(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 September 2012 (07.09.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/117036 A1

(51) International Patent Classification:
G01N 21/33 (2006.01)

(21) International Application Number:
PCT/EP2012/053481

(22) International Filing Date:
29 February 2012 (29.02.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/447,941	1 March 2011 (01.03.2011)	US
11168005.4	29 May 2011 (29.05.2011)	EP
1200031.1	3 January 2012 (03.01.2012)	GB

(71) Applicant *(for all designated States except US)*: **TRINEAN NV** [BE/BE]; Dulle-Grietlaan 17bus3, B-9050 Gentbrugge (BE).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **BOONEFAES, Tom** [BE/BE]; Duinhoekstraat 127/112, B-8660 De Panne (BE).

(74) Agents: **WAUTERS, Davy** et al.; DenK iP bvba, Pastoor Ceulemansstraat 3, B-3191 Schiplaken (Hever) (BE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

(54) **Title**: DNA AND/OR RNA DETERMINATION FROM UV-VIS SPECTROPHOTOMETER DATA

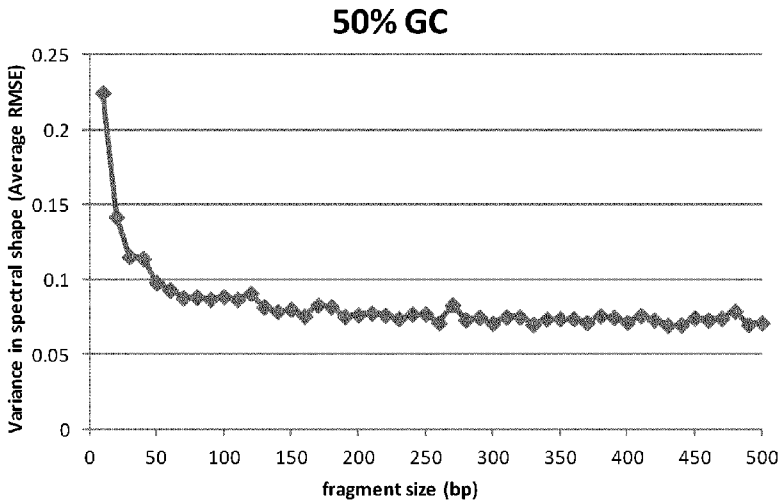


FIG. 6

(57) **Abstract**: A method for analyzing UV-VIS spectrophotometer data of a sample comprising at DNA and/or RNA is described. The method comprises receiving UV-VIS spectrophotometer data, fitting the UV-VIS spectrometer data taking into account at least one spectrum representative for a base pair being any of more of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil, and deriving from the fitting a quantification of an amount of DNA and/or RNA.

WO 2012/117036 A1

DNA and/or RNA determination from UV-VIS spectrophotometer data

Field of the invention

The invention relates to the field of sample characterisation. More particularly, the present invention relates to methods and systems for analyzing samples using recorded spectra, such as for example for analyzing samples comprising DNA and/or RNA.

Background of the invention

Although numerous analysis techniques for qualification and quantification of samples exists, only few analysis techniques are as simple to perform, fast and accurate as spectrophotometry. One example of spectrophotometry is UV-VIS absorbance spectroscopy. During such experiments, samples are irradiated with UV-VIS radiation of different wavelengths, the radiation remaining after passage through the sample is detected and the absorbance at different wavelengths is determined. As particular components will show a particular absorbance at particular wavelengths, such a particular absorbance profile can be used as a fingerprint which allows, upon comparison with reference spectra, to identify the components. When more complex samples are studied, the absorbance features in the spectrum can be more broad, rendering the interpretation of spectra substantially more difficult.

A number of different spectral analysis techniques have been described in the past. Some methods are provided allowing determining the concentration or relative concentration of components of a sample based on the approximation of the measured spectrum with a linear combination of standard, also referred to as reference, spectra of the individual components. Approximation of the measured spectrum with reference spectra may be performed using minimization techniques for fitting the reference spectra to the absorption spectra, such as for example least squares regression or least absolute deviations regression. Such methods also may be iteratively applied.

More complex methods also are known, such as methods making use of derivatives of a spectrum, as amongst others the latter may allow to increase the insensitivity to

wavelength shifts. One example of a method using derivatives is the analysis of a protein mixture using the second derivative spectrum of the sample. Another known method is based on matrix computation for determining the concentration of a number of chemical components using a number of predetermined spectral data of each chemical component, whereby the number of predetermined spectral data is larger than the number of chemical components to be determined. Some of the known methods are based on cross-correlation of samples and reference spectra, both being weighted by mean transmittance per wavelength. Particular applications for analyzing samples with known composition have been described, such as for example analysis of the composition of a protein using electromagnetic spectroscopy, analysis of RNA wherein samples are chemically split in AC and GU fractions and the AC and GU composition is then determined based on absorption at 260nm.

Furthermore, also some methods are known that deal with unknown component spectra. Methods are known wherein the separation of unknown component spectra from samples is based on constrained linear or quadratic programming and sample transforms, i.e. wavelet transforms. Other methods make use of approximated component spectra by point-wise cross-correlation of absorbance and making use of known concentration. Several techniques make use of principal component analysis (PCA) for obtaining diagnostically relevant reference spectra from a database of spectra associated with disease states whereby correlation with reference spectra is used to aid in the diagnosis of a new sample.

Spjotvoll et al. describe in Technometrics **24** (1982) a technique for determining spectra and concentrations of two-constituent mixtures based on a least squares estimator. A method is discussed for identification of two unknown chemical compounds and an estimation of their proportion in a set of unknown mixtures of the two compounds. It is based on a least-squares fit and principle component analysis for separation whereby additional constraints are introduced such as the sum of concentrations being equal to unity and spectra and concentrations being non-negative.

The use of UV-VIS spectrophotometer data for DNA and/or RNA quantification or analysis still has not found its breakthrough due to the lack of an accurate and efficient analysis technique for the obtained results.

Therefore, there is need for a method allowing analysis of samples comprising DNA and/or RNA components, particularly using UV-VIS spectroscopy.

Summary of the invention

It is an object of embodiments of the present invention to provide good methods and systems for analyzing UV-VIS spectroscopy spectra of samples and good corresponding methods and systems for characterizing samples comprising DNA and/or RNA. The sample advantageously comprises double stranded DNA, although the technique also can be used for samples comprising only RNA. The UV-VIS spectroscopy spectra may be UV-VIS absorbance spectra.

It is an advantage of embodiments according to the present invention that systems and methods are provided allowing to assist in PCR analysis, e.g. without the need for purifying the reaction mixture, even if the reaction mixture is a mixture of 4dNTP's, primer oligo's, polymerase-enzyme and DNA in the sample.

It is an advantage that methods and systems according to embodiments of the present invention allow for obtaining relevant and conclusive results.

The above objective is accomplished by a method and device according to the present invention.

The present invention relates to a computer-implemented method for analyzing UV-VIS spectrophotometer data of a sample comprising at least DNA and/or RNA, the at least DNA and/or RNA advantageously being at least double stranded DNA, the method comprising receiving UV-VIS spectrophotometer data, fitting the UV-VIS spectrometer data taking into account at least one spectrum representative for a base pair being any of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil (AU), and deriving from the fitting a quantification of an amount of DNA and/or RNA, the deriving being advantageously deriving from the fitting a quantification of an amount of double stranded DNA. It is an advantage of embodiments of the present invention that UV-VIS spectrophotometer data can be used for

characterizing, without a lot of prior knowledge, a DNA and/or RNA containing sample. Quantification may be quantification of an amount of double stranded DNA (dsDNA).

Said fitting may comprise taking into account a set of reference spectra representative for two distinct base pair contents. The spectra representative for distinct base pair contents may be representative for base pair contents differing at least 20%, advantageously at least 30%, more advantageously at least 40%, e.g. at least 50%.

Said fitting may comprise taking into account a set of two reference spectra representative of two distinct GC contents.

Said fitting may comprise taking into account a set of exactly two reference spectra representative for two distinct GC contents.

Said fitting may comprise fitting UV-VIS spectrophotometer data of a sample comprising DNA and/or RNA having at least 50 base pairs, advantageously at least 100 base pairs. It is an advantage according to embodiments of the present invention that a fit can be obtained using a restricted number of reference spectra for base pairs, resulting in accurate and efficient fitting.

The sample may comprise DNA and the fitting may comprise using a combination of a set of reference spectra representative for base pairs and reference spectra being representative for separate nucleobases.

Said fitting using a combination of reference spectra representative for base pairs and reference spectra being representative for separate nucleobases may comprise fitting using spectra being the difference spectra between reference spectra representative for base pairs and reference spectra being representative for separate nucleobases. It is an advantage of embodiments according to the present invention that for the analysis of a DNA comprising sample, the AT and GC balance is automatically taken into account. It is an advantage of embodiments of the present invention that only the formed double stranded DNA is quantified and not a mixture of the dsDNA and the formed amplicons, which may occur e.g. in PCR.

The difference spectra between reference spectra representative for base pairs and reference spectra being representative for separate nucleobases may be spectra obtained by subtracting from a given reference spectrum representative for a given base pair, the reference spectra being representative for the corresponding separate nucleobases.

The set of spectra representative for base pairs may consist of two reference spectra representative for samples having two distinct, e.g. two extreme, GC contents.

The reference spectra representative of separate nucleobases may be reference spectra for adenine, guanine, thymine, cytosine and/or uracil.

The method may be adapted for analyzing a PCR reaction of a sample comprising DNA, whereby the deriving may comprise determining an amount of double stranded DNA formed during or after a PCR reaction. It is an advantage of embodiments according to the present invention that analysis can be done of a PCR reaction of a DNA-comprising-sample using UV-VIS spectrometry, without the need for first purifying the reaction mixture formed during the PCR reaction.

Fitting furthermore may comprise taking into account a spectral component for a contamination component.

Taking into account a spectral component for a contamination component may comprise taking into account a spectral component e.g. for one or more of trimethylglycine, guanidinethiocyanate, phenol, ethylenediaminetetraacetic acid, proteins or hemoglobine. It is an advantage of embodiments according to the present invention that compensation for one of the most important contaminants in the PCR reaction can be obtained.

Fitting UV-VIS spectrophotometer data of a sample may comprise fitting UV-VIS spectrophotometer data of a mixture comprising deoxynucleotide Triphosphates, primer-oligo's and polymerase and DNA. It is an advantage of embodiments according to the present invention that the UV-VIS spectrophotometer data directly recorded on the PCR starting mixture can be used, although it has a high optical density value at 260nm, thus avoiding the need for processing, e.g. purifying, the PCR starting mixture before characterisation.

Fitting furthermore may comprise taking into account one or more spectral components representative for scattering in the sample.

Fitting furthermore may comprise taking into account one or more spectral components representative for turbidity in the sample.

The present invention also relates to a computer-implemented system for analyzing UV-VIS spectrophotometer data of one or more samples comprising DNA and/or RNA, advantageously one or more samples comprising at least double stranded DNA, the system comprising an input means for receiving said UV-VIS spectrophotometer data, a processing means programmed for fitting the UV-VIS spectrometer data taking into account at least one spectrum representative for a base pair being any of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil (AU), and deriving from the fitting a quantification of an amount of DNA and/or RNA, advantageously an amount of double stranded DNA, and an output means for outputting a quantification of an amount of DNA and/or RNA, advantageously the amount of double stranded DNA, for the one or more samples under study.

Said processing means may be adapted for performing fitting as described in any of the above mentioned methods.

The present invention also relates to the use of a method as described above for quantifying an amount of DNA and/or RNA.

The present invention furthermore relates to a computer program product for, if implemented on a processing unit, performing a method as described above.

The present invention also relates to a data carrier medium comprising such a computer program product and/or to the transmission thereof over a wide or local area network.

The above described aspect may be part of, may correspond with, may comprise the steps or features of or may make use of the following aspects of the present invention.

The present invention also relates to a method for analyzing UV-VIS spectrophotometer data of one or more samples , the one or more samples consisting of a number of constituents, the method comprising obtaining prior

information for the one or more samples regarding their constituents, obtaining UV-VIS spectrophotometer data for the one or more samples, defining a number of overlapping components contributing in the UV-VIS spectrophotometer data, the number of overlapping components comprising one or more components assigned to known constituents of the one or more samples and the number of components comprising one or more components that cannot be assigned to known constituents of the one or more samples, using the prior information for the one or more samples regarding their constituents and using the UV-VIS spectrophotometer data, estimating the constituents composition and the component contributions to the UV-VIS spectrophotometer data for the number of components for the one or more samples by minimizing a residue between the UV-VIS spectrophotometer data and a fit based on said constituent composition and said component contributions, thus obtaining information regarding the one or more components that cannot be assigned to known constituents of the one or more samples. It is an advantage of embodiments according to the present invention that the analysis method allows for accurate detection of contributions to the spectrum that stem from modified constituents, interaction of different constituents or contaminations in the one or more samples.

Estimating the constituents composition and the component contributions to the UV-VIS spectrophotometer data for the number of components for the one or more samples by minimizing a residue between the UV-VIS spectrophotometer data and a fit based on said constituent composition and said component contributions may be performed iteratively until a residue smaller than a predetermined level is obtained. It is an advantage of embodiments according to the present invention that using an iterative method an increased accuracy for the fit and the corresponding constituent composition and component contribution can be obtained.

The method furthermore may comprise cross-checking the prior information based on any or more of the constituents composition, the component contribution or the residue between the UV-VIS spectrophotometer data and the fit. It is an advantage of

embodiments that a cross-check regarding the obtained results can be performed using the prior information regarding the one or more samples.

