Volume 5 Number 6 June 1978 Nucleic Acids Research

A simple method for the computation of first neighbour frequencies of DNAs from CD spectra

Christian Marck[+a] and Wilhelm Guschlbauer[++]

[+]Service de Biophysique and [++]Service de Biochimie, Département de Biologie, Centre d'Etudes Nucléaires de Saclay, 91190 Gif-sur-Yvette, France

Received 11 April 1978

ABSTRACT

A procedure for the computation of the first neighbour frequencies of DNA's is presented. This procedure is based on the first neighbour approximation of Gray and Tinoco. We show that the knowledge of all the ten elementary CD signals attached to the ten double stranded first neighbour configurations is not necessary. One can obtain the ten frequencies of an unknown DNA with the use of eight elementary CD signals corresponding to eight linearly independent polymer sequences. These signals can be extracted very simply from any eight or more CD spectra of double stranded DNA's of known frequencies. The ten frequencies of a DNA are obtained by least square fit of its CD spectrum with these elementary signals. One advantage of this procedure is that it does not necessitate linear programming, it can be used with CD data digitalized using a large number of wavelengths, thus permitting an accurate resolution of the CD spectra. Under favorable case, the ten frequencies of a DNA (not used as input data) can be determined with an average absolute error < 2%. We have also observed that certain satellite DNA's, those of *Drosophila virilis* and *Callinectes sapidus* have CD spectra compatible with those of DNA's of quasi random sequence ; these satellite DNA's should adopt also the B-form in solution.

INTRODUCTION

The dependence of CD spectra on base composition has been extensively studied during the past ten years. Felsenfeld and Hirschmann[1] have first investigated linear dependence of UV absorption spectra over GC content ; these authors have concluded to the insufficiency of such an hypothesis. ORD and CD studies of Samejima and Yang ,[2] Zimmer and Luck [3], Courtois *et al.*[4] have demonstrated that a linear relation links the specific rotation at 290 nm and the GC content and that a similar relation takes place at p^H for the specific rotation at 260 nm. Gratzer *et al.* [5] have investigated a quadratic dependence on GC content and have shown that this hypothesis was still imperfect. Further studies by Bernardi and Timasheff [6], Wells *et al.* [7] have shown that DNA's of same GC content but of different sequence exhibit different physico-chemical properties. This demonstrated that the CD spectrum of DNA depends upon the base interactions as well as upon the contributions of the bases themselves.

The most sophisticated interpretation of the variations of the DNA CD spectra due to base composition has been presented by Gray and Tinoco in 1970 [8]. These authors have shown that any sequence dependent property of a polynucleotide chain can be expressed as a linear function of the properties of a limited number of polynucleotides of simple sequence. If only first neighbour interactions contribute to the formation of the total CD signal, the CD signal of any double stranded polymer can be expressed as a sum of eight elementary CD signals corresponding to simple sequences. From this, Gray and Tinoco have then predicted that it should be possible to estimate the nearest neighbour frequencies of a DNA only from its CD spectrum. Allen *et al.* [9] have first presented a numerical application of the first neighbour hypothesis. Further work of Gray *et al.* [10], Allen and Daub [11], Gray and Gall [12], Gray and Skinner [13] using the original computer procedure of Allen *et al.* [9] had shown the difficulty to obtain accurately the frequencies of independent test case DNA's. This was due partly to difference in the geometric configurations of the DNA's and polymers used as input data. This difficulty has been partially overcomed by a suggestion of Arnott [14] : polynucleotides with repetitive trimer sequences should be better suited as library spectra than those containing dimer sequences.

In this paper, we first present the analytical procedure we have used for the computation of first neighbour frequencies. This procedure is simpler, both in its principle and its numerical application, than that previously presented by Allen *et al.* [9] Examples of first neighbour frequency determinations are given. Some problems concerning the numerical applications and the development of the first neighbour formalism are discussed.

DATA

CD spectra were recorded as previously described[15]. In order to improve signal to noise ratio, four scans were performed for DNA samples and for base lines. DNA samples were dialysed three times against 0.15 M NaCl, 10^{-3} M Tris, pH 7.0. The following DNA's were used : *Tetrahymena pyriformis* 24% GC, *Clostridium perfringens* 28% GC, *Haemophilus influenzae* 39% GC, *Bacillus subtilis* 44% GC, Wheat germ 45% GC, Phage λ 50% GC, *Serratia marcescens* 59% GC, *Micrococcus lysodeikticus* 72% GC.

Poly (dGC).poly(dGC) was a commercial sample from Boehringer, Mannheim. The CD spectrum used for poly(dG).poly(dC) was that determined for the pure double stranded complex in previous work (see ref. 15 for details). Other CD spectra were taken from literature : *Drosophila virilis* satellite DNA's from Gray and Gall[12], crab satellite DNA's from Gray and Skinner[13].

ANALYTICAL PROCEDURE

Definitions : According to the nearest neighbour hypothesis of Gray and Tinoco[8] (first neighbour approximation), the CD spectrum of any DNA can be expressed as a linear sum of ten elementary CD signals. These signals are each attached to one of the ten possible base paired first neighbour configurations. We will consider in the present demonstration that all DNA CD spectra obey ideally this hypothesis. Then, any CD spectrum is written :

$$s = \sum_{i=1}^{10} T_i . f_i, \quad \text{or in matrix form } s = T.f \qquad (1)$$

any set of m CD spectra is written :

$$S = T.F \qquad (2)$$

S is an (n x m) matrix, every column of which is the CD spectrum of a different DNA (these spectra are of course to be recorded at the same n wavelengths and under the same experimental conditions). The column vectors of the T matrix (n x 10) are the elementary CD signals of the 10 base paired first neighbours [9] :

$$T_{\substack{AA\\TT}},\ T_{\substack{AT\\TA}},\ T_{\substack{TA\\AT}},\ T_{\substack{AC\\TG}},\ T_{\substack{CA\\GT}},\ T_{\substack{AG\\TC}},\ T_{\substack{GA\\CT}},\ T_{\substack{GC\\CG}},\ T_{\substack{CG\\GC}},\ T_{\substack{GG\\CC}}$$

The F matrix has therefore to consist obligatorily of the corresponding frequencies i.e. base paired first neighbour configurations frequencies ($f_{\substack{XY\\X'Y'}}$: double stranded frequency) :

$$f_{\substack{AA\\TT}},\ f_{\substack{AT\\TA}},\ f_{\substack{TA\\AT}},\ f_{\substack{AC\\TG}},\ f_{\substack{CA\\GT}},\ f_{\substack{AG\\TC}},\ f_{\substack{GA\\CT}},\ f_{\substack{GC\\CG}},\ f_{\substack{CG\\GC}},\ f_{\substack{GG\\CC}}$$

