Fully automated generation, recording, storing, and processing of complex chromatograms demand sophisticated laboratory instrumentation and computer implementation. In the high-resolution liquid-chromatographic system for body fluids analyses we use in our clinical laboratory, the chromatograph is interfaced to a 2114 Hewlett-Packard minicomputer that accomplishes, in an on-line mode, analog-to-digital conversion, reductive averaging of sampled data, transformation from transmittance to absorbance units, and storage of the digitized spectrum on a peripheral disk. The information on disk is eventually transferred to a library of spectral files resident on a direct-access disk associated with an IBM 360/Model 65 computer. These data files serve as the input source for program COCOA, written in Fortran IV for the IBM 360 computer, which provides the following processing: (a) iterative polynomial smoothing; (b) sectionally linear baseline tracking; (c) peak and envelope detection and delineation; (d) least-squares gaussian resolution of up to five component peaks in an envelope; and (e) location, size, and shape quantification of peaks. The capability and merits of program COCOA as an off-line procedure on a maxicomputer for the analysis of chromatograms are discussed.

Additional Keyphrases Fortran IV • off-line computerization of chromatograms • uv-absorbing components of physiological fluids • drug metabolism • gas–liquid chromatography • peak distinction and measurement • baseline tracking

The use of computer technology to automate the acquisition, storage, and analysis of chromatographic spectra has heralded significant advances in the science of chromatography. The challenge to exploit such technology in realizing desired performance criteria has spawned various operating styles, which are exemplified in on-line (1, 14), off-line (7), and interactive processing (4), and a concomitant plethora of digital algorithms for peak picking and quantification. On-line processing has the advantage of producing the analysis on a small computer as the spectrum is generated. Off-line processing capitalizes on the accessibility of stored data, thereby allowing greater flexibility in programming effectiveness. Recently, an interactive methodology has been developed that permits dynamic user–computer interplay in eliciting the analysis.

In our clinical laboratories a high-resolution anion-exchange chromatography system for quantitative determination of ultraviolet-absorbing molecular constituents of physiological fluids is used to study drug metabolism. The prototype for this liquid chromatograph was developed (11–15) at the Oak Ridge National Laboratories. Our instrument is being used for human urine specimens to study the metabolic fate of drugs; it is called the body fluids analyzer (BFA). The BFA chromatograph measures two channels of ultraviolet transmittance, at 254 and 280 nm, with a spectrometer designed for continuous monitoring of ultraviolet-absorbing materials as they are eluted from the anion-exchange column. The complexity of the chromatographic spectrum produced (Figure 1) is a result of: (a) total elution time, which is about 40 h; (b) the appearance of 50 to 75 peaks, of which several are incompletely separated; (c) instability in the baseline; and (d) the presence of random background noise. Automation of the BFA was undertaken with the idea that data were to be collected by a dedicated, real-time, multiprogramming minicomputer and analyzed off-line by means of a maxicomputer.

Data Collection and Preprocessing

The chromatograph of the BFA is interfaced to a Hewlett-Packard 2114B minicomputer with 8K words of core. The spectrometer analog signals are amplified to augment resolution independent of the BFA recorder span and offset and modified by an L-C filter that suppresses high-frequency noise. Analog-to-digital conversion is effected at a data rate of 10 Hz and reductive averaging of 150 digitized points is applied to produce a
data point every 15 s. On transformation from transmittance to absorbance units, the sampled spectrum is stored on the removable pack of a disk unit with as many as 9600 data points allotted for each ultraviolet channel. From the disk a digital plot can be made and the data transferred via punched paper tape to a library of spectral files resident on a direct-access disk associated with an IBM 360/MP 65 computer. In addition to handling the data from the BFA, the computer renders on-line processing for six single-channel AutoAnalyzers and two gas-liquid chromatographs under the scrutiny of an interactive "debug" facility.

**Chromatographic Analysis**

The complex pattern from the BFA chromatograph has induced us to develop an initial software package for off-line analysis on a large-scale computer that offered the advantages of high-level programming and core-resident data. On this basis, program COCOA (an acronym for computerized chromatographic analysis), written in Fortran IV for the IBM 360 computer, incorporates the following capabilities: (a) iterative polynomial smoothing of the spectrum; (b) sectionally linear tracking of the baseline; (c) peak and envelope detection and delineation; (d) least-squares gaussian resolution of as many as five component peaks in an envelope; and (e) calculation of the location, size, and shape of peaks. With the spectral data core-resident, program COCOA requires 180K of memory and from 10 to 15 min of CPU (central processing unit) time.

