

(19) AUSTRALIAN PATENT OFFICE

(54) Title
Identifying statistically linear data

(51)⁶ International Patent Classification(s)
G06F 19/00 (2006.01) 19/34
C12P 19/34 (2006.01) 20060101ALI2008041
G11C 17/00 (2006.01)^{7BHUS} **G11C**
G06F 19/00 17/00
20060101AFI2008041 20060101ALI2008041
^{7BHUS} **C12P** ^{7BHUS}
PCT/US2006/018549

(21) Application No: 2006247597

(22) Application Date: 2006 .05 .12

(87) WIPO No: W006/124673

(30) Priority Data

(31) Number	(32) Date	(33) Country
60/681,182	2005 .05 .13	US
11/432,856	2006 .05 .11	US

(43) Publication Date : 2006 .11 .23

(71) Applicant(s)
Bio-Rad Laboratories, Inc.

(72) Inventor(s)
Lerner, Jeffrey

(74) Agent/Attorney
Griffith Hack, 167 Eagle Street, Brisbane, QLD, 4000

(56) Related Art
LIVAK et al.
DHEDA et al.

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
23 November 2006 (23.11.2006)

PCT

(10) International Publication Number
WO 2006/124673 A2(51) International Patent Classification:
G06F 19/00 (2006.01)(21) International Application Number:
PCT/US2006/018549

(22) International Filing Date: 12 May 2006 (12.05.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/681,182 13 May 2005 (13.05.2005) US
11/432,856 11 May 2006 (11.05.2006) US(71) Applicant (for all designated States except US):
BIO-RAD LABORATORIES, INC. [US/US]; 1000
Alfred Nobel Drive, Hercules, CA 94547 (US).

(72) Inventor; and

(75) Inventor/Applicant (for US only): LERNER, Jeffrey
[US/US]; 1004 9th Avenue, Oakland, CA 94606 (US).(74) Agents: RACZKOWSKI, David et al.; Two Embar-
cadero Center, Eighth Floor, San Francisco, CA 94111
(US).(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AI, AG, AL, AM,AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN,
CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI,
GB, GD, GE, GH, GM, GR, GU, HD, IL, IN, IS, JP, KE,
KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV,
LY, MA, MD, MG, MK, MN, MW, MX, MY, NA, NG, NI,
NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG,
SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, YU, ZA, ZM, ZW.(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,
ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT,
RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA,
GN, GQ, GW, ML, MR, NE, SN, TD, TG).

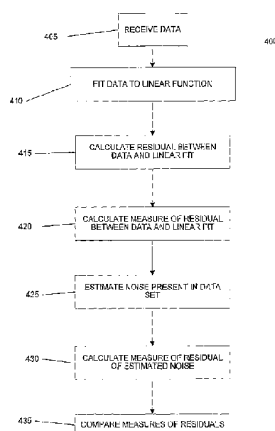
Declaration under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

— without international search report and to be republished
upon receipt of that reportFor two-letter codes and other abbreviations, refer to the "Guid-
ance Notes on Codes and Abbreviations" appearing at the begin-
ning of each regular issue of the PCT Gazette.

(54) Title: IDENTIFYING STATISTICALLY LINEAR DATA



(57) Abstract: Methods, apparatus, and systems are provided for processing a data set having noise to determine whether the data set exhibits statistically linear behavior. A true data signal is calculated based on local properties of the data, and an estimate of the noise in the data is calculated from the true data signal. A measure of the estimated noise is then compared to properties of a linear fit to the data set.

WO 2006/124673 A2

IDENTIFYING STATISTICALLY LINEAR DATA**CROSS-REFERENCES TO RELATED APPLICATIONS**

5 [0001] This application claims the benefit of U.S. Provisional Application 60/681,182,
filed May 13, 2005, entitled "SYSTEMS AND METHODS FOR IDENTIFYING
STATISTICALLY FLAT DATA IN A DATA SET," which is related to, and
concurrently filed with U.S. Provisional Application Serial No. 60/680,765 entitled
DETERMINATION OF BASELINE END CYCLE IN REAL TIME PCR USING
DERIVATIVE PEAK ANALYSIS (Attorney Docket No. 002558-072200US), the
10 disclosures of which are incorporated herein by reference in their entirety. The present
application is also related to the following commonly-assigned concurrently filed U.S.
Patent Applications: No. _____ (Attorney Docket No. 002558-072210US), entitled
"BASELINING AMPLIFICATION DATA." The respective disclosures of these
applications are incorporated herein by reference for all purposes.

15

FIELD OF THE INVENTION

[0002] The present invention relates generally to data processing systems and methods,
and more particularly to systems and methods for identifying statistically linear data in
a data set of an amplification process, such as polymerase chain reaction (PCR).

20

BACKGROUND OF THE INVENTION

[0003] Many experimental processes exhibit amplification of a quantity. For example,
in PCR, the quantity may correspond to the number of parts of a DNA strand that have
been replicated, which dramatically increases during an amplification stage or region.
Other experimental processes exhibiting amplification include bacterial growth
25 processes. The quantity is detected from an experimental device via a data signal,
whose data points are analyzed to determine information about the amplification. As
part of the data analysis, it is important to know if amplification has potentially
occurred; otherwise, effort might be wasted on analyzing non-amplifying data. If the
data is statistically linear, then amplification has not occurred.

[0004] Ideally, the data from the amplification detection device would be a monotonic and continuous signal, thus one could easily identify whether the data, or portions thereof, has statistically linear behavior. However, the signal from the amplification device typically contains noise, thus making identifying a behavior of the signal difficult. The noise manifests itself in each data point in the signal from the device having random fluctuations that occur on top of the true signal, e.g. the actual number of DNA strands. Thus, the data requires processing to allow for identifying of linear behavior.

[0005] A typical prior method for processing data to determine if it is statistically linear is with a linear least squares (LSQ) fit. The correlation value of the LSQ fit can be used to determine whether there is an adequate fit. By standard convention, a correlation value of 0 is related to a bad fit, thus the data is not linear, and a value of 1 suggests a good fit for linearity. The problem is that in the presence of noise, the correlation value can be close to 0 or 1 for data that looks statistically linear. Additionally, the correlation value does not correspond to a physical value that may provide additional insight and efficacy. Thus, the correlation value is not an acceptable criterion, particularly for data that can be extremely noisy.

