of eukaryotic organisms, Britten's group discovered that reassociation for eukaryotes occurs over a much wider range in Cot than for prokaryotes/viruses/organelles. Careful study of HAP-based Cot curves prepared for complete eukaryotic genomes led Britten and coworkers to conclude that eukaryotic Cot curves are amalgams of two or more second-order subcurves known as "components." The resulting function is a mixture model, that is, a convex combination of the component subcurves (Fig. 1). Each component of a eukaryotic Cot curve can be described by its Cot½, its k, the fraction (f) of the genome for which it accounts, and its kinetic complexity (KnCx). KnCx is a Cot curve-based estimate of sequence complexity, that is, the total amount of novel sequence information in a component [8,10]. Despite the fact that eukaryotic genomes are typically hundreds to thousands of times larger than prokaryotic genomes, the fastest reassociating components of eukaryotes typically possess Cot½ values smaller than those of nonrepetitive genomes, a finding that led Britten and Kohne to conclude that eukaryotic genomes contain significant quantities of repetitive DNAs [1]. In eukaryotic Cot curves, the slowest reassociating component often represents reassociation of single-copy sequences, and the k of this component can be used to estimate genome size by comparison with the k and genome size of *Escherichia coli* [17]. The functional form of a second-order kinetic reaction for the ith component in a multicomponent Cot curve is $f_i(1 + k_i\mathrm{Cot})^{-1}$ [16].

The second technique used to generate Cot data is the S1 nuclease digestion method, which is centered on the ability of S1 nuclease to preferentially digest ssDNA in mixtures of dsDNA and ssDNA [18]. A DNA sample reassociated to a given Cot can be treated with S1 nuclease to eliminate ssDNA, and a comparison of the amount of DNA in the sample prior to denaturation and after S1 nuclease digestion can be used to determine the fraction of ssDNA at a given Cot (see section A of the supplementary material). Cot curves generated from S1 nuclease-derived Cot points deviate from second-order kinetics but can be roughly described by the form $f_i(1 + k_i\mathrm{Cot})^{-0.44}$ [16].

To facilitate analysis of Cot data, several computer programs have been developed to partially automate nonlinear least squares regression analysis and calculate values that can be used in biological comparisons. The first reassociation kinetics program was FINGER [2], which was subsumed by NNNBAT [16] and COTFIT [19]. These FORTRAN programs, all of which are rooted in the algorithm for least squares estimation of nonlinear parameters of Marquardt [20], originally were designed to run on mainframe computers or minicomputers. Later, Green and coworkers [21] developed a program that employed the derivative-free unconstrained optimization "pattern search" algorithm [22], specifically for use on an Apple II microcomputer. This program was not actually developed for DNA reassociation research but could ostensibly be adapted for HAP- or S1 nuclease-based Cot analysis. Of the aforementioned programs, NNNBAT has been used in the vast majority of Cot studies performed since 1980, including our research [10,14].

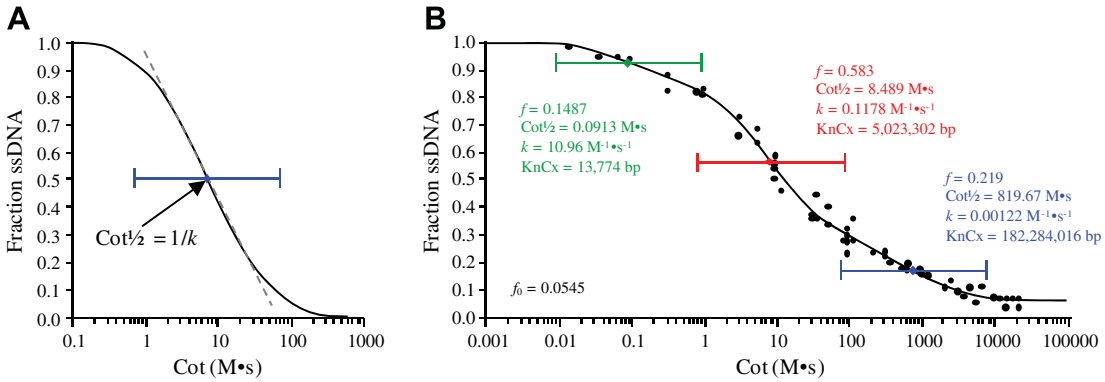


Fig. 1. Second-order reassociation kinetics. (A) An ideal second-order kinetics DNA reassociation reaction. Note that the central two-thirds of the curve is nearly linear (dotted line). The Cot½ is marked by a blue diamond, while the "two Cot decade region" is demarcated by brackets flanking the Cot½. Whereas the ideal curve starts at complete ssDNA and ends with complete reassociation, actual Cot curves generated from experimental data will rarely, if ever, start or end at these values due to DNA foldback or damaged DNA, respectively. (B) Multicomponent Cot curve for the eukaryote *Nicotiana tabacum*. Data were extracted from Zimmerman and Goldberg [25]. The CotQuest best fit curve for these HAP-generated data has three components indicated by the colors green, red, and blue.

NNNBAT and other existing Cot analysis tools require users to input guesses as to the final values of various parameters. These programs take these initial values and use them to converge on a solution that minimizes the least squares deviation of the function. Note that the output of a particular run is not the actual best fit for the dataset; rather, it is simply the best fit given the particular starting parameter guesses. Consequently, users must test a variety of different input value sets in attempting to find the neighborhood of the best fit solution. Once this neighborhood is reached (as is evidenced by relatively low goodness-of-fit and mean-squared-error values), minor changes in the input (guess) parameters should have little effect on the output. This process is time-consuming and can occasionally lead to suboptimal results. Most notably, all "parameter guessing" approaches are susceptible to local minima or maxima (over the parameter space) that can result in erroneous fits. Moreover, in our experience, we have found highly divergent fits to have nearly identical goodness-of-fit and mean-squared-error values, making model selection more arbitrary than convincing.