A computerized analysis of complex chromatographic spectra has two formidable requirements: a pattern recognition algorithm that will discern peaks and envelopes of peaks, and the resolution of overlapping peaks. The first requirement is complicated by noise and by shifts and trends in the baseline. To meet the second requirement successfully, strongly overlapping peaks must be discriminated, the peak profile appropriately idealized mathematically, and the potential shift in the profile with time must be dealt with.

**Program Processing Logic**

The five objectives incorporated into COCOA are realized by means of seven basic processing steps. In the order of the flow of the program logic, the processing procedure develops through the following sequence.

**Smoothing.** To decrease spurious noise in the spectrum beyond that decrease achieved by the analog filter, least-squares polynomial smoothing can be applied according to a choice of two schemes: a second-degree polynomial fit from five points or a fourth-degree polynomial fit from nine points. The smoothing algorithms are based upon the convenient forms tabulated by Savitsky and Golay (10). The smoothing operation is iterated by specifying as an input parameter the number of successive smoothing applications to be made. With the progress made in controlling noise originating from the instrumentation, the BFA spectral data currently appear relatively noise-free, so that digital filtering is not used (or used only once). The exception has been in the case of spikes, which resist smoothing without a marked effect on the legitimate data. Such spikes are treated as nuisance elements even though they may be identified as peaks; they prove troublesome only when they are superimposed on the tracing of legitimate peaks. Digital printer plots for both the smoothed and unsmoothed spectra are program output options.

**Critical point array.** The pattern-recognition algorithm in COCOA by which the baseline and the outline of peaks and envelopes are delineated is based on the construction of a so-called critical point (CP) array. A spectral data point qualifies for the CP array when the values of the first or second derivatives in its neighborhood satisfy certain conditions relative to a "zero zone"—a band, symmetrically centered on the time axis, that is constructed for both the first and second derivatives of the spectrum by means of input threshold parameters delimiting the width of the band. With the computation of the first and second derivatives via a nine-point quadratic polynomial scheme (10), the derived values of successive pairs of points are examined for their placement relative to the preset "zero zone." A CP is identified when it and its predecessor have first or second derivatives that traverse the zone boundaries. Each of the six possible cases that can arise where successive derivative values straddle a zone boundary is coded with integers 11 through 16 according to the pattern illustrated in Figure 2. The status of the neutral derivative for a CP is coded by -1, 0, or 1—depending upon whether the value falls below, within, or above the "zero zone." The serial order plus the derivative codes for each CP are stored sequentially in the CP array. If an absence or an excess of points is found
for the CP array, the thresholds are automatically diminished or enhanced, respectively, and the construction of the array is reattempted, with a maximum of 10 such attempts permitted. The CP array is optionally printed out and has proven valuable in helping select the various input parameters.

**Baseline tracking.** The baseline of the pattern is identified through a selection of CP’s that pinpoint the salient features in the lower envelope, viz., minima, peak starts, and peak ends. The derivative codes that distinguish these lower envelope markers are included in Figure 3. The baseline is thereby represented by a set of baseline reference points (BRP) from which the height of the baseline at any time is obtained via linear interpolation with the nearest flanking BRP. This approach is tantamount to a sectionally linear construction of the baseline that is pivoted on special markers located in the lower envelope of the spectrum. The selection of markers is effected iteratively from the CP array in accordance with specified input parameters. A first-approximation baseline is established by determining the levels at the beginning and at the end of the spectrum through averaging of a prescribed spectral segment. The CP array is then searched for minima, peak starts, and peak ends, which, when they fall below a selected threshold erected parallel to the current baseline, are subsumed into the BRP array. When the entire spectrum is scanned in this way, the procedure can be repeated an indicated number of times, with the updated baseline array.

**Peak and envelope detection.** The next phase of the program logic consists in the pattern recognition of peaks and envelopes of fused peaks and their processing and quantification. The CP array is serially searched for a potential peak start, and, when one is detected, testing is activated for identification of consequent peak-delineation points of eight possible types, which are coded as displayed in Figure 3. If another peak start is found before any further delineation points are recorded, the earlier peak start is rejected. In addition to satisfying the code criteria, delineation points are confined to those points that are elevated above a prescribed noise level erected over the local linear baseline segment, except for the peak start and end that are to be embedded in the noise zone. The imposition of the noise zone makes it possible to discriminate between a peak end and a valley between interacting peaks. Information on the delineation of the peak or envelope and the bracketing baseline reference levels and locations is printed out with each complex detected.