Said prior information may comprise reference contributions in UV-VIS spectrophotometer data for components assigned to known constituents and said estimating the constituents composition and the component contributions may comprise estimating the component contributions to the UV-VIS spectrophotometer for all components based on the reference contributions and determining an estimate for the constituents composition based on minimization of the residue between the UV-VIS spectrophotometer and the fit based on said estimated component contributions to the UV-VIS spectrophotometer data for all components. It is an advantage of embodiments according to the present invention that accurate composition information of the one or more samples can be obtained based on reference spectra of constituents present in the one or more samples, taking into account the presence of unknown contributions in the UV-VIS spectrophotometer data.

Said prior information may comprise constituent composition information for the one or more samples for the known constituents and said estimating the constituents composition and the component contributions may comprise estimating the constituent composition for all constituents based on the prior information on the constituent composition information for the one or more samples for the known constituents, and determining an estimate for the component contribution to the UV-VIS spectrophotometer data based on minimization of the residue between the UV-VIS spectrophotometer data and the fit based on said estimated constituent composition for all constituents. It is an advantage of embodiments according to the present invention that accurate composition information of the one or more samples can be obtained based on reference spectra of constituents present in the one or more samples, taking into account the presence of unknown contributions in the UV-VIS spectrophotometer data.

The constituent composition and the component contributions may be interrelated by a model defining weighing factors for component contributions as function of the

constituent composition and wherein determining an estimate for the constituent composition or for the component contribution comprises taking into account the weighing factors. It is an advantage of embodiments according to the present invention that the composition of constituents in the samples can be taken into account.

Analysis may be performed on spectrophotometer data for a plurality of samples, and said iteratively estimating may be performed for those samples having the smallest residue.

The components that cannot be assigned to known constituents of the one or more samples may be representative for any or more of modified constituents, neighbouring effects of different constituents or contaminations in the one or more samples. Another component that advantageously can be taken into account by components that cannot be assigned to known constituents of the one or more samples may be representative for turbidity caused by concentration or aggregation effects. The method may comprise outputting a contribution related to such turbidity. It is an advantage of embodiments of the present invention that good and/or accurate quantification of components in samples showing high turbidity can be obtained, e.g. based on prior knowledge of reference spectra of components in the sample. Another component that advantageously can be taken into account by components that cannot be assigned to known constituents of the one or more samples may be representative for scatter by residual microparticles. Such particles may be non-intentionally present in the sample, such as for example introduced by earlier processing of the sample. In one example, such residual particles may be magnetical micro-particles. The method may comprise outputting a contribution related to such scattering from residual particles. It is an advantage of embodiments of the present invention that good and/or accurate quantification of components in samples comprising residual particles can be obtained, e.g. based on prior knowledge of reference spectra of components in the sample.

The method may comprise using spectrophotometric data of a plurality of samples, wherein the samples comprises at least one sample wherein at least one unknown

constituent is present and at least one sample wherein the at least one unknown constituent is absent.

The number of components may be determined by component analysis.

The number of components may be determined based on principal component analysis.

The present invention also relates to a system for analyzing UV-VIS spectrophotometer data of one or more samples , the one or more samples consisting of a number of constituents, the system comprising an input means for obtaining prior information for the one or more samples regarding their constituents and for obtaining UV-VIS spectrophotometer data for the one or more samples, a processing means programmed for defining a number of components contributing in the UV-VIS spectrophotometer data, the number of components comprising one or more components assigned to known constituents of the one or more samples and the number of components comprising one or more components that cannot be assigned to known constituents of the one or more samples, the processing means furthermore being programmed for using the prior information for the one or more samples regarding their constituents and using the UV-VIS spectrophotometer data, estimating the constituents composition and the component contributions to the UV-VIS spectrophotometer data for the number of components for the one or more samples by minimizing a residue between the UV-VIS spectrophotometer data and a fit based on said constituent composition and said component contributions, thus obtaining information regarding the one or more components that cannot be assigned to known constituents of the one or more samples.

The present invention also relates to the use of a method as described above for identifying and/or quantifying contaminants in DNA and/or RNA samples. Said contaminants may be protein contamination. The contaminants may be PCR-inhibiting contaminants.

The present invention also relates to the use of a method as described above for quantifying an amount of double-stranded DNA in a mixture of double-stranded DNA and single-stranded RNA and/or DNA.

The present invention also relates to the use of a method as described above for obtaining the composition of a polynucleotide or protein.

The present invention also relates to the use of a method as described above for quantification of the modification efficiency in a sample.

The present invention also relates to the use of a method as described above for validating a sample composition or constituent thereof.

The present invention also relates to a computer program product for, if implemented on a processing unit, performing a method as described above. It also relates to a data carrier storing the computer program product and transmission of the computer program product over a network.

It is an advantage of embodiments of the present invention that a powerfull data analysis of UV-VIS spectroscopy spectra is obtained.

It is an advantage of embodiments according to the present invention that good quantification can be obtained, both for routine applications and more complex applications.

It is an advantage of embodiments according to the present invention that based on prior knowledge in one or more samples, accurate reference spectra can be obtained for components in the sample.

It is an advantage of embodiments according to the present invention that based on prior knowledge of reference spectra of one or more components in the one or more samples, accurate concentration information can be obtained for components in the sample.

It is an advantage that methods and systems according to embodiments of the present invention allow for obtaining relevant and conclusive results.

Particular and preferred aspects of the invention are set out in the accompanying independent and dependent claims. Features from the dependent claims may be combined with features of the independent claims and with features of other dependent claims as appropriate and not merely as explicitly set out in the claims.

These and other aspects of the invention will be apparent from and elucidated with reference to the embodiment(s) described hereinafter.

Brief description of the drawings

FIG. 1 illustrates a flow chart of a method for analyzing spectra according to an embodiment of the present invention.

FIG. 2 illustrates an example of data flow as can be obtained using embodiments of the present invention.

FIG. 3 illustrates an example of data flow as can be obtained using embodiments of the present invention.

FIG. 4 illustrates the different spectral shapes for free nucleotides, nucleotides in single stranded DNA and double stranded DNA for adenine and thymine (top graph) and for guanine and cytosine (bottom graph), as can be used according to embodiments of the present invention.

FIG. 5a to FIG. 5f illustrates simulation results illustrating the spectral variation for different lengths of base pair strands as function of the GC content, as can be used according to embodiments of the present invention.

FIG. 6 is another illustration of the spectral variation as function of the length of strands of base pairs, as can be used according to embodiments of the present invention.

FIG. 7a, FIG. 7b, FIG. 8a and FIG. 8b illustrate measurement results obtained with different fitting techniques for distinguishing between a sample comprising double stranded DNA and a sample not comprising double stranded DNA, as can be used in embodiments of the present invention.

FIG. 9 illustrates spectroscopic spectra of mixed RNA and DNA, as can be analysed using methods according to embodiments of the present invention.

FIG. 10 to FIG. 11 illustrate deconvolution examples of two samples comprising RNA, analysed using methods according to embodiments of the present invention.

The drawings are only schematic and are non-limiting. In the drawings, the size of some of the elements may be exaggerated and not drawn on scale for illustrative purposes.

Any reference signs in the claims shall not be construed as limiting the scope.

In the different drawings, the same reference signs refer to the same or analogous elements.

Detailed description of illustrative embodiments

The present invention will be described with respect to particular embodiments and with reference to certain drawings but the invention is not limited thereto but only by the claims. The drawings described are only schematic and are non-limiting. In the drawings, the size of some of the elements may be exaggerated and not drawn on scale for illustrative purposes. The dimensions and the relative dimensions do not correspond to actual reductions to practice of the invention.

Furthermore, the terms first, second and the like in the description and in the claims, are used for distinguishing between similar elements and not necessarily for describing a sequence, either temporally, spatially, in ranking or in any other manner. It is to be understood that the terms so used are interchangeable under appropriate circumstances and that the embodiments of the invention described herein are capable of operation in other sequences than described or illustrated herein.

Moreover, the terms top, under and the like in the description and the claims are used for descriptive purposes and not necessarily for describing relative positions. It is to be understood that the terms so used are interchangeable under appropriate circumstances and that the embodiments of the invention described herein are capable of operation in other orientations than described or illustrated herein.

It is to be noticed that the term “comprising”, used in the claims, should not be interpreted as being restricted to the means listed thereafter; it does not exclude other elements or steps. It is thus to be interpreted as specifying the presence of the stated features, integers, steps or components as referred to, but does not preclude the presence or addition of one or more other features, integers, steps or components, or groups thereof. Thus, the scope of the expression “a device comprising means A and B” should not be limited to devices consisting only of components A and B. It means that with respect to the present invention, the only relevant components of the device are A and B.

Reference throughout this specification to "one embodiment" or "an embodiment" means that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present invention. Thus, appearances of the phrases "in one embodiment" or "in an embodiment" in various places throughout this specification are not necessarily all referring to the same embodiment, but may. Furthermore, the particular features, structures or characteristics may be combined in any suitable manner, as would be apparent to one of ordinary skill in the art from this disclosure, in one or more embodiments.

Similarly it should be appreciated that in the description of exemplary embodiments of the invention, various features of the invention are sometimes grouped together in a single embodiment, figure, or description thereof for the purpose of streamlining the disclosure and aiding in the understanding of one or more of the various inventive aspects. This method of disclosure, however, is not to be interpreted as reflecting an intention that the claimed invention requires more features than are expressly recited in each claim. Rather, as the following claims reflect, inventive aspects lie in less than all features of a single foregoing disclosed embodiment. Thus, the claims following the detailed description are hereby expressly incorporated into this detailed description, with each claim standing on its own as a separate embodiment of this invention.

Furthermore, while some embodiments described herein include some but not other features included in other embodiments, combinations of features of different embodiments are meant to be within the scope of the invention, and form different embodiments, as would be understood by those in the art. For example, in the following claims, any of the claimed embodiments can be used in any combination.

Where in embodiments of the present invention reference is made to spectrophotometer data, reference is made to a reflectance, transmittance or absorbance intensity at a given wavelength or as function of wavelength. Advantageously, the method can be applied to spectrophotometer data being a

spectrophotometer spectrum, indicating an intensity as function of wavelength. In advantageous embodiments, the data are absorbance data, representative of the absorption that has occurred in the one or more samples.

Embodiments of the present invention are directed to methods and systems for analyzing UV-VIS spectrophotometer data. Where reference is made to UV-VIS, reference is made to a wavelength region having an upper wavelength in the region 700nm to 1100nm and a lower wavelength in the region 150nm to 300nm.

Where in embodiments of the present invention reference is made to components contributing to the UV-VIS spectrophotometer data, reference is made to components to an intensity fraction for one or more wavelengths in the spectrophotometer data, e.g. spectrophotometer spectrum. Such a component may be directly assignable to a constituent in the one or more samples or may be not or not directly assignable to a constituent in the one or more samples.

Where in embodiments of the present invention reference is made to components that cannot be assigned to known constituents of the one or more samples, reference is made to components that do not correspond with or for which it is not yet known that these correspond with a physical-chemical absorption characteristic of the constituents of the one or more samples. The components thus do not necessarily correspond to a physical-chemical absorption characteristic of the constituents. The components do not need to correspond with constituents having a significant contribution, e.g. absorbance, within the wavelength region under study.

Where in embodiments of the present invention reference is made to constituents of one or more samples, reference is made to biological or non-biological elements such as chemical elements that are present in the sample. Such constituents can for example be one of the following or elements or groups of elements thereof : biopolymers, small organic molecules, metabolites such as glucose or ethanol, proteins, peptides, nucleic acid segments, oligonucleotides, oligonucleotides containing one or more modifications such as for example fluorescent dye labels or chromophores, chemical linker modifications such as aldehyde, phosphate, amine- or thiol-modifications, affinity-labels such as biotin,

reactive components such as enzymes (ALP, HRP), alternative bases such as deoxyuracil, deoxyinosine, phosphothiolates, locked nucleic acids, polynucleotides, nucleotides, oligos, peptide nucleic acids, antibodies and variations thereof such as for example nanobodies or alphabodies, micro- and nanoparticles either intentionally or unintentionally added, magnetic or non-magnetic particles, lipids, micelles, vesicles, viral particles, metal-complexes, interacallating dyes and polymers such as polythiophene-derivatives, chromogens such as p-NPP, tetrazolium salts, staining reagents such as Coomassie, methyl blue, eosin, molecules such as small molecule pharmaceuticals, antibiotics or drugs, molecules with a regulatory effect in enzymatic processes such as promoters, activators, inhibitors, or cofactors, viruses, bacteria, cells, cell components, cell membranes, spores, DNA, RNA, micro-organisms and fragments and products. Such constituents also can be for example chemical elements, such as chemical elements present in a pharmaceutical or chemical product.

Where in embodiments according to the present invention reference is made to overlapping components, reference is made to the fact that the component contribution to the spectrum of such components is significant at the same wavelength or in the same wavelength range.

Where in embodiments of the present invention reference is made to turbidity, reference is made to cloudiness or haziness of a fluid caused by individual particles that are generally invisible to the naked eye.

In a first aspect, the present invention relates to a method for analyzing UV-VIS spectrophotometer data of one or more samples. A typical disadvantage of UV-VIS spectrophotometer data of complex data is that different contributions e.g. corresponding with different constituents in the sample or with unknown constituents, will result in overlapping contributions. This results in broad-banded spectra, whereby the different contributions are not as such identifiable from the spectra in the form of shoulders or peaks. Embodiments of the present invention allows analyzing of UV-VIS spectrophotometer data of such complex samples resulting in spectra with broad band absorption built up by overlapping contributions.