Such frequencies are, of course, related to the usual single stranded first neighbour frequencies (f_{XY} : single stranded frequency) defined by Josse *et al.*[16] The relations between the two sets of frequencies are : [9]

a) for the non-autocomplementary doublets (X and Y complementary to X' and Y', respectively)

$$f_{\substack{X\,Y\\X'Y'}} = 2\,f_{XY} = 2\,f_{Y'X'} = f_{X\,Y} + f_{Y'X'} \tag{3}$$

b) for the autocomplementary doublets

$$f_{\substack{X\,X'\\X'X}} = f_{X\,X'} \tag{4}$$

Both sets of frequencies sum to 1. Josse *et al.*[16] have shown that the single stranded frequencies are constrained by the four relations :

$$f_{AT} + f_{AC} + f_{AG} = f_{TA} + f_{CA} + f_{GA} \tag{5}$$

$$f_{GC} + f_{GA} + f_{GT} = f_{CG} + f_{AG} + f_{TG} \tag{6}$$

$$f_{TA} + f_{TC} + f_{TG} = f_{AT} + f_{CT} + f_{GT} \tag{7}$$

$$f_{CG} + f_{CA} + f_{CT} = f_{GC} + f_{AC} + f_{TC} \tag{8}$$

The origin of these relations is the first topologic rule of DNA : both strands consist each of a linear array of bases. The second rule, which is the Watson-Crick base pairing, makes that only two, over the four relations (5) through (8), are independent[16] ; say relations (5) and (6). As a consequence of the third topological rule of DNA, the opposite polarity of the two strands , two analogous independent relations can be written with the double stranded frequencies[8] :

$$f_{\substack{AT\\TA}} + 1/2\ f_{\substack{AC\\TG}} + 1/2\ f_{\substack{AG\\TC}} = f_{\substack{TA\\AT}} + 1/2\ f_{\substack{CA\\GT}} + 1/2\ f_{\substack{GA\\CT}} \tag{9}$$

$$f_{\substack{GC\\CG}} + 1/2\ f_{\substack{GA\\CT}} + 1/2\ f_{\substack{GT\\CA}} = f_{\substack{CG\\GC}} + 1/2\ f_{\substack{AG\\TC}} + 1/2\ f_{\substack{TG\\AC}} \tag{10}$$

As a consequence, the various lines of the F matrix (eq.(2)) are also constrained by the relations (9) and (10), whatever are the sequences of the DNA's used (provided that these DNA's are of infinite length or circular), since their frequencies always obey eq.(5) through (10). Therefore, although the physical definition of the T matrix is clear-cut, we are unable to compute it [17] (i.e. obtain separately the ten elementary CD signals) by solving eq.(2) since the order of the F matrix is only eight (ten frequencies minus two relations).

Mathematical procedure

Two frequencies, say $f_{\substack{AT\\TA}}$ and $f_{\substack{GC\\CG}}$, can be chosen dependent and expressed as a function of the other ones according to eq.(9) and (10) :

$$f_{\substack{AT\\TA}} = f_{\substack{TA\\AT}} - 1/2\ f_{\substack{AC\\TG}} + 1/2\ f_{\substack{CA\\GT}} - 1/2\ f_{\substack{AG\\TC}} + 1/2\ f_{\substack{GA\\CT}} \qquad (11)$$

$$f_{\substack{GC\\CG}} = f_{\substack{CG\\GC}} - 1/2\ f_{\substack{AC\\TG}} + 1/2\ f_{\substack{CA\\GT}} + 1/2\ f_{\substack{AG\\TC}} - 1/2\ f_{\substack{GA\\CT}} \qquad (12)$$

In matrix form, this means that any f column vector can be written as a product :

$$f = Z.f' \qquad (13)$$

Any F matrix can also be written :

$$F = Z.F' \qquad (14)$$

with :

$$Z = \begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1/2 & +1/2 & -1/2 & +1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1/2 & +1/2 & +1/2 & -1/2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{vmatrix} \qquad (15)$$

The F' matrix so defined is a subset matrix of F, the frequencies $f_{\substack{AT\\TA}}$ and $f_{\substack{GC\\CG}}$ being missing. Let us define a T' matrix as :

$$T' = T.Z \qquad (16a)$$

Let T'_0 through T'_7 be the columns vectors of T' : eq.(16a) define these vectors as :

$$T'_0 = T_{\substack{AA\\TT}},\quad T'_1 = T_{\substack{TA\\AT}} + T_{\substack{AT\\TA}},\quad T'_2 = T_{\substack{AC\\TG}} - 1/2.T_{\substack{AT\\TA}} - 1/2\ T_{\substack{GC\\CG}},$$

$$T'_3 = T_{\substack{CA\\GT}} + 1/2\ T_{\substack{AT\\TA}} + 1/2\ T_{\substack{GC\\CG}},\quad T'_4 = T_{\substack{AG\\TC}} - 1/2\ T_{\substack{AT\\TA}} + 1/2\ T_{\substack{GC\\CG}}, \qquad (16b)$$

$$T'_5 = T_{\substack{GA\\CT}} + 1/2\ T_{\substack{AT\\TA}} - 1/2\ T_{\substack{GC\\CG}},\quad T'_6 = T_{\substack{CG\\GC}} + T_{\substack{GC\\CG}},\quad T'_7 = T_{\substack{GG\\CC}}.$$

Equation (2) can now be written using eq.(14) and (16a) :

$$\begin{aligned} S &= T.F = T.Z.F' \\ S &= T'.F' \end{aligned} \qquad (17)$$

T' and F' are (n x 8) and (8 x m) matrices, respectively. Let us consider an input set, S, made of eight or more CD spectra of DNA's, the frequencies of which are known. If m = 8, F' is a square matrix and T' is obtained by :

$$T' = S.F'^{-1} \qquad (18a)$$

If we use more than eight CD spectra for the input set, T' is obtained by :

$$T' = S.F'^{t}.(F'.F'^{t})^{-1} \qquad (18b)$$

The T' matrix so obtained is sufficient for all further computations :
1) the unknown frequencies of a DNA, the CD spectrum of which is known, are defined by eq. (1). This equation can be written :

$$\begin{aligned} s &= T.f = T.Z.f' \\ s &= T'.f' \end{aligned} \qquad (19)$$

Therefore the least square fit of the s spectrum by the column vectors of T' gives f', i.e. the 8 frequencies chosen independent :

$$f' = (T'^{t}.T')^{-1}.T'^{t}.s \qquad (20)$$

The two missing frequencies are obtained from eq. (13).

2) Equation (19) can also be used to obtain the CD spectrum corresponding to any DNA sequence.