Although modern statistical software packages allow users an alternate means of conducting nonlinear least squares regression analysis of Cot data, these programs do not return the key biological data derived from curve fitting. Such values can be calculated, but they require a firm understanding of the mathematics and statistics of DNA reassociation kinetics, something that is rare even among those who have performed Cot analyses in the past. Moreover, these programs still require users to input parameter guesses.

Here we report a new nonlinear regression procedure for analysis of HAP- and S1 nuclease-derived Cot data. This procedure, which is implemented by using our freely downloadable program suite CotQuest in association with the statistical software "gold standard" SAS, eliminates the need for user-supplied starting values and, thus, circumvents problems associated with parameter guessing. A novel algorithm steps through a sequence of progressively more complex models, using the results from one model as the starting values for the next model. The program returns biologically relevant values for each model—most notably, each component's genome fraction (f), Cot½, k, and KnCx—as well as a graphical display of its Cot curve. For each model, the program yields a corrected Akaike's information criterion (AICc) value, which serves as the principal model selection statistic. In addition, CotQuest generates a variety of qualitative and quantitative model assessment and selection diagnostics, including residual analyses (in graphical format) and mean squared errors to assist investigators in making an educated choice between models with highly similar AICc values.

Materials and methods

NLIN options

Curve fitting was performed using a novel algorithm (see below) and the NLIN (nonlinear regression) module in SAS. We note that NLIN provides an option to specify the numerical search algorithm; the choices are Gauss (Gauss–Newton), Marquardt, Newton, and gradient. SAS documentation identifies gradient as the least robust method, and the Newton algorithm consistently produced poorer goodness-of-fit scores with our datasets (data not shown), so only the Gauss and Marquardt algorithms were included in development of our algorithm.

Algorithm

The general functions for Cot curves produced by HAP chromatography and S1 nuclease digestion are shown in Eqs. (1) and (2), respectively:

$$y = f_0 + \sum_{1\le i\le m} f_i(1 + k_i x)^{-1} \tag{1}$$

$$y = f_0 + \sum_{1\le i\le m} f_i(1 + k_i x)^{-0.44}, \tag{2}$$

where x denotes the "Cot value" and y denotes the proportion of ssDNA in the sample. For the derivation of these functions, see Britten and Kohne [1], Pearson and coworkers [16], or (more generally) Érdi and Tóth [23]. Here f_0 represents the "final unreassociated fraction" because y declines to f_0 as $x \to \infty$. In regression parlance, f_0 is the intercept because if $f_1 = \ldots = f_m = 0$, then $y = f_0$. The terms $f_1, \ldots, f_m$ are the fractions associated with each of the m components of the mixture, and $k_1, \ldots, k_m$ are the corresponding reassociation rates. Note that the fractions need not sum to 1 even including f_0. Given data (x_i,y_i), $i = 1, \ldots, n$, we need to fit Eq. (1) or (2) to the data via nonlinear regression for any given order m. Our algorithm for accomplishing this goal is based on some simple ideas:

- Empirical experience shows that f_0 is often less than 0.10 (10%), and almost certainly less than 0.25 (25%), so the search for f_0 need not cover the entire interval (0,1).
- All f_i ($i = 0, \ldots, m$) are bounded between 0 and 1, so an equally spaced grid in some subinterval of (0,1) represents a reasonable search pattern.
- As with all mixture models, the function is nonidentifiable with respect to permutation of the parameter indexes. For example, Function (1) with $m = 2$, $f_0 = 0.1$, $f_1 = 0.2$, $k_1 = 3$, $f_2 = 0.6$, and $k_2 = 10$ is the same as Function (1) with $m = 2$, $f_0 = 0.1$, $f_1 = 0.6$, $k_1 = 10$, $f_2 = 0.2$, and $k_2 = 3$; both are equivalent to $y = 0.1 + 0.2(1 + 3x)^{-1} + 0.6(1 + 10x)^{-1}$. We can exploit this fact to restrict our search (see below). We have essentially no *a priori* knowledge about the k_i except that they are positive. Therefore, it is plausible to use a logarithmic search grid in, say, powers of 10: $\ldots, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}$, and so forth. Empirical experience shows that possible values may extend fairly far down into the negative powers; we return to this below. The nonidentifiability mentioned above can also be used to narrow the search for the k_i.
- It is easier to fit a lower order model (smaller m) than a higher order model. Furthermore, it is reasonable to suppose that certain (approximate) relationships will hold between the parameter values for lower and higher order models.
- In its current implementation, the algorithm fits only $m = 1$ through $m = 4$, although higher orders could in principle be computed.

The model-fitting algorithm begins with $m = 1$ (Model 1). In this case, we search for f_0 across a grid from 0.01 to 0.25 in steps of 0.05, and we search for f_1 in a grid from 0.01 to 0.96 in steps of 0.05; the logarithmic grid for k_1 ranges from 10^{-7} to 10^{2}. Next, we move from Model 1 to Model 2 (i.e., $m = 2$). We now search for f_0 within the 95% confidence interval for f_0 given by Model 1 using standard error increments. These are t distribution-based intervals, so they consist of the parameter estimate ± approximately 2·SE, and it is reasonable to take five steps: (approximately) est $- 2 \cdot$ SE, est $- 1 \cdot$ SE, est, est $+ 1 \cdot$ SE, and est $+ 2 \cdot$ SE (thereby keeping the combinatorial explosion of the grid under control). We do the same for f_1 and k_1. The parameters for the second component must now be estimated. For f_2, note that $0 \leqslant f_0 + f_1 + f_2 \leqslant 1$; hence, $f_2 \leqslant 1 - (f_0 + f_1)$. Therefore, we search for f_2 along a grid ranging from one order of magnitude less than (the apparent values of) f_0 and f_1 up to 1 minus the sum of the lower 95% confidence bounds for f_0 and f_1. For k_2, we again have no *a priori* knowledge, so we use the full logarithmic grid. Fits for Model 3 ($m = 3$) and Model 4 ($m = 4$) are extensions of the same logic used in fitting Model 2. Note that the lower bound for f_{m+1} is decreased by one order of magnitude with each pass because

the highest indexed component may be assumed to be the smallest (under the nonidentifiability mentioned above).