Methods according to embodiments of the present invention can be used for analysis of one or more samples. Whereas analysis of a single sample is possible, the method is especially suitable for analyzing spectrophotometer data of multiple samples. The one or more samples comprise a number of constituents. The methods according to embodiments of the present invention are especially suitable for analyzing data for samples having two or more, advantageously three or more constituents. According to embodiments of the present invention, the methods are especially suitable for analyzing UV-VIS spectrophotometer data of a sample comprising at least DNA and/or RNA. The method according to embodiments of the present invention, being a computer-implemented method, comprises receiving UV-VIS spectrophotometer data, fitting the UV-VIS spectrometer data taking into account at least one spectrum representative for a base pair being any of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil, and deriving from the fitting a quantification of an amount of DNA and/or RNA. Receiving UV-VIS spectrophotometer data may comprise obtaining pre-stored UV-VIS spectrophotometer data or measuring UV-VIS spectrophotometer data for the one or more samples comprising DNA and/or RNA. Such UV-VIS spectrophotometer data typically may be a spectrophotometer spectrum, although information at one or more individual wavelengths or wavelength ranges also may be used. The sample comprising DNA and/or RNA may comprise single stranded DNA, double stranded DNA (dsDNA) and/or RNA.

According to embodiments of the present invention, one step of the method comprises fitting the UV-VIS spectrometer data taking into account at least one spectrum representative for a base pair being any of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil. In some embodiments according to the present invention exactly two spectra representative for a base pair being any of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil are used. More particularly, in some embodiments use will be made of two reference spectra representative for samples having distinct GC contents. In other exemplary embodiments, such spectral components are used in combination with other spectra, or are combined with other spectra so that new spectra are obtained that can be

used for the fitting. An example of different spectral forms wherein the nucleotides adenine, thymine, guanine and cytosine can occur is shown in FIG. 4, illustrating the occurrence from adenine and thymine on the one hand (drawing at top) and of guanine and cytosine on the other hand (drawing at bottom) as respectively a combination of free nucleotides (A+T curve 402; G+C curve 412), as occurring in single stranded DNA (A+T curve 404; G+C curve 414) and occurring as base pairs in double stranded DNA (A+T curve 406; G+C curve 416). The different spectral shapes can clearly be identified. Where in embodiments according to the present invention reference is made to a spectrum representative for a base pair, reference is made to the spectrum where nucleotides occur as base pairs in double stranded DNA or to a spectrum comprising spectral features for the situation where nucleotides occur as base pairs in double stranded DNA.

The method according to embodiments of the present invention also comprises deriving from the fitting a quantification of an amount of DNA and/or RNA. Deriving a quantification of an amount of DNA and/or RNA may for example comprise deriving an amount of single stranded DNA, an amount of double stranded DNA and or an amount of RNA. Embodiments of the present invention may be adapted for determining an amount of double stranded DNA, and deriving therefrom an amount of single stranded DNA or an amount of RNA.

By way of illustration, embodiments of the present invention not being limited thereto, a number of particular embodiments will further be described below.

In a first set of particular embodiments according to the present invention, a computer-implemented method as described above is disclosed, whereby fitting of UV-VIS spectrophotometer data of a sample comprising DNA and/or RNA comprises fitting the UV-VIS spectrophotometer data using exactly two reference spectra representative for base pairs. In the present embodiment, these two reference spectra are spectra representative for two samples having distinct particular base pair content, such as for example GC content. Typically, in such embodiments one reference spectrum will be representative of a sample having a higher GC content, whereas the other reference spectrum will be representative of a sample having a

lower GC content. Where reference is made to distinct GC contents, the difference in content may be for example at least 20%, advantageously at least 30%, more advantageously at least 40%, e.g. at least 50%. Such embodiments may be especially suitable for fitting samples comprising DNA and/or RNA strands having at least 50 base pairs, advantageously at least 100 base pairs. Efficient fitting thus can be obtained. The latter is based on the surprising observation that the spectral variation for DNA with randomly generated sequences having a sufficient length can be fully explained taking into account the GC content, and can be accurately fit using two spectra representative for samples having distinct GC content.

In particular embodiments according to the present invention, a method for analyzing a polymerase chain reaction (PCR) based on the above described computer-implemented method is disclosed. For a PCR reaction, typically one starts from a mixture comprising deoxynucleotide Triphosphates, primer-oligo's, polymerase and a certain amount of DNA. The dNTP's in the mixture typically have a high optical density at 260nm obscuring the contribution of the primers, target and amplicon, thus rendering it difficult to analyse such samples with respect to specific discrimination between the different components. Using embodiments according to the present invention will allow to evaluate and/or analyse the PCR reaction, even without purifying the reaction mixture, i.e. without separating components in the reaction mixture. Using embodiments according to the present invention, the amount of double stranded DNA that is formed, which is typically relatively small, can be accurately quantified, as well as the loss of deoxyribonucleotide triphosphate (dNTP). The formed amplicon can differ substantially in GC content, dependent on the final sequence. The varying shape of the double stranded DNA spectrum, being dependent on the GC content as indicated above, can be fitted using embodiments of the present invention as illustrated in the current description. Furthermore, in some embodiments, also the shape dependency on the mixture of free dNTP's can be taken into account. This shape dependency can e.g. be caused by the degree of consumption or incorporation of A+T and G+C nucleotides.

In a second set of particular embodiments, a method for analyzing a polymerase chain reaction (PCR) is described, wherein for fitting two reference spectra are used for distinct samples having distinct base pair content, such as for example distinct GC

content, in combination with reference spectra for separate nucleobases. The latter reference spectra can for example be reference spectra for respectively A&T and G&C components, being combinations of individual spectra for nucleobases and not being the spectrum of dinucleotides. By making use of these reference spectra an appropriate fit can be obtained. Additionally also other components can be taken into account for the fitting, such as for example a component for scattering or turbidity, a component for a contamination such as for example for one or more of trimethylglycine, guanidinethiocyanate, phenol, ethylenediaminetetraacetic acid, proteins or hemoglobin, etc.

In a third set of particular embodiments, a method for analyzing a polymerase chain reaction (PCR) is described, wherein for fitting use is made of a spectrum of the sample before PCR action has taken place and two reference spectra being difference spectra expressing the incorporation of nucleobases, e.g. A&T and G&C nucleotides. Such difference spectra are obtained as the difference of the base pair spectrum of base pairs, e.g. AT respectively GC in a double stranded molecule, on the one hand and the spectrum obtained for separate nucleobases, e.g. nucleotide mixtures of A&T respectively G&C on the other hand.

By way of illustration, embodiments of the present invention not being limited thereby, simulation results are provided, illustrating how the above described deconvolution techniques can advantageously be used for analyzing samples comprising DNA.

A first set of simulation results illustrates that for long double stranded DNA, all spectral differences can be explained using two spectral components, e.g. an AT spectral component and a GC spectral component. The set of simulation results comprises fitting for double stranded DNA molecules with sizes ranging from 10 base pairs to 500 base pairs, with increments of 10 base pairs. The GC content for these particles ranges from 20% to 80% with increments of 10%. For each class of samples, 200 random sequences were generated.

Thereafter a prediction of the absorption spectra was made for all sequences using a method as described in "Predicting ultraviolet spectrum of single stranded and double stranded deoxyribonucleic acids", Biophysical Chemistry 133 (2008) 66-70 by Tataurov et al. The different spectra are shown in FIG. 5 for a selection of different lengths of the double stranded DNA. FIG. 5(a), 5(b), 5(c), 5(d), 5(e), 5(f) illustrate spectra for dsDNA having a length of respectively 10, 20, 50, 100, 200 and 500 base pairs.

For each class, the spread of the 200 spectra in each class was studied and it was identified whether or not a dependency as function of different GC content could be seen. It can be seen that for short double stranded DNA molecules, a large variation is present. No clear dependency of the GC content can be derived for the short double stranded DNA molecules, and it is concluded that the effect of the GC content is obscured by variation of the nearest-neighbour content. On the contrary, for the long double stranded DNA molecules, the variance in spectral shape for the 200 spectra in each class can mainly be explained by variance in GC content. In other words, the variance in spectral shape can substantially be explained by a single variable, being the GC content. By way of illustration, for double stranded DNA with a length of 500 base pairs, the possibility for fitting all these different spectra with one of two reference spectra having a particular GC content (one with high GC content and one with low GC content) is illustrated.

FIG. 6 illustrates for a given GC content (in the present case being 50% of GC content), the variance in spectral shape between different spectra as function of the fragment size of the double stranded DNA. It can be seen that the variance is large for fragment sizes between 1 and 50 base pairs. For fragment sizes of 50 base pairs or larger, e.g. 100 base pairs or larger, the variance in spectral shape is substantially lower, again indicative of the fact that for longer fragment sizes (although randomly generated), the spectral shape of the absorption spectrum is characteristic for a given GC content.

By way of illustration, embodiments of the present invention not being limited thereby, an example of the application of an analysis method for analyzing a PCR measurement according to an embodiment of the present invention is discussed

below with reference to FIG. 7a and FIG. 7b. FIG. 7a illustrates the output for analysis of an absorption spectrum of a sample whereby no double stranded DNA is formed and FIG. 7b illustrates the output for analysis of an absorption spectrum of a sample where double stranded DNA is formed. Although samples where PCR is used typically have a high optical density, the spectra still can be analysed, as shown in FIG. 7a and FIG. 7b.

The deconvolution method used is based on the use of two reference spectra for different GC concentrations (representative for basepairs; referred to as dsDNA_AT or A=T and dsDNA_GT or G≡C), e.g. extreme GC concentrations, and using two reference spectra for respectively A&T and G&C components (being combinations of individual spectra and not being the spectrum of dinucleotides; referred to as dNTPs_AT and dNTPs_GC). By making use of these reference spectra, in the present example, also taking into account other components such as for example a component for turbidity, an appropriate fit can be obtained, and it can indeed be distinguished whether or not double stranded DNA is formed by a PCR reaction. From the fitted concentrations of AT and GC double stranded DNA content, a quantification also can be performed. The measured spectrum 702 and the best fitted spectrum 704 coincide substantially. The dNTPS_AT component 706 and the DNTPS_GC component 708 are also indicated as well as the dsDNA_AT component 710, the dsDNA_GC component 712, a residue component 714 and a component of other known contributing components 716. The last four are nearly not contributing.

By way of illustration, embodiments of the present invention not being limited thereby, another example of the application of an analysis method for analyzing a PCR measurement according to an embodiment of the present invention is discussed below with reference to FIG. 8a and FIG. 8b. FIG. 8a and FIG. 8b illustrate the output for analysis of an absorption spectrum of a sample whereby no double stranded DNA is formed (FIG. 8a) and a sample where double stranded DNA is formed (FIG. 8b). Although samples where PCR is used typically have a high optical density, the spectra still can be analysed, as shown in FIG. 8a and FIG. 8b.

The deconvolution method used is based on the use of a spectrum of the sample before PCR action has taken place (Referred to as PCR mix) and two reference spectra being difference spectra expressing the building in of A&T and G&C. Such difference spectra for AT and GC are obtained as the difference of the spectrum obtained for separate nucleotide mixtures of AT respectively GC and the basepair spectrum of AT respectively GC in a double stranded configuration, i.e. as basepair. The measured spectrum 802 and the best fitted spectrum 804 are shown. For FIG. 8a, the other components and residue are substantially equal to 0. For FIG. 8b, furthermore the AT incorporation component 806, the GC incorporation component 808 and a PCR mix component 810 are also indicated. The other components and residue are substantially equal to 0.

By making use of these reference spectra, in the present example, optionally also taking into account other components such as for example a component for turbidity, an appropriate fit can be obtained, and it can indeed be distinguished whether or not double stranded DNA is formed by a PCR reaction. From the fitted concentrations of AT and GC double stranded DNA content, a quantification also can be performed. It is an advantage of embodiments according to the present invention that the result of the fit is a measure for the formed double stranded DNA and therefore allows a better quantification of the PCR reaction, than in the case of a method providing as end result the combination of the original double stranded DNA and the formed amplicons. It is furthermore an advantage of embodiments according to the present invention that automatically the AT and GC content balance is taken into account.

The above embodiments illustrate that by using deconvolution methods that taking into account the GC spectrum allows proper analysis of double stranded DNA and further even allows analysis of PCR reactions.

By way of illustration, embodiments of the present invention not being limited thereby, another example of the application of an analysis method for analyzing a sample comprising DNA and RNA according to an embodiment of the present invention is discussed below with reference to FIG. 9. FIG. 9 illustrates a set of samples comprising samples with different DNA to RNA ratios. The DNA/RNA ratios

varied from 4:1 to 1:4. It can be seen that the spectral contributions of DNA and RNA significantly overlap. When the set of samples with the predetermined DNA/RNA ratios are deconvoluted using methods according to embodiments of the present invention, DNA/RNA ratios corresponding with the predetermined DNA/RNA ratios are found within experimental error, illustrating features and advantages of methods according to embodiments of the present invention.

Further by way of illustration, the deconvolution of two RNA comprising samples is illustrated in FIG. 10 and 11. FIG. 10 illustrates the deconvolution of a sample comprising both RNA, RNA nucleotides and DNA, whereby the measured spectrum 1002 is indicated, the best fit 1004 is indicated, the RNA component 1006 is indicated, the DNA component 1008 is indicated, an RNA nucleotides component 1010 is indicated, as well as a residue fraction 10012 is indicated. It could be seen that the residue fraction is smaller than 3%, indicative of the fit being accurate. The inset illustrates alternative measurement results based on electrophoretic separation and fluorescence analysis and confirms the presence of RNA (the RIN score of 10 corresponding with high integrity of RNA), as detected using the deconvolution technique according to an embodiment of the present invention. FIG. 11 illustrates the deconvolution of another sample comprising RNA, whereby the measured spectrum 1002 is indicated, the best fit 1004 is indicated, the RNA component 1006 is indicated, a Guanidinium thiocyanate component is also indicated, labeled as Guanidinium component 1014 , and the residue 10012 also is indicated. The fraction of DNA and Protein is negligible. It could be seen that the residue fraction is only 0.83%, being indicative of a sufficiently good fit. The inset illustrates alternative measurement results based on electrophoretic separation and fluorescence analysis and confirms the quality of the sample RNA (the RIN score of 10) corresponding with high integrity of RNA, as detected using the deconvolution technique according to an embodiment of the present invention.

