Sequence Analysis of Oligodeoxyribonucleotides by Mass Spectrometry. 2. Application of Computerized Pattern Recognition to Sequence Determination of Di-, Tri-, and Tetranucleotides[†]

D. R. Burgard, S. P. Perone,* and J. L. Wiebers*

ABSTRACT: A novel strategy for the sequence analysis of oligodeoxyribonucleotides has been devised which is based upon the analysis of intact underivatized oligonucleotides by mass spectrometry followed by interpretation of the mass-spectral data by computerized pattern-recognition techniques. The pyrolytic and electron-impact conditions of the mass spectrometer permit the cleavage of oligonucleotides of varying chain length and composition, yielding reproducible fragmentations and characteristic *m/e* values which can be used to reveal purine and/or pyrimidine base sequence information. The selection of optimum features (which are the ratios of peak heights of specific ions, or the linear combination of such ratios) has been done by an interactive feature selection method employing multidimensional *k* nearest-neighbor analysis and two-dimensional feature-space plots (nonlinear mappings) of the mass-spectral data. Features have been found which allow 100% classification accuracy in predicting the 5′ and 3′ terminus of all of the dinucleotides commonly found in DNA. Other specific features have been found which indicate adjacent nucleotides within a tetranucleotide. Knowledge of the adjacent nucleotide pairs present, in conjunction with the information as to the 3′ or 5′ position of the residues in each pair, permits the reconstruction of the sequence of the tetranucleotide.

The method described in the preceding paper of this issue for distinguishing sequence isomers of dinucleoside monophosphates suggested that mass-spectral data for molecules of greater complexity could be interpreted by applying more sophisticated mathematical analysis. The use of the technique of computerized pattern recognition to solve multivariant problems in chemistry has been reviewed recently (Kowalski, 1975). This review emphasizes how powerful this approach can be in the determination of molecular structural features directly from spectral data. Some examples of the application of pattern recognition to mass-spectral data are the studies of Isenhour and Jurs (1971), Justice and Isenhour (1974), Kowalski and Bender (1972b), Tunnicliff and Wadsworth (1973), and Schechter and Jurs (1973). To our knowledge, the application of pattern recognition to the problem of sequence analysis of biological molecules has not been investigated previously. We have recently reported results which demonstrate that the technique can be used advantageously, however, to reveal sequence information on small oligonucleotides (Perone et al., 1975; Wiebers et al., 1975a,b).

The term "pattern recognition" is multifaceted and pattern recognition methods have been applied to numerous and varied problems (Andrews, 1972). For the work reported here, however, pattern recognition can be defined in simplified terms as a mathematical way of categorizing a set of observed data as a member of the "class" to which it belongs. This is accomplished by the selection of certain "features" that can distinguish between the different classes. For example, the mass spectra of two dinucleotide sequence isomers are superficially alike; however, there are subtle and obscure differences in the spectra which are presumably due to variations in fragmentation directly related to the structures of the two compounds. These differences can be identified within a complex data set by the application of pattern recognition.

As Kowalski (1975) has cogently pointed out, computerized pattern-recognition methods can be used to extend the ability of human pattern recognition. However, these methods should rely heavily on graphics for the presentation of results of complicated systems, since humans have a well-developed visual ability to recognize pictorial patterns and deviations from patterns. Consequently, there should be a strong interactive role of the scientist with the computer in data analysis. In our study, we have exploited this interactive approach. In addition, we have tried to utilize our experience and expertise in the interpretation of mass spectra of nucleic acids, and our knowledge of the structure and chemistry of oligonucleotides, to make educated guesses as to which ions in a spectrum are likely to yield useful information. These selected ions are then used to generate the features for the pattern-recognition analysis of the data. This communication presents evidence that certain features from mass spectra of oligonucleotides which have been selected (1) from structural considerations, and (2) by empirical feature-selection methods, can be used to determine the sequence of the purine and pyrimidine bases in dinucleotides, and that these features may be useful for the sequence analysis of longer-chain oligonucleotides.

Materials and Methods

Materials

Oligonucleotides. Deoxyribodinucleotides, the defined sequence deoxyribo- tri-, tetra-, hexa-, and octanucleotides, the deoxyribopentanucleotide with a 3′-ribo terminus, the deoxyribodinucleoside diphosphates with 3′-terminal phosphate, the cyclic deoxyribomonophosphate, and deoxyribodinucleotides with 5′-5′ linkage were all lyophilized ammonium salts and were obtained from Collaborative Laboratories, Waltham, Mass. Data sheets from Collaborative Laboratories indicated

[†] From the Department of Biological Sciences (J.L.W.), and the Department of Chemistry (D.R.B. and S.P.P.), Purdue University, West Lafayette, Indiana 47907. *Received May, 14, 1976.* This investigation was supported by Research Grants BMS-74-22213 and MPS 74-12762 from the National Science Foundation, and by the Office of Naval Research.

TABLE I: Composition of Features Found Useful for Class Separation of d-ApC and d-CpA.