In order to illustrate this, let us consider the following sequence (assumed to be infinitely repetitive) :

$$\begin{array}{c} \ldots\overrightarrow{\mathrm{A\ A\ T\ C\ A\ G}}\ldots \\ \ldots\underleftarrow{\mathrm{T\ T\ A\ G\ T\ C}}\ldots \end{array}$$

The double stranded first neighbour frequencies of this polymer are :

$$f_{\substack{AA\\TT}} = 1/6,\ f_{\substack{AT\\TA}} = 1/6\ ;\ f_{\substack{CA\\GT}} = 1/6,\ f_{\substack{AG\\TC}} = 1/6,\ f_{\substack{GA\\CT}} = 2/6$$

all other frequencies being equal to 0. According to eq.(19), the CD spectrum of this polymer is :

$$s_{\substack{\mathrm{AATCAG}\\ \mathrm{TTAGTC}}} = 1/6\ T'_0 + 1/6\ T'_3 + 1/6\ T'_4 + 2/6\ T'_5 \qquad (21)$$

Replacing the T'_i vectors by their expression as a function of the T_i vectors given by eq.(16b) leads to :

$$s_{\substack{\mathrm{AATCAG}\\ \mathrm{TTAGTC}}} = 1/6 \left[T_{\substack{\mathrm{AA}\\ \mathrm{TT}}} + \left(T_{\substack{\mathrm{CA}\\ \mathrm{GT}}} + 1/2\ T_{\substack{\mathrm{AT}\\ \mathrm{TA}}} + 1/2\ T_{\substack{\mathrm{GC}\\ \mathrm{CG}}} \right) + \left(T_{\substack{\mathrm{AG}\\ \mathrm{TC}}} - 1/2\ T_{\substack{\mathrm{AT}\\ \mathrm{TA}}} \right. \right.$$

$$\left. \left. + 1/2\ T_{\substack{\mathrm{GC}\\ \mathrm{CG}}} \right) + 2 \left(T_{\substack{\mathrm{GA}\\ \mathrm{CT}}} + 1/2\ T_{\substack{\mathrm{AT}\\ \mathrm{TA}}} - 1/2\ T_{\substack{\mathrm{GC}\\ \mathrm{CG}}} \right) \right]$$

$$s_{\substack{\mathrm{AATCAG}\\ \mathrm{TTAGTC}}} = 1/6 \left(T_{\substack{\mathrm{AA}\\ \mathrm{TT}}} + T_{\substack{\mathrm{AT}\\ \mathrm{TA}}} + T_{\substack{\mathrm{CA}\\ \mathrm{GT}}} + T_{\substack{\mathrm{AG}\\ \mathrm{TC}}} + 2\ T_{\substack{\mathrm{GA}\\ \mathrm{CT}}} \right)$$

We obtain for $s_{\substack{\mathrm{AATCAG}\\ \mathrm{TTAGTC}}}$ the same expression than that given by eq.(1), despite the fact that $f_{\substack{\mathrm{AT}\\ \mathrm{TA}}}$ does not appear in F' and f' and is therefore left unused in eq.(21). Let us now examine the reverse procedure, i.e. how to obtain the frequencies of this polymer, if its spectrum is known. The fitting of $s_{\substack{\mathrm{AATCAG}\\ \mathrm{TTAGTC}}}$ by the T' matrix according to eq.(20) will give for the frequencies appearing in f' the values :

$$f_{\substack{\mathrm{AA}\\ \mathrm{TT}}} = 1/6,\ f_{\substack{\mathrm{CA}\\ \mathrm{GA}}} = 1/6,\ f_{\substack{\mathrm{AG}\\ \mathrm{TC}}} = 1/6,\ f_{\substack{\mathrm{GA}\\ \mathrm{CT}}} = 2/6$$

and zero for $f_{\substack{\mathrm{TA}\\ \mathrm{AT}}}$, $f_{\substack{\mathrm{AC}\\ \mathrm{TG}}}$, $f_{\substack{\mathrm{CG}\\ \mathrm{GC}}}$ and $f_{\substack{\mathrm{GG}\\ \mathrm{CC}}}$. The frequency $f_{\substack{\mathrm{AT}\\ \mathrm{TA}}}$ is missing, but it can be obtained from eq.(11)

$$f_{\substack{\mathrm{AT}\\ \mathrm{TA}}} = 0 + 0 + 1/12 - 1/12 + 2/12 = 1/6$$

Similarily eq.(12) gives $f_{\substack{\mathrm{GC}\\ \mathrm{CG}}}$

$$f_{\substack{\mathrm{GC}\\ \mathrm{CG}}} = 0 + 0 + 1/12 + 1/12 - 2/12 = 0$$

Practical computations

Any set of two frequencies chosen dependent leads to the same numerical result ; however, once a set has been chosen, it must be kept unchanged throughout all further computations. We use $f_{\substack{AT\\TA}}$ and $f_{\substack{GC\\CG}}$ as the dependent frequencies, which define the Z matrix as given in eq. (15). The computation proceeds in three steps :

1) Computation of the T' matrix : The F matrix has to be built up according to the rules given by eq.(3) and (4). $f_{\substack{AT\\TA}}$ and $f_{\substack{GC\\CG}}$ are left unused ; as a result, unless these two frequencies are 0 for a given DNA, the corresponding column of F' will not sum to 1. Given a set, S, of m CD spectra ($m \geqslant 8$) of DNA's of known frequencies, T' is obtained according to eq.(18a) or (18b), respectively. In order to avoid computer rounding-off errors, the expression $F'^t.(F'.F'^t)^{-1}$ in eq.(18b) must be computed before its multiplication with S.

2) Computation of the frequencies from a CD spectrum : the eight independent frequencies are obtained according to eq.(20) and the ten frequencies are obtained by the product $f = Z.f'$ (eq.(13)) ; therefore the scalar product matrix $(T'^t.T')$ needs to be computed only once for a given input set S.

3) Computation of a CD spectrum from the frequencies : The f' column vector is built up as indicated for the F' matrix (see 1) above) ; the desired CD spectrum is given by the product : $s = T'.f'$ (eq.(19)).

The computations are performed in our laboratory on a Digital PDP-12 computer. Core memory is 12K 12 bit words ; programs are written in machine language. The computation of the T' matrix, plus $(T'^t.T')^{-1}$, takes around 3 minutes for 240 wavelengths, the computation of the 10 unknown frequencies of a DNA takes 30 seconds. A 9 significative digits precision (36 bits mantissa) is used for all matrix computations ; this provides a precision of 10^{-5} for the recomputed frequencies. A 10-inch TV screen is used for on-line display of the CD spectra and a remote 7004B Hewlett-Packard X-Y recorder for drawing the figures.