Embodiments of the present invention furthermore may be part of, may correspond with, may comprise one or more steps or features of or may make use of one or more of the following aspects, also being embodiments of the present invention.

By way of illustration, an exemplary flow chart of a method according to an embodiment of the present invention, comprising standard and optional steps is shown in FIG. 1, and will be further described with reference thereto below.

The method 100 according to embodiments of the present invention comprises obtaining 110 prior information for the one or more samples regarding their constituents. Such prior information for the one or more samples regarding their constituents may comprise in some embodiments one or more reference spectra. Such prior information alternatively or in addition thereto also may comprise expected composition information, such as for example expected concentrations, expected ratios between different constituents, etc. One example of such composition information may for example be the presence of a known polynucleotide, whereby the known polynucleotide is a known sequence of nucleotides such that known ratios of constituents, in the present example being nucleotides, are available. In some embodiments the prior information may be obtained based on the samples under study. In some embodiments the prior information may be previously stored information, such as for example previously stored reference spectra for constituents. In some embodiments a combination of such prior information also can be used.

The method also comprises obtaining 120 UV-VIS spectrophotometer data for the one or more samples. Such UV-VIS spectrophotometer data typically may be a spectrophotometer spectrum, although information at one or more individual wavelengths or wavelength ranges also may be used.

The method according to the present invention also comprises defining 130 a number of overlapping components contributing in the UV-VIS spectrophotometer data, the number of overlapping components comprising one or more components assigned to known constituents of the one or more samples and the number of components comprising one or more components that cannot be assigned to known constituents of the one or more samples. The components according to embodiments of the present invention expressing the contribution to the UV-VIS spectrophotometer data

may in some embodiments comprise a spectral contribution with a shape differing from a Gaussian, Lorentzian or a mixture thereof. The shape of the spectral contribution may be a more broad contribution. Such one or more components that cannot be assigned to known constituents of the one or more samples may be as described above, such as for example modified constituents, neighboring effects between two constituents or contaminants. Defining the number of overlapping components may be based on a predetermined algorithm, a neural network, etc. The method may make use of a component analysis for defining the number of overlapping components, such as for example using a principal component analysis technique. In some embodiments the number of independent components can be obtained by putting a large number of absorbance spectra of the sample into a matrix and calculating a rank of the matrix.

The method according to the present invention furthermore comprises, using the prior information for the one or more samples regarding their constituents and using the UV-VIS spectrophotometer data, estimating 140 the constituents composition and the component contributions to the UV-VIS spectrophotometer data for the number of components for the one or more samples by minimizing a residue between the UV-VIS spectrophotometer data and a fit based on said constituent composition and said component contributions. For minimizing a residue different techniques may be employed such as for example a least square method, iterative error minimization, error entropy minimization, weighted error minimization,...

In some embodiments, the prior information comprises reference contributions in the UV-VIS spectrophotometer data, such as for example reference spectra, for components assigned to known constituents. Estimating the constituents composition and the component contributions then may comprise estimating the component contributions to the UV-VIS spectrophotometer data for all components based on the reference contributions. Based on the component contributions, an estimate of the constituent composition is then determined based on the minimization of the residue between the UV-VIS spectrophotometer and the fit based

on the estimated component contributions to the UV-VIS spectrophotometer data for all components.

In some embodiments, the prior information comprises constituent composition information for the one or more samples for the known constituents. Estimating the constituents composition and the component contributions then comprises estimating the constituent composition for all constituents based on the prior information on the constituent composition information for the one or more samples for the known constituents, and determining an estimate for the component contribution to the UV-VIS spectrophotometer data based on minimization of the residue between the UV-VIS spectrophotometer data and the fit based on said estimated constituent composition for all constituents.

In some embodiments a mixture of prior information is obtained, whereby the mixture of information allows for determining an estimate of the constituents composition and the component contributions.

The constituent composition and the component contributions may be interrelated by a model defining weighing factors for component distributions as function of the constituent composition. Such weighing factors may be taken into account when determining an estimate for the constituent composition or the component contribution.

The method also may comprise repeating 150 the estimating step and fitting step for further minimizing the total residue by iteratively applying these steps. Such iteration process may be performed until the remaining residue between the UV-VIS spectrophotometer data and the fit based on the constituent composition is smaller than a predetermined value, or until a maximum number of iteration steps would be reached. The predetermined value referred to may be based on predetermined rules, based on a neural network, based on predetermined algorithms, based on information regarding the one or more samples, etc. In some embodiments, minimization of the residue may only be performed for those samples that have the smallest residue, allowing obtaining more accurate results and/or quicker convergence.

The method allows providing as an output information regarding the composition of the sample and/or information regarding the spectral contribution of constituents to the absorbance spectra. In some examples, the obtained information may allow deriving a reference spectra for a constituent, a modified constituent, a contaminant, a neighbouring effect, ... The latter can be used for setting up a library of different reference spectra compatible with different constituents or effects thereof.

In one embodiments, the method also comprises using the residue of each sample or information related thereto to check the available a priori information. In other words, the results obtained using the method can be used for cross-checking the available a priori information.

The method according to embodiments of the present invention may especially be suitable for analyzing a plurality of samples, whereby in one part of the samples the unknown component is present and in another part of the samples the unknown component is absent.

The method according to embodiments of the present invention may be implemented as software as well as hardware.

By way of illustration, embodiments of the present invention not being limited thereto, an example of how data can be combined and obtained is illustrated in FIG. 2. FIG. 2 illustrates how, according to embodiments of the present invention, based on spectrophotometer data and composition information (e.g. obtained based on prior knowledge), component contributions can be determined, e.g. in the shape of reference spectra by minimizing the residue. Based on the obtained component contributions and using the spectrophotometer data, improved estimates of the composition can be obtained. FIG. 3 illustrates a further embodiment of the present invention, whereby upon estimation of the composition information or the component contribution information, it can be decided to add an additional component or remove a component by evaluating the component contribution data obtained e.g. in the light of the expected results in view of the prior data, or it can be decided to remove a sample in the evaluation based on an evaluation of the

composition information obtained e.g. in the light of the expected results in view of the prior data.

The method may be suitable for extracting quantitative information about one or more constituents in the sample and the corresponding component contribution in the UV-VIS spectrum. In some embodiments, corresponding reference spectra of known and unknown components suspected to the present in the sample, can be used for fitting the UV-VIS spectrophotometer data. The component contributions may not be limited to the reference spectra of the individual pure components, but may be extended with corrective spectral functions of components that do not necessarily correspond to a physic-chemical absorption characteristic of the component, nor does the component itself needs to show significant absorption characteristics within the wavelength region under study. In one example, one of the component contributions may be a contribution corresponding with turbidity caused by concentration or aggregation effects. In one example, one of the component contributions may be a contribution corresponding with scatter caused by residual particles, e.g. particles present due to pre-processing of the sample.

By way of illustration, embodiments of the present invention not being limited thereto, an algorithm comprising a number of particular steps will be described illustrating an exemplary way of implementing the method according to embodiments of the present invention or that can be used in embodiments for fitting in methods for analyzing samples comprising DNA and/or RNA. The method will be described using a particular mathematical matrix formalism, although embodiments of the present invention are not limited thereto another mathematical formalism also may be used. The example illustrates the methods used for deconvolution of spectrophotometric data. The principles apply for different kinds of samples.

For illustrating the algorithm, a number of definitions first are provided :

A_i : this is a measurement (expressed in absorbances, OD10mm) where the subscript i denotes the wavelengths (ranging from 220nm to 360nm in 1nm steps, in case of nucleic acids)

Q_j : this matrix contains the relative presences of the different known and unknown constituents, interaction effects, modifications or contaminations of the mixture. In the present example these can be e.g. the nucleotides A,G,C and T but also modifications (like labels) or dinucleotides such as for example AG, because some dinucleotide spectra differ from the combined spectra of their constituents. In the current algorithm dinucleotides are treated identical to modifications and contaminants. j is the number of subcomponents in the mixture.

R_{ij} : the reference spectra. These are the (known) spectra of the nucleotides and the spectra of the modifications, dinucleotides and contaminants. With i the wavelengths and j the subcomponents.

When working with stored reference spectra, R_j , the coefficients of each reference spectrum, Q_{ij} , can be the unknown.

When measuring known components, Q_{ij} is already specified and the reference spectra R_j can be derived.

In reality, an iterative procedure will be used where in a first step, R_j is determined and in subsequent steps Q_{ij} is found and again and again.

The measurements, relative presences and contributions (in the present example being known reference spectra) are given as

$R_{ij}Q_j = A_i$ or $R\,Q = A$ [1]

The definitions in case a plurality of samples are measured result in the measurement matrix becoming

A_{ik} : this is a measurement (expressed in absorbances, OD10mm) where the subscript i denotes the wavelengths (ranging from 220nm to 360nm in 1nm steps, in case of nucleic acids) and k the number of the measurements.

the coefficient matrix becoming

Q_{jk} : this matrix contains the coefficients (the relative presences) of every subcomponent of the mixture. Here subcomponents can be the nucleotides A,G,C and T but also modifications (like labels) and dinucleotides (e.g. AG) and

contaminants. Again j is the number of subcomponents, and k the number of measurements.

Resulting in the following equation expressing the relation between the components (in the present example known from prior info and being reference spectra R) the coefficient matrix (in the present example expressing the constituent composition) and the measurement matrix.

$$R_{ij}Q_{jk} = A_{ik} \quad \text{or} \quad R\,Q = A \qquad [2]$$

In the present example, the available prior information corresponds with the reference spectra for the constituents in the sample. In such a case both the components contribution information R and the spectrophotometer data A in equation [2] are known. To determine the sample composition Q first left-multiply the equation with the component contribution information R transposed and then left-multiply both sides of the equation with the inverse of the matrix product of the component contribution R transposed and the component contribution R

$$\left(R^T.R\right)^{-1}.R^T.R.Q = \left(R^T.R\right)^{-1}.R^T.A \qquad [3]$$

$$Q = \left(R^T.R\right)^{-1}.R^T.A = pinv(R).A \qquad [4]$$

This is in fact the equation for the *matrix linear regression* or *linear least squares.* The expression in the right side before the spectrophotometer data A is known as the pseudo-inverse of the matrix R and can be for example calculated using singular value decomposition.

By way of illustration, also an example for the inverse situation will de discussed, whereby constituents composition is known and whereby the reference spectra need to be determined.

Thus, the inverse problem where the coefficients of the subcomponents are known, the measurements are known and the aim is to obtain the reference spectra, is discussed below.

By transposing the entire equation [1] the following relation is obtained.

$$Q^T.R^T = A^T \quad [5]$$

and the reference spectra can be found by

$$R^T = pinv(Q^T).A^T \quad [6]$$

In the above two examples, the situation of known constituents is shown.

The above algorithm indicates features and advantages that can be obtained using a method as described in an embodiment of the present invention.

In one aspect, the present invention also relates to a system for analyzing UV-VIS spectrophotometer data of one or more samples comprising DNA and/or RNA, the system comprising an input means for receiving said UV-VIS spectrophotometer data, a processing means programmed for fitting the UV-VIS spectrometer data taking into account at least one spectrum representative for a base pair being any of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil, and deriving from the fitting a quantification of an amount of DNA and/or RNA, and an output means for outputting a quantification of an amount of DNA and/or RNA for the one or more samples under study. Further optional features and advantages may correspond with components for performing the functionality of the steps expressed in the methods described above. Furthermore, features and advantages of further systems or components thereof described below, also being aspects of the present invention, also may be partially or fully incorporated in systems for analyzing UV-VIS spectrophotometer data of one or more samples comprising DNA and/or RNA referred to above. Such further systems may be for example be a system for performing analysis of spectrophotometer data of one or more samples consisting of a number of constituents. The system comprises an input means for obtaining prior information for the one or more samples regarding their constituents and for obtaining UV-VIS spectrophotometer data for the one or more samples. The latter may be an input port for receiving prior information and spectrophotometer data. Alternatively the input means may also comprise a measurement system for recording UV-VIS spectrophotometer data such as for example a spectrophotometer. The system furthermore comprises a processing means programmed for defining a

number of overlapping components contributing in the UV-VIS spectrophotometer data, the number of components comprising one or more components assigned to known constituents of the one or more samples and the number of components comprising one or more components that cannot be assigned to known constituents of the one or more samples. Such a processing means may be for example a CPU although embodiments of the present invention are not limited thereto. The processing means according to embodiments of the present invention furthermore is programmed for using the prior information for the one or more samples regarding their constituents and using the UV-VIS spectrophotometer data, estimating the constituents composition and the component contributions to the UV-VIS spectrophotometer data for the number of components for the one or more samples by minimizing a residue between the UV-VIS spectrophotometer data and a fit based on said constituent composition and said component contributions, thus obtaining information regarding the one or more components that cannot be assigned to known constituents of the one or more samples. Such information may be putted out using an outputting means, such as e.g. a memory, a display, a printer or a plotter.

In another aspect, the present invention relates to a method for extracting component contributions, e.g. reference spectra for particular constituents of the sample of for contributions of such constituents being modified, disturbed or being in a particular environment. The reference spectra can be extracted using methods as described above. Such a method for extracting component contributions may correspond with a method for analyzing UV-VIS spectrophotometer data of a sample comprising DNA and/or RNA as described above.

In one aspect, the present invention also relates to the use of a method as described above for particular applications. One of these applications may be identifying and/or quantifying contaminants in DNA or RNA samples. The contaminatns may be protein contamination. The contaminants in one application may be PCR-inhibiting contaminants. Another application envisaged is the quantification of an amount of

double-stranded DNA in a mixture of double-stranded DNA and single-stranded RNA or DNA. Still a further application is determination of the composition of a polynucleotide or protein. A further application is the quantification of the modification efficiency in a sample. The method also may be applied for validating a sample composition or a constituent thereof or group of constituents thereof. By extension the method allows to extract more detailed information about the composition of a component, in particular of polymers, and more specifically in biopolymers such as polynucleotides and proteins. Inversely, the method also allows to extract quantitative information of polymers based on the a priori knowledge of the composition of the known components, rather than the knowledge of the spectrum of the component.