Feature No.	Feature Composition (Ratios of *m/e* Values)
1	110/81
2	188/81
3	311/81
4	111/81
5	160/172
6	160/233
7	160/374
8	296/374

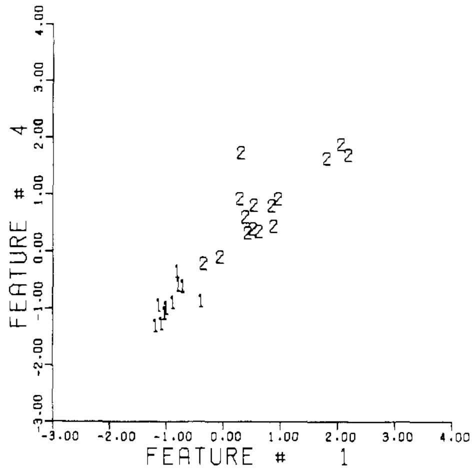

FIGURE 1: Feature space plot of features 1 and 4 (Table I) for d-ApC (class 1) and d-CpA (class 2).

the purity, based upon (1) R_f values from two paper chromatographic systems, (2) the characterization of the products following degradation by spleen or snake venom phosphodiesterase, and (3), when applicable, the characteristics of the products formed by removal of the 5′-terminal phosphate by bacterial alkaline phosphatase. Purity of the compounds was checked in our laboratory by thin-layer chromatography of the dinucleotides. For the longer oligonucleotides, it was important to check not only the purity, but to know that they were intact compounds at the time of mass-spectral analysis. This was done by using an ion-exchange chromatographic system (Ho and Gilham, 1973) modified to separate oligodeoxyribonucleotides of chain length up to 12 residues. The integrity of the compounds was confirmed with the exception of the larger oligonucleotides of repeating sequence, $(pApC)_4$, $(pCpA)_4$, $(pCpG)_4$, and $(pGpC)_4$, which gave three peaks in the chromatograms. In these cases, the major peak at the correct chain-length position was isolated, and the mass-spectral analysis was done on that material. For the oligonucleotides of defined sequence, the mass spectrum itself served as a check on purity, especially in those cases where protecting groups were used in the synthesis of the compounds. A mass-spectral study (Wiebers, 1976, manuscript in preparation) of nucleotides containing the common protecting groups used in the chemical synthesis of oligodeoxyribonucleotides has demonstrated that the presence of protecting groups on oligonucleotides is easily detected by mass-spectral analysis. No such contamination was evident in the compounds obtained from Collaborative Laboratories.

Computer Facilities. A Hewlett-Packard 2100S 16-bit digital computer with 32K words of memory was used. Peripheral devices included a teletype, high-speed paper tape punch, high-speed paper tape photoreader, digital storage oscilloscope, 2.2 million word moving head disk, Calcomp plotter, and a Centronics 306 line printer. All programs were written in disk-based Hewlett-Packard Fortran IV. Programs are available from the authors upon request.

Methods

Sample Preparation and Mass-Spectral Analysis. These procedures were as described in the preceding paper of this issue.

Nearest-Neighbor Pattern Classification. The *k*-nearest-neighbor (*k*NN) method of pattern recognition provides that an unknown pattern is classified according to a majority vote of its *k* nearest neighbors in *d*-dimensional feature space. Computationally, the Euclidean (or any other metric) distance between the single *d*-dimensional point representing the pattern under study and all other pattern points in *d* space must be calculated to find the *k* nearest neighbors. For a given unknown, the calculation may be summarized as in eq 1

$$\text{Distance}_j = \left[\sum_{i=1}^{n} (U_i - X_{ij})^2 \right]^{1/2} \qquad (1)$$

where n = number of dimensions, X_j = jth spectra in the data set, and U = unknown. For the work reported here, unknowns were obtained by use of the leave-one-out procedure (Kowalski and Bender, 1972b) and classification of the unknown was based upon its nearest neighbor ($k = 1$). The *k*NN method has an advantage over other classification algorithms, such as the learning machine, in that with the *k*NN procedure a large number of classes can be analyzed simultaneously.

Interactive Feature-Selection Method. This method (Pichler and Perone, 1974) utilizes an operator interactive approach in which a large number of potentially useful features for classification of patterns can be screened for the most relevant ones. A systematic approach is employed which includes one-dimensional (one feature) *k* nearest-neighbor analysis of all patterns. Subsequently, a computerized trial and error procedure is employed to find the best combination of a minimum number of features for accurate classification using the multidimensional *k*NN method. Two-dimensional feature-space plots (Kowalski, 1975) were used for verification and interpretation of the results obtained from *k*NN analysis.