RESULTS

Using the procedure described, we have first extracted the T' matrix from the spectra of seven DNA's and two polymers : poly(dGC).poly(dGC) and poly(dG).poly(dC). These 9 DNA's make up input set I. As independent test cases, we then tried to estimate the frequencies of the satellite DNA's of *Drosophila virilis*. The three CD spectra were quite well reproduced (fig.1),

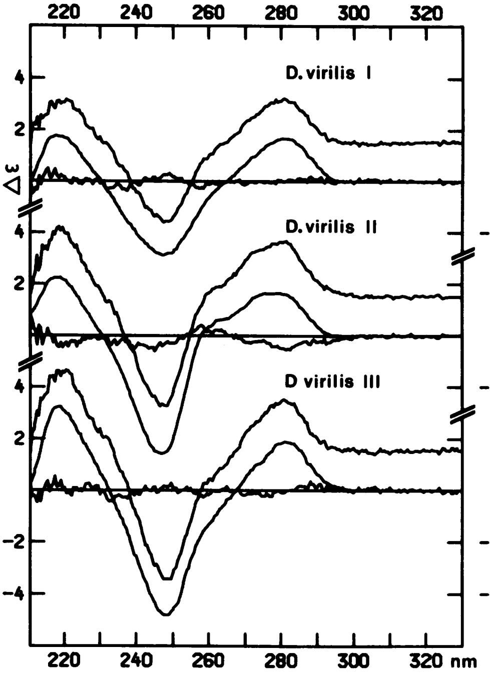

Fig. 1. CD spectra of *Drosophila virilis* satellite DNA I, II and III reconstructed by least square fit using the T' matrix extracted from set I (two polymers and seven DNA). Fits are moved up for convenience; difference between spectra and fits are also plotted. Original CD spectra are taken from ref. 12 .

the frequencies, however, were not all correct (Table I). Some negative frequencies were found, up to - 0.05, in place of nul frequencies (this accounts for the anomalously low GC content found for satellite I and III). Neverthe- ness, the obtained frequencies for satellite III were in closer agreement than those determined by Gray and Gall[12]. It must be noted that, most of the time, the frequencies of two opposite first neighbour configurations, say $\frac{AC}{TG}$ and $\frac{CA}{GT}$, deviate from the correct values, one by excess, the other one by defect. We attributed this to the fact that, possibly, the CD contributions of two such first neighbour configurations were incorrectly differenciated since their frequencies are most of time very similar for natural DNA's, or, whenever significantly different, they represent only a very low fraction of the total frequencies.

We therefore reversed the problem by including the three *Drosophila* satellite DNA's CD spectra into the input set and removing three DNA's. The four non satellite DNA's retained were chosen so that they represented an

Table I : First neighbour frequencies obtained for *Drosophila virilis* satellite DNA I, II and III using the T' matrix extracted from set I . The two GC content values are computed from the "found" and "true" frequencies, respectively. Third column is the difference "found" minus "true".

	D. virilis I			*D. virilis II*			*D. virilis III*		
	found	true[18]		found	true[18]		found	true[18]	
(dAA).(dTT)	.3276	.2867	. 409	.3709	.2867	. 842	.4313	.4286	. 27
(dAT).(dAT)	.1235	0	.1235	.1429	.1429	0	.1559	.1429	. 130
(dTA).(dTA)	. 924	.1429	-. 505	.1486	.2867	-.1381	.1325	.1429	-. 104
(dAC).(dGT)	.1286	.2867	-.1581	. 554	.1429	-. 875	. 825	.1429	-. 604
(dCA).(dTG)	.1686	.1429	. 257	. 527	0	. 527	.1260	.1429	-. 169
(dAG).(dCT)	. 978	.1429	-. 451	.1080	.1429	-. 349	. 754	0	. 754
(dGA).(dTC)	.1201	0	.1201	. 991	0	. 991	. 789	0	. 789
(dGC).(dGC)	-. 129	0	-. 129	-. 65	0	-. 65	-. 36	0	-. 36
(dCG).(dCG)	-. 217	0	-. 217	-. 96	0	-. 96	-. 236	0	-. 236
(dGG).(dCC)	-. 240	0	-. 240	. 386	0	-. 386	-. 552	0	-. 552
GC %	19.89	28.67		18.00	14.29		9.90	14.29	

about equally spaced range of GC content : *Tetrahymena pyriformis* (24%), *Haemophilus influenzae* (38%), phage λ (50%), *Micrococcus lysodeikticus* (72%)(Fig. 2) Table II gives the reobtained frequencies for the nine DNA's of the input set II. The agreement between the input and re-computed frequencies is exeptionnaly good for the three satellite DNA's ; this is not surprising since most of the information for the computation of the CD signals of $\substack{AA\\TT}$, $\substack{AT\\TA}$, $\substack{TA\\AT}$, $\substack{AC\\TG}$, $\substack{CA\\GT}$ and $\substack{AG\\TC}$ come from these DNA's. For the four other DNA's the obtained frequencies fall very close to the input frequencies, the maximal discrepency was .0134 for $f_{\substack{TA\\AT}}$ in *Haemophilus influenzae.*

As independent test cases, we have then tried to estimate first the frequencies of four non satellite DNA's, the GC contents of which fall between those of the DNA's of the input set. These test DNA's were : *Clostridium perfringens* (28% GC), *Bacillus subtilis* (44% GC), wheat germ (45% GC) and *Serratia marcescens* (57% GC). For all four DNA's the fits were reasonably correct (fig. 3) and the frequencies were obtained with less than 0.01 absolute error (65%), or less than 0.02 error (20%) or less than 0.03 error (15%) (table III). Only one negative frequency, $f_{\substack{GC\\CG}}$ = - 0.003, is obtained for *Clostridium perfringens*, this, however, corresponds only to an absolute error of - 0.008. The computed frequencies obey the general rules $f_{\substack{AT\\TA}} > f_{\substack{TA\\AT}}$ and $f_{\substack{GC\\CG}} > f_{\substack{CG\\GC}}$; for *Clostridium perfringens* the first rule does not hold, and the value found for $f_{\substack{TA\\TA}}$ is off by 0.03 ; this is the largest error found for the frequencies of these four DNA's.