In still another aspect, embodiments of the present invention also relate to computer-implemented methods for performing at least part of the methods for analyzing spectrophotometer data. Embodiments of the present invention also relate to corresponding computer program products. The methods may be implemented in a computing system. They may be implemented as software, as hardware or as a combination thereof. Such methods may be adapted for being performed on computer in an automated and/or automatic way. In case of implementation or partly implementation as software, such software may be adapted to run on suitable computer or computer platform, based on one or more processors. The software may be adapted for use with any suitable operating system such as for example a Windows operating system or Linux operating system. The computing means may comprise a processing means or processor for processing data. According to some embodiments, the processing means or processor may be adapted for analyzing spectra according to any of the methods as described above or extracting reference spectra according to any of the methods as described above. Besides a processor, the computing system furthermore may comprise a memory system including for example ROM or RAM, an output system such as for example a CD-rom or DVD drive or means for outputting information over a network. Conventional computer

components such as for example a keyboard, display, pointing device, input and output ports, etc also may be included. Data transport may be provided based on data busses. The memory of the computing system may comprise a set of instructions, which, when implemented on the computing system, result in implementation of part or all of the standard steps of the methods as set out above and optionally of the optional steps as set out above. The obtained results may be outputted through an output means such as for example a plotter, printer, display or as output data in electronic format.

Further aspect of embodiments of the present invention encompass computer program products embodied in a carrier medium carrying machine readable code for execution on a computing device, the computer program products as such as well as the data carrier such as dvd or cd-rom or memory device. Aspects of embodiments furthermore encompass the transmitting of a computer program product over a network, such as for example a local network or a wide area network, as well as the transmission signals corresponding therewith.

Claims

1. A computer-implemented method for analyzing UV-VIS spectrophotometer data of a sample comprising DNA and/or RNA, the method comprising
   - receiving UV-VIS spectrophotometer data,
   - fitting the UV-VIS spectrometer data taking into account at least one spectrum representative for a sample having a base pair content, the base pair being any of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil (AU), and
   - deriving from the fitting a quantification of an amount of DNA and/or RNA.
2. A computer-implemented method according to claim 1, wherein said fitting comprises taking into account a set of reference spectra representative for two distinct base pair contents.
3. A computer-implemented method according to claim 2, wherein said fitting comprises taking into account a set of two reference spectra representative of two distinct GC contents.
4. A computer-implemented method according to claim 3, wherein said fitting comprises taking into account a set of exactly two reference spectra representative for two distinct GC contents.
5. A computer-implemented method according to any of the previous claims, wherein said fitting comprises fitting UV-VIS spectrophotometer data of a sample comprising DNA and/or RNA having at least 50 base pairs, advantageously at least 100 base pairs.
6. A computer-implemented method according to any of the previous claims, the sample comprising DNA,
   wherein said fitting comprises using a combination of a set of reference spectra representative for base pairs and reference spectra being representative for separate nucleobases.
7. A computer-implemented method according to claim 6, wherein said fitting using a combination of reference spectra representative for base pairs and reference spectra being representative for separate nucleobases comprises fitting using

spectra being the difference spectra between reference spectra representative for base pairs and reference spectra being representative for separate nucleobases.

8. A computer-implemented method according to claim 7, wherein the difference spectra between reference spectra representative for base pairs and reference spectra being representative for separate nucleobases are spectra obtained by subtracting from a given reference spectra representative for a given base pair, the reference spectra being representative for the corresponding separate nucleobases.

9. A computer-implemented method according to any of claims 6 to 8, wherein the reference spectra representative of separate nucleobases are reference spectra for adenine, guanine, thymine, cytosine and/or uracil.

10. A computer-implemented method according to any of the previous claims, the method being adapted for analyzing a PCR reaction of a sample comprising DNA, wherein said deriving comprises determining an amount of double stranded DNA formed during or after a PCR reaction.

11. A computer-implemented method according to claim 10, wherein fitting furthermore comprises taking into account a spectral component for a contamination component.

12. A computer-implemented method according to claim 11, wherein taking into account a spectral component for a contamination component comprises taking into account a spectral component for one or more of trimethylglycine, guanidinethiocyanate, phenol, ethylenediaminetetraacetic acid, proteins sodium azide or hemoglobine..

13. A computer-implemented method according to any of claims 10 to 12, wherein fitting UV-VIS spectrophotometer data of a sample comprises fitting UV-VIS spectrophotometer data of a mixture comprising deoxynucleotide Triphosphates, primer-oligo's, polymerase and DNA.

14. A computer-implemented method according to any of the previous claims, wherein fitting furthermore comprises taking into account one or more spectral components representative for scattering in the sample.
15. A computer-implemented method according to any of the previous claims, wherein fitting furthermore comprises taking into account one or more spectral components representative for turbidity in the sample.
16. A computer-implemented method according to any of the previous claims, wherein the DNA and/or RNA is at least an amount of double stranded DNA.
17. A computer-implemented system for analyzing UV-VIS spectrophotometer data of a sample comprising DNA and/or RNA, the system comprising
    - an input means for receiving said UV-VIS spectrophotometer data,
    - a processing means programmed for fitting the UV-VIS spectrometer data taking into account at least one spectrum representative for a base pair being any of more of adenine-thymine (AT) or guanine-cytosine (GC) or adenine-uracil, and deriving from the fitting a quantification of an amount of DNA and/or RNA, and
    - an output means for outputting a quantification of an amount of DNA and/or RNA for the sample under study.
18. A computer-implemented system according to claim 17, wherein said processing means is adapted for performing fitting as described in any of the methods of claims 1 to 16.
19. Use of a method according to any of claims 1 to 16 for quantifying an amount of DNA and/or RNA.
20. A computer program product for, if implemented on a processing unit, performing a method according to any of claims 1 to 16.
21. A data carrier medium comprising a computer program product according to claim 20.
22. Transmission of a computer program product according to claim 20 over a wide or local area network.

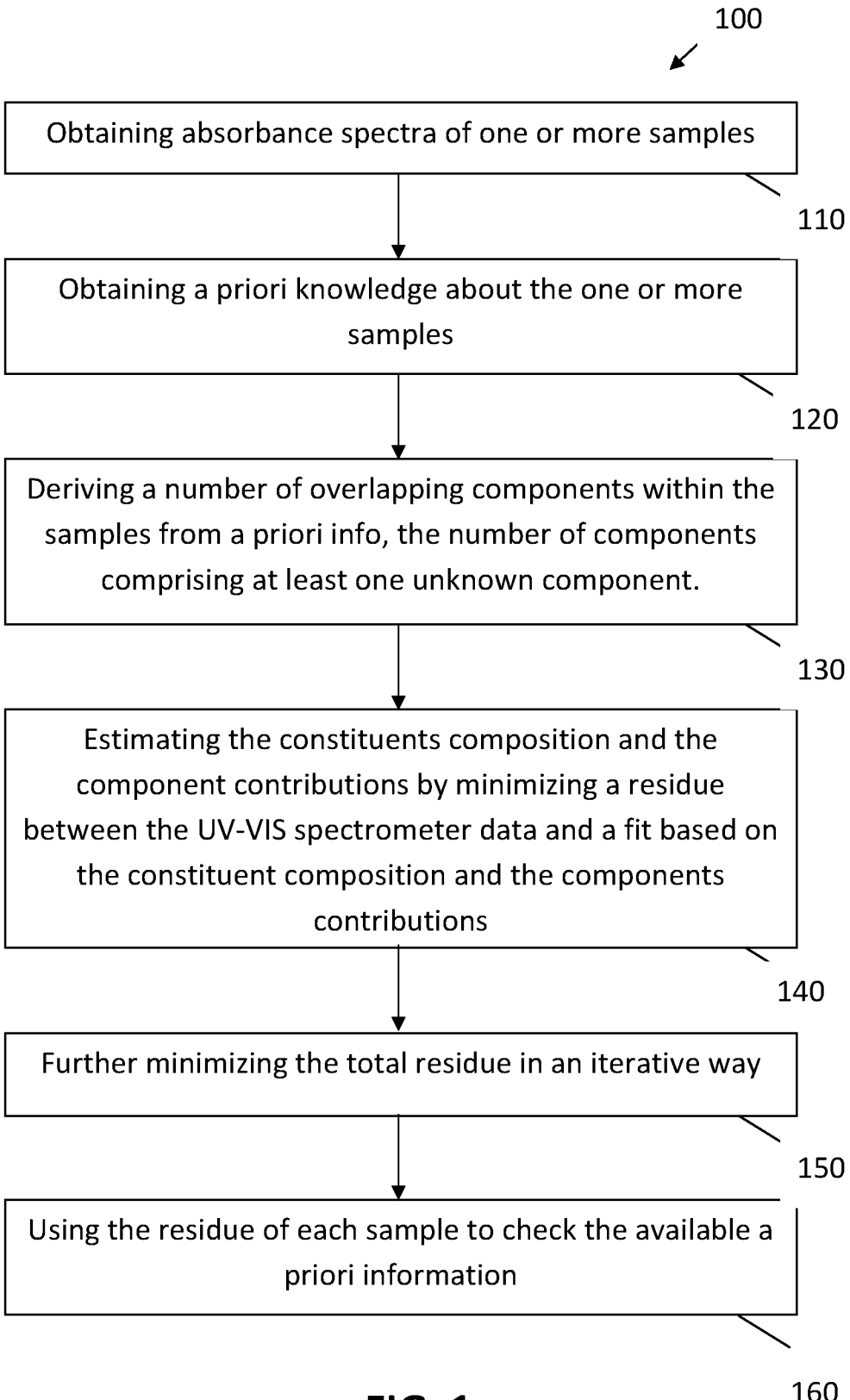

100
Obtaining absorbance spectra of one or more samples
110
Obtaining a priori knowledge about the one or more samples
120
Deriving a number of overlapping components within the samples from a priori info, the number of components comprising at least one unknown component.
130
Estimating the constituents composition and the component contributions by minimizing a residue between the UV-VIS spectrometer data and a fit based on the constituent composition and the components contributions
140
Further minimizing the total residue in an iterative way
150
Using the residue of each sample to check the available a priori information
160

FIG. 1

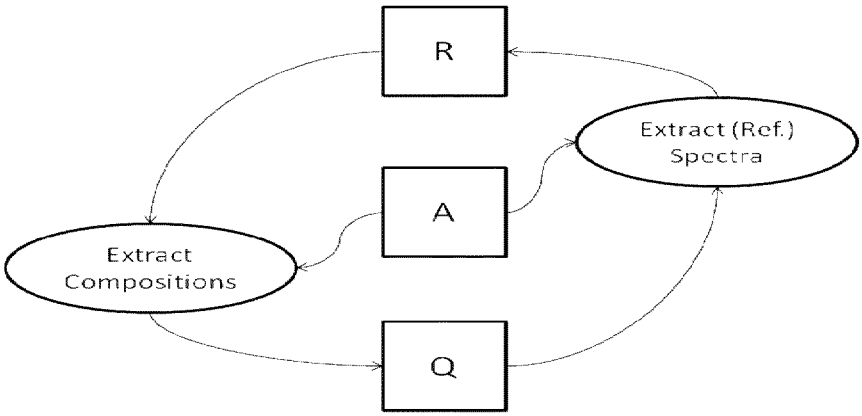

R
Extract (Ref.) Spectra
A
Extract Compositions
Q

FIG. 2

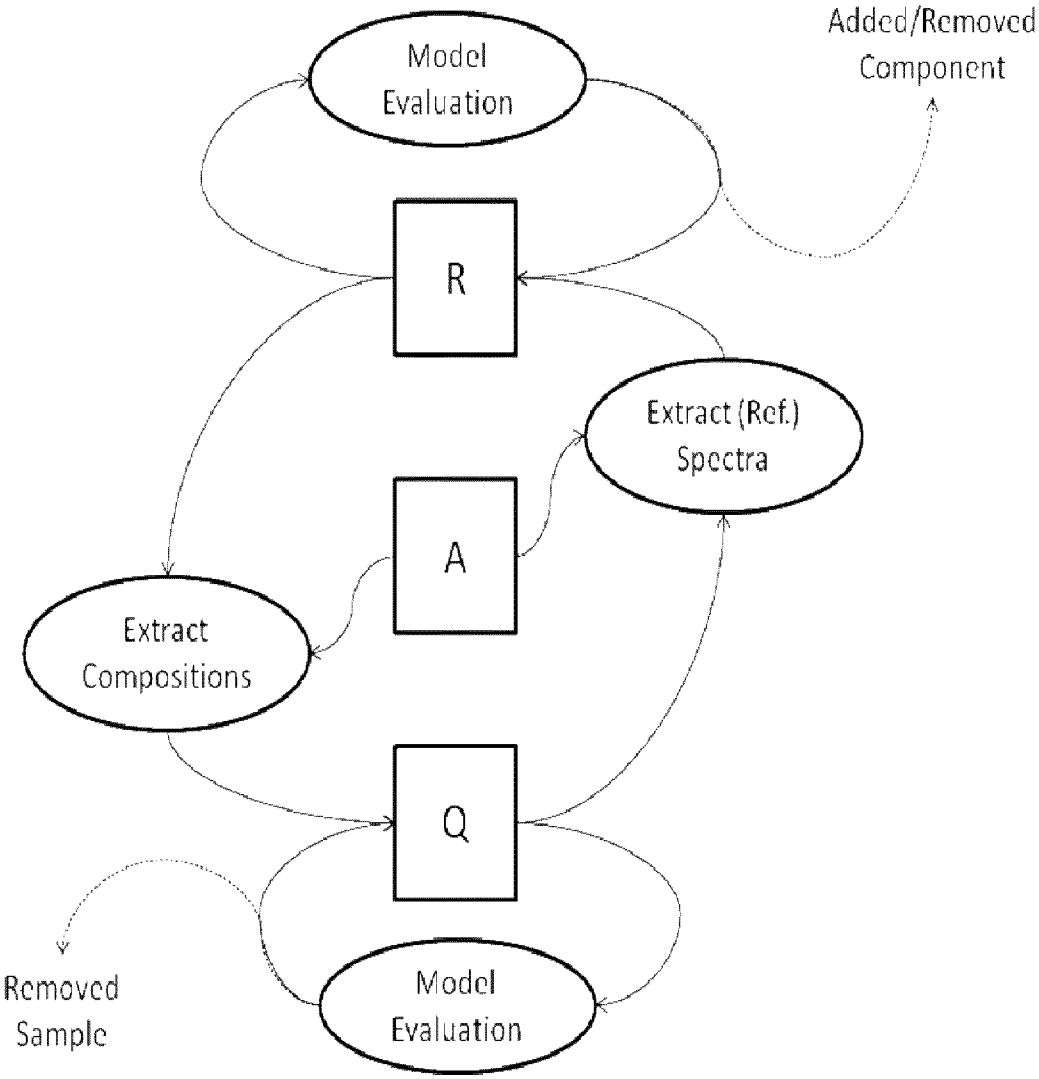

Added/Removed Component
Model Evaluation
R
Extract (Ref.) Spectra
A
Extract Compositions
Q
Removed Sample
Model Evaluation