The interactive feature-selection method was applied to this study on nucleotides in the following manner. Initially, mass spectra from multiple analyses on each compound were recorded, and the peak heights of all ions over background were obtained on ions from mass 100 to 320. Certain ions were then selected for the generation of features for the pattern-recognition study. This selection for the initial studies was based upon previous knowledge that certain ions had known structural significance. Previous work (Wiebers and Shapiro, 1977) had shown that these ions were derived from specific purine or pyrimidine bases, or from these bases plus part of the deoxyribose or phosphate moieties, and that they had been useful in distinguishing sequence isomers of deoxyribodinucleoside monophosphates. Ratios of the peak heights for these selected ions were then calculated and used as features for pattern-recognition analysis. For other studies, the feature-selection method was more empirical in that ratios of every peak height from mass 40 through 312 to every other peak height for spectra of the 16 different dinucleotides were calculated. Those ratios for a particular compound that had values which exceeded by 100 or more the values for the other 15 dinucleotides were selected as possible features.

Once a set of features had been selected by either method,

TABLE II: Results of Multidimensional *k*NN Analysis of Data from d-ApC, d-CpA, d-pApC, and d-pCpA.

	Total Predictive Ability (%)	
Feature Combination[a]	d-ApC, d-CpA	d-ApC, d-CpA, d-pApC, d-pCpA
5, 7	100	50
5, 8	100	50
6, 7	100	83
7, 8	100	50
5, 6, 7	100	50
5, 6, 8	100	50
5, 7, 8	100	50
6, 7, 8	100	50

[a] See Table I for *m/e* composition.

the appropriate standardization of variables (autoscaling (Kowalski and Bender, 1972b)) was done, and the data was subjected to one-dimensional *k*NN analysis. Those features that showed the best classification accuracy for one-dimensional *k*NN analysis were then used in a two-dimensional *k*NN analysis. If the desired accuracy (100%) was not obtained by two-dimensional analysis, those features that showed the greatest predictive accuracy were used in combination in a three-dimensional *k*NN analysis, and only those combinations of features which showed 100% predictive accuracy were retained. Two-dimensional feature space plots (Kowalski, 1975) of the data were then made for verification of class separation. As mentioned by Kowalski, the two-dimensional feature space plots obtained by nonlinear mappings from higher dimensional feature space contain a finite amount of error and should be used only as qualitative indications of the distribution of the data points in higher dimensional space.

The entire feature selection process can be summarized by noting that the criterion for selecting pertinent features is one of maintaining those features which have the largest interclass variance. Since some features vary considerably more than others for the different classes, the result at the conclusion of the process is fewer features with the greatest discriminatory value.

Results

Pattern-Recognition Analysis of Deoxyribodinucleoside Monophosphates. To determine the feasibility of the pattern-recognition approach to the sequence analysis of oligonucleotides, an initial study was done on the *m/e* features that had previously been used (Wiebers and Shapiro, 1977) to differentiate the sequence isomers d-ApC and d-CpA. Twenty-six patterns, 11 spectra of d-ApC, and 15 spectra of d-CpA were used. Thirty-five *m/e* values were selected based on the results of the earlier work. All of the possible ratios for the 35 ions were obtained. One-, two-, and three-dimensional *k*NN analysis indicated that the eight features shown in Table I could be used to distinguish between the isomers d-ApC and d-CpA. The two-dimensional plot shown in Figure 1 demonstrates how a combination of features 1 and 4 can be used to distinguish between the two isomers.

Pattern-Recognition Analysis of Deoxyribodinucleotides. Multiple mass spectra of the 16 deoxyribodinucleotides (d-pXpY) commonly found in DNA were obtained. This set of 65 patterns was used as the data base for the following studies to determine what relationships existed between the different isomers.

Comparison to Deoxyribodinucleoside Monophosphates. The mass spectra for the deoxyribodinucleotide isomers were very similar in appearance to those of the corresponding deoxyribodinucleoside monophosphate isomers. Pattern-recognition methods were used to compare the multiple spectra for d-pApC, d-pCpA, d-pApT, and d-pTpA to those of the corresponding deoxyribodinucleoside monophosphates.

TABLE III: Features Found Useful for Class Separation of d-ApC, d-CpA, d-pApC, and d-pCpA.

Feature No.	Feature Composition (Ratios of *m/e* Values)	Feature Combinations which Gave 100% Predictive Accuracy[a]
24	215/266	*24, 49, 90
40	172/271 + 186/271 + 215/271	24, 86, 90
49	215/252 + 215/266 + 215/295	24, 87, 92
63	191/172	*40, 49, 90
69	191/215	49, 69, 86
79	228/266	49, 87, 92
81	271/266	63, 79, 90
86	228/266 + 242/266 + 271/266	69, 81, 90
87	228/295 + 242/295 + 271/295	69, 86, 90
90	271/252 + 271/266 + 271/295	79, 81, 92
92	215/186	

[a] The combinations marked with an asterisk are those that contain primarily sequence information.