The crab satellite DNA's studied by Gray and Skinner[13] provided a fur-

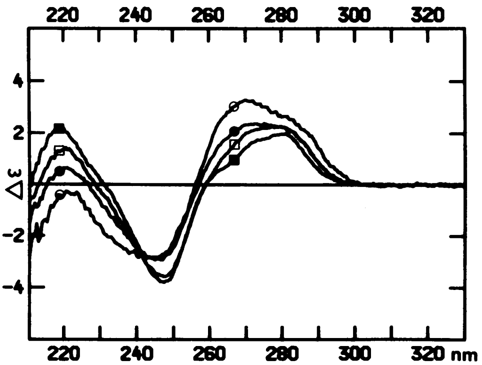

Fig. 2. CD spectra of the four quasi random sequence DNA's of input set II (see Table II)))

■ *Tetrahymena pyriformis*

□ *Haemophilus influenzae*

● *Phage*

○ *Micrococcus lysodeikticus*

ther test of the T' matrix obtained from set II. We have found that only the CD spectrum of *Callinectes sapidus* was correctly reproduced. (Fig. 4) The three lower frequencies were found for the configurations $\substack{GA\\CT}$, $\substack{GC\\CG}$ and $\substack{CG\\GC}$. These configurations are also absent in the three *Drosophila virilis* satellite DNA's. The 0.043 value found for $f_{\substack{GG\\CC}}$ may be not significative. Gray and Skinner[13] have obtained 0.061 for $f_{\substack{GC\\CG}}$. It is interesting to note that, although the *Callinectes sapidus* spectrum ressembles much more the spectrum of *Tetrahymena pyriformis* than those of *Drosophila virilis* satellite DNA's, the first neighbour analysis indicates a distribution of frequencies reminiscent of that of these satellite DNA's i.e. the absence of $\substack{GC\\CG}$ and $\substack{CG\\GC}$. Gray and Skinner[13] had also pointed out the similarity of the CD spectrum of *C. Sapidus* DNA with that of poly (dAAT).poly(dATT) ; this is also reflected in the computed frequencies (Table III) : a large amount of $\substack{AA\\TT}$, about as abundant as $\substack{AT\\TA}$ and $\substack{TA\\AT}$ together. If we assume that any value above 2% is significant, one can make an assumption concerning the dimer composition of this DNA, similar as it is done in amino acid analyses. This would yield : $\substack{AA\\TT}$: 15, $\substack{AT\\TA}$: 7, $\substack{TA\\AT}$: 9, $\substack{AC\\TG}$: 5, $\substack{CA\\GT}$: 3, $\substack{AG\\TC}$: 3, $\substack{GA\\CT}$: 1, $\substack{GG\\CC}$: 2. Satellite DNA's are frequently composed of partially or completely repetitive sequences [12,13]; we have attempted to arrange these dimers in a repetitive sequence of 3 x 15 base pairs, with only one transition $\substack{G\\C} \rightarrow \substack{A\\T}$ every third block. Evidently within each of the three blocks of 14 bases, no unique sequence can be given. One possible sequence would be :

```
 45 1             15               30               45
 gGTAAAAAATATACAG A TAAAAAATATACAG G TAAAAAATATACAG G T
 cCATTTTTTATATGAC T ATTTTTTATATGTC C ATTTTTTATATGTC C A
```

or a sequence rich in $\substack{AATA\\TTAT}$ sequences as suggested by Gray and Skinner[13] :

Table II : Recomputed first neighbour frequencies of the nine DNA making up set II

	D. virilis I			*D. virilis II*			*D. virilis III*		
	found	true[18]		found	true[18]		found	true[18]	
(dAA).(dTT)	.2864	.2867	-. 3	.2857	.2867	-. 10	.4289	.4286	. 3
(dAT).(dAT)	-. 5	0	-. 5	.1435	.1429	. 6	.1427	.1429	-. 2
(dTA).(dTA)	.1429	.1429	0	.2850	.2867	-. 17	.1438	.1429	. 9
(dAC).(dGT)	.2865	.2867	-. 2	.1422	.1429	-. 7	.1432	.1429	. 3
(dCA).(dTG)	.1426	.1429	-. 3	. 6	0	. 6	.1423	.1429	-. 6
(dAG).(dCT)	.1428	.1429	-. 1	.1421	.1429	-. 8	. 3	0	. 3
(dGA).(dTC)	-. 1	0	-. 1	. 6	0	. 6	-. 8	0	-. 8
(dGC).(dGC)	-. 5	0	-. 5	. 3	0	. 3	-. 3	0	-. 3
(dCG).(dCG)	. 0	0	. 0	. 3	0	. 3	-. 4	0	-. 4
(dGG).(dCC)	. 0	0	. 0	-. 2	0	-. 2	. 3	0	. 3
GC %	28.53	28.67		14.31	14.29		14.21	14.29	

	Tetrahymena pyriformis			*Haemophilus influenzae*			Phage λ		
	found	true[19]		found	true[16]		found	true[16]	
(dAA).(dTT)	.3282	.3280	. 2	.2302	.2320	-. 18	.1460	.1460	. 0
(dAT).(dAT)	.1313	.1330	-. 17	. 981	. 950	. 31	. 650	. 680	-. 30
(dTA).(dTA)	.1315	.1270	. 45	. 596	. 730	-. 134	. 576	. 470	. 106
(dAC).(dGT)	. 714	. 700	. 14	. 918	. 970	-. 52	.1135	.1090	. 45
(dCA).(dTG)	. 850	. 890	-. 40	.1442	.1340	. 102	.1317	.1400	-. 83
(dAG).(dCT)	.1031	.1010	. 21	. 919	. 990	-. 71	.1150	.1090	. 60
(dGA).(dTC)	. 891	. 930	-. 39	.1166	.1060	. 106	.1116	.1200	-. 84
(dGC).(dGC)	. 197	. 200	-. 3	. 574	. 530	. 44	. 758	. 790	-. 32
(dCG).(dCG)	. 59	. 80	-. 21	. 436	. 380	. 56	. 650	. 690	-. 40
(dGG).(dCC)	. 348	. 330	. 18	. 666	. 730	-. 64	.1189	.1130	. 59
GC %	23.47	23.75		38.98	38.20		49.55	50.00	

	Micrococcus lysodeikticus			poly(dGC).poly(dGC)			poly(dG).poly(dC)		
	found	true[20]		found	true		found	true	
(dAA).(dTT)	. 283	. 280	. 3	. 0	0	. 0	. 0	0	. 0
(dAT).(dAT)	. 222	. 220	. 2	. 0	0	. 0	. 0	0	. 0
(dTA).(dTA)	. 72	. 90	-. 18	. 0	0	. 0	. 0	0	. 0
(dAC).(dGT)	.1102	.1110	-. 8	. 0	0	. 0	. 0	0	. 0
(dCA).(dTG)	.1066	.1050	. 16	. 0	0	. 0	. 0	0	. 0
(dAG).(dCT)	. 979	. 990	-. 11	. 0	0	. 0	. 0	0	. 0
(dGA).(dTC)	.1315	.1300	. 15	. 0	0	. 0	. 0	0	. 0
(dGC).(dGC)	.1259	.1240	. 19	.5000	.5000	. 0	. 0	0	. 0
(dCG).(dCG)	.1445	.1440	. 5	.5000	.5000	. 0	. 0	0	. 0
(dGG).(dCC)	.2256	.2270	-. 14	. 0	0	. 0	1.000	1.000	. 0
GC %	71.92	71.75		100.0	100.0		100.0	100.0	

```
   45 1               15                30                45
G G TAATAACAAATAAG A TAATAACAAATAAG G TAATAACAAATAAG G T
C C ATTATTGTTTATTC T ATTATTGTTTATTC C ATTATTGTTTATTC C A
```

Both sequences would satisfy the distribution listed in Table III (given under the heading "true"). It would be interesting whether the c_0.t value of *C. sapidus* satellite DNA would reflect the above suggestion.