FIG. 3

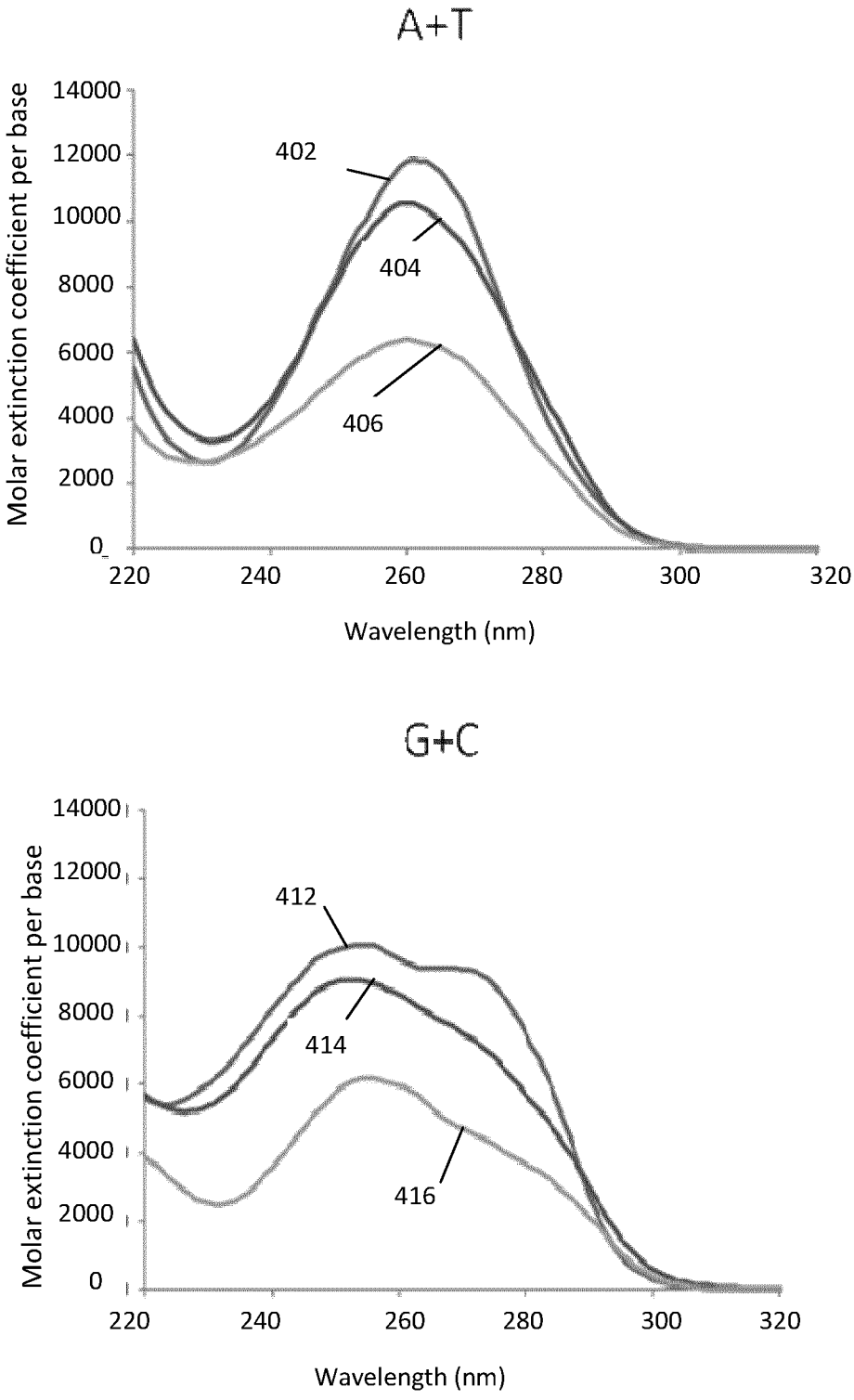

A+T
Molar extinction coefficient per base
14000
12000
10000
8000
6000
4000
2000
0
220
240
260
280
300
320
Wavelength (nm)
402
404
406
G+C
Molar extinction coefficient per base
14000
12000
10000
8000
6000
4000
2000
0
220
240
260
280
300
320
Wavelength (nm)
412
414
416

FIG. 4

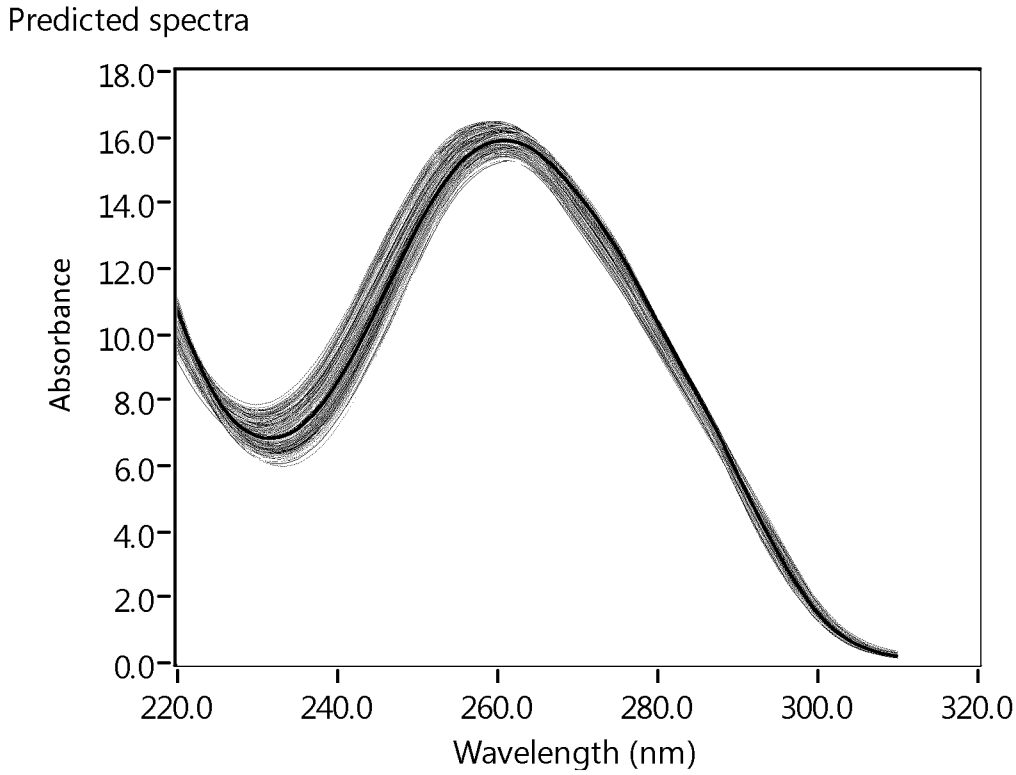

Predicted spectra
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Absorbance
220.0
240.0
260.0
280.0
300.0
320.0
Wavelength (nm)

FIG. 5a

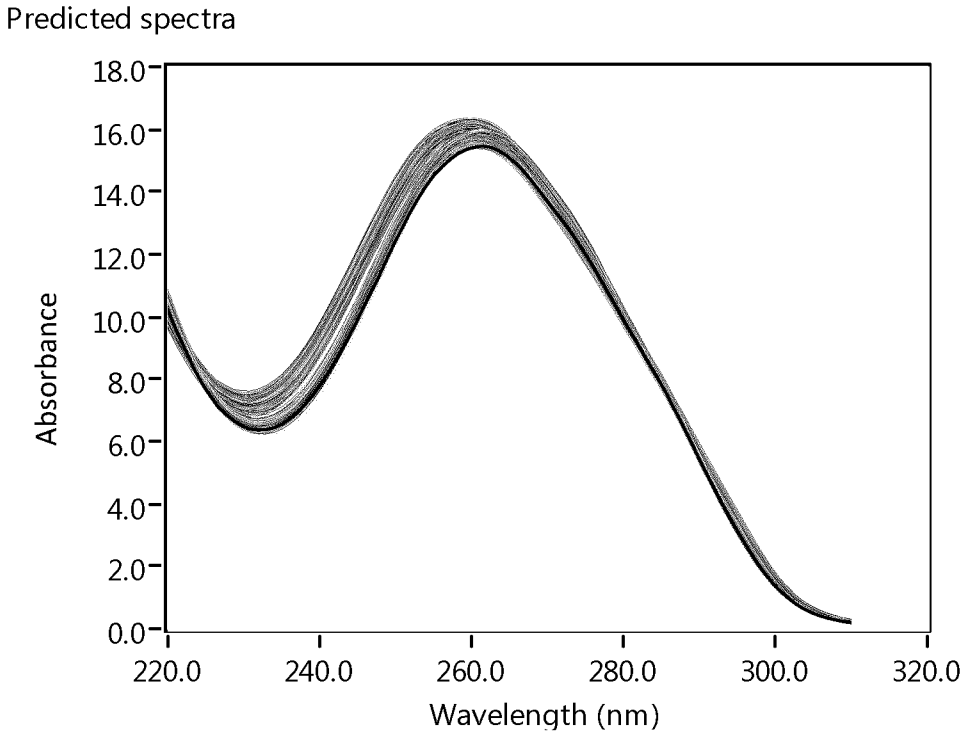

Predicted spectra
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Absorbance
220.0
240.0
260.0
280.0
300.0
320.0
Wavelength (nm)

FIG. 5b

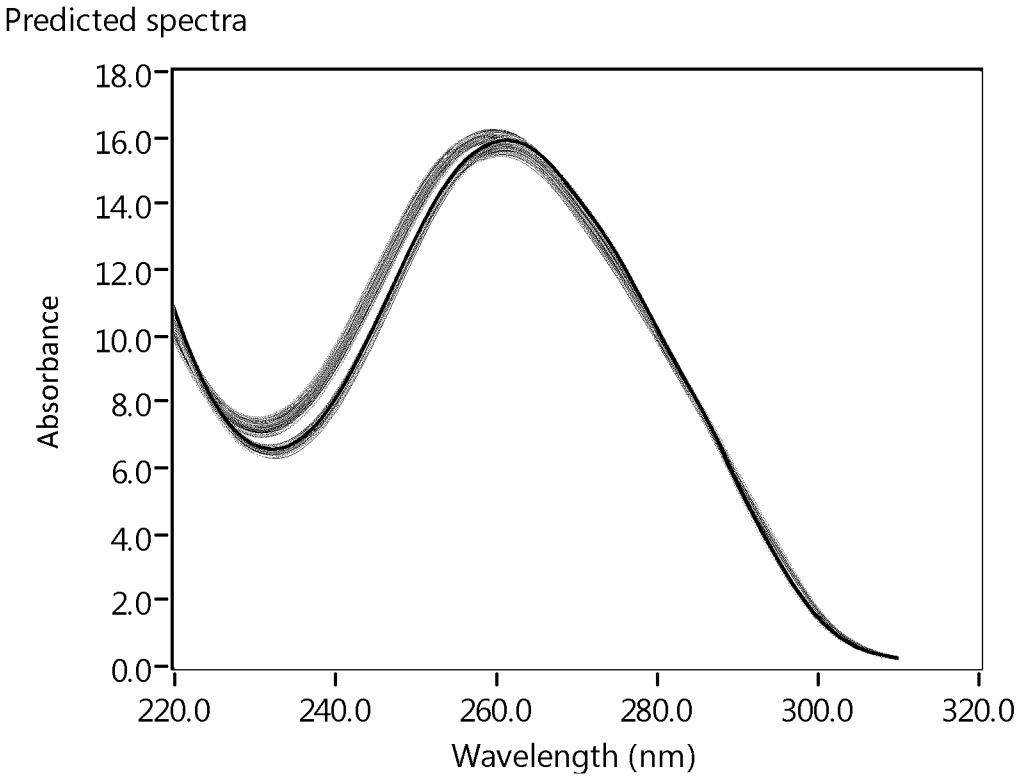

Predicted spectra
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Absorbance
220.0
240.0
260.0
280.0
300.0
320.0
Wavelength (nm)