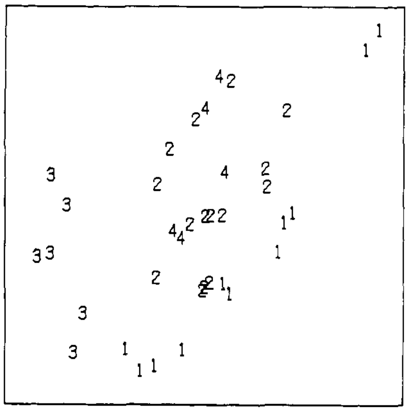

FIGURE 2: Nonlinear mapping of features 40, 49, 90 (Table III) for d-ApC (class 1), d-CpA (class 2), d-pApC (class 3), and d-pCpA (class 4).

Initially, the features found useful for separation of d-ApC and d-CpA were used for comparison to d-pApC and d-pCpA. As indicated by the results in Table II, these features were not able to correctly classify the compounds. Therefore, other features were selected by the processes described under Methods, and multidimensional *k*NN analysis of these new features yielded ten combinations which gave 100% predictive accuracy (Table III). A nonlinear mapping of the three-dimensional feature space to two dimensions for the combination of features 40, 49, and 90 is shown in Figure 2. This feature combination, as well as the combination of features 24, 49, and 90, contains information related to the sequence of the compounds only. The other combinations listed in Table III contain information about both the sequence and the number of phosphates present in the compounds, as indicated by the feature space plot (nonlinear mapping) in Figure 3.

TABLE IV: Features Found Useful for Classification of d-ApT, d-TpA, d-pApT, and d-pTpA.

Feature No.	Feature Composition (Ratios of *m/e* Values)	Combinations which gave 100% Predictive Accuracy[a]
4	215/172	*4, 10, 21
10	206/240	4, 10, 23
21	117/252	4, 10, 27
23	110/252	4, 10, 29
27	186/266	4, 10, 31
29	215/266	4, 10, 35
31	117/295	*4, 21, 23
35	206/295	*4, 21, 27
		*4, 21, 29
		*4, 21, 31
		*4, 21, 35
		4, 23, 27
		*4, 23, 29
		4, 23, 31
		*4, 23, 35
		4, 31, 35
		*10, 21, 27
		*10, 23, 27
		*21, 23, 27
		*21, 27, 35
		*23, 27, 31
		*23, 27, 35

[a] The combinations marked with an asterisk are those that contain primarily sequence information.

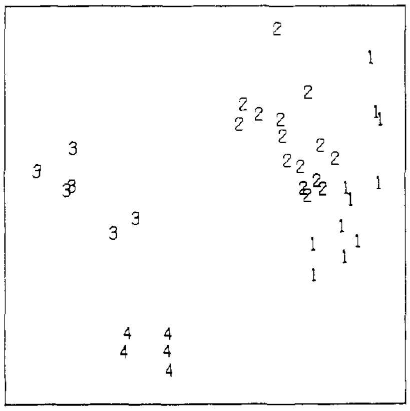

FIGURE 3: Nonlinear mapping of features 69, 86, 90 (Table III) for d-ApC (class 1), d-CpA (class 2), d-pApC (class 3), and d-pCpA (class 4).

Comparison of d-ApT, d-TpA, d-pApT, and d-pTpA yielded similar results. The features and combinations that yielded 100% predictive accuracy are given in Table IV. The combinations marked with an asterisk are those that contain primarily sequence information, while the others contain information about sequence and the number of phosphates present in the compounds.

The feature combinations that contain sequence information only would be the most useful type of features for sequence analysis of oligonucleotides. However, since so many combinations indicated that there is a significant difference between the mass spectra of deoxyribodinucleoside monophosphates and the corresponding deoxyribodinucleotide isomers, the presence of the extra phosphate must alter the volatility of the nucleoside to which it is attached. The deoxyribodinucleotides were believed to be more representative of the spectra for oligonucleotides, and subsequent selection of features for the other isomers was done for the deoxyribodinucleotides only.

TABLE V: Features Found Useful for Classification of Dinucleotides.

Feature No.	Feature Composition (Ratios of *m/e* Values)
4	191/117
6	231/117
8	162/126
10	191/126
12	231/126
13	311/126
18	311/162
19	191/186
21	231/186
26	231/215
29	191/117 + 231/117
34	191/117 + 311/162
36	191/117 + 231/186
48	162/126 + 311/126
53	191/126 + 231/126
54	191/126 + 311/126
55	191/126 + 311/162
56	191/126 + 191/186
58	191/126 + 231/215

TABLE VI: Feature Combinations Yielding 100% Predictive Accuracy for Classification of Dinucleotide Sequence Isomers.