Fixing the lowest k

If the genome size (1C DNA content) of a particular organism has been determined independently (e.g., via flow cytometry), the genome size value can be used to fix the k of the slowest reassociating component. The rationale behind fixing this parameter is discussed in the Results and Discussion below. With regard to the algorithm, the lowest k for each of the aforementioned models (Models 1–4) is set equal to $G \div (G_{\text{coli}} \cdot k_{\text{coli}})$, where G is the genome size of the organism for which the Cot curve has been prepared, G_{coli} is the genome size of *Escherichia coli* (i.e., 4,639,221 bp [24]), and k_{coli} is the empirically determined reassociation rate for *E. coli* (0.22 M^{-1} s^{-1} [25]). Each model is then refit using the fixed value while letting all other parameters float within their ranges as described above. We refer to the model $m = 1$ with a fixed lowest k as Model 1F, the model $m = 2$ with a fixed lowest k as Model 2F, and so forth. All models with the fixed parameter still require their preceding (lower order) models to run without fixed parameters so as to facilitate bounds selection.

AIC and AICc

Akaike's information criterion (AIC) is a widely used model selection statistic, and in our research AIC with a second-order correction for small sample sizes (AICc) is used as the primary model selection criterion. AIC is defined as $2p + n \ln(\text{RSS}/n)$, where p is the number of parameters, n is the number of observations, and RSS is the sum of squares of residuals. AICc is defined as $\text{AIC} + [2p(p+1)/(n-p-1)]$ [26]. If genome size is provided by users, our algorithm calculates AIC and AICc for each of the eight models (Models 1, 1F, 2, 2F, 3, 3F, 4, and 4F) generated for a particular dataset. If genome size is unknown or not provided by users, AIC and AICc values are calculated for the four "nonfixed" models (Models 1, 2, 3, and 4). In general, the model with the lowest AICc is considered the most accurate fit of the data.

Residual analysis

In instances where two or more models produce identical (or nearly identical) AICc values, residual graphs and statistics can be used as a secondary measure of model appropriateness. In classical linear and nonlinear regression analysis, the standard assumption is that the "errors" (vertical deviations of the observed data points from the true regression function) are independent, identically distributed (hence homoscedastic or equal-variance) random variables following the normal (Gaussian) distribution. To assess these assumptions, standard practice is to examine several graphical displays based on the residuals [27]. Our program includes four graphs for residual analysis of each fitted model (i.e., for each order m and each m with a fixed lowest k). To assess normality, we provide a normal probability plot (which should appear to be linear if the errors are normally distributed) and a histogram of the residuals with a fitted kernel density estimate and normal curve. The latter display also includes the A^2 (Anderson–Darling) and W^2 (Cramér–von Mises) tests for normality [28]. Both of these tests are based on the squared difference between the normal and actual distributions; hence, the lower their values, the closer the residuals follow the normal distribution. These criteria can be useful when visual inspection of graphs does not reveal the superiority of one model over another. The P values of the A^2 and W^2 criteria are also generated; higher P values indicate failure to reject (agreement with) the null hypothesis of normality. We also provide plots of the residuals versus the fitted values and of residuals versus the order of the data; these plots should look like random noise with no evident pattern. The aforementioned displays permit assessment of the assumptions of normality, homoscedasticity, and (to some degree) independence.

Because the abscissa of the Cot data plot is a multiple of time, it might appear to be reasonable to regard a Cot dataset as a time series and, hence, to test formally for autocorrelation (a specific form of dependence) using the Durbin–Watson test [27]. We have not included this test in our program for a variety of reasons, with the chief reason being that Cot data are typically not obtained as a time series (the separate points are obtained from independent experiments) and, hence, a positive finding of autocorrelation from the Durbin–Watson test is likely to be a false positive. Our empirical experience bears this out; Durbin–Watson tests for autocorrelation in Cot data usually are negative, inconclusive, or weakly positive (data not shown).

Outlier detection using the ROUT algorithm

Our program provides a separate routine for outlier detection or "nomination." A number of methods have been proposed for outlier nomination, but we use a recent innovation of Motulsky and Brown [29] called ROUT for "robust regression and outlier removal." We have implemented their method exactly as given in their original article, so we refer the reader to that publication for details. Essentially, we fit the desired functional model (Cot curve) to the data, but under the assumption that the error terms follow a Cauchy (Lorentz) distribution rather than the normal (Gaussian) distribution. This fit is robust (insensitive) to outliers. We then compute the residuals from the resulting fitted curve, normalize them, and convert each to a P value that measures its distance from the center of a suitable t distribution. These P values are then adjusted using the false discovery rate (FDR) adjustment (see Ref. [29]), and observations with FDR-adjusted P values less than 0.01 (a heuristic but reasonable cutoff) are flagged as outliers.

It is important to integrate the ROUT algorithm with the correct model (best m) because underfitted models produce too many false positives and overfitted models produce too many false negatives. However, choosing the correct model with which to examine a dataset is complicated by the fact that inclusion or exclusion of outliers may influence which m is deemed to be optimal. Consequently, we used the following strategy, with the caveat that all outlier nomination techniques have their shortcomings. In brief, a dataset was evaluated in toto using our Cot analysis algorithm (with the Marquardt numerical fitting option). The m producing the best fit (lowest AICc) was then used in ROUT-based detection of outliers in that dataset. Outliers flagged by ROUT were deleted, and the ROUT algorithm appropriate for the m was rerun on the amended dataset. It is possible that such a process could iterate repeatedly, deleting more and more points; however, in our experience, no further outliers are found after one or two iterations. The dataset from which outliers were removed was then analyzed using our nonlinear regression algorithm.

