Methods for Automatic Nucleotide-Sequence Analysis

MULTICOMPONENT SPECTROPHOTOMETRIC ANALYSIS OF MIXTURES OF NUCLEIC ACID COMPONENTS BY A LEAST-SQUARES PROCEDURE

By SHEILA LEE, D. McMULLEN, G. L. BROWN

Medical Research Council Biophysics Research Unit, Biophysics Department

AND A. R. STOKES

Physics Department, King's College, London, W.C. 2

(Received 13 May 1964)

1. A theoretical analysis of the errors in multicomponent spectrophotometric analysis of nucleoside mixtures, by a least-squares procedure, has been made to obtain an expression for the error coefficient, relating the error in calculated concentration to the error in extinction measurements. 2. The error coefficients, which depend only on the 'library' of spectra used to fit the experimental curves, have been computed for a number of 'libraries' containing the following nucleosides found in s-RNA: adenosine, guanosine, cytidine, uridine, 5-ribosyluracil, 7-methylguanosine, 6-dimethylaminopurine riboside, 6-methylaminopurine riboside and thymine riboside. 3. The error coefficients have been used to determine the best conditions for maximum accuracy in the determination of the compositions of nucleoside mixtures. 4. Experimental determinations of the compositions of nucleoside mixtures have been made and the errors found to be consistent with those predicted by the theoretical analysis. 5. It has been demonstrated that, with certain precautions, the multicomponent spectrophotometric method described is suitable as a basis for automatic nucleotide-composition analysis of oligonucleotides containing nine nucleotides. Used in conjunction with continuous chromatography and flow chemical techniques, this method can be applied to the study of the sequence of s-RNA.

The sequential order of nucleotides in ribonucleic acids is of considerable interest with respect to the mechanism of protein synthesis and the nature of the genetic code. The determination of this order, however, is a formidable problem because of the large number of different components. Viral RNA and messenger RNA molecules consist of chains several thousand nucleotides long, composed of four different nucleotides, adenylic acid, guanylic acid, cytidylic acid and uridylic acid. Transfer RNA (s-RNA) is only about 80 nucleotides long, but contains a number of minor nucleotide components in addition to the four main nucleotides. Recent fractionation studies by Apgar, Holley & Merrill (1962) and by Goldstein, Bennett & Craig (1964) indicate that there may be more than 40 different amino acid-specific s-RNA molecules in s-RNA preparations from yeast and *Escherichia coli*. The development of rapid automatic methods for nucleotide-sequence analysis applicable to small amounts of material is clearly desirable, as the

present methods using paper electrophoresis and chromatography are unduly time-consuming and laborious for extensive studies of sequences in different s-RNA and viral RNA compounds [see reviews by Smith (1955), Dekker (1960) and Brown (1963)].

Each nucleoside component of the nucleic acids has a distinct absorption spectrum in the ultraviolet by which the nucleosides may be identified and their concentrations measured. Theoretically, a mixture of nucleosides can be analysed from its absorption spectrum, without separation of the individual components, if the spectra of the pure components are known and are sufficiently different from each other. The absorption spectrum of nucleoside mixtures can be readily determined in a flowing stream; therefore spectrophotometric methods are particularly suitable as a basis for automatic procedures and can be combined with continuous chemical methods (Skeggs, 1957) to give the fundamental determinations of composition and end group. An automatic

system of this type has been developed by the authors for the analysis of nucleic acids and briefly described (Brown, Lee & McMullen, 1963; Brown, Lee, McMullen & Stokes, 1964). The purpose of this paper is to present the details of a spectrophotometric method used by us and to demonstrate the validity of multicomponent spectrophotometric analysis under practicable conditions for the analysis of mixtures of nucleosides, especially those found in s-RNA compounds.

In principle, it is possible to determine the concentrations of n components in a mixture from the extinction measured at n wavelengths if the components have different absorption spectra. In practice, as n becomes large, the results are increasingly affected by experimental error when the spectra of the different components overlap. Sternberg, Stillo & Schwendeman (1960) first applied curve-fitting by the least-squares method to increase the accuracy of spectrophotometric analysis of multicomponent systems of this nature, and their method was later used by Reid & Pratt (1960) and Vasilenko, Kamzolova & Knorre (1962) for the analysis of mixtures of the four main components of ribonucleic acids. However, Zscheile, Murray, Baker & Peddicord (1962) demonstrated that when the spectra of some of the components are so similar that there is a large degree of linear relationship between them, i.e. one curve can be made up from the sum of the other curves in non-zero proportions, the system of simultaneous equations becomes unstable and the precision of the calculated concentrations is low. With nucleotide mixtures containing adenylic acid, cytidylic acid, uridylic acid and guanylic acid in hydrochloric acid at pH 1.1, Zscheile *et al.* (1962) found that poor results were obtained owing to the large degree of linear relationship between the spectra of adenylic acid and uridylic acid. They calculated the correlations between pairs of spectra for different wavebands and found that there was an approximate relationship between these correlations and the precision of the calculations for the nucleotide concentrations.