FIG. 5c

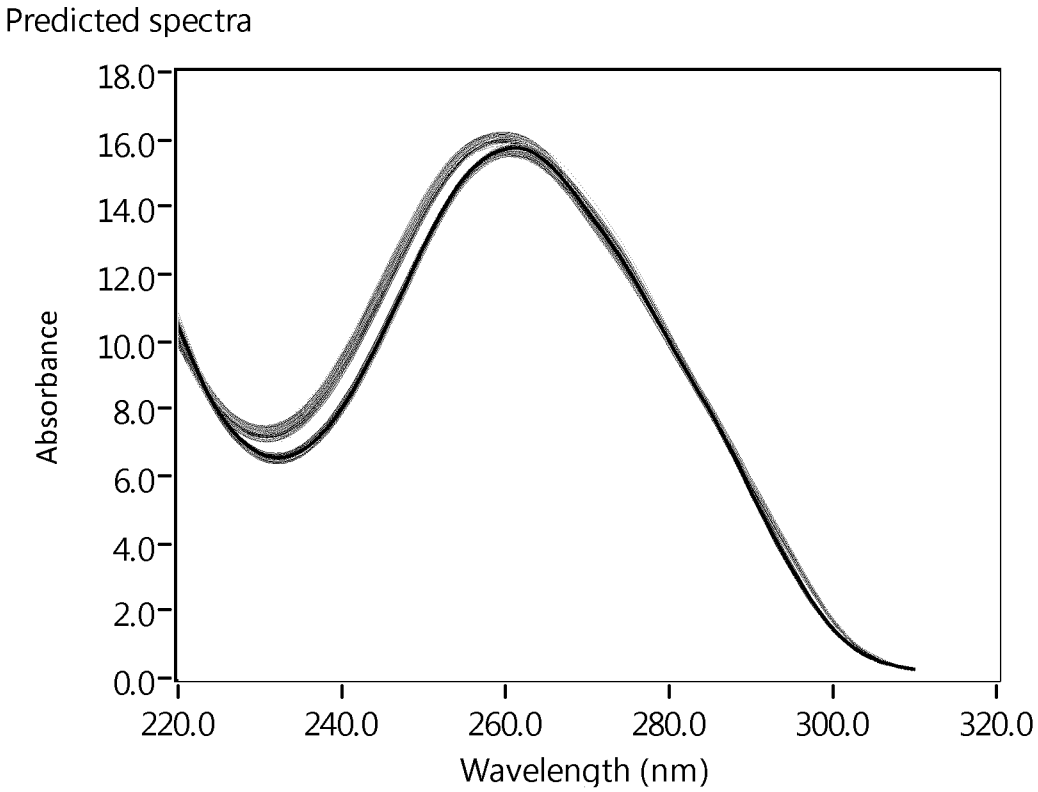

Predicted spectra
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Absorbance
220.0
240.0
260.0
280.0
300.0
320.0
Wavelength (nm)

FIG. 5d

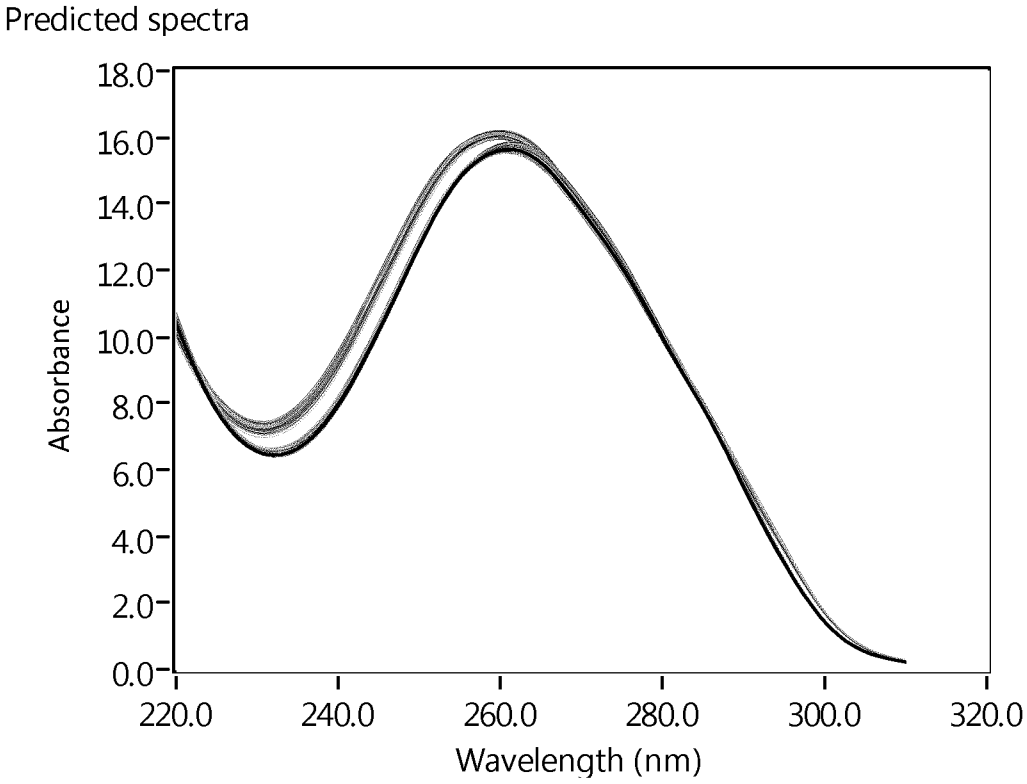

Predicted spectra
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Absorbance
220.0
240.0
260.0
280.0
300.0
320.0
Wavelength (nm)

FIG. 5e

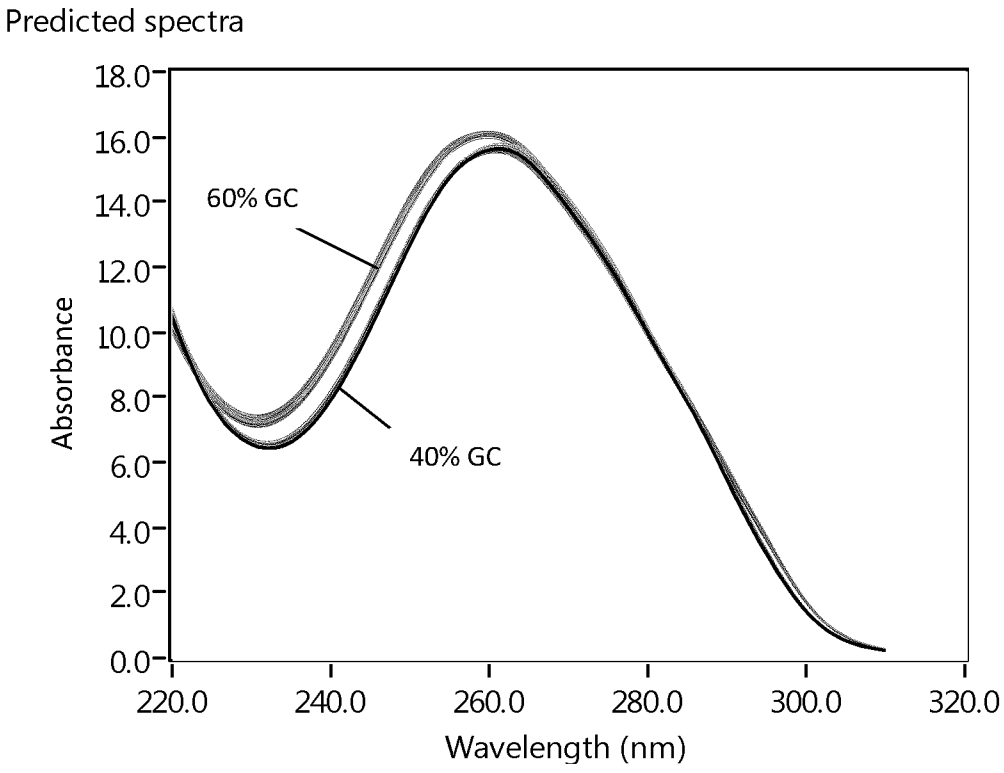

Predicted spectra
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
Absorbance
60% GC
40% GC
220.0
240.0
260.0
280.0
300.0
320.0
Wavelength (nm)

FIG. 5f

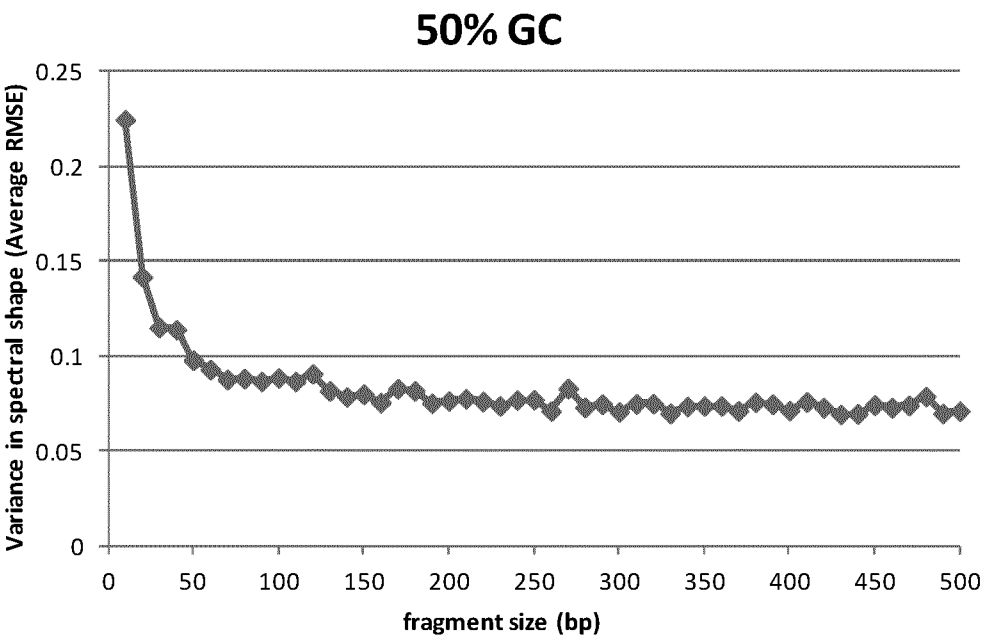

50% GC
Variance in spectral shape (Average RMSE)
0.25
0.2
0.15
0.1
0.05
0
0
50
100
150
200
250
300
350
400
450
500
fragment size (bp)

FIG. 6

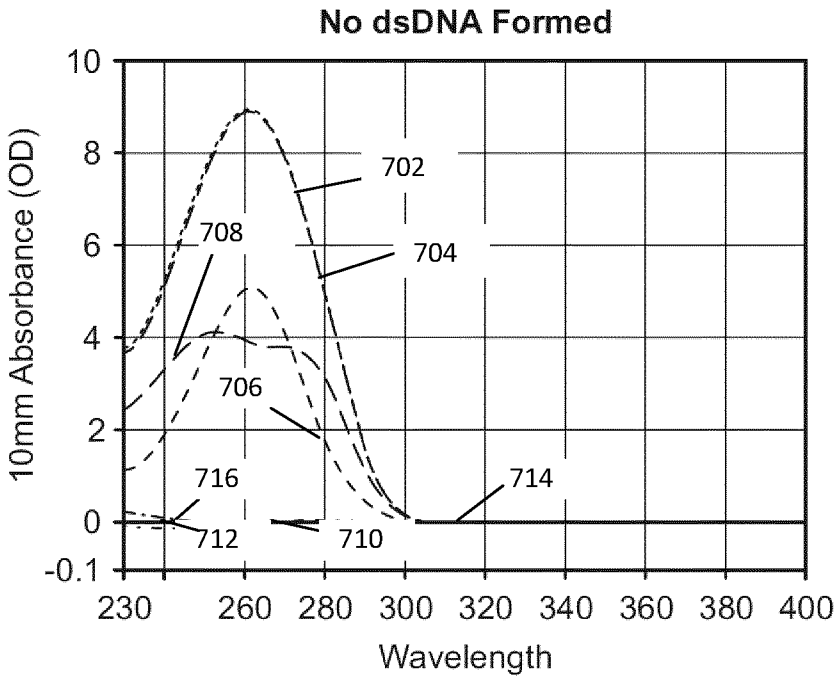

No dsDNA Formed
10mm Absorbance (OD)
10
8
6
4
2
0
-0.1
230
260
280
300
320
340
360
380
400
Wavelength
702
704
706
708
710
712
714
716

FIG. 7a

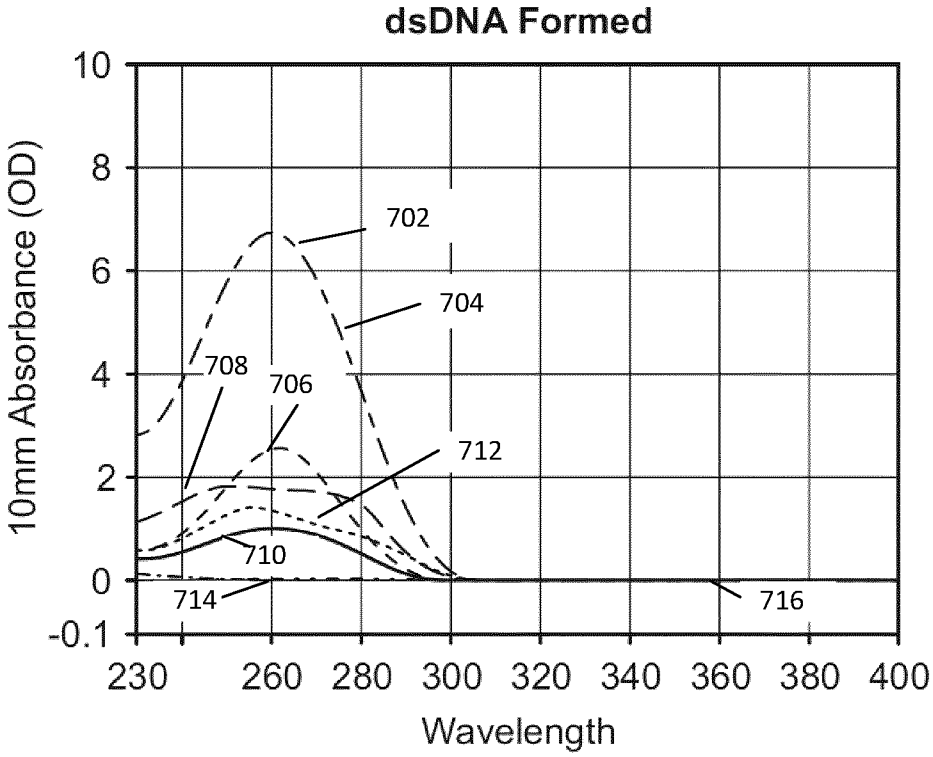

dsDNA Formed
10mm Absorbance (OD)
10
8
6
4
2
0
-0.1
230
260
280
300
320
340
360
380
400
Wavelength
702
704
708
706
712
710
714
716