[0006] Therefore it is desirable to provide systems and methods for processing a data set having noise, and for identifying whether the dataset is statistically linear, that overcome the above and other problems.

[0006A] It is to be understood that, if any prior art publication is referred to herein, such reference does not constitute an admission that the publication forms a part of the common general knowledge in the art, in Australia or any other country.

BRIEF SUMMARY OF THE INVENTION

[0006B] In one aspect, the invention provides a method of processing a data set to determine whether the data set exhibits statistically linear behavior, the method comprising:

receiving an original set of data points having a signal component and a noise component, each data point representing a physical quantity of a substance during an amplification process;

fitting the original data set to a linear function;

5 calculating a residual between the original data set and the fitted linear function;

calculating a first measure of the residual between the original data set and the fitted linear function;

estimating the noise component present in the original data set, by:

10 i) calculating a smoothed data set by determining a set of smoothed data points, wherein a value of a smoothed data point is based on values of a plurality of original data points that are local to that smoothed data point; and

ii) calculating a residual between the smoothed data set and the original data set;

15 calculating a second measure of the residual between the smoothed data set and the original data set; and

comparing the first measure to the second measure to determine whether the original data set exhibits statistically linear behavior, wherein the method is implemented using a processor.

20

[0006C] In another aspect, the invention provides a tangible information storage medium storing a plurality of instructions adapted to direct an information processing device to perform an operation of processing data to determine whether the data exhibits linear behavior, the operation comprising the steps of:

25 receiving an original set of data points having a signal component and a noise component, each data point representing a physical quantity of a substance during an amplification process;

fitting the original data set to a linear function;

30 calculating a residual between the original data set and the fitted linear function;

calculating a first measure of the residual between the original data set and the fitted linear function;

3A

estimating the noise component present in the original data set, by:

i) calculating a smoothed data set by determining a set of smoothed data points, wherein a value of a smoothed data point is based on values of a plurality of original data points that are local to that smoothed data point; and

5 ii) calculating a residual between the smoothed data set and the original data set;

calculating a second measure of the residual between the smoothed data set and the original data set; and

10 comparing the first measure to the second measure to determine whether the original data set exhibits statistically linear behavior.

[0006D] In yet another aspect, the invention provides a PCR detection system comprising:

15 a detector for producing an original set of data points having a signal component and a noise component, wherein the data represents a PCR amplification curve; and

logic for processing data to determine whether the data exhibits linear behavior, by:

20 fitting the original data set to a linear function; calculating a residual between the original data set and the fitted linear function;

calculating a first measure of the residual between the original data set and the fitted linear function;

25 estimating the noise component present in the original data set, by:

i) calculating a smoothed data set by determining a set of smoothed data points, wherein a value of a smoothed data point is based on values of a plurality of original data points that are local to that smoothed data point; and

30 ii) calculating a residual between the smoothed data set and the original data set;

calculating a second measure of the residual between the smoothed data set and the original data set; and

comparing the first measure to the second measure to determine whether the original data set exhibits statistically linear behavior.

5 [0007] Accordingly, embodiments of the present invention provide methods and systems directed to processing data to determine whether the data exhibits statistically linear behavior. Statistically linear data means that the data generally does not curve downward or upward or otherwise display amplification. Such data typically appears to be roughly linear with a large noise signal superposed upon it. The data may be received from real-time PCR processes or other processes exhibiting amplification or growth.

10 [0008] According to one exemplary embodiment of the present invention, a method of processing data is provided. The method typically includes receiving an original set of data points having a signal component and a noise component. The original data set is fit to a linear function. In one aspect, the fit is accomplished by calculating a linear least squares fit to the data set. The method also includes calculating a residual between the original data set and the linear fit, and calculating a measure of the residual between the original data set and the linear fit. In one aspect, the measure is a standard deviation.

20 [0009] The method according to an embodiment also typically includes estimating the noise component present in the data set by calculating a smoothed data set and calculating the residual between the smoothed data set and the original data set. A smoothed data point is based on values of original data points that are local to that smoothed data point. In one aspect, a low pass filter is used to calculate the smoothed data set. Exemplary low pass filters include a Savitzky-Golay filter, a digital filter, or digital smoothing polynomial filter. In another aspect, a value of a smoothed data point is an average of original data points within a window around the smoothed data point.

25 [0010] The method according to an embodiment also typically includes calculating a measure of the residual of the estimated noise, and comparing the measures to determine whether the original data set exhibits statistically linear behavior. The comparing may include calculating a ratio of the first and second measure to determine

if the ratio is smaller or greater than a pre defined value. In one aspect, the pre-defined value is of order 1.

5 [0011] In preferred aspects, the method is implemented in a processor, such as a processor in a stand-alone computer, a network attached computer or a data acquisition device such as a real-time PCR machine. One example of a real-time PCR machine is the iCycler iQ System provided by Bio-Rad Laboratories.

10 [0012] According to another exemplary embodiment of the present invention, an information storage medium having a plurality of instructions adapted to direct an information processing device to perform an operation of processing data to determine whether the curve exhibits linear behavior is provided. In one aspect, the information storage medium is a RAM or ROM unit, hard drive, CD, DVD or other portable medium.

15 [0013] According to another exemplary embodiment of the present invention, a PCR detection system is provided. The PCR detection system includes a detector for producing an original set of data points having a signal component and a noise component and includes logic for processing data to determine whether the data exhibits linear behavior.