To investigate the probable error in the calculated concentrations for a larger number of components it is essential to be able to calculate the effect of linear relationships between the spectra of groups of components on the precision of the results. In the present paper this problem has been investigated theoretically and experimentally for mixtures including a number of the minor components of s-RNA and the conditions for satisfactory analyses have been established. The methods have been developed by us for the specific purpose of combining them with continuous chemical procedures to form a system for the automatic determination of nucleotide sequences.

MATERIALS AND METHODS

Chemicals. The ribonucleosides of adenine, guanine, cytosine and uracil were obtained from the California Corp. for Biochemical Research, Los Angeles, U.S.A. Samples of thymine riboside, 6-dimethylaminopurine riboside, 6-methylaminopurine riboside and 7-methylguanosine were a gift from Dr C. B. Reese. The pseudouridine, which was the gift of Dr W. E. Cohn, was further purified by chromatography on Whatman no. 1 paper in butan-1-ol-water (21:4, v/v) according to Michelson & Cohn (1962).

Spectrophotometry. The ultraviolet-absorption spectra of solutions were obtained with a modified Perkin-Elmer UV137 spectrophotometer by using matched stoppered 1 cm. silica cells. The spectrophotometer was fitted with a punched paper tape digital readout. A readout potentiometer, attached to the shaft of the optical wedge, gave 10 mv for an extinction 1.5. The voltage was measured by a Digital Measurement (Farnborough, Hants) voltmeter (type 2020). The output from the digital voltmeter was fed into a Creed Paper Tape Punch, no. 25, through a Punch Drive Unit (Digital Measurements, no. 5021) to give extinction coefficients to three significant figures. The wavelength drum of the spectrophotometer was fitted with a commutator which actuated the punch at 2 μ intervals throughout the wavelength scan. The commutator was manufactured to an accuracy of better than 0.4% and was aligned relative to the wavelength drum at 260 μ to $\pm 0.1 \mu$. The optical and digital systems were adjusted by using manufacturers' extinction standards to give extinction coefficients to an absolute accuracy of better than ± 0.008 extinction unit and to a reproducibility of ± 0.002 extinction unit. The spectrophotometer was also fitted with a repetitive scan controller, enabling spectra to be repeatedly taken at regular intervals over selected wavelength ranges.

Ribonucleosides were dissolved in 0.01 N-HCl and in carbon dioxide-free 0.01 N-NaOH, adjusted to pH 12.0 with N-NaOH, at concentrations of approximately 0.1 mM. Test mixtures of ribonucleosides were made up by mixing, in simple ratios, standard volumes of stock solutions, using pipettes with a tolerance of 1% or less.

Molar extinction coefficients were taken from the literature, and the following values were used in the results presented below. Nucleosides in 0.01 N-HCl: adenosine, 14.6×10^3 at 257 μ (Beaven, Holiday & Johnson, 1955); uridine, 10.1×10^3 at 262 μ (Fox & Shugar, 1952); cytidine, 13.4×10^3 at 280 μ (Fox & Shugar, 1952); guanosine, 12.2×10^3 at 256.5 μ (Beaven *et al.* 1955); pseudouridine, 7.9×10^3 at 262 μ (Shapiro & Chambers, 1961). Nucleosides at pH 12.0: adenosine, 14.9×10^3 at 259.5 μ (Beaven *et al.* 1955); uridine, 7.45×10^3 at 262 μ (Fox & Shugar, 1952); guanosine, 11.3×10^3 at 264 μ (Beaven *et al.* 1955); cytidine, 9.1×10^3 at 271 μ (Fox & Shugar, 1952); pseudouridine, 7.8×10^3 at 286 μ (Michelson & Cohn, 1962); thymine riboside, 11.8×10^3 at 268 μ (Littlefield & Dunn, 1958); 6-methylaminopurine riboside, 19.7×10^3 at 266 μ (Littlefield & Dunn, 1958); 6-dimethylaminopurine riboside, 18.6×10^3 at 276 μ (Littlefield & Dunn, 1958); 7-methylguanosine is unstable at pH 12.0 and the extinction coefficient of the degradation product was taken to be 12.3×10^3 at 265.5 μ , as found from the absorption spectrum of the product at pH 12.0 relative to the spectrum of 7-methylguanosine at pH 8.7, and from the molar extinction coefficient of 7-methylguanosine at pH 8.7 (Lawley & Brookes, 1963).

THEORY

Let the absorption spectrum of a nucleoside component be sampled at r wavelengths, to give a set of r molecular extinctions, ϵ_i , ($i = 1, r$): For every component that might be in a mixture there will be such a set of extinctions, and together they will form a defining array for the mixture, which is called the 'library'. If there are n such components then the library can be written as an $r \times n$ array ϵ_{ij} ($i = 1, r$; $j = 1, n$).