As mentioned in a previous section, we do not obtain individually all the elementary CD signals, but only eight linear combinations of them ; nevertheless the spectrum of a given polymer can be computed. This spectrum would be that of the corresponding polymer, if it were in the same structure as the DNA's of the input set. For instance, poly(dA).poly(dT) differs from the usual polymer spectrum, and is also very different from that computed by Allen et

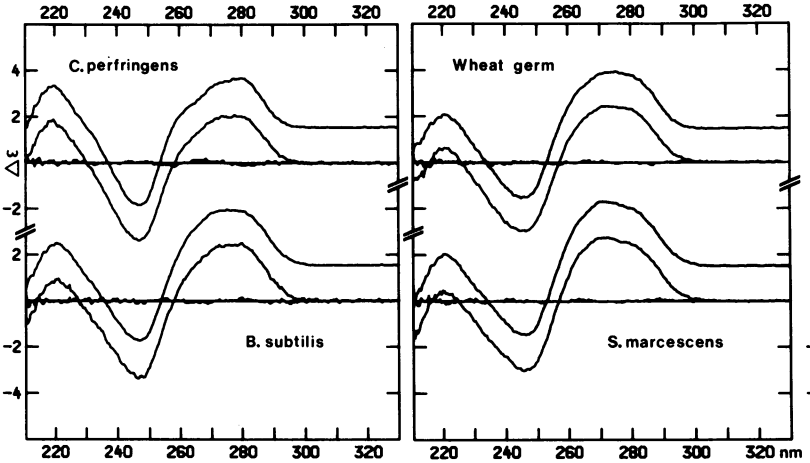

Fig. 3. CD spectra of four quasi random sequence DNA's reconstructed by least square fit using the T' matrix extracted from set II (see Table II). Computed frequencies are given in Table III.

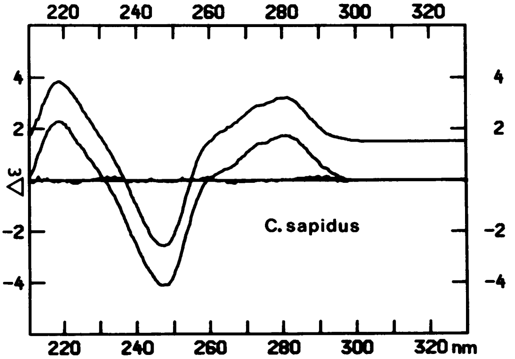

Fig. 4. CD spectrum of *Callinectes sapidus* satellite DNA reconstructed by least square fit using the T' matrix extracted from set II (see Table II). Computed frequencies are given in Table III. Original CD spectrum is taken from ref. 13.

Table III : Computed first neighbour frequencies of five test case DNA's using the T' matrix extracted from set II (see Table II)

	Clostridium perfringens			*Bacillus subtilis*			wheat germ		
	found	true[20]		found	true[19]		found	true[16]	
(dAA).(dTT)	.2858	.2789	. 78	.1806	.1870	-. 64	.1391	.1610	-. 219
(dAT).(dAT)	.1209	.1210	-. 1	. 887	. 800	. 87	. 696	. 750	-. 54
(dTA).(dTA)	. 971	.1240	-. 269	. 556	. 520	. 26	. 682	. 580	. 102
(dAC).(dGT)	. 728	. 790	-. 62	. 907	. 960	-. 53	.1187	.1090	. 97
(dCA).(dTG)	.1186	. 930	. 256	.1358	.1350	. 8	.1278	.1380	-. 102
(dAG).(dCT)	.1219	.1240	-. 21	.1068	.1150	-. 82	.1272	.1270	. 2
(dGA).(dTC)	.1235	.1020	. 215	.1300	.1320	-. 20	.1208	.1330	-. 122
(dGC).(dGC)	. 191	. 230	-. 39	. 679	. 610	. 69	. 660	. 500	. 160
(dCG).(dCG)	-. 30	. 50	-. 80	. 569	. 500	. 69	. 582	. 390	. 192
(dGG).(dCC)	. 434	. 520	-. 86	. 881	. 920	-. 39	.1042	.1100	-. 52
GC %	27.78	27.90		44.45	44.20		47.58	45.25	

	Serratia marcescens			*Callinectes sapidus*		
	found	true[20]		found	true (see text)	
(dAA).(dTT)	.1215	.1090	. 125	.3407	.3333	. 73
(dAT).(dAT)	. 548	. 590	-. 42	.1504	.1556	-. 51
(dTA).(dTA)	. 287	. 350	-. 63	.1896	.2000	-. 104
(dAC).(dGT)	. 876	.1040	-. 164	.1048	.1111	-. 63
(dCA).(dTG)	.1239	.1360	-. 121	. 648	. 667	-. 18
(dAG).(dCT)	.1247	.1020	. 227	. 642	. 667	-. 25
(dGA).(dTC)	.1406	.1170	. 236	. 257	. 222	. 35
(dGC).(dGC)	. 992	.1050	-. 58	. 82	0	. 82
(dCG).(dCG)	. 891	. 960	-. 69	. 89	0	. 89
(dGG).(dCC)	.1300	.1380	-. 80	. 427	. 444	-. 17
GC %	55.67	56.80		18.95	17.78	

al.[9]. The spectra obtained for poly(dAC).poly(dGT) and poly(dAG).poly(dCT) (not shown) ressemble much more the actual polymer spectra ; the main features of the spectra are found at the same wavelengths.

DISCUSSION

Which kind of DNA is the most suitable to make up the input set ? All DNA's used to make up the input set have to be in the same structure, so that the geometry and consequently the CD signal of all first neighbour base paired subunits are the same. It should be apparently advisable to use only natural DNA's of quasi random sequence. If we use only such DNA's, the amount of information present in the input set will be not large enough : quasi-random DNA's have their frequencies of opposite first neighbour $(\frac{\mathrm{AC}}{\mathrm{TG}}$ and $\frac{\mathrm{CA}}{\mathrm{GT}})$ always very close to each other. If these frequencies were rigorously equal, then the order of linear dependence would be no more 8 but 6 (eq. (9) and (10) become identities and four relations of the type $f_{\substack{AC\\TG}} = f_{\substack{CA\\GT}}$ appear). It is therefore hopeless to obtain a correct determination of the eight dependent frequencies, if we use only DNA's of quasi random sequence. On the other hand, if we use for the input set, DNA's of highly non random sequence i.e. polymers the repetitive unit of which contains only a few bases, say up to three, we have to deal with a structure problem : most of polymers of simple sequence have their CD spectra not consistent among each other[10,11], nor consistent with those of DNA's of random sequence[9-13]. For these reasons, a set such as the one we have used (set II), made of both satellite DNA's and DNA's of quasi random sequence can give good results for the computation of the frequencies of other quasi random sequence DNA's.