20 [0014] Reference to the remaining portions of the specification, including the drawings and claims, will realize other features and advantages of the present invention. Further features and advantages of the present invention, as well as the structure and operation of various embodiments of the present invention, are described in detail below with respect to the accompanying drawings. In the drawings, like reference numbers indicate identical or functionally similar elements.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0015] Figure 1 illustrates an example of a PCR amplification curve.
- [0016] Figure 2 illustrates a real-time PCR data set exhibiting noise and statistically linear behavior.
- [0017] Figure 3 illustrates a real-time PCR data set exhibiting noise and amplification.
- [0018] Figure 4 illustrates a method of processing a data set to determine whether the data set exhibits statistically linear behavior according to an embodiment of the present invention.
- [0019] Figure 5A illustrates a linear fit to a data set exhibiting statistically linear behavior.
- [0020] Figure 5B illustrates a linear fit to a data set exhibiting amplifying behavior.
- [0021] Figure 6A illustrates a smoothed data set of real-time PCR data according to an embodiment of the present invention.
- [0022] Figure 6B illustrates an estimated noise of real-time PCR data according to an embodiment of the present invention.
- [0023] Figure 7 illustrates a system that processes real-time PCR data according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

- [0024] The present invention provides techniques for processing a data set and identifying whether the data set is statistically linear, as well as distinguishing such a linear data set from a data set containing an amplification signal. In preferred aspects, the present invention is particularly useful for processing data from PCR growth or amplification processes to identify and remove statistically linear data prior to further analysis of the data. It should be appreciated, however, that the teachings of the present invention are applicable to processing any data set or curve that may include noise, and particularly curves that should otherwise exhibit growth or amplification such as a bacterial growth process.

[0025] Figure 1 shows an example of a PCR curve 100, where intensity values 110 vs. cycle number 120 are plotted for a typical PCR process. The values 110 may be any physical quantity of interest, and the cycle number may be any unit associated with time or number of

steps in the process. Such amplification curves typically have a linear region 130 followed by an amplification region 140 and then by an asymptotic region 150, as shown in FIG. 1. There also might be additional types of behavior such as downward curving data. An amplification region may have exponential, sigmoidal, high order polynomial, or other type of logistic function or logistic curve that models growth.

[0026] To understand the experimental process involved, it is important to identify the position and shape of amplification region 140. For example, in a PCR process, it may be desirable to identify the onset of amplification, which occurs at the end of the baseline region (linear region 130). A step in identifying the location is to identify if a possible amplification region even exists, as a PCR process may not show any amplification. However, since real-time PCR data has noise, the identification of whether the data set might exhibit amplification, or equivalently that it is not statistically linear, can be difficult.

[0027] For example, Figure 2 illustrates a linear region 230 of a real-time PCR curve 200 made from a data set with data points 240 that include a signal and noise. Note that even for devices that produce a constant signal, this data must be broken into data points for analysis. The noise causes the fluctuations in the data points. Overall, the data is generally moving upward (i.e. positive slope) in a linear fashion. However, as curve 200 is very non-linear from point to point, the generally linear behavior cannot be determined by directly analyzing curve 200 at any one point along the curve. A direct analysis of curve 200 would falsely determine that the data does not exhibit statistically linear behavior. Embodiments of the present invention effectively determine whether data exhibits statistically linear behavior.

[0028] Additionally, it is important to differentiate data curves having linear behavior and data curves having amplifying behavior. Figure 3 illustrates a real-time PCR curve 300 that exhibits amplification. Initially, the data exhibits linear behavior in region 330 and in later cycles there is amplification in region 340. Embodiments of the present invention robustly and with consistent accuracy differentiate between PCR curve 200 having only linear behavior and PCR curves possibly having an amplifying region, such as PCR curve 300.

[0029] Figure 4 illustrates a method 400 of processing data to determine whether the data exhibits statistically linear behavior according to an embodiment of the present invention. The data set is composed of data points and represents a curve having a signal component and a noise component.

[0030] In step 405, the data set is first collected or received. The data set may be received through many mechanisms. For example, the data set may be acquired by a processor (executing instructions) resident in a PCR data acquiring device such as an iCycler iQ device or similar PCR analysis device. The data set may be provided to the processor in real time as the data is being collected, or it may be stored in a memory unit or buffer and provided to the processor after the experiment has been completed. Similarly, the data set may be provided to a separate system such as a desktop computer system via a network connection (e.g., LAN, VPN, intranet, Internet, etc.) or direct connection (e.g., USB or other direct wired or wireless connection) to the acquiring device, or provided on a portable medium such as a CD, DVD, floppy disk or the like to a stand-alone computer system. After the data set has been received or acquired, the data may be analyzed.

[0031] In step 410, a linear fit to the data set is calculated. Typically, a fit defines a merit function S that measures the agreement or difference between the data set and the fit, where small values of the merit function typically represent better parameters for the fit. For example, in a linear least squares fit, the merit function is the squares of the difference between the data values Y and the fit function $f(x_i)$, where for N data points $S \equiv \sum_{i=1}^N (Y_i - f(x_i))^2$. In a PCR process, Y is the data intensity and x is the cycle number. Figure 5A shows a linear fit 510 of the PCR curve 200. Figure 5B shows a linear fit 550 of PCR curve 300.

[0032] Merit functions may include different weight contributions or normalization factors to the merit function for different data points. Merit functions may also scale data point values or take a function of data points before a difference is taken. The difference may be taken between the data at one x value and $f(x)$ at a different x value. For example, a term in the merit function may represent the length of a line from data curve to the linear fit, such that the line is perpendicular to the linear fit. This occurs at a difference cycle number unless the linear fit has a slope of 0. One skilled in the art will recognize the many different merit functions that could be used.

[0033] In step 415, a residual R between the data and the linear fit is calculated. The residual R is a set of values corresponding to an error in the data points from the linear fit. For example, the residual may be the difference in the linear fit value and the actual data point for each cycle number, giving $R_i = Y_i - f(x_i)$, which is a standard form of the residual.

In some embodiments, the residual is related to the values used to determine the merit function of the linear fit. In other embodiments, the residual is a different value. In Figure 5A, errors 520 are used to calculate values of the residual R between curve 200 and linear fit 510. In Figure 5B, errors 560 are used to calculate values of the residual R between curve 300 and linear fit 550.

[0034] In step 420, a measure σ_1 of the residual between the data and the linear fit is calculated. The measure is a single value made from the set of values that are the residual.