Program descriptions

The algorithms described above were coded as SAS scripts. The scripts produce a series of graphs, statistics, and HTML data pages. We refer to the suite of SAS scripts and associated files as CotQuest. CotQuest is available in two downloadable variations: CotQuestU and CotQuestG. CotQuestU (the U stands for universal) includes the SAS scripts, HTML report viewer files, sample datasets, sample output report files, and a detailed user's guide. CotQuestG (the G stands for graphical user interface or GUI) contains everything found in CotQuestU as well as a Windows GUI application that leads users through the analysis process. CotQuestG's GUI

(CotQuestG.exe) asks users a series of questions about their data and their analysis goals. These data are fed into the SAS scripts so that users do not need to manually edit any SAS code. The GUI program for CotQuestG requires .NET Framework 2.0 (or higher), which is included in Windows XP with current updates and Windows Vista (otherwise it can be downloaded from Microsoft's website). We recommend that those users who are unfamiliar or uncomfortable with SAS scripting use CotQuestG if possible. However, the user's guides included with CotQuestU and CotQuestG should, in association with this article, permit users to conduct their Cot analyses even if they have never used SAS or CotQuest before.

For each analysis, the CotQuest programs produce an HTML file (Report.htm) from which users can view a summary of the results and, through hyperlinks, access a variety of data, statistics, and analysis files (Fig. 2). The AICc for each model is displayed in the second column of the report page table (Fig. 2A). If a model produces a converged fit but exhibits more than 50% component overlap, the percentage of overlap will be displayed in red text. If a model fails to converge, in part or whole, the words "Partially Converged" or "Failed to Converge" are shown in red. Models that cannot produce fits without producing an illogical parameter value (e.g., $f_i > 1$) are indicated by the text "This model is invalid" (e.g., Model 4F in Fig. 2A).

Datasets and analysis criteria

We tested our program on Cot data from eight species (Table 1). Each dataset was evaluated using both Gauss and Marquardt numerical search algorithms. As described above, datasets were tested for outliers using the appropriate ROUT script. Datasets for two of the species possessed outliers; thus, these datasets were evaluated with the outliers included and the outliers removed. All datasets are included in the Sample Data folder in the downloadable CotQuestU and CotQuestG packages. Those datasets that have been published or are extracted from published works are also found in section B of the supplementary material.

To permit comparisons between CotQuest and NNNBAT, each dataset was also analyzed using the latter program [16]. NNNBAT, which is available at http://faculty.virginia.edu/wrpearson/fasta/other, starts with initial parameter guesses provided by users. Because the number of different sets of parameter guesses is unlimited, we used the same approach to generate initial parameter guesses for each dataset (see section C of the supplementary material for details). NNNBAT results for a given dataset and model were used as starting parameter guesses in a second NNNBAT analysis. If the goodness-of-fit value obtained for the second iteration decreased by at least 5% compared with that of the first analysis, a third analysis was conducted using the values generated in the second fit as starting parameters. This general procedure was continued until goodness-of-fit values did not change by more than 5% between successive iterations.

Results and discussion

Format of results

Both the CotQuestG and CotQuestU programs generate a report page (Report.htm) from which all data associated with an analysis can be accessed. An example report page generated in analysis of the sorghum dataset is shown in Fig. 2A. In the figure, Model 3F provides the lowest AICc (−438.65). Although Models 3 and 4 are tied for the next lowest AICc (−436.04), the 100% overlap between the highly repetitive and moderately repetitive components in Model 4 indicate that this model is not valid (hence the "Partially Converged" warning message). From the Reports.htm page, each Cot curve (e.g., Fig. 2B) can be viewed by clicking on the link below its corresponding model name (Fig. 2A), whereas residual analysis graphs (e.g., Fig. 2C and D) and associated data can be accessed through the hyperlinks in the far right column of the table. Standard SAS output pages, additional graphs, and statistical details can be accessed through the links found beneath the output table (Fig. 2A).

Effect of fixing the lowest k

As mentioned previously, a Cot curve often can be used to make an estimate of genome size. This is typically done using the formula $G = (G_{\text{coli}} \cdot k_{\text{coli}}) \div k$, where k is the reassociation rate of the sole component in prokaryotes/viruses/organelles or the slowest reassociating component in eukaryotes [1,25]. However, Cot analysis is arguably not the best means of obtaining genome size values; indeed, a Cot-based estimate of genome size may vary considerably from published values obtained via more direct means (e.g., flow cytometry). Consequently, it is common to fix the lowest k based on a published genome size value to both compensate for potential error in a Cot analysis and possibly attain a better curve fit. The reasons for the inaccuracy in estimating genome size based on slowest reassociating k value are likely rooted in the relatively large number of experimental steps in a Cot analysis (vs. a procedure aimed directly at determining genome size) and the cumulative effects of minor lab/person/organism-specific variations on final results. Ideally, each laboratory conducting a Cot analysis for a particular organism should also prepare a Cot curve for *E. coli* and, in calculating genome size, use the k_{coli} value as determined in that laboratory rather than a published k_{coli} value [25]. However, this best-practice procedure is not practiced very often. Moreover, determining the k of the single-copy component of large genomes (>1 Gb) may be complicated by uncertainty regarding an organism's ploidy level [25]. When possible, additional sources of data should be employed in making interpretations based on the k value of the slowest reassociating component.