Isomers	Combinations
d-pApC, d-pCpA	29, 48, 56; 34, 48, 54; 36, 48, 53; 36, 48, 54; 36, 48, 55; 36, 48, 56; 36, 48, 58
d-pApT, d-pTpA	4, 12, 18; 4, 18, 21; 6, 8, 12; 6, 8, 13; 6, 8, 19; 6, 8, 21; 6, 8, 26; 6, 10, 13; 6, 10, 21; 8, 13, 19; 8, 13, 26; 8, 18, 21; 10, 13, 19; 10, 18, 21; 12, 18, 19; 18, 19, 21
d-pApG, d-pGpA	36, 48, 56; 36, 48, 58; 36, 55, 56; 36, 55, 58; 55, 56, 58
d-pCpG, d-pGpC	4, 18, 21; 6, 8, 12; 6, 8, 19; 6, 8, 21; 6, 8, 26; 6, 10, 21
d-pCpT, d-pTpC	6, 8, 12; 6, 8, 13; 6, 10, 13; 8, 10, 12
d-pGpT, d-pTpG	4, 12, 18; 6, 8, 12; 6, 8, 13; 6, 8, 19; 6, 10, 13; 8, 13, 19; 19, 13, 19; 12, 18, 19
d-pCpC, d-pGpG, d-pTpT, d-pApA	4, 12, 18; 4, 18, 21; 6, 8, 12; 6, 8, 13; 6, 8, 19; 6, 8, 21; 6, 8, 26; 6, 10, 13; 6, 10, 21; 8, 12, 19; 8, 12, 26; 8, 13, 19; 8, 13, 26; 8, 18, 21; 10, 12, 19; 10, 12, 26; 10, 13, 19; 12, 18, 19; 18, 19, 21

Selection of Features for All Deoxyribodinucleotide Isomers. The results of the previous studies indicated that features could be found that would separate each pair of sequence isomers. However, if the features are selected for each pair of sequence isomers individually, they would probably not be generally applicable to other isomer pairs. In order to minimize the total number of features (and the total number of ions needed from each mass spectrum selection of features for all the isomers was done simultaneously.

The ratios of the peak height for each ion (*m/e* 40 through *m/e* 312) to the peak heights for all of the other ions within each spectrum were calculated for all of the 16 different deoxyribodinucleotides. It was found in many cases that the largest two values for a particular ion ratio were for sequence isomers. For example, the ratio of *m/e* 111 to *m/e* 46 was higher for d-pTpC and d-pCpT than for any of the other 14 dinucleotides. Other ratios were found that had the largest

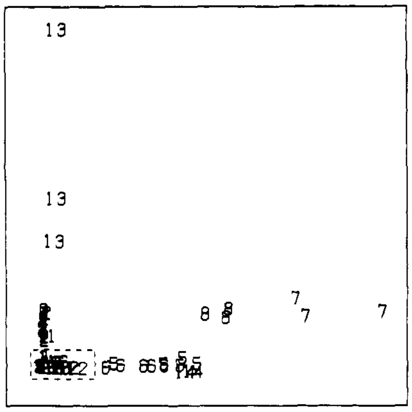

FIGURE 4: Nonlinear mapping of features 6, 8, 12 (Table V) for d-pApC (class 1), d-pCpA (class 2), d-pApG (class 5), d-pGpA (class 6), d-pCpG (class 7), d-pGpC (class 8), d-pCpC (class 13), and d-pGpG (class 14). Classes 3, 4, 9, 10, 11, 12, 15, 16 are enclosed by (- - -). See Figure 5.

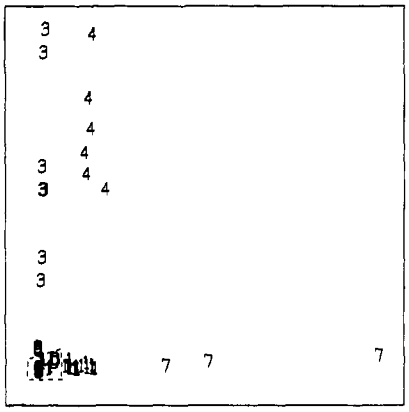

FIGURE 6: Nonlinear mapping of features 8, 12, 19 (Table V) for d-pApC or d-pCpA (class 1), d-pApG or d-pGpA (class 3), d-pCpG or d-pGpC (class 4), d-pCpC (class 7), and d-pGpG (class 8). Classes 2, 5, 6, 9, 10 are enclosed by (- - -). See Figure 7.

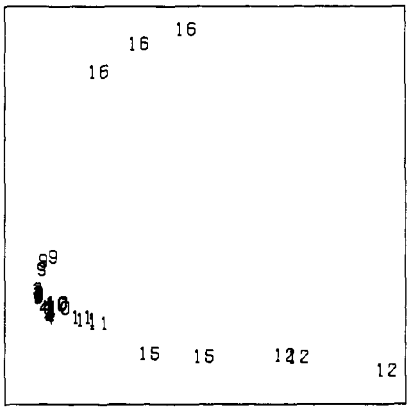

FIGURE 5: Expanded scale for Figure 4 (features 6, 8, 12; Table V) showing separation of d-pApT (class 3), d-pTpA (class 4), d-pCpT (class 9), d-pTpC (class 10), d-pGpT (class 11), d-pTpG (class 12), d-pTpT (class 15), and d-pApA (class 16).