In one embodiment, the residual is a standard deviation, giving $\sigma_1 = \sqrt{\frac{1}{N} \sum_{i=1}^N R_i^2}$. Some embodiments may have a weighting value for each value of the residual, and other embodiments may subject each residual value or all residual values to additional or other functions. One skilled in the art will recognize the many different measures that could be used.

[0035] In step 425, an estimated noise component present in the data set is calculated. The data is presumed to consist of two components, a true signal and noise. Thus, the noise is the difference between the true signal and the actual data point. However, the true signal can never be directly measured as noise is always added or present when a signal is detected.

[0036] The true signal is estimated as a smoothed data set composed of smoothed data points. Figure 6A shows a smoothed data set 670 of PCR curve 300. A value of a smoothed data point is based on a function G of a plurality of original data points that are local to that smoothed data point. The term local relates to how far away the x value of the data points are from the data point being calculated. For example, a point may be local to another point if they differ by a preset number (window) of cycles. A window of three and five cycles has proved adequate, but other windows may be used, such as 10 or 20 cycles or more. A variable window value may also be used, i.e. each smoothed data point may be calculated with a different window. Additionally, a window having fractions of a cycle may be used, for example where fractional data points are interpolated. A window may also not be symmetric around a data point, i.e. one point before and three points after that data point may be used. A point ceases to be local once the difference in the x value approaches the total scale used, i.e. total number of cycles.

[0037] In one embodiment, the function G is a moving average or low pass filter. For example, the function G may take an average of the original data points within a prescribed

number of cycles, e.g. a centered mean. Thus in one embodiment, $G(x_L) = \frac{1}{2K+1} \sum_{i=L-K}^{L+K} Y_i$,

where L is the index of the smoothed data point being calculated and K is the window used.

[0038] Also, in step 425, a residual between the smoothed data and the original data is calculated. This residual is defined to be the estimated noise. The residual between the
 5 smoothed data and the original data may be defined in the same manner as the residual between the original data and the linear fit, or the residuals may be defined in a different manner. Figure 6B shows an estimated noise component 680 associated with PCR curve 300 and smoothed data set 670. A superposition of noise component 680 on signal 670 gives the data curve 300.

10 [0039] In step 430, a measure σ_2 of the residual between the smoothed data and the original data is calculated. The σ_2 value is used as a measure of the amplitude of intrinsic noise. In one embodiment σ_2 is a standard deviation. The measures σ_1 and σ_2 may be defined in a similar or different fashion.

[0040] In step 435, the first measure σ_1 is compared to the second measure σ_2 to determine
 15 whether the data set exhibits linear behavior. In one embodiment, a ratio of σ_1 and σ_2 is taken. If the ratio is smaller or greater than a pre-defined value then the data is determined to exhibit linear behavior. For example, if σ_1/σ_2 is less than a value of order one, e.g. 1.5, the data is determined to be linear. Equivalently, the expression $\sigma_1 < c_0 * \sigma_2$ may be used. This expression states that the measure of the difference between the data and a linear fit must be
 20 less than a constant times the measure of the estimated noise present in the data. In some embodiments, the value of c_0 may vary.

[0041] The constant c_0 is related to the fact that the definition of noise, as well as other values, is not unique. The value for c_0 may be obtained by examining large numbers of data sets to obtain a reasonable value for this number. Studies have indicated a value of 1.5 works
 25 well for the constant (c_0), when a standard deviation of a standard residual is used. When other residuals and measures of the residual are used, other values might be more suitable. In general, a value of c_0 on the order of 1 should work well.

[0042] Once the data has been identified as statistically flat (linear), e.g. not curving downward or upward or otherwise displaying amplification, the data may be discarded from
 30 further analysis.

[0043] In certain aspects, code and instructions for controlling a processor to implement the data processing techniques of the present invention are stored on a computer-readable or information storage medium such as a RAM or ROM unit, hard drive, CD, DVD or other portable medium.

5 [0065] Figure 7 illustrates a system 700 according to one embodiment of the present invention. The system as shown includes a sample 705, such as bacteria or DNA, within a sample holder 710. A physical characteristic 715, such as a fluorescence intensity value, from the sample is detected by detector 720. A signal 725, including a noise component, is sent from detector 720 to logic system 730. The data from signal 725 may be stored in a
10 local memory 735 or an external memory 740 or storage device 745. In one embodiment, an analog to digital converter converts an analog signal to digital form.

[0066] Logic system 730 may be, or may include, a computer system, ASIC, microprocessor, etc. It may also include or be coupled with a display (e.g., monitor, LED display, etc.) and a user input device (e.g., mouse, keyboard, buttons, etc.). Logic system 730
15 and the other components may be part of a stand alone or network connected computer system, or they may be directly attached to or incorporated in a thermal cycler device. Logic system 730 may also include optimization software that executes in a processor 750.

[0067] According to one embodiment, logic system 730 includes instructions for processing data and identifying statistically flat data. The instructions are preferably
20 downloaded and stored in a memory modules 735, 740, or 745 (e.g., hard drive or other memory such as a local or attached RAM or ROM), although the instructions can be provided on any software storage medium such as a floppy disk, CD, DVD, etc. It should be understood that computer code for implementing aspects of the present invention can be implemented in a variety of coding languages such as C, C++, Java, Visual Basic, and others,
25 or any scripting language, such as VBScript, JavaScript, Perl or markup languages such as XML. In addition, a variety of languages and protocols can be used in the external and internal storage and transmission of data and commands according to aspects of the present invention.

[0044] It will be appreciated that the process described herein is illustrative and that
30 variations and modifications are possible. Steps described as sequential may be executed in parallel, order of steps may be varied, and steps may be modified or combined.

[0045] While the invention has been described by way of example and in terms of the specific embodiments, it is to be understood that the invention is not limited to the disclosed embodiments. To the contrary, it is intended to cover various modifications and similar arrangements as would be apparent to those skilled in the art. Therefore, the
5 scope of the appended claims should be accorded the broadest interpretation so as to encompass all such modifications and similar arrangements.