If we choose the conditions such that there are no interactions between the components, then the absorption spectra, E_i , of the mixture will be theoretically a linear superposition of the library components' spectra, the constants being the concentrations C_j ($j = 1, n$) of the components.

$$C_1 \epsilon_{i1} + C_2 \epsilon_{i2} + C_3 \epsilon_{i3} + \dots C_n \epsilon_{in} = E'_i \quad (1)$$

In practice the calculated extinction E'_i is not exactly equal to the measured spectrum of the mixture, as it is subject to random experimental errors. By the method of least-squares values of the concentrations, C_j , are calculated so as to minimize the divergence S between the observed and the calculated spectra:

$$S = \sum_{i=1}^r (E_i - E'_i)^2 \quad (2)$$

where all terms are given equal weight as the errors of measurement are practically constant throughout the range of wavelengths considered.

Substituting for E'_i in eqn. (2), and differentiating with respect to the k th concentration C_k , ($1 \leq k \leq n$) after rearrangement gives n simultaneous equations for the n unknown values C_j :

$$\sum_{j=1}^n C_j \sum_{i=1}^r \epsilon_{ij} \epsilon_{ik} = \sum_{i=1}^r E_i \epsilon_{ik} \quad (3)$$

or, in matrix notation

$$\mathbf{B}\mathbf{c} = \tilde{\mathbf{\epsilon}}\mathbf{E}, \quad (4)$$

where $\mathbf{\epsilon}$ is the matrix of elements ϵ_{ij} , $\tilde{\mathbf{\epsilon}}$ is its transpose whose ji element is ϵ_{ij} , and $\mathbf{B} = \tilde{\mathbf{\epsilon}}\mathbf{\epsilon}$,

$$\text{i.e.} \quad B_{jk} = \sum_{i=1}^r \epsilon_{ij} \epsilon_{ik}$$

thus $\mathbf{B}$ is a square matrix with $n \times n$ elements, and $B_{jk} = B_{kj}$. The solution of these equations is then given by

$$\mathbf{c} = \mathbf{M}\mathbf{E} \quad (5)$$

$$\text{where} \quad \mathbf{M} = \mathbf{B}^{-1} \tilde{\mathbf{\epsilon}} \quad (6)$$

If the elements of $\mathbf{M}$ are m_{ij} , then

$$c = \sum_{i=1}^r m_{ji} E_i \quad (7)$$

This mathematical procedure corresponds to the determination of the concentrations of the components in the mixture by fitting the experimental curve E_i with the library of the spectra of pure components. This library should contain all the components in the mixture.

From standard error theory the standard deviation of the calculated concentration c_j is given by

$$\sigma_{c_j}^2 = \sum_{i=1}^r \left(\frac{\partial c_j}{\partial E_i} \right)^2 \sigma_{E_i}^2 \quad (8)$$

where σ_{E_i} is the standard deviation of E_i . Assuming that all the extinction measurements E_i have the same standard deviation σ_E , we have

$$\sigma_{c_j}^2 = \sum_{i=1}^r \left(\frac{\partial c_j}{\partial E_i} \right)^2 \sigma_E^2 \quad (9)$$

or, from eqn. (7),

$$\sigma_{c_j} = P_j \sigma_E \quad (10)$$

where P_j , the error coefficient for the j th component, is defined by

$$P_j^2 = \sum_{i=1}^r m_{ji}^2 \quad (11)$$

P_j therefore is a measure of the error in the calculated concentration of the component j , for a given error in the measurement of extinction coefficients, and is a function only of the set of library spectra used in the calculation.

P_j varies from a minimum value of $\left[\sum_{i=1}^r \epsilon_{ij}^2 \right]^{-1/2}$, the value

when there is a single component in the library, or when there is no linear relationship between the spectrum of the j th component and those of any other components, to infinity when there is an exact linear relationship between several of the components, i.e.

$$\sum_j a_j \epsilon_{ij} = 0 \quad \text{for all } i$$

where the a_j values are constants which are non-zero only for the related components. This implies that the matrix $\mathbf{B}$ has zero determinant, so at least some of the elements of $\mathbf{B}^{-1}$, and therefore of $\mathbf{M}$, are infinite. Physically it means that the spectrum of one of the components can be exactly balanced by a summation of the spectra of other components of the related group. The ratio of P_j to $\left[\sum_{i=1}^r \epsilon_{ij}^2 \right]^{-1/2}$ can be

taken as a measure of the degree of linear relationship in a set of spectra. A study of the decrease in the values of P_j when a single component is removed from the library indicates the effect of that component on the accuracy of the determination of the other components in the library due to the degree of linear relationship.

Computation. The library spectra of the pure components used in the computations were the mean of at least three spectra corrected with solvent blanks. The compositions of mixtures were computed from their ultraviolet-absorption spectra by using a programme with the following features. Experimental curves, averaged if desired, were corrected, using an experimental solvent blank, and the concentrations and percentage compositions of the mixture were computed from eqn. (5), by using a specified library, which was selected from the complete library for each computation. The calculated absorption spectra were also printed out, together with a fit parameter,

$$\frac{\left[\sum_i (E_i - E'_i)^2 \right]^{1/2}}{\sum_i E_i}$$

which indicates the difference between the calculated and experimental absorption spectra. If any of the calculated concentrations were negative the computation could be repeated with a library comprising only the positive-valued components. All computations were carried out on an IBM 7090 digital computer.