For each of the sample datasets (Table 1), we analyzed the data using both nonfixed and fixed lowest k values for each m; indeed, if a CotQuest user provides the program with a genome size value, it will automatically generate both standard and fixed model fits (Fig. 2A). Fixing the lowest k usually leads to a slight deviation from the optimal statistical fitting and RSS increases slightly (data not shown). On the other hand, it follows from the definitions of AIC and AICc that decreasing p by 1 lowers their values slightly. Therefore, fixing the lowest k may or may not improve AICc depending on which of these two effects is greater. As shown in Fig. 3, in some instances the best fit model for a species' Cot curve is standard, whereas in others it is fixed.

Numerical search algorithm selection

To examine the effect of the underlying numerical search algorithm on model selection, we analyzed each of the eight data sets using both Gauss (Gauss–Newton) and Marquardt fitting algorithms. For seven of the eight species, the Gauss and Marquardt algorithms resulted in selection of the same "optimal" model (i.e., the model with the lowest AICc), and in most instances the results produced using the two algorithms were identical across most models. However, for pine and tobacco datasets, one algorithm was able to reach convergence for a suboptimal model, whereas its counterpart could not (see section D of the supplementary material).

The onion dataset was the only one in which Gauss and Marquardt algorithms generated different optimal model selections. We attribute this to two interacting features that appeared

A

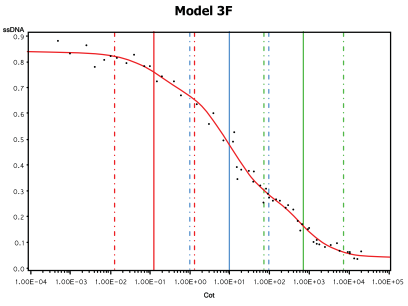
CotQuest

Species/Sample: Sorghum bicolor BTx623 Date: 06/16/2008
Algorithm: Marquardt Number of Cot points: 62
Input file: C:\CotQuestG\Sample Data\sorghum.cot

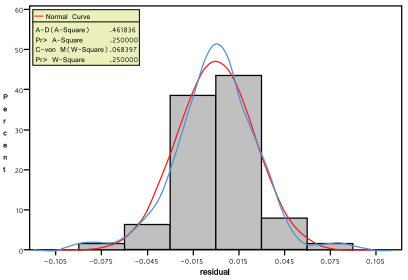
	AICc, A-sq, W-sq, Convergence	Component	Fraction	Kinetic Complexity	*k*	Cot½	Residual Analysis
Model 1 Cot curve	-344.655, 0.97148860, 0.15991545, Converged NLIN output	Reassociated	0.6537	9757434	0.0680	14.7059	Plots | Data
		Unreassociated	0.3806				
Model 1F Cot curve	-224.184, 1.51459463, 0.25986591, 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	
		Unreassociated	0.0493				
		Fraction overlap. HR-SL: 0%					
Model 2F Cot curve	-414.521, 0.55600124, 0.08845857, Converged NLIN output	Highly Repetitive	0.4998	2385035	0.2127	4.7015	Plots | Data
		Single\Low	0.2695	195667024	0.001398	715.3076	
		Unreassociated	0.0323				
		Fraction overlap. HR-SL: 0%					
Model 3 Cot curve	-436.04, 0.48570032, 0.07194388, Converged NLIN output	Highly Repetitive	0.1458	18755	7.8907	0.1267	Plots | Data
		Mod. Repetitive	0.4142	3958691	0.1062	9.4162	
		Single\Low	0.238	163222973	0.00148	675.6757	
		Unreassociated	0.0405				
		Fraction overlaps. HR-MR: 6.45%, MR-SL: 7.21%					
Model 3F Cot curve	-438.65, 0.46183585, 0.06839740, Converged NLIN output	Highly Repetitive	0.1472	19350	7.7214	0.1295	Plots | Data
		Mod. Repetitive	0.416	4075676	0.1036	9.6525	
		Single\Low	0.2359	171272175	0.001398	715.3076	
		Unreassociated	0.0393				
		Fraction overlaps. HR-MR: 6.38%, MR-SL: 6.51%					
Model 4 Cot curve	-436.04, 0.48570864, 0.07194535, Partially converged NLIN output	Very Highly Rep.	0.1458	18755	7.8907	0.1267	Plots | Data
		Highly Repetitive	0.0619	591605	0.1062	9.4162	
		Mod. Repetitive	0.3523	3367086	0.1062	9.4162	
		Single\Low	0.238	163222973	0.00148	675.6757	
		Unreassociated	0.0405				
		Fraction overlaps. VHR-HR: 6.45%, HR-MR: 100%, MR-SL: 7.21%					
Model 4F	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

B



C



D

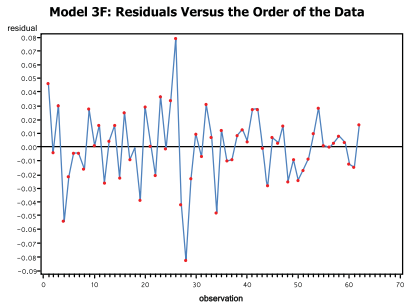


Fig. 2. Results screen shots. (A) Report.htm page autogenerated by CotQuest in an analysis of sorghum Cot data. The Report.htm page possesses summary data and links to graphs, statistics, and data pages. (B) Best fit Cot curve. (C) Histogram of residuals compared with normal distribution and kernel density distribution estimate. (D) Residuals versus the order of the data.

only in the onion dataset. First, there are five points with essentially identical *y* axis values at the left end of the curve, indicating no detectable reassociation at the first five Cot values. Second, there is a complex system of nonmonotonic residuals at the right

Table 1
Sources of data used in evaluating CotQuest.