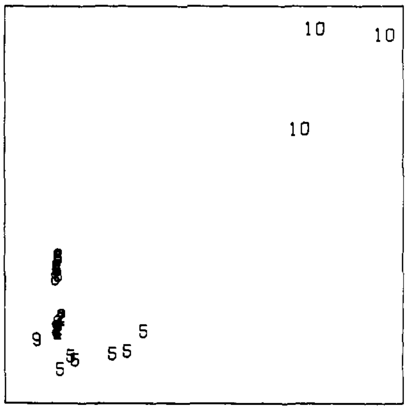

FIGURE 7: Expanded scale for Figure 6 (features 8, 12, 19; Table V) showing separation of d-pApT or d-pTpA (class 2), d-pCpT or d-pTpC (class 5), d-pGpT or d-pTpG (class 6), d-pTpT (class 9), d-pApA (class 10).

values for d-pTpT, d-pGpG, d-pApA, or d-pCpC. One-, two-, and three-dimensional *k*NN analysis was then applied to these features.

Different combinations of these features allowed 100% predictive accuracy for all of the isomers except d-pApC and d-pCpA and d-pGpA and d-pApG. Linear combinations of these features were made and different combinations of these new features allowed the separation of these last two pairs of isomers. The features are given in Table V and the combinations that allow separation of the different isomers are given in Table VI. Although no single combination of features could *simultaneously* classify all 16 isomers, only six features are required for classification of all the isomers. For example, a combination of features 6, 8, and 12 can correctly classify 12 of the isomers, as demonstrated in Figures 4 and 5. Note that with these features all the different isomer pairs cluster in different regions of space and that the isomers that cannot be separated (pApC vs. pCpA and pGpA vs. pApG) by these features cluster together. A combination of features 36, 48, and 56 can be used to correctly classify these last two sequence isomer pairs.

Features Indicating Adjacent Nucleotides. Four combinations of the features listed in Table V were found to be useful for identification of adjacent nucleotides only. For example, an AC (or CA), AT (or TA), AG (or GA), TG (or GT), GC (or CG), CT (or TC), AA, TT, GG, or CC linkage can be identified using the feature combinations 8, 12, 19; 8, 12, 26; 10, 12, 19; or 10, 12, 26. Although these combinations do not contain sequence information, they may prove to be useful for identification of adjacent nucleotides. A feature-space plot for a combination of features 8, 12, and 19 is shown in Figures 6 and 7.

Analysis of Deoxyribooligonucleotides Containing only Two Residue Types. The feature combinations that indicate adjacent nucleotides were used for a comparison of the deoxyribodinucleoside monophosphates d-ApC, d-CpA, d-ApT, d-TpA; the 16 dinucleotides; and the oligonucleotides shown in Table VII. A feature-space plot for features 8, 12, and 19 is shown in Figures 8 and 9. The values plotted are the average of the multiple spectra for each compound (class).

The dinucleoside monophosphate compounds cluster in the same region as the corresponding dinucleotides. The differences attributed to the presence of the extra phosphate that were noted before are small for these features compared to the differences between the isomer pairs. The oligonucleotides containing repeating sequences of dinucleotides also cluster in the same region as the corresponding dinucleotides. For example, $(pApC)_5$ (class 5) and $(pCpA)_4$ (class 6) are in the same cluster as the dinucleoside monophosphates d-ApC and d-CpA (classes 1,3) and the dinucleotides d-pApC and d-pCpA (classes 2,4). The oligonucleotides containing more than one pair of isomers lie somewhere between the clusters for the

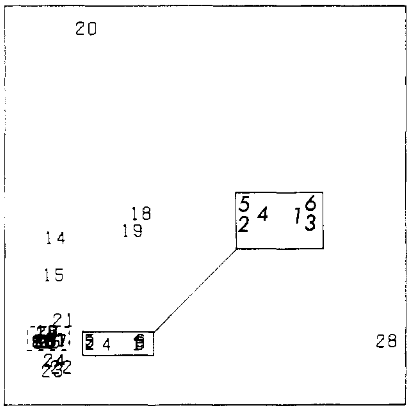

FIGURE 8: Nonlinear mapping of features 8, 12, 19 (Table V) for d-ApC (class 1), d-pApC (class 2), d-CpA (class 3), d-pCpA (class 4), d-(pApC)$_5$ (class 5), d-(pCpA)$_4$ (class 6), d-pApG (class 14), d-pGpA (class 15), d-pCpG (class 18), d-pGpC (class 19), d-(pGpC)$_4$ (class 20), d-(pCpG)$_4$ (class 21), d-pCpT (class 22), d-pTpC (class 23), d-pCpTpT (class 24), d-pCpC (class 28). Classes 7-13, 16, 17, 25-27, 29-31 are enclosed by (- - -). See Figure 9.