FIG. 7b

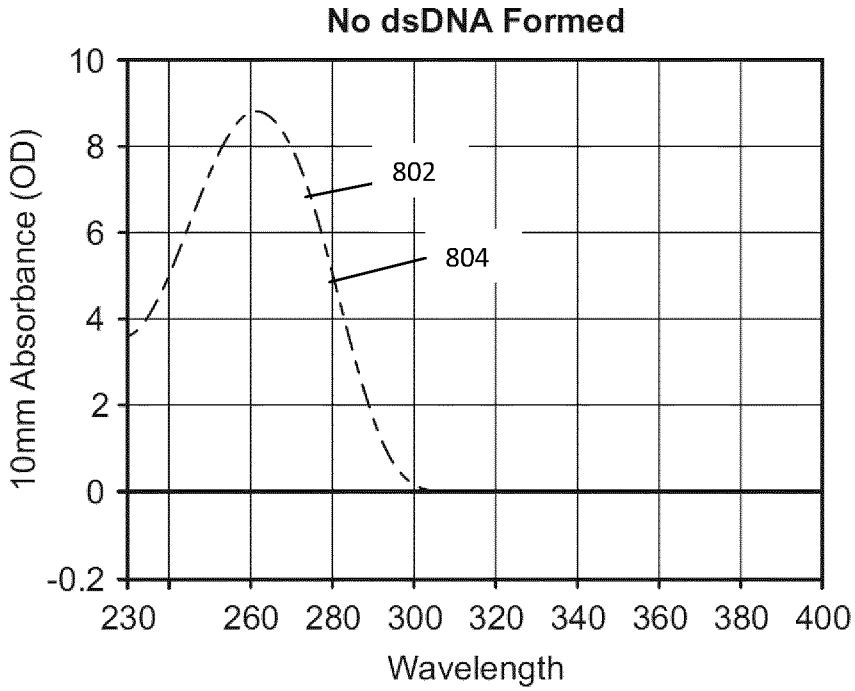

No dsDNA Formed
10mm Absorbance (OD)
10
8
6
4
2
0
-0.2
230
260
280
300
320
340
360
380
400
Wavelength
802
804

FIG. 8a

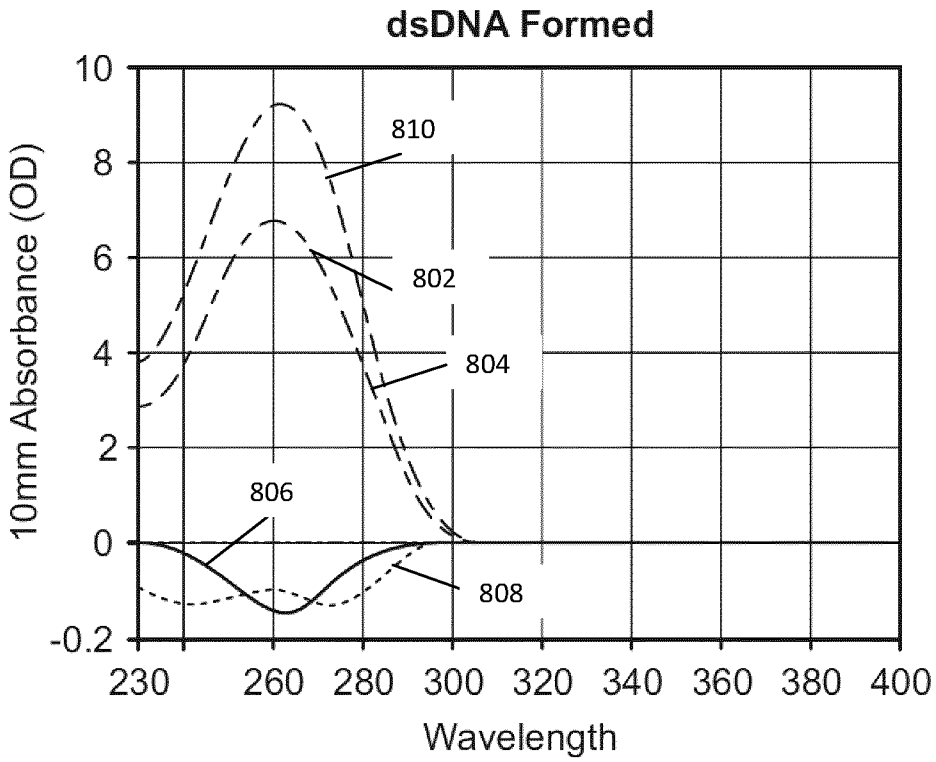

dsDNA Formed
10mm Absorbance (OD)
Wavelength
810
802
804
806
808
10
8
6
4
2
0
-0.2
230
260
280
300
320
340
360
380
400

FIG. 8b

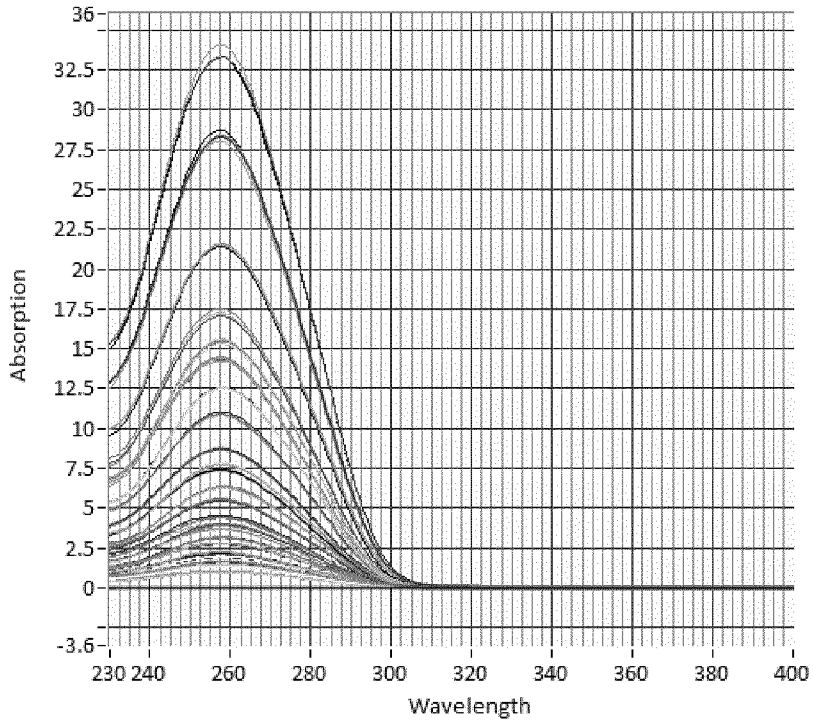

Absorption
Wavelength
36
32.5
30
27.5
25
22.5
20
17.5
15
12.5
10
7.5
5
2.5
0
-3.6
230 240
260
280
300
320
340
360
380
400

FIG. 9

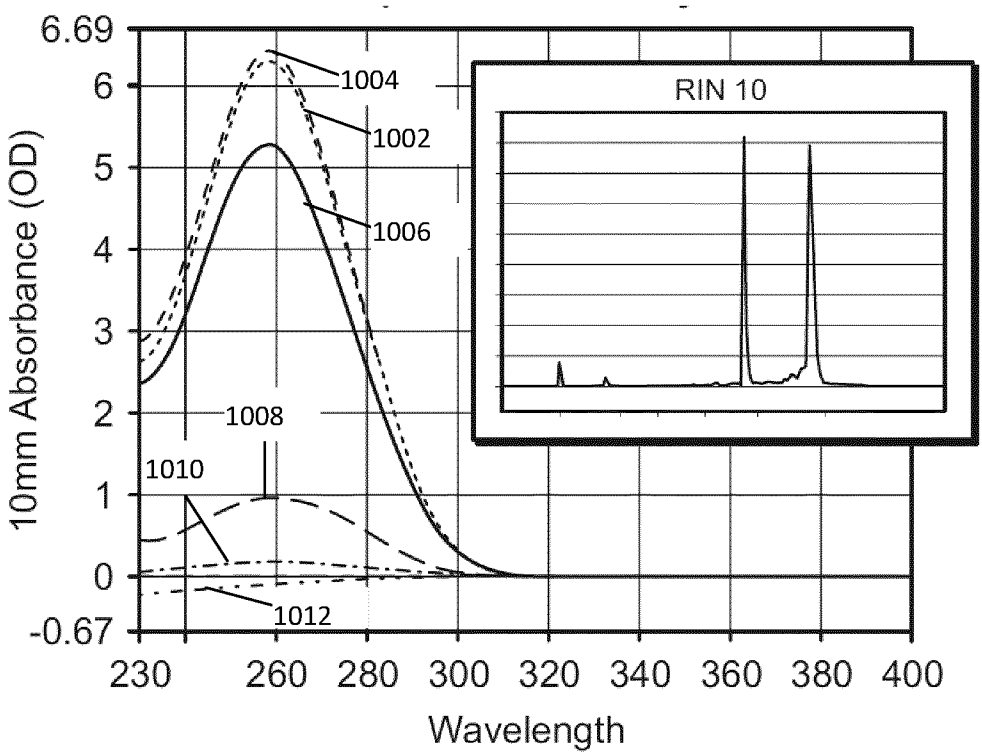

6.69
6
5
4
3
2
1
0
-0.67
10mm Absorbance (OD)
230
260
280
300
320
340
360
380
400
Wavelength
1004
1002
1006
1008
1010
1012
RIN 10

FIG. 10

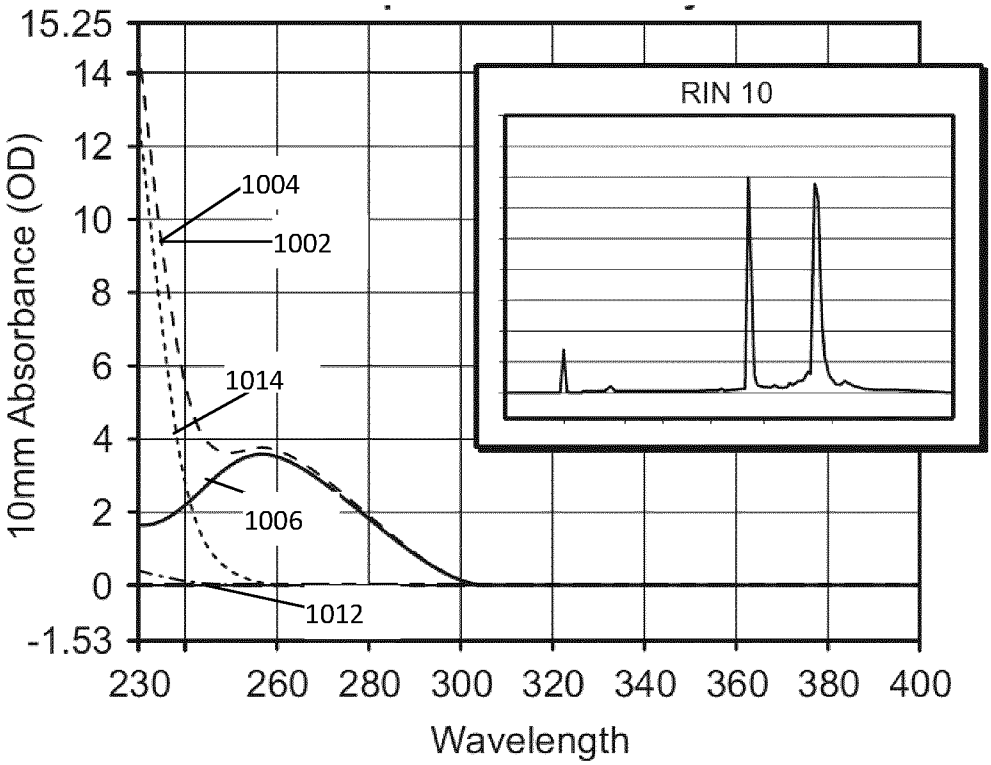

15.25
14
12
10
8
6
4
2
0
-1.53
10mm Absorbance (OD)
230
260
280
300
320
340
360
380
400
Wavelength
1004
1002
1014
1006
1012
RIN 10

FIG. 11

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2012/053481

A. CLASSIFICATION OF SUBJECT MATTER
INV. G01N21/33
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data, INSPEC, COMPENDEX, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 7 786 295 B2 (HEATH ELLEN M [US] ET AL) 31 August 2010 (2010-08-31) column 10, line 64 - column 11, line 11	1-22
A	EP 1 524 514 A1 (ZEON CORP [JP]) 20 April 2005 (2005-04-20) paragraph [0045]	1-22

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 June 2012	28/06/2012

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Hoogen, Ricarda

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2012/053481

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 7786295 B2	31-08-2010	NONE	
EP 1524514 A1	20-04-2005	CN 1662807 A	31-08-2005
		EP 1524514 A1	20-04-2005
		US 2006096884 A1	11-05-2006
		WO 2004001397 A1	31-12-2003