Dataset	Scientific name	Source of Cot data	Genome size (1C) (Mb)	Genome size reference
E. coli[a]	*Escherichia coli*	[1]	4.64	[24]
Sorghum	*Sorghum bicolor*	[10]	735	Bennett and Leitch (2004)[c]
Tomato	*Solanum lycopersicum*	[14]	950	[30]
Cow[a]	*Bos taurus*	[31]	3100	[32]
Tobacco[a]	*Nicotiana tabacum*	[25]	5730	Bennett and Leitch (2004)[c]
Bald cypress	*Taxodium distichum*	MGEL[b]	9750	Murray et al. (2004)[d]
Onion[a]	*Allium cepa*	[33]	16400	Bennett and Leitch (2004)[c]
Pine	*Pinus taeda*	MGEL[b]	21658	Murray et al. (2004)[d]

[a] The Cot curve for this organism was digitized and saved in a PDF file. We then used the measurement tool in Adobe Acrobat 8.0 to determine the relative x and y coordinates of each Cot point, which were subsequently converted into fraction ssDNA and log Cot values, respectively.
[b] Mississippi Genome Exploration Laboratory (D. G. Peterson et al., unpublished results).
[c] M. D. Bennett, I. J. Leitch, Angiosperm DNA C-values database (release 5.0, December 2004), http://www.kew.org/cvalues/homepage.html.
[d] B. G. Murray, I. J. Leitch, M. D. Bennett, Gymnosperm DNA C-values database (release 3.0, December 2004), http://www.rbgkew.org.uk/cval/homepage.html.

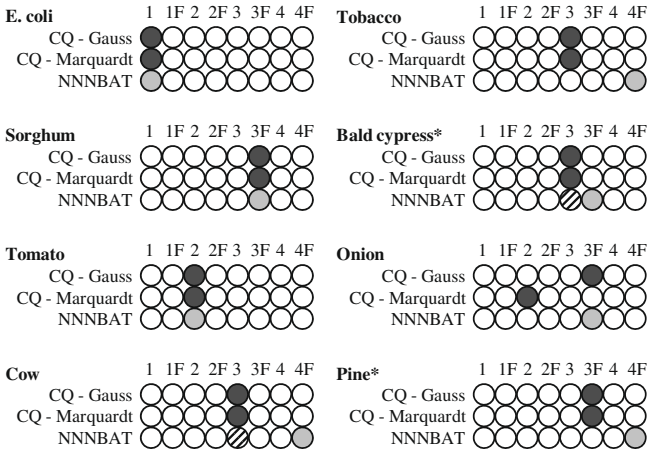


Fig. 3. Model selection by CotQuest (CQ) using Gauss and Marquardt algorithms versus NNNBAT analysis. For each dataset, the CQ best fit model (i.e., the model with the lowest AICc) is indicated by black shading. For all datasets except onion, both the Gauss and Marquardt algorithms resulted in selection of the same best fit model. Likewise, the model yielding the lowest AICc after NNNBAT analysis (as performed according to section C of the supplementary material) is shaded in gray. When the model identified through NNNBAT analysis differed from that selected by CQ, the CQ–Marquardt best fit parameters were used as starting values for an additional NNNBAT analysis. In two instances, the CQ–Marquardt values resulted in a new lowest AICc with a concomitant change in best model (diagonal striping) to the one selected by at least one of the CQ analyses. Two datasets (*) were shown to contain significant outliers. Removal of the outliers followed by reanalysis did not change model selection.

end of the curve. This combination may present challenges to the model fitting procedure. The onion optimal Gauss fit, Model 3F, has a slightly lower AICc (−366.788) than the optimal Marquardt fit, Model 2 (−365.668). With regard to A^2 and W^2 values, there is a split between the two numerical search algorithms, with one having the better A^2 and the other having the better W^2 (see section D of the supplementary material). In addition, the Cot curves and the residual plots are not noticeably different between the two numerical search algorithms. Thus, based on its lower AICc, the Gauss Model 3F would probably be the best choice. Our observations suggest value in evaluating data using both the Gauss and Marquardt options.

Effect of outlier removal

All of the sample datasets were tested for significant outliers using the strategy described in Materials and Methods. One outlier was detected in the pine dataset, whereas five outliers were detected in the bald cypress data. Removal of the putative outlier from the pine data and a second round of ROUT analysis revealed no additional outliers, and no additional outliers were detected in the second ROUT analysis of the bald cypress data. CotQuest analysis of the original and amended datasets did not result in a change in the best fit model for pine or bald cypress. Although the output values in the pine analyses were essentially identical for the datasets with and without the outlier removed, there were slight differences in the output values for the bald cypress datasets (see section E of the supplementary material). In general, we recommend that users conduct outlier removal only if (i) those points that are signaled as outliers can be identified as anomalous on scientific grounds or (ii) a reasonable model fit cannot be obtained using the complete dataset.

Comparison of CotQuest with NNNBAT

To test the efficacy of the CotQuest programs in comparison with the current Cot analysis standard, NNNBAT, we analyzed each of our test datasets (Table 1) using the NNNBAT program. NNNBAT does not actually provide users with the best fit of their data; rather, it provides them with the best fit for a given model based on the parameter guesses entered. Because the potential parameter combinations are endless, we used a heuristic approach to select reasonable parameter guesses for each model (see section C of the supplementary material). Each model (1, 1F, 2, 2F, 3, 3F, 4, and 4F) needed to be tested independently, and results for a model needed to be used as starting values for at least one additional test (to see whether goodness of fit changed by > 5%).[3] In general, it took approximately, 1.5 h to complete the NNNBAT analysis for one dataset using the approach given in section C of the supplementary material.