RESULTS

Theoretical errors

Minor components of s-RNA. The spectra of nucleosides and their mixtures have been studied, rather than those of nucleotides, to avoid difficulties due to the small differences between the spectra of isomeric nucleotides. Also, in the automatic analytical method based on this work, it was convenient to measure the chain length of the dephosphorylated oligonucleotides from a phosphate analysis after degradation to nucleosides and free phosphate. Preliminary experiments indicated that the inclusion of extinction data obtained at wavelengths below 230 $m\mu$ introduced large errors due to traces of impurities, and all spectra used here were measured between 230 and 328 $m\mu$.

The most abundant nucleoside in s-RNA compounds, after the four common nucleosides adenosine, guanosine, cytidine and uridine, is 5-ribosyluracil or pseudouridine (Davis & Allen, 1957; Cohn, 1959), which occurs to the extent of about 4%. It is therefore necessary in investigating the conditions for maximum theoretical accuracy in the spectrophotometric analysis of oligonucleotides from s-RNA to discover the effect of including the spectrum of pseudouridine in the library of nucleosides used. Table 1 shows the theoretical error coefficients of concentrations determined by using libraries of adenosine, uridine, guanosine and cytidine with and without pseudouridine at different pH values. The pH values used, 2.0 and 12.0, were those at which the spectra of the components were relatively stable to small changes in pH (Cohn, 1960) and where the use of buffers, which absorb

ultraviolet light between 230 and 240 $m\mu$, could be avoided.

From Table 1 it can be seen that the error coefficients of the four-component acid and alkali libraries are 10–15 times as large as the error coefficients of a single component calculated from its own library spectrum. They are, however, comparable with each other. The average error in the calculated concentration corresponding to an error of 0.01 in the extinction coefficient is 2.8 μM for the alkaline four-component library, and 2.3 μM for the acid four-component library. When pseudouridine is included in the libraries, the error coefficients of the four main components increase, but this increase, for the alkaline library, is only about 20%, and the error coefficient of pseudouridine is only about 12% of that obtained with the acid library. With the acid library the error coefficients of uridine and pseudouridine are high because of the high degree of linearity between their spectra. The most satisfactory pH for the spectrophotometric analysis of oligonucleotides from s-RNA is, then, pH 12.0. This library of spectra is shown in Fig. 1.

Minor methylated components. In addition to pseudouridine, s-RNA contains a number of minor components, nucleosides with methylated bases, which total a few per cent of the RNA (Smith & Dunn, 1959). Any one particular molecular species of amino acid transfer RNA is unlikely to contain more than a fraction of all the possible methylated nucleosides.

A study has therefore been made of the effect on the accuracy of spectrophotometric analysis of including some minor bases in the library of spectra. Some of the methylated nucleosides are unstable in alkali conditions, the bases undergoing rapid intra-

Table 1. Concentration error coefficients, P_i , for libraries containing the spectra of adenosine, uridine, guanosine, cytidine and 5-ribosyluracil

The error coefficients, P_i , for each component were calculated according to eqn. (11) and are the ratios of the theoretical standard deviations of concentration calculated with the specified sets of library curves to the standard deviations in extinction coefficients. Each spectrum used in the calculations consisted of 50 extinction values taken at 2 $m\mu$ intervals between 230 and 328 $m\mu$ of a solution of pure nucleosides at 0.1 mM in either 0.01 N-HCl or 0.01 N-NaOH adjusted to pH 12.0.

	P_i (10 μM)				
	Adenosine	Uridine	Guanosine	Cytidine	5-Ribosyluracil
Acid:					
Single-component library	2.0	3.0	1.9	2.0	2.3
Four-component library	31	39	15	8	
Five-component library	32	85	16	8	64
Alkali:					
Single-component library	2.0	3.2	2.1	2.4	3.2
Four-component library	22	30	31	29	
Five-component library	24	34	34	37	7

Fig. 1. Ultraviolet spectra of adenosine (A), uridine (B), cytidine (C), guanosine (D) and 5-ribosyluracil (E), in 0.01 N-NaOH, adjusted to pH 12, at a concentration of 0.1 mM.