[0045A] In the claims which follow and in the preceding description of the invention, except where the context requires otherwise due to express language or necessary implication, the word "comprise" or variations such as "comprises" or "comprising" is
10 used in an inclusive sense, i.e. to specify the presence of the stated features but not to preclude the presence or addition of further features in various embodiments of the invention.

CLAIMS

1. A method of processing a data set to determine whether the data set exhibits statistically linear behavior, the method comprising:
 - receiving an original set of data points having a signal component and a noise component, each data point representing a physical quantity of a substance during an amplification process;
 - fitting the original data set to a linear function;
 - calculating a residual between the original data set and the fitted linear function;
 - calculating a first measure of the residual between the original data set and the fitted linear function;
 - estimating the noise component present in the original data set, by:
 - i) calculating a smoothed data set by determining a set of smoothed data points, wherein a value of a smoothed data point is based on values of a plurality of original data points that are local to that smoothed data point; and
 - ii) calculating a residual between the smoothed data set and the original data set;
 - calculating a second measure of the residual between the smoothed data set and the original data set; and
 - comparing the first measure to the second measure to determine whether the original data set exhibits statistically linear behavior, wherein the method is implemented using a processor.
2. The method of claim 1, wherein fitting the original data set includes using a least squares fit.
3. The method of claim 1 or claim 2, wherein calculating the smoothed data set comprises using a low pass filter.
4. The method of any one of the preceding claims, wherein a value of a smoothed data point is an average of the original data points within a window around the smoothed data point.

5. The method of claim 4, wherein the window is five units.
6. The method of any one of the preceding claims, wherein the first measure and the second measure are each a standard deviation.
7. The method of any one of the preceding claims, wherein
5 comparing includes calculating a ratio of the first and second measure to determine if the ratio is smaller or greater than a pre-defined value.
8. The method of claim 7, wherein the pre-defined value is of order
1.
9. The method of any one of the preceding claims, wherein the data
10 represents a PCR amplification curve.
10. The method of any one of the preceding claims, wherein the processor is integrated in one of a stand alone computer system, a networked computer system or a real-time PCR machine.
11. A tangible information storage medium storing a plurality of
15 instructions adapted to direct an information processing device to perform an operation of processing data to determine whether the data exhibits linear behavior, the operation comprising the steps of:
receiving an original set of data points having a signal component and a noise component, each data point representing a physical quantity of a substance during
20 an amplification process;
fitting the original data set to a linear function;
calculating a residual between the original data set and the fitted linear function;
calculating a first measure of the residual between the original data set
25 and the fitted linear function;
estimating the noise component present in the original data set, by:
i) calculating a smoothed data set by determining a set of smoothed data points, wherein a value of a smoothed data point is based on values of a plurality of original data points that are local to that smoothed data point; and

ii) calculating a residual between the smoothed data set and the original data set;

calculating a second measure of the residual between the smoothed data set and the original data set; and

5 comparing the first measure to the second measure to determine whether the original data set exhibits statistically linear behavior.

12. The information storage medium of claim 11, wherein fitting the original data set includes using a least squares fit.

10 13. The information storage medium of claim 11 or claim 12, wherein calculating the smoothed data set comprises using a low pass filter.

14. The information storage medium of any one of claims 11 to 13, wherein a value of a smoothed data point is an average of the original data points within a window around the smoothed data point.

15 15. The information storage medium of any one of claims 11 to 14, wherein the first measure and the second measure are each a standard deviation.

16. The information storage medium of any one of claims 11 to 15, wherein the comparing includes calculating a ratio of the first and second measure to determine if the ratio is smaller or greater than a pre-defined value.

20 17. The information storage medium of any one of claims 11 to 16, wherein the data represents a PCR amplification curve.

18. A PCR detection system comprising:

a detector for producing an original set of data points having a signal component and a noise component, wherein the data represents a PCR amplification curve; and

25 logic for processing data to determine whether the data exhibits linear behavior, by:

fitting the original data set to a linear function;

- calculating a residual between the original data set and the fitted linear function;
- calculating a first measure of the residual between the original data set and the fitted linear function;
- 5 estimating the noise component present in the original data set, by:
- i) calculating a smoothed data set by determining a set of smoothed data points, wherein a value of a smoothed data point is based on values of a plurality of original data points that are local to that smoothed data point; and
- 10 ii) calculating a residual between the smoothed data set and the original data set;
- calculating a second measure of the residual between the smoothed data set and the original data set; and
- comparing the first measure to the second measure to determine
- 15 whether the original data set exhibits statistically linear behavior.
19. The PCR detection system of claim 18, wherein fitting the original data set includes using a least squares fit.
20. The PCR detection system of claim 18 or claim 19, wherein calculating the smoothed data set comprises using a low pass filter.
- 20 21. The PCR detection system of any one of claims 18 to 20, wherein a value of a smoothed data point is an average of the original data points within a window around the smoothed data point.
22. The PCR detection system of any one of claims 18 to 21, wherein the first measure and the second measure are each a standard deviation.
- 25 23. The PCR detection system of any one of claims 18 to 22, wherein the comparing includes calculating a ratio of the first and second measure to determine if the ratio is smaller or greater than a pre-defined value.

24. A method of processing a data set to determine whether the data set exhibits statistically linear behaviour substantially as hereinbefore described in the detailed description with reference to any of the accompanying Figures 4, 6A, 6B and 7.

25. A tangible information storage medium substantially as
5 hereinbefore described in the detailed description with reference to any of the accompanying Figures 4, 6A, 6B and 7.

26. A PCR detection system substantially as hereinbefore described in the detailed description with reference to any of the accompanying Figures 4, 6A, 6B and 7.