As shown in Fig. 3, for only three of the eight datasets, NNNBAT analysis resulted in selection of the same best fit model (using the strategy in section C of the supplementary material) as CotQuest. In tobacco, the NNNBAT best fit model was Model 4F, although CotQuest indicated that Model 3 was best. In cow and pine, NNNBAT favored Model 4F, although this model does not reach convergence using CotQuest, indicating that NNNBAT's error checking mechanisms are insufficient. For bald cypress and cow, the NNNBAT best fit models differed from those selected by CotQuest, but use of the CotQuest best fit values as starting values for NNNBAT analysis substantially improved the NNNBAT fits and changed the NNNBAT best fit models to the same ones as selected by CotQuest. For the onion dataset, where a difference was noted between best fit

[3] NNNBAT can theoretically fit data with one, two, three, four, and five components ($m \leqslant 5$), whereas CotQuest is designed to perform fits for $m \leqslant 4$. However, in our experience, even the largest eukaryotic genomes cannot be resolved into more than two or three components without conducting additional reassociation kinetics experiments (e.g., "mini-Cot" analyses) [2,25]. CotQuest could be adapted for $m \geqslant 5$, but this would substantially and unnecessarily increase run time.

models selected using CotQuest Gauss and Marquardt numerical search algorithms, the NNNBAT best fit model was the same as the CotQuest Gauss best fit model. This observation is interesting because NNNBAT uses the Marquardt algorithm to conduct numerical searches; thus, one might predict that NNNBAT would select the same best fit model as CotQuest using the Marquardt search algorithm.

The advantages of CotQuest over NNNBAT are fairly clear.[4] In a single run, CotQuest produces a best fit curve for each model (if possible) and provides a model selection statistic (AICc) to help users choose an optimal model. In contrast, NNNBAT produces a fit for a given model based on starting parameter guesses; the quality of the fit depends on the proximity of the parameter guesses to a true best fit solution. For each model, multiple rounds of guess testing are required to try to avoid fits biased by local minima or maxima over the parameter space. Typically, results from NNNBAT analyses of different models must be consolidated (e.g., in a spreadsheet) for comparison. Moreover, NNNBAT does not provide a model selection statistic, although AICc can be calculated from NNNBAT results. As shown in Fig. 3, these differences between CotQuest and NNNBAT can result in selection of suboptimal models based on NNNBAT output even when relatively logical starting guesses are used (see section C of the supplementary material). In addition, the residual/statistical analyses and graphing capabilities of CotQuest distinguish it from NNNBAT, whereas the GUI associated with CotQuestG and the Report.htm pages generated by both CotQuestG and CotQuestU provide user-friendliness not found in NNNBAT.

SAS versus R

There is an increasing tendency for programmers in the biosciences to use freeware in development of their applications, a trend we strongly endorse. In statistics, the freeware program R is often used. However, in developing CotQuest, we chose to use SAS because it is far better supported, documented, benchmarked, and disseminated than is R. Moreover, SAS costs only a small amount (typically ~$80 U.S.) to academic site license users. R may be suitable for test applications or small-scale implementations, but SAS-based CotQuest, especially in its GUI version, is immediately usable (with no programming) by biologists with minimal statistical expertise, and it provides reliable results and detailed analytical and graphical reports in a user-friendly format while exploiting the full power of the numerical and statistical routines already implemented in SAS. However, to facilitate use of the CotQuest algorithm with other platforms, we have added a comprehensive listing of pseudo-code for all of our algorithms to the freely downloadable materials available on our website (www.mgel.msstate.edu/tools.htm).

Conclusions

We have presented CotQuest, a freely available program that can be used in association with SAS to perform nonlinear regression analysis of DNA reassociation kinetics data. CotQuest represents a major improvement in Cot analysis methods by implementing a novel algorithm that eliminates the need for input of parameter guesses. CotQuest greatly surpasses the automation, statistical robustness, and user-friendliness of existing Cot analysis programs, and it should be of use to anyone conducting Cot analyses or conducting any other line of research involving nonlinear regression. The CotQuest scripts and detailed documentation are available at www.mgel.msstate.edu/tools.htm.

[4] It should be noted that NNNBAT also includes subprograms designed for fitting first-order kinetic reactions, fitting reactions where the apparent order is unknown, and determining tracer reaction rates in tracer/driver mixtures [16]. We have not equipped CotQuest with these functions because we do not use them. However, the CotQuest code could be easily adapted to perform these additional functions, and we would be happy to help interested parties make the necessary code changes.

Acknowledgments

We thank Harvey Motulsky for consultation regarding the ROUT algorithm. This work was supported, in part, by the National Science Foundation (DBI-0421717 to D.G.P.), the U.S. Department of Agriculture (CSREES 2006-34506-17290 and ARS-58-6402-7-241 to D.G.P.), and the Mississippi Corn Promotion Board (to D.G.P.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ab.2009.03.007.