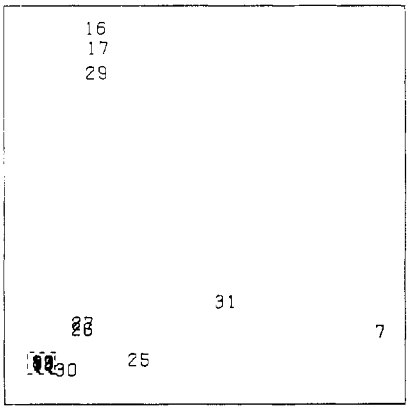

FIGURE 9: Expanded scale for Figure 8 (features 8, 12, 19; Table V) showing relative positions of d-pApApApC (class 7), d-pApGpA (class 16), d-pApGpApApGpA (class 17), d-pCpTpTpT (class 25), d-pGpT (class 26), d-TpG (class 27), d-pGpG (class 29), d-pTpT (class 30), and d-pApA (class 31). d-ApT (class 8), d-pApT (class 9), d-TpA (class 10), d-pTpA (class 11), d-(pTpA)$_3$ (class 12), and d-pTpTpApA (class 13) are enclosed by (- - -).

TABLE VII: Oligonucleotides Compared to Dinucleoside Monophosphates and Dinucleotides.

d-pApGpA
d-pApGpApApGpA
d-pCpTpT
d-pCpTpTpT
d-pTpTpApA
d-(pTpA)$_3$
d-pApApApC
d-(pCpA)$_4$
d(pApC)$_5$
d-(pCpG)$_4$
d-(pGpC)$_4$

TABLE VIII: Oligonucleotides of Defined Sequence Analyzed by Mass Spectrometry.

Trinucleotides	Tetranucleotides		
d-pCpTpT	d-pCpTpTpT	d-(pCpA)$_2$	d-(pGpT)$_2$
d-pApTpC	d-pApApApG	d-(pApC)$_2$	d-(pCpG)$_2$
d-pTpCpG	d-pTpTpApA	d-(pGpA)$_2$	d-(pGpC)$_2$
d-pApGpA	d-pApApApC	d-(pApG)$_2$	d-(pTpT)$_2$
		d-(pCpT)$_2$	d-(pApA)$_2$
d-GpCpA	d-ApTpGpC	d-(pTpC)$_2$	d-(pGpG)$_2$
		d-(pTpG)$_2$	d-(pCpC)$_2$
Pentanucleotide			
d-pCpApCpA-rpU			
Hexanucleotides		Octanucleotides	
d-pCpGpApTpGpC		d-(pCpA)$_4$	
d-pApGpApApGpA		d-(pApC)$_4$	
d-(pTpA)$_3$		d-(pCpG)$_4$	
d-(pApT)$_3$		d-(pGpC)$_4$	
d-ApTpGpCpApT			

corresponding pairs. For example, pApApApC (class 7) is between the cluster containing pApC (class 2) and pApA (class 31). The poor clustering for (pCpG)$_4$ and (pGpC)$_4$ may be due to the impurities mentioned under Materials.

Three-dimensional *k*NN analysis of the data for the feature combinations given above did not yield any combinations with 100% predictive accuracy when the compounds were classed according to adjacent nucleotides (i.e., ApC, CpA, pApC, pCpA, (pApC)$_5$, and (pCpA)$_4$ were considered to be one class, etc.). However, 97% predictive accuracy was obtained for a combination of features 8, 12, and 19. Predictive accuracy for other feature combinations was generally above 90%.

Mass Spectra of Defined Sequence Oligodeoxyribonucleotides. Oligonucleotides of varying chain length and composition have been analyzed by mass spectrometry (Table VIII). Reproducible spectra from these compounds show ions which clearly define the components of the oligonucleotide (as discussed in the preceding paper of this issue). The presence or absence of a phosphate moiety at the 5′ terminus does not appear to affect the principal fragmentation of the molecule. In contrast, no spectrum was obtained when the pentanucleotide d-pCpApCpA-rU, which is a hybrid molecule with a ribonucleotide at the 3′ terminus, was analyzed, and the molecule appeared to be as inert upon mass-spectrometry analysis as ribooligonucleotides are. This observation led to a mass-spectral study of some nucleotide compounds of other structural types—e.g., d-ApTp, d-TpAp, d-TpTp, d-ApAp, and thymidine cyclic 3′,5′-phosphate. All of these compounds are nearly inert under the mass-spectrometric conditions described. Observation of the sample tubes after these compounds have been heated in the direct probe in the spectrometer showed that the samples are charred and presumably pyrolyzed. However, they are not susceptible to electron-impact fragmentation, or, perhaps they are not pyrolyzed in a way that gives products that can be fragmented by electron impact. In contrast, dinucleotides of the pXpY variety are easily fragmented, as are dinucleotides with the unnatural 5′-5′ phosphodiester linkage, for example, d-TppC, d-TppT, and d-CppC. These results suggest that a free 3′-hydroxyl group is a requirement for the mass-spectral fragmentation of an oligodeoxyribonucleotide.