Fig. 2. Ultraviolet spectra of 6-methylaminopurine riboside (A), 6-dimethylaminopurine riboside (B), thymine riboside (C) and the 7-methylguanosine derivative (D) in 0.01 N-NaOH, adjusted to pH 12.0, at a concentration of 0.1 mM. The spectrum of the 7-methylguanosine derivative was taken after 15 min. in 0.01 N-NaOH at pH 12.0, after which the spectrum remained unchanged for 15 hr.

molecular rearrangement or degradation at high pH values which involve changes in their ultraviolet absorption spectra (Brooks & Lawley, 1960; Lawley & Wallick, 1957). For 7-methylguanosine it was found that, on adjusting a solution to pH 12.0, the spectrum changed rapidly, but within 15 min. this change was complete, and the spectrum was found to be relatively stable for 15 hr. This spectrum was similar to the spectrum of 2-amino-6-hydroxy-5-methylformamido-4-ribosylaminopyrimidine when adjusted to pH 8.7 (Lawley & Brookes, 1963). The spectrum of the degradation product at pH 12.0 was used, then, as a library curve for 7-methylguanosine for estimating this nucleoside in the mixtures. Thymine riboside, 6-methylaminopurine riboside, 6-dimethylaminopurine riboside and 7-methylguanosine were used in the investigation of the effect of including the spectra of minor nucleosides in the library. The spectra, at pH 12.0, of these minor components which were used as library spectra in the computation are shown in Fig. 2. The calculated error coefficients for libraries containing the minor methylated components of s-RNA are shown in Table 2. For the complete nine-component library the accuracy of the determination of uridine and thymine is very low, so that the library is not suitable for the final determination of the composition of a mixture of nucleosides, but can be clearly used for the identification of one of the components in an end-group procedure. The high values of the error coefficients are due to the linear relationship between some of the nine spectra. Comparison of the error coefficients of the nine-component library with those of the eight-component libraries in Table 2 shows that removal of thymine riboside from a library sharply increases the accuracies of the adenosine and uridine determinations and the removal of 6-dimethylaminopurine riboside improves that of 5-ribosyluracil strikingly because of the high degree of linear relationship between these groups of spectra. The removal of both

Table 2. Concentration error coefficients, P_i , for libraries containing the spectra of seven to nine nucleosides of transfer ribonucleic acid

The error coefficients, P_i , were calculated as in Table 1 using spectra taken at pH 12.0. For 7-methylguanosine the spectrum was taken 15 min. after adjustment of the pH to 12.0, and is the spectrum of the alkali-degradation product. A, Adenine; U, uridine; G, guanosine; C, cytidine; ψ U, 5-ribosyluracil; 7-MG, 7-methylguanosine; 6-MA, 6-methylaminopurine riboside; DMA, 6-dimethylaminopurine riboside; T, thymine riboside.

	P_i (10 μ M)								
	A	U	G	C	ψ U	7-MG	6-MA	DMA	T
Single-component library	2.0	3.2	2.1	2.4	3.2	2.3	1.4	—	1.9
Seven-component library	25	68	42	38	9	30	18	—	—
Eight-component library	76	367	46	41	61	40	—	63	142
Eight-component library	88	309	48	45	14	45	63	—	146
Eight-component library	29	127	42	63	74	31	91	93	—
Nine-component library	91	370	49	65	83	45	101	101	153

thymine riboside and 6-dimethylaminopurine riboside from the nine-component library gives a set of error coefficients which were comparable with those obtained with a five-component library. Although it is evident from Table 2 that some seven- and eight-component libraries can be used for the analysis of mixtures, in general the accuracy of determination of a given component increases as the number of components in the library decreases. From Table 2 it can also be seen that even with a nine-component library some of the components in a mixture can be determined with a useful accuracy in the presence of other components which cannot be determined very accurately.

Experimental mixtures

Mixtures of the major nucleoside components. One of the practical limits on nucleotide-sequence analysis is the separation of complex oligonucleotide mixtures. The most successful procedure at present available for this purpose, that of Tener, Khorana, Markham & Pol (1958) using elution from DEAE-cellulose columns with and without 7 M-urea, will not fractionate oligonucleotides containing more than nine nucleotides. The limit of accuracy required for a simple automatic analysis of oligonucleotides is, then, that necessary for the analysis of oligonucleotides containing nine nucleotides.

Table 3 shows the analysis of experimental mixtures of nucleosides simulating digests of some oligonucleotides with nucleotide chain lengths of seven to nine. Although the differences between the spectra are not very large the computed concentrations are sufficiently accurate to determine the composition and chain length of the corresponding oligonucleotides if the approximate chain length is known to within two nucleotides. Standard deviations of the calculated compositions, expressed in volume units in the Table for clarity, are, in concentration units, 0.2 μM for adenosine, 0.4 μM for uridine, 0.5 μM for guanosine and 0.5 μM for cytidine and the concentrations for all the components were approximately 20 μM . The root mean square of the deviation of the calculated concentrations from the experimental concentrations is a measure of the experimental accuracy of the results and the values are 0.5 μM for adenosine, 1.3 μM for uridine, 1.4 μM for guanosine and 1.1 μM for cytidine. Assuming an error of 0.005 in the measurements of the extinction coefficients, the concentration errors predicted from the error coefficients are 1.1 μM for adenosine, 1.4 μM for uridine, 1.6 μM for guanosine and 1.4 μM for cytidine. The experimental errors are slightly smaller than these predicted errors but both vary in the same manner from one component to another.