10

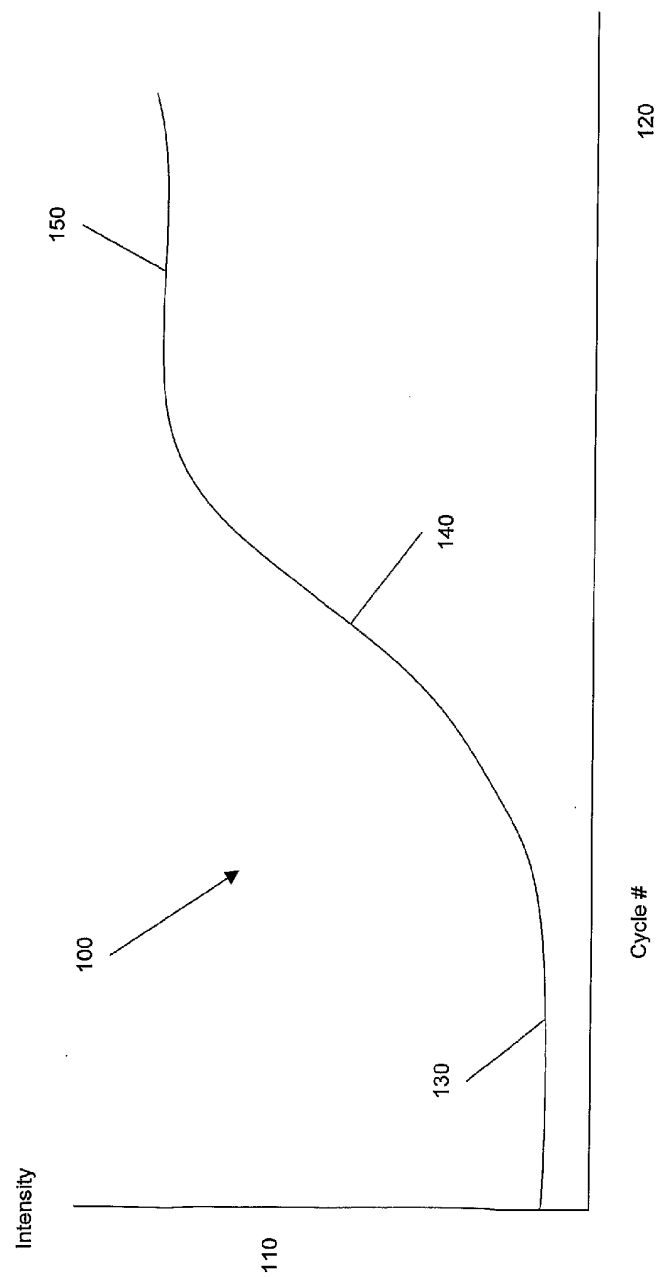


FIG. 1

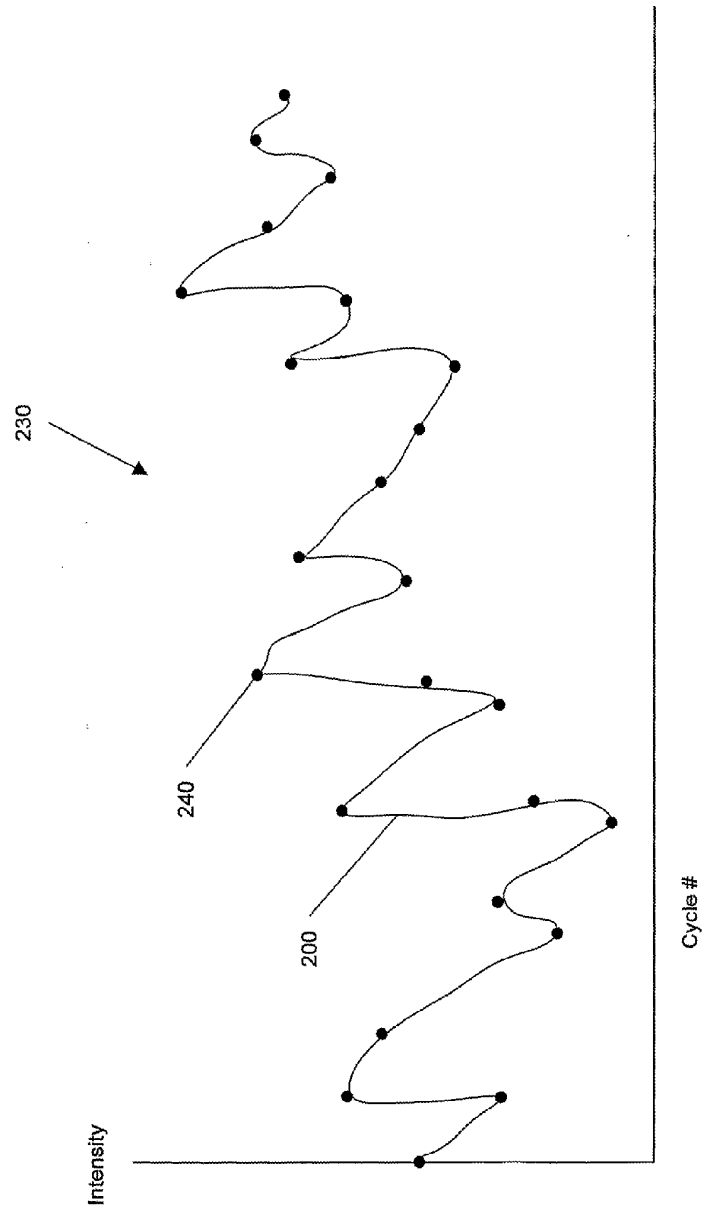


FIG. 2

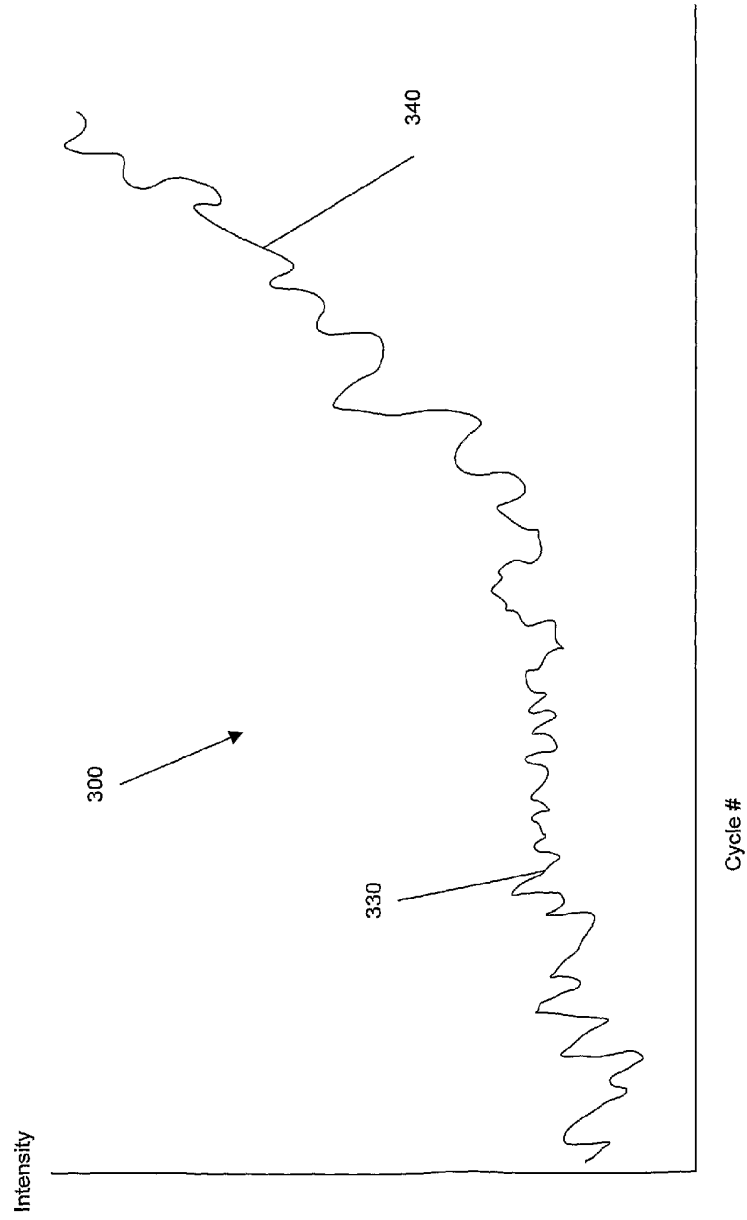


FIG. 3