References

[1] R.J. Britten, D.E. Kohne, Repeated sequences in DNA: Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms, Science 161 (1968) 529–540.
[2] R.J. Britten, D.E. Graham, B.R. Neufeld, Analysis of repeating DNA sequences by reassociation, Methods Enzymol. 29 (1974) 363–418.
[3] J. Rimpau, D. Smith, R. Flavell, Sequence organisation analysis of the wheat and rye genomes by interspecies DNA/DNA hybridisation, J. Mol. Biol. 123 (1978) 327–359.
[4] J. Rimpau, D.B. Smith, R.B. Flavell, Sequence organisation in barley and oats chromosomes revealed by interspecies DNA/DNA hybridisation, Heredity 44 (1980) 131–149.
[5] V. Torsvik, J. Goksoyr, F.L. Daae, High diversity in DNA of soil bacteria, Appl. Environ. Microbiol. 56 (1990) 782–787.
[6] R. Sandaa, V. Torsvik, V. Enger, F.L. Daae, T. Castberg, D. Hahn, Analysis of bacterial communities in heavy metal-contaminated soils at different levels of resolution, FEMS Microbiol. Ecol. 30 (1999) 237–251.
[7] R.B. Goldberg, From Cot curves to genomics: How gene cloning established new concepts in plant biology, Plant Physiol. 125 (2001) 4–8.
[8] D.G. Peterson, S.R. Wessler, A.H. Paterson, Efficient capture of unique sequences from eukaryotic genomes, Trends Genet. 18 (2002) 547–550.
[9] M.B. Soares, M.F. Bonaldo, P. Jelene, L. Su, L. Lawton, A. Efstratiadis, Construction and characterization of a normalized cDNA library, Proc. Natl. Acad. Sci. USA 91 (1994) 9228–9232.
[10] D.G. Peterson, S.R. Schulze, E.B. Sciara, S.A. Lee, J.E. Bowers, A. Nagel, N. Jiang, D.C. Tibbitts, S.R. Wessler, A.H. Paterson, Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery, Genome Res. 12 (2002) 795–807.
[11] Y. Yuan, P.J. SanMiguel, J.L. Bennetzen, High-Cot sequence analysis of the maize genome, Plant J. 34 (2003) 249–255.
[12] T. Wicker, J.S. Robertson, S.R. Schulze, F.A. Feltus, V. Magrini, J.A. Morrison, E.R. Mardis, R.K. Wilson, D.G. Peterson, A.H. Paterson, R. Ivarie, The repetitive landscape of the chicken genome, Genome Res. 15 (2005) 126–136.
[13] D. Lamoureux, D.G. Peterson, W. Li, J.P. Fellers, B.S. Gill, The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum*L.), Genome 48 (2005) 1120–1126.
[14] D.G. Peterson, W.R. Pearson, S.M. Stack, Characterization of the tomato (*Lycopersicon esculentum*) genome using in vitro and in situ DNA reassociation, Genome 41 (1998) 346–356.
[15] M.G. Murray, R.E. Cuellar, W.F. Thompson, DNA sequence organization in the pea genome, Biochemistry 17 (1978) 5781–5790.
[16] W.R. Pearson, E.H. Davidson, R.J. Britten, A program for least squares analysis of reassociation and hybridization data, Nucleic Acids Res. 4 (1977) 1727–1737.
[17] D.G. Peterson, Reduced representation strategies and their application to plant genomes, in: K. Meksem, G. Kahl (Eds.), The Handbook of Genome Mapping: Genetic and Physical Mapping, Wiley–VCH, Weinheim, Germany, 2005, pp. 307–335.
[18] M.J. Smith, R.J. Britten, E.H. Davidson, Studies on nucleic acid reassociation kinetics: Reactivity of single-stranded tails in DNA–DNA renaturation, Proc. Natl. Acad. Sci. USA 72 (1975) 4805–4809.
[19] R.F. Murphy, W.R. Pearson, J. Bonner, Computer programs for analysis of nucleic acid hybridization, thermal denaturation, and gel electrophoresis data, Nucleic Acids Res. 6 (1979) 3911–3921.
[20] D.W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, J. Soc. Industr. Appl. Math. 11 (1963) 431–441.
[21] S. Green, J.K. Field, C.D. Green, R.J. Beynon, A microcomputer program for analysis of nucleic acid hybridization data, Nucleic Acids Res. 10 (1982) 1411–1420.
[22] R. Hooke, T.A. Jeeves, "Direct search" solution of numerical and statistical problems, J. Assoc. Comput. Mach. 8 (1961) 212–229.

[23] P. Érdi, J. Tóth, Mathematical Models of Chemical Reactions, Princeton University Press, Princeton, NJ, 1989.
[24] F.R. Blattner, G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, Y. Shao, The complete genome sequence of *Escherichia coli* K-12, Science 277 (1997) 1453–1474.
[25] J.L. Zimmerman, R.B. Goldberg, DNA sequence organization in the genome of *Nicotiana tabacum*, Chromosoma 59 (1977) 227–252.
[26] K.P. Burnham, D.R. Anderson, Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Springer, New York, 2002.
[27] G.A.F. Seber, C.J. Wild, Nonlinear Regression, John Wiley, New York, 1989.
[28] R.B. D'Agostino, M.A. Stephens, Goodness-of-Fit Techniques, Marcel Dekker, New York, 1986.
[29] H.J. Motulsky, R.E. Brown, Detecting outliers when fitting data with nonlinear regression: A new method based on robust nonlinear regression and the false discovery rate, BMC Bioinformatics 7 (2006) 123.
[30] E. Asamizu, Tomato genome sequencing: Deciphering the euchromatin region of the chromosome 8, Plant Biotechnol. 24 (2007) 5–9.
[31] R.J. Britten, E.H. Davidson, Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty, Q. Rev. Biol. 46 (1971) 111–133.
[32] W.M. Snelling, R. Chiu, J.E. Schein, M. Hobbs, C.A. Abbey, D.L. Adelson, J. Aerts, G.L. Bennett, I.E. Bosdet, M. Boussaha, R. Brauning, A.R. Caetano, M.M. Costa, A.M. Crawford, B.P. Dalrymple, A. Eggen, A. Everts-van der Wind, S. Floriot, M. Gautier, C.A. Gill, R.D. Green, R. Holt, O. Jann, S.J. Jones, S.M. Kappes, J.W. Keele, P.J. de Jong, D.M. Larkin, H.A. Lewin, J.C. McEwan, S. McKay, M.A. Marra, C.A. Mathewson, L.K. Matukumalli, S.S. Moore, B. Murdoch, F.W. Nicholas, K. Osoegawa, A. Roy, H. Salih, L. Schibler, R.D. Schnabel, L. Silveri, L.C. Skow, T.P. Smith, T.S. Sonstegard, J.F. Taylor, R. Tellam, C.P. Van Tassell, J.L. Williams, J.E. Womack, N.H. Wye, G. Yang, S. Zhao, A physical map of the bovine genome, Genome Biol. 8 (2007) R165.
[33] S.M. Stack, D.E. Comings, The chromosomes and DNA of *Allium cepa*, Chromosoma 70 (1979) 161–181.