Mixtures including the minor nucleoside components. It can be seen from Tables 1 and 2 that the

Table 3. *Spectrophotometric analysis of four-component nucleoside mixtures*

The mixtures were made by mixing appropriate volumes of stock solutions of the four major nucleosides in 0.01 N-NaOH adjusted to pH 12.0, in ratios simulating digests of oligonucleotides with chain lengths of seven to nine. The stock solutions were 0.0980 mM-adenosine, 0.0871 mM-uridine, 0.1044 mM-guanosine and 0.0820 mM-cytidine. The true and calculated compositions were expressed in terms of volumes of the stock solutions. The compositions are expressed relative to one component and were calculated by the least-squares method by using a library of the spectra of the four components. The fit parameter was calculated from the expression:

$$\frac{\left[\sum_{i=1}^r (E_i - E'_i)^2 \right]^{1/2}}{\sum_{i=1}^r E_i}$$

where E_i was the experimental extinction at wavelength i and E'_i was the calculated extinction at that wavelength. A, Adenosine; U, uridine; G, guanosine; C, cytidine.

	A	U	G	C	10 ⁴ × Fit parameter
Composition	2	2	2	2	
Calc. composition	2	2.00	2.15	1.95	15
Composition	1	2	2	2	
Calc. composition	1	2.12	1.98	1.91	17
Composition	2	1	2	2	
Calc. composition	1.92	1	1.83	1.82	12
Composition	2	2	1	2	
Calc. composition	1.83	2.04	1	1.72	11
Composition	2	2	2	3	
Calc. composition	1.96	2.24	2	3.16	14
S.D.	0.02	0.04	0.04	0.05	

error σ_{c_j} in the calculated concentration, c_j , of a particular nucleoside in general decreases as the number of components in the library utilized for the calculation decreases, as would be expected from the fact that the degree of linear relationship tends to decrease as the number of components in the library decrease. The precision of the calculated concentrations can therefore be increased by recalculating concentrations with a restricted library in which components which are not present in the mixture are omitted from the library. For the analysis of oligonucleotide digests, a simple procedure for eliminating absent components from the library has been devised. This procedure can be derived from an examination of the calculated error coefficients and a knowledge of the chain length of

the oligonucleotide. In the digest of an oligonucleotide with a chain length n , the smallest possible concentration of a single component j is C/n or zero, where C is the total nucleoside concentration. If a component has a small calculated value c_j , there is an 86% probability that the true value is less than $c_j + 1.5\sigma_{c_j}$, where σ_{c_j} is given by the product of the j th error coefficient and the standard deviation of the extinction coefficient. Therefore if $c_j + 1.5\sigma_{c_j} < C/n$, then the true concentration of j is zero with an 86% probability. Therefore if $c_j < (C/n) - 1.5\sigma_{c_j}$, then there is an 86% probability that the j th component should not be in the library. A recalculation without this component should, then, give more accurate values for the concentrations of the components of the mixture.

Table 4. *Spectrophotometric analysis of mixtures and of some methylated components with a library of nine nucleosides*

The mixtures were made by mixing the appropriate volumes of stock solutions of known concentrations at pH 12.0. All solutions containing 7-methylguanosine were left for at least 15 min. before the spectrum was measured. The compositions were calculated by using the least-squares procedure and a library of nine-component spectra for each mixture. If any of the calculated concentrations were negative, the calculations were automatically repeated with a modified library from which these components were deleted. The root-mean square deviations were calculated for the nine-component-library results. The theoretical errors were calculated from the error coefficients for the nine-component library assuming an error of 0.01 in the measurement of extinction coefficients. Superscript figures 1 and 2 in the Table indicate that the calculation shown is the first or the second calculation (with negative concentrations rejected). A, Adenosine; U, uridine; G, guanosine; C, cytidine; ψ U, 5-ribosyluracil; 7-MG, 7-methylguanosine; 6-MA, 6-methylaminopurine riboside; DMA, 6-dimethylaminopurine riboside; T, thymine riboside.

	10 ⁴ × Concn. (M)									10 ⁴ × Fit parameter
Composition	A	U	G	C	ψ U	7-MG	6-MA	DMA	T	
Composition	0	0	0	0	0	0.97	0	0	0	
Calc. composition ¹	0.01	-0.08	0.02	0.00	0.01	0.99	-0.01	-0.00*	0.02	8
Calc. composition ²	-0.01		0.01	0.01	-0.00	0.98			-0.01	8
Composition	0	0	0	0	0	0	0.58	0	0	
Calc. composition ¹	0.02	-0.05	-0.02	-0.03	-0.02	0.02	0.53	0.03	0.03	3
Calc. composition ²	-0.01					0.01	0.58	-0.00	-0.01	4
Composition	0	0	0	0	0	0	0	0.60	0	
Calc. composition ¹	0.02	-0.09	0.01	-0.01	0.01	-0.00	0.00	0.58	0.05	4
Calc. composition ²	0.00		0.00		0.00		0.00	0.60	-0.00	5
Composition	0	0	0	0	0	0	0	0	0.84	
Calc. composition ¹	0.01	-0.02	-0.00	0.01	0.01	0.01	-0.00	-0.00	0.83	4
Calc. composition ²	0.00			0.00	0.00	0.01			0.82	4
Composition	0	0	0	0	0	0.24	0.15	0.15	0.21	
Calc. composition ¹	0.06	-0.22	0.03	0.04	0.06	0.21	0.14	0.08	0.31	8
Calc. composition ²	0.01			-0.01	0.02	0.20	0.12	0.13	0.23	10
Composition	0.20	0.20	0.21	0.17	0.24	0	0	0	0	
Calc. composition ¹	0.09	0.50	0.11	0.29	0.31	-0.11	0.21	-0.12	-0.18	12
Calc. composition ²	0.22	0.16	0.18	0.22	0.24		-0.00			18
Composition	0.24	0.24	0.26	0.20	0	0	0	0	0	
Calc. composition ¹	0.10	0.58	0.17	0.30	0.10	-0.11	0.25	-0.15	-0.18	15
Calc. composition ²	0.24	0.27	0.24	0.20	0.00		0.00			21
Root mean square deviation	0.08	0.21	0.06	0.07	0.06	0.07	0.13	0.08	0.11	
Theoretical error	0.09	0.37	0.05	0.07	0.08	0.05	0.10	0.10	0.15	