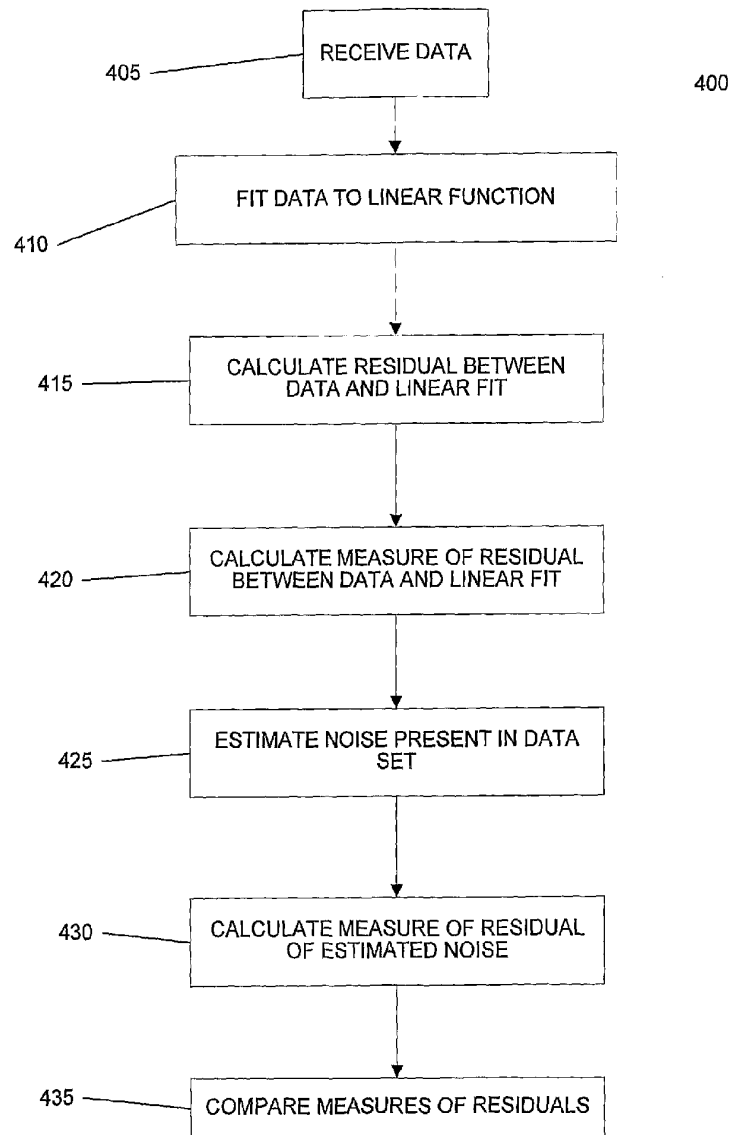


FIG. 4

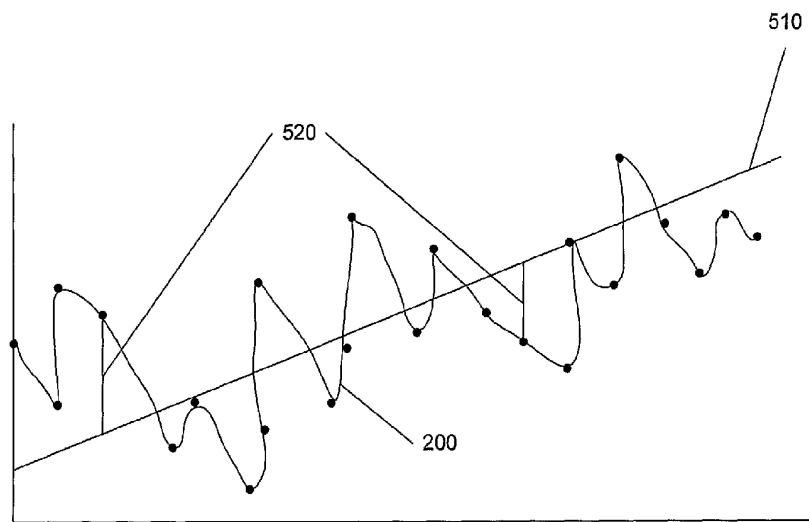


FIG. 5A

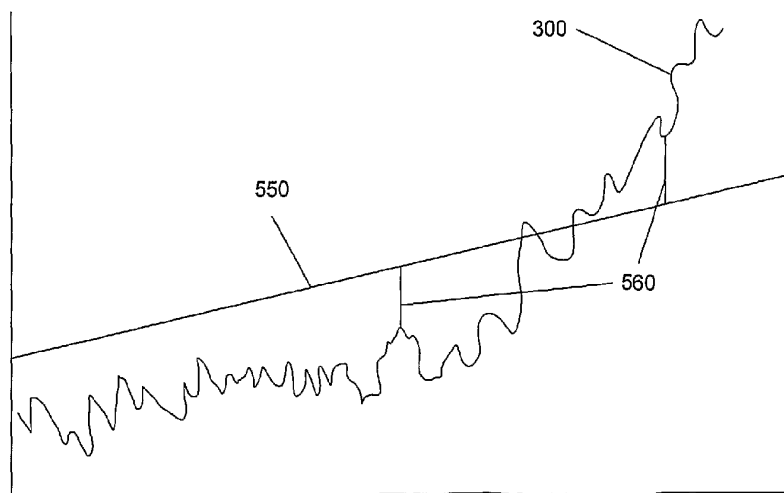


FIG. 5B