Discussion

The results clearly indicate that computerized pattern-recognition analysis of mass-spectral data can be used to distinguish sequence isomers of dinucleoside monophosphates and dinucleotides. Application of the technique to molecules of longer chain length appears to be a promising approach to sequence determination. It seems reasonable to expect that the mass spectrum for a particular oligonucleotide will differ from the mass spectra for all other oligonucleotides, due to differences in primary structure and chain length. Even the 14 sequence isomers of tetranucleotides containing two nucleotide components, for example, should have mass-spectral characteristics which differ slightly for each isomer and which reflect the different sequential structures. The utilization of pattern-recognition analysis of the mass-spectral data should permit the identification of such obscure discriminatory information.

The experimental results reported in this communication are initial steps for the design of a sequencing algorithm which will ultimately encompass all of the pertinent features needed (1) to define the components of an oligonucleotide, (2) to define the adjacent pairs of nucleotides within the oligonucleotide, (3) to determine the 5′ or 3′ position of the nucleotides within the adjacent pairs, and (4) to specify the 5′ and 3′ terminals of the whole molecule.

A number of problems need to be solved before such a scheme can be perfected. One of these is to be able to identify the terminals of the whole molecule without resorting to derivatization procedures. Theoretically, this information should reside in the spectrum, since the terminal dinucleotides are different structurally from the rest of the oligonucleotide. A search for features which reflect the identity of the terminals is underway in our laboratory.

The main problem to be resolved is the question of how large a molecule can actually be sequenced by applying pattern-recognition methods to the mass-spectral data obtained from an intact molecule. The answer is probably related to how the oligonucleotides are fragmented in the mass spectrometer. The results shown in Figures 8 and 9 indicate that the oligonucleotides containing only two types of residues yield spectra very similar to those for the dinucleotides. In addition, the results described for deoxyribooligonucleotides with a 3′-terminal phosphate or with a 3′-terminal ribonucleotide, both of which rendered the oligonucleotide unsusceptible to mass-spectral fragmentation, lead to the speculation that a 3′-hydroxyl moiety is required for the thermal cleavage to begin and that perhaps a progressive fragmentation (not unlike enzymatic cleavage) from the 3′ terminus occurs. If this is the case, the products actually undergoing electron-impact fragmentation are dinucleotides. This is currently being investigated in our laboratory by studies on mixtures of model dinucleotides, as well as on mixtures of dinucleotides from SP3 DNase cleavage of oligonucleotides, which is known to result in predominantly dinucleotide products (Aposhian et al., 1970). Comparison of the mass-spectral data from mixtures of dinucleotides containing the same residues as intact oligonucleotides should give further insight into the fragmentations occurring in the mass spectrometer.

Acknowledgments

The authors express their gratitude to Dr. Nancy Ho for purification of some of the oligonucleotide samples. We also thank Carl Hines for expert technical assistance and Phil Gaarenstroom for developing the nonlinear mapping programs. J. L. W. expresses appreciation to Dr. David Filmer and Phil Hess for help in the computer programming for the ion ratio analysis of the dinucleotides, and to Dr. H. Lee Weith for helpful discussions on the research.

References

Andrews, H. C. (1972), Introduction to Mathematical Techniques in Pattern Recognition, New York, N.Y., Wiley-Interscience, Chapter 1.

Aposhian, H. V., Friedman, N., Nishihara, M., Heimer, E. P., and Nussbaum, A. L. (1970), *J. Mol. Biol. 49,* 367.

Ho, N. W., Y., and Gilham, P. T. (1973), *Biochim. Biophys. Acta 308,* 53.

Isenhour, T. L., and Jurs, P. C. (1971), *Anal. Chem. 43,* 20A.

Justice, J. B., and Isenhour, T. L. (1974), *Anal. Chem. 46,* 223.

Kowalski, B. R. (1975), *Anal. Chem. 47,* 1152A.

Kowalski, B. R., and Bender, C. F. (1972a), *Anal. Chem. 44,* 1405.

Kowalski, B. R., and Bender, C. F. (1972b), *J. Am. Chem. Soc. 94,* 5632.

Perone, S. P., Burgard, D. R., and Wiebers, J. L. (1975), Abstracts of the 30th Annual Northwest Regional American Chemical Society Meeting, Honolulu, Hawaii.

Pichler, M. A., and Perone, S. P. (1974), *Anal. Chem. 46,* 1790.

Schechter, J., and Jurs, P. C. (1973), *Appl. Spectrosc. 27,* 30.

Tunnicliff, D. D., and Wadsworth, P. A. (1973), *Anal. Chem. 45,* 12.

Wiebers, J. L., Burgard, D. R., and Perone, S. P. (1975a), Abstracts, American Society for Mass Spectrometry, 23rd Annual Conference on Mass Spectrometry and Allied Topics, Houston, Texas, Abstract T-17.

Wiebers, J. L., Shapiro, J. A. (1977), *Biochemistry 16* (preceding paper in this issue).

Wiebers, J. L., Shapiro, J. A., Perone, S. P., and Burgard, D. R. (1975b), Abstracts of the First Chemical Congress of the North American Continent, Mexico City, Abstract 62.