We have performed the orthogonalisation [21] of the spectra making up the set II. The 8th orthogonal form was weak, but still significative, while the 9th one was only random noise. In any first neighbour analysis computation, it is advisable to check first whether the order of linear dependence experimentally determined by orthogonalisation equals the order of dependence predicted by the first neighbour hypothesis.

The mathematical interpretation of the first neighbour formalism of Gray and Tinoco. The first neighbour hypothesis must be interpreted as follows : A set of m (m > 10) CD spectra, s_j, digitalized at n wavelengths

$$s_j = \sum_{i=1}^{10} T_i . f_{ij}$$

would define a vector sub-space E (10 dimensions), one base of which is the T matrix. This base is the most interesting since it is the actual base i.e. the base every element of which has a physical meaning : the 10 first neighbour interaction CD signals. In fact the frequencies f_{ij} are, for all DNA's, constrained by the two same linear relations (eq.(9) and (10)) ; therefore the s_j vectors define only a sub-space E' (8 dimensions) of E

$$s_j = \sum_{i=1}^{8} T'_i . f'_{ij}$$

one base of this sub-space E' is the T' matrix. T' and T are linked by the relations listed as eq. (16b).

It is very important to point out that this does not have as a consequence, nor necessitates, that the T matrix itself, would be singular, i.e. that there would exist some linear relations between the T_i vectors as it has been assumed by Allen *et al.*[9] (this problem is discussed below). On the contrary, it should be clear that, although the order of the actual physical basis T is 10, the order 8 basis T' is sufficient to express, or to fit, any double stranded DNA CD spectrum because :

1) T' contains in itself the constraint relations : the T_i' vectors are linear combinations of the T_i vectors, these combinations being a consequence of the constraint relations over the frequencies (see eq.(17)) ;

2) the frequencies of the DNA the CD spectrum of which is expressed, or fitted, obey also the same constraint relations.

It must be noted that the various column vectors T_i' of the T' matrix can be regarded as the CD spectra of eight "polymers" of simple sequence ; several of these polymers have no physical equivalent since some frequencies are negative. Nevertheless, if one examines the frequencies of these polymers, it can be noted that : 1) the eight sequences are linearly independent (i.e. the frequencies in any one polymer cannot be expressed as a linear combination of the frequencies in the other seven[8]) ; 2) although some frequencies are negative, all sequences obey the constraints equations (5) through (10). These polymers therefore fulfill the requirements of the first neighbour hypothesis of Gray and Tinoco[8] : they make up a set of eight polymers of independent sequence which is sufficient to obtain the non-independent frequencies of a DNA from the analysis of its CD spectrum.

The problem of second and higher neighbour interactions. This problem has been frequently evoked : incorrect first neighbour spectral agreement has been often attributed to the presence of significative second (or higher) neighbour interactions. First, it must be noted that both the mathematical treatment we propose and that presented by Allen *et al.*[9] do not fulfill exactly the first neighbour hypothesis of Gray and Tinoco. Equation (18a), or (18b), which are used to compute the T' matrix shows that *no part* of the input spectra is neglected. If there are higher than first neighbour interactions, their contributions will be averaged and show up as additional - but indiscernable - terms in the column vectors of the T matrix (and consequently of the T' matrix too). In the limit of exactness of the results obtained in the present study, we cannot attribute the errors in the recomputed frequencies to the presence of second and higher neighbour interactions. We must conclude on the correctness of the first neighbour hypothesis of Gray and Tinoco[8] for all practical purposes.

Another reason that could explain a poor first neighbour spectral agreement is a difference in the geometric configuration of the various DNA's considered. Under favorable case, one can decide which of these two phenomena second neighbour interactions or different geometry, is responsible for the imperfect result obtained. Let us consider, as an example, an equation proposed by Arnott[14] to express the CD spectrum of *Drosophila virilis* ; this satellite DNA approximates to poly (dAATAT).poly(dATATT)[18].

$$ {}^{s}\!\begin{smallmatrix}\text{AATAT}\\ \text{TTATA}\end{smallmatrix} = \frac{3}{5}\; {}^{s}\!\begin{smallmatrix}\text{AAT}\\ \text{TTA}\end{smallmatrix} + \frac{2}{5}\; {}^{s}\!\begin{smallmatrix}\text{AT}\\ \text{TA}\end{smallmatrix} $$

If one lists the first, second and third neighbour frequencies of these DNA's, one can check that the above equation fulfills the first, second and third neighbour hypothesis. Therefore, even third neighbour interactions, if present, would be taken into account. Then, if the spectrum obtained for ${}^{s}\!\begin{smallmatrix}\text{AATAT}\\ \text{TTATA}\end{smallmatrix}$ by the above linear combination is not correct, we have to consider that we deal with three DNA's of different geometry.

Comparison of the analytical method used with that of Allen *et al.*[9]. The procedure we use for the computation of the first neighbour frequencies is different from that presented by Allen *et al.*[9] These authors make use of two reentrant conditions over the columns of the T matrix :

$$ {}^{T}\!\begin{smallmatrix}\text{AT}\\ \text{TA}\end{smallmatrix} + {}^{T}\!\begin{smallmatrix}\text{AC}\\ \text{TG}\end{smallmatrix} + {}^{T}\!\begin{smallmatrix}\text{AG}\\ \text{TC}\end{smallmatrix} = {}^{T}\!\begin{smallmatrix}\text{TA}\\ \text{AT}\end{smallmatrix} + {}^{T}\!\begin{smallmatrix}\text{CA}\\ \text{GT}\end{smallmatrix} + {}^{T}\!\begin{smallmatrix}\text{GA}\\ \text{CT}\end{smallmatrix} \qquad (22) $$

$$ {}^{T}\!\begin{smallmatrix}\text{GC}\\ \text{CG}\end{smallmatrix} + {}^{T}\!\begin{smallmatrix}\text{GA}\\ \text{CT}\end{smallmatrix} + {}^{T}\!\begin{smallmatrix}\text{GT}\\ \text{CA}\end{smallmatrix} = {}^{T}\!\begin{smallmatrix}\text{CG}\\ \text{GC}\end{smallmatrix} + {}^{T}\!\begin{smallmatrix}\text{AG}\\ \text{TC}\end{smallmatrix} + {}^{T}\!\begin{smallmatrix}\text{TG}\\ \text{AC}\end{smallmatrix} \qquad (23) $$

There is no physical evidence that could explain why these relations would hold. There is also no justification to deduce these relations from the constraint relations (5) and (6) as done by Allen *et al.*[9] Equations (5) and (6) result of a topological constraint : both strands of DNA are linear arrays of bases, the number of time a given base is preceeded by any ot the three others equals the number of time the same base is followed by any of the three others[16]. Equations (22) and (23) do not exist, the 10 elementary CD signals are to be considered as being linearly independent, unless it could be demonstrated that some relations link these signals. In a detailed demonstration[22], we have shown that, unfortunatly, the first neighbour hypothesis does not allow an experimental verification of linear constraint over elementary CD signals.