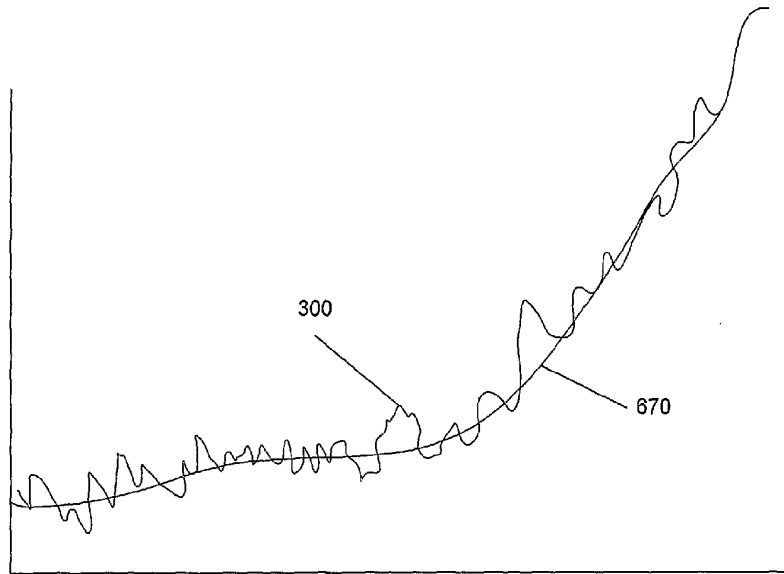


FIG. 6A

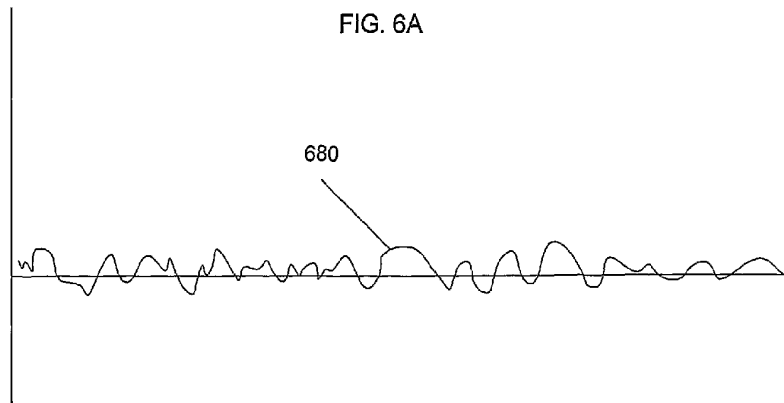


FIG. 6B

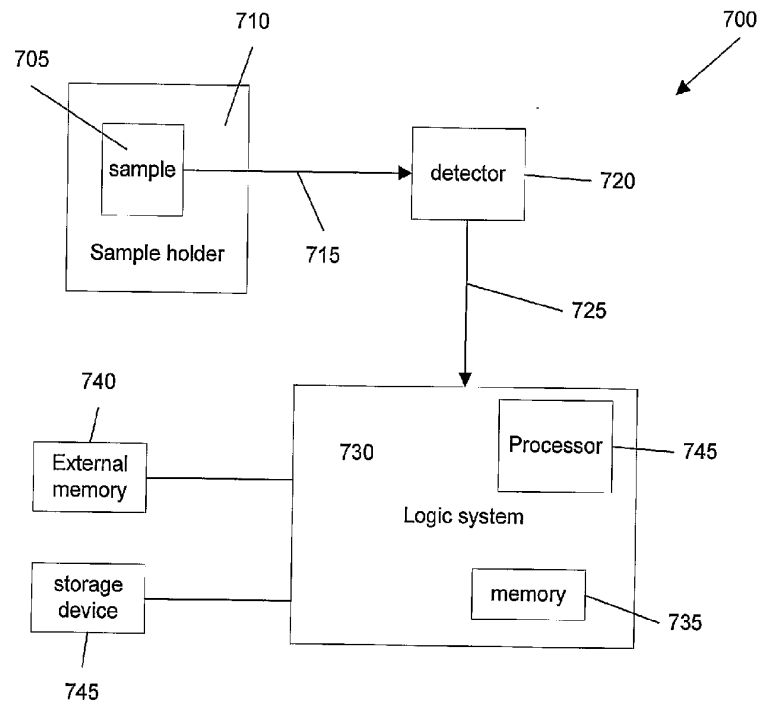


FIG.7