* Negative-calculated concentrations with absolute values of less than 0.005 are represented as -0.00.

If n is 1, say, and $C \approx 1 \times 10^{-4}$, then $(C/n) - 1.5\sigma_c$ is much greater than zero for all components for a nine-component library, for an error of 0.01 in the measurement of extinction coefficients. Similarly, if n is 9, $(C/n) - \sigma_c$ is greater than zero for the average standard deviation σ_c , giving a 68% probability that the true concentration is zero, for the same error in extinction coefficients. If the conservative limit for c_i of zero is taken and all negative-valued components are rejected from the library for the recalculation, then improved results can be obtained as shown in Table 4. The first four calculations show that any of the methylated nucleosides studied can be identified unequivocally with an unrestricted library. However, the method of recalculation with rejection of negative components clearly improves the accuracy of the calculated concentration of the component present and assists the elimination of apparent traces of the absent components. In the fifth calculation, in which a mixture of four methylated nucleosides was analysed, the recalculation with rejection of negative components was necessary to obtain reasonable values for the composition of the mixture. Calculations six and seven also show the increase in accuracy when the library restriction procedure is applied to mixtures of the main nucleosides with and without 5-ribosyluracil, starting with a nine-component library. It can be seen that the root mean square of the deviations of the calculated concentrations from the true values vary in a manner comparable with that of the theoretical errors.

DISCUSSION

In the application of spectrophotometric analysis to the analysis of the nucleoside mixtures that might be encountered in the sequence analysis of s-RNA molecules, the main problem is to select practical conditions under which the degree of linear relationships between the spectra of the pure components are minimized so that their concentrations and mixtures can be determined with maximum accuracy. The accuracy under these conditions must be sufficient for the determination of compositions of oligonucleotides. The calculations of the error coefficients, P , readily enable the degrees of linearity and the accuracies obtained with libraries of spectra taken at different values to be compared and the best conditions selected. With the additional requirements that, at the pH chosen, the spectra must be relatively stable to small changes in pH, and that ultraviolet-absorbing buffers must not be present, it is clear from Table 1 that the most suitable solvent for the determination of the compositions from the spectra is 0.01 N-sodium hydroxide at pH 12.0. At acid pH values the spectrum of 5-ribosyluracil is linearly related to that of uridine.

It is clear, however, from a study of the error coefficients, that some components can be determined accurately in the presence of other components whose accuracy is low. It is therefore possible to determine the concentration of each component under conditions which give the highest accuracy for that component, and combine the results for all the components to give the best composition of the mixture. For instance, cytidine and guanosine can be determined most accurately at pH 2.0 with a five-component library, whereas adenosine, uridine and 5-ribosyluracil are best determined at pH 12.0.

The presence of small amounts of methylated components in the s-RNA compounds raises the problem of the effect of including these components in the library. The values of the error coefficients for each component in libraries containing different methylated nucleosides show clearly that the errors in the method increase with the number of components. The increase is very marked when components are included in the library whose spectra are partially linearly related to one of the major components. It is, then, very desirable to restrict the number of methylated components in a particular library to a minimum to ensure adequate accuracy in the calculation. This can readily be done with s-RNA because the methylated components can be specifically labelled with $^{14}\text{CH}_3$ groups (Fleissner & Borek, 1962) and the presence of a methylated component in an oligonucleotide detected by a radioactive monitor. Most oligonucleotides in digests of s-RNA will not contain them, and in their absence a five-component library can be used. The other method of restricting the library is the automatic elimination of negative-valued components from the library in a recalculation in which the accuracy will be correspondingly high. It is clearly essential to begin with a library which includes at least all the nucleosides present in the mixture. With the above procedure it is possible to make an initial calculation with an extensive library and then refine the solution by reducing the number of library components to a minimum. For automatic identification of a single component split off in an end-group procedure an extensive library can be used satisfactorily without refinement. For mixtures the fit parameter provides a convenient check on faulty data, system faults and incorrect library selection, as the value of the fit parameter under these circumstances in general becomes very much larger than the average value characteristic of a particular library. Large negative values for the calculated concentration of components also indicates too small a library or serious errors in the system. Components present in the mixture but not included in the library can also often be detected as localized discrepancies between the calculated and experimental spectra.