**Single peak processing.** The single peak is discriminated from an envelope of peaks (i.e., two or more unresolved peaks) by the total of maxima plus shoulder points yielding one or less. In contrast to a multiple-peak envelope, it is processed as such, with no mathematical fit of the peak profile being attempted. With adjustment for the baseline over which it is supported, the single peak is quantitatively assessed for location, size, and shape. The maximum height and the location of the maximum (or mode) are interpolated by means of a quadratic fit to the three-point crest of the peak. The area is calculated by means of a Newton–Cotes quadrature scheme of order 4 (8), and the same method is applied in estimating some of the higher central moments of the peak distribution as defined by the formulas

\[ m = \int p(t) \, dt / \int p(t) \, dt \]

and

\[ m_i = \int (t - m)p(t) \, dt / \int p(t) \, dt, \quad i = 2, 3, 4 \]

where \( p \) represents the peak curve, \( t \) the time, and \( T \) the interval over which the peak is defined. The first moment \( m \), which is the mean, has been proposed as a more appropriate index of the elution time rather than the mode (7). From the higher central moments an index of skew, or asymmetry, viz.

\[ \text{skew} = m_3 / m_2^{1/2} \]
and an index of excess or "peakedness," viz. 

\[ \text{excess} = (m_2/m_4) - 3 \]

can be used to quantitatively characterize the shape of the peak (5). Moreover, the second central moment, or variance, provides a measure of the spread or dispersion of the peak and can be used, together with the maximum height and area, to serve to quantitatively characterize the size of the peak. The skew and excess, which vanish for a gaussian peak, give an indication of the departure from the common gaussian idealization for chromatographic peaks. Moreover these indexes can be used to detect extensively overlapped peaks that present no apparent graphical evidence of fusion (6). Figure 4 gives a representative printout from program cocoa for a typical peak in a BFA chromatograph. Note the disparity between mode and mean and the nonvanishing character of the skew and excess.

Envelope resolution and processing. When an envelope of overlapping peaks has been discerned, the number of component peaks is taken to be the total number of maxima plus shoulder points detected. With these points as bench marks, the maximum height and its position, together with the half-maximum spread of each component, are estimated and delivered as initial guesses into a subprogram of cocoa that attempts an iterative, least-squares fit of a linear combination of gaussian peaks to the envelope (corrected for the baseline) based on the Gauss–Newton local linearization procedure (9). To promote and stabilize the calculations, we carry out the fitting in three phases: (a) the first phase imposes the side constraint of maintaining the modes fixed and exploits the use of a damping factor (2, 3), (b) the constraint on the modes is released, and (c) the damping factor is set to zero. The number of iterations in each phase and the value of the damping factor are input choices. Figure 5 illustrates a printer plot generated by the peak stripping subprogram. In addition, a plot of residuals can also be obtained that might offer some insight relative to systematic differences between the observed and predicted envelopes. When the envelope is resolved, the parameter estimates—mode, maximum, half-maximum spread, and area and variance of the gaussian components—are printed out.

**Rescue operation.** A special subroutine, the function of which is referred to as a "rescue operation," handles anomalous peaks or envelopes that are rejected by the normal processing means. A dubious peak that is delineated only by peak start and peak end is relegated to the rescue subroutine that endeavors to evaluate maximum, mode, and area. Also, envelopes of more than five component peaks and envelopes for which the stripping algorithm has failed are handled through the rescue operation by providing estimates of the areas of the component peaks. From a computation of the total area under the complex and over the baseline, the area of each constituent peak is determined as a proportion of the total area based on height and spread of the peak.

**Discussion**

Program cocoa, written in Fortran IV for an IBM 360 computer, furnishes a fairly sophisticated off-line approach to the processing of ultraviolet chromatograph patterns generated by the BFA. The complexities of such patterns—unsteady baseline, background noise, multiply convoluted peaks, slowly developing small peaks, and nonstationary, asymmetric peak formation—warrant such an approach. To handle these complexities more effectively, additional capabilities in the software package that ought to be incorporated in future updated versions include:

- polynomial spline tracking of the baseline to provide a smoothly developing baseline contour
- a more robust algorithm for detection of peaks and the recognition of valid shoulder points
- least-squares polynomial fit of a peak crest to interpolate for the peak maximum
- a choice of symmetrical and asymmetrical functional forms with which to idealize peak profile
- the