It has been recently attempted to obtain informations on the first neighbour specificities of actinomycin-DNA binding with the use of the nearest neighbour hypothesis. Allen *et al.*[23,24] have examined the perturbations of the various elementary CD signals due to the specific binding of actinomycin on certain first neighbour configurations. However, since eq. (22) and (23) do not hold, one does not obtain the actual elementary CD signals, but only linear combinations of them[22]. Furthermore, it is incorrect to assume, *a priori*, that the perturbations of the elementary CD signals due to actinomycin binding would also obey eq.(22) and (23)[25]. The first neighbour hypothesis cannot give access to any information linked to *one* first neighbour configuration, except $\frac{\mathrm{AA}}{\mathrm{TT}}$ and $\frac{\mathrm{GG}}{\mathrm{CC}}$.

CONCLUSION

The main problem encountered in the practical application of the first neighbour formalism does not reside in the mathematical treatment which is in fact very simple, but in the first neighbour hypothesis itself. As a matter of fact, the linear dependence is hardly reached, even when using 240 wavelengths per spectrum. If we want to use the first neighbour hypothesis to obtain even more accurately the first neighbour frequencies of quasi-random sequence DNA's, we have to hope that improved instrumentation may help us. On the other hand the first neighbour hypothesis seems to open great possibilities in the field of satellite DNA's; such DNA's have their frequencies highly different from the random values and therefore provide a high amount of information as input data . When more sequences of satellite DNA's become available, one can hope that the frequencies, or even the sequence in favourable cases, could be obtained from the CD data analysis according to the first neighbour hypothesis.

ADDITION (30 march 1978).

After termination of this manuscript, two papers by Gray *et al.*[26,27] appeared concerning very similar work as reported here. These authors [26] use the usual analytical procedure first described by Allen *et al.*[9] to compute matrix T (m x 10, i.e. the 10 individual CD contributions) using the constraint relations of eq.(22) and (23). It is noteworthy that they state about this matrix T : "The ten spectral components are convenient for computation, *but have no direct physical significance*" [26].

The results Gray *et al.* [26,27] present are considerable improvement over their previous papers[9-12], but the average absolute error of nearest neighbour frequencies of natural DNA's is between 3 and 7 percent, with peaks exceeding 10 percent, although the spectral fits are reasonably close.

The results of these two papers do not change our conclusions about the feasibility of computing nearest neighbour frequencies without knowing the T matrix[17].

[a] Collaborateur temporaire de thèse du CEA. Present address : Max-Planck Institut für experimentelle Medizin, Abt. Chemie, D-34-Göttingen, G.F.R. This paper will be part of the "Thèse de Doctorat d'Etat-ès-Sciences Naturelles" to be submitted to Université Paris VII.

Acknowledgements : We wish to thank the following colleages for gifts of DNA samples : G. Bernardi (Paris), D. Hayes (Paris), L. Hirschbein (Orsay), M. Leng (Orléans), A. Levine (Gif) and S. Litvak (Bordeaux). Special thanks are due to Mrs. P. Sentenac (Orsay), Mrs E. Abillon and M.L. Brehamet (Saclay) for stimulating discussions concerning the mathematical aspects of this paper.

REFERENCES

1) Felsenfeld, H. & Hirschman, G. (1966) J. Mol. Biol. 13, 407-427.
2) Samejima, T. & Yang, J.T. (1965) J. Biol. Chem. 240, 2094-2100.
3) Zimmer, C., Luck, G., Venner, H, & Frič, J. (1968) Biopolymers 6, 563-574.
4) Courtois, Y., Fromageot, P. and Guschlbauer, W. (1968) European J. Biochem. 6, 493-501.
5) Gratzer, W.B., Hill, L.R., and Owen, R.J. (1970) European J. Biochem. 15, 209-214.
6) Bernardi, G. & Timascheff, S.N. (1970) J. Mol. Biol. 48, 43-52.
7) Wells, R.D., Larson, J.E., Grant, R.C., Shortle, B.E. & Cantor, C.R. (1970) J. Mol. Biol. 54, 465-497.
8) Gray, D.M., & Tinoco, I. Jr. (1970) Biopolymers, 9, 223-244.
9) Allen, F.S., Gray, D.M., Roberts, G.P. & Tinoco, I. Jr. (1972) Biopolymers 11, 853-879.
10) Gray, D.M., Ratliff, R.L. & Williams, D.L. (1973) Biopolymers 12, 1233-1245.
11) Allen, F.S. & Daub, G.W. (1974) Biopolymers, 13, 241-255.
12) Gray, D.M. & Gall, J.G. (1974) J. Mol. Biol. 85, 665-679.
13) Gray, D.M. & Skinner, D.M. (1974) Biopolymers 13, 843-852.
14) Arnott, S.(1975) Nucleic Acids Res. 2, 1493-1502.
15) Marck, C. & Thiele, D. (1978) Nucleic Acids Res. 5, 1017-1028..
16) Josse, J., Kaiser, A.D. & Kornberg, A. (1961) J. Mol. Biol. 236, 864-875.
17) Marck, C. & Guschlbauer, W. (1978) Comptes Rendus Acad. Sci. (Paris), 286D, 713-716.
18) Gall, J.G. & Atherton, D.A. (1974) J. Mol. Biol. 85, 633-664.
19) Swartz, M.N., Trautner, T.A. & Kornberg, A. (1962) J. Biol. Chem. 237, 1961-1967.
20) Russel, G.J., McGeoch, D.J. Elton, R.A. & Subak-Sharpe, J.H. (1973) J. Molec. Evol. 2, 277-292.
21) Marck, C., Schneider, C. & Brehamet, L. (1978) Biopolymers 17, in press.
22) Marck, C. (1978) Comptes Rendus Acad. Sci. (Paris), 286D
23) Allen, F.S., Moen, R.P. & Hollstein, U. (1976) J. Am. Chem. Soc. 98, 864-865.
24) Allen, F.S., Jones, M.B. & Hollstein, U. (1977) Biophys. J. 20, 69-78.
25) Marck, C. & Guschlbauer, W. (1978) Biophys. J. 21, in press.
26) Gray, D.M., Hamilton, F.D. & Vaughan, M.R. (1978) Biopolymers, 17, 85-106.
27) Gray, D.M., Lee, C.S. & Skinner, D.M. (1978) Biopolymers 17, 107-114.