The above analysis shows that multicomponent spectrophotometric analysis using the least-squares curve-fitting procedure is sufficiently accurate to analyse nucleoside mixtures corresponding to digests of those oligonucleotides that can be readily fractionated from digests of s-RNA compounds. The method can clearly be used either with an automatic end-group procedure to identify the split-off products or to analyse the composition of oligonucleotides in a flowing stream. The accuracy obtainable approaches that of paper electrophoresis or chromatography in which the components are physically separated and the concentrations determined by a single spectrophotometric measurement at the peak wavelength. The average error coefficients for this single measurement for the five main nucleosides in s-RNA is approximately 0.1 mM, whereas the corresponding figure for the analysis using a library containing five components is 0.25 mM. When the errors in the chromatographic method due to differential elution from the paper, with losses often varying from 2 to 5%, are considered, the accuracy of the spectrophotometric analysis compares favourably with the chromatographic method. For speed, convenience and suitability for automatic operation the spectrophotometric method is superior. In practice it may be useful to use chromatography qualitatively, to check the components present in a number of pooled samples, to ensure that the library used includes all the components present, and yet restrict it to the minimum number necessary to maintain sufficient accuracy for the intended purpose.

We thank Professor Sir John Randall and Professor M. H. F. Wilkins for their interest and encouragement, Dr W. E. Cohn and Dr C. B. Reese for gifts of nucleosides, and Dr D. B. Dunn for useful discussion. We are grateful to IBM (U.K.) Ltd. for providing computing facilities under their Endowed Research Time project.

REFERENCES

- Apgar, J., Holley, R. W. & Merrill, S. H. (1962). *J. biol. Chem.* **237**, 796.
- Beaven, G. H., Holiday, E. R. & Johnson, E. A. (1955). In *The Nucleic Acids*, vol. 1, p. 493. Ed. by Chargaff, E. & Davidson, J. N. New York: Academic Press Inc.
- Brookes, P. & Lawley, P. D. (1960). *J. chem. Soc.* p. 539.
- Brown, G. L., Lee, S. & McMullen, D. W. (1963). *Biochem. J.* **89**, 2p.
- Brown, G. L., Lee, S., McMullen, D. W. & Stokes, A. R. (1964). *Biochem. J.* **92**, 4p.
- Brown, G. L. (1963). *Progr. nucleic Acids*, **2**, 259.
- Cohn, W. E. (1959). *Biochim. biophys. Acta*, **32**, 569.
- Cohn, W. E. (1960). *J. biol. Chem.* **235**, 1488.
- Davis, F. F. & Allen, F. W. (1957). *J. biol. Chem.* **227**, 907.
- Dekker, C. A. (1960). *Annu. Rev. Biochem.* **29**, 453.
- Fleissner, E. & Borek, E. (1962). *Proc. nat. Acad. Sci., Wash.*, **48**, 1199.
- Fox, J. J. & Shugar, D. (1952). *Biochim. biophys. Acta*, **9**, 369.
- Goldstein, J., Bennett, T. P. & Craig, L. C. (1964). *Proc. nat. Acad. Sci., Wash.*, **51**, 119.
- Lawley, P. D. & Brookes, P. (1963). *Biochem. J.* **89**, 127.
- Lawley, P. D. & Wallick, C. A. (1957). *Chem. & Ind.* p. 633.
- Littlefield, J. W. & Dunn, D. B. (1958). *Biochem. J.* **70**, 642.
- Michelson, A. M. & Cohn, W. E. (1962). *Biochemistry*, **1**, 490.
- Reid, J. C. & Pratt, A. W. (1960). *Biochem. biophys. Res. Commun.* **3**, 337.
- Shapiro, R. & Chambers, R. W. (1961). *J. Amer. chem. Soc.* **83**, 3921.
- Skeggs, L. T. (1957). *Amer. J. clin. Path.* **28**, 311.
- Smith, J. D. (1955). In *The Nucleic Acids*, vol. 1, p. 267. Ed. by Chargaff, E. & Davidson, J. N. New York: Academic Press Inc.
- Smith, J. D. & Dunn, D. B. (1959). *Biochem. J.* **72**, 294.
- Sternberg, J. C., Stillo, H. S. & Schwendeman, R. H. (1960). *Analyt. Chem.* **32**, 84.
- Tener, G. P., Khorana, H. G., Markham, R. & Pol, E. H. (1958). *J. Amer. chem. Soc.* **80**, 6223.
- Vasilenko, S. K., Kamzolova, S. G. & Knorre, D. G. (1962). *Biokhimiya*, **27**, 1942.
- Zscheile, F. P., Murray, H. C., Baker, G. A. & Peddicord, R. G. (1962). *Analyt. Chem.* **34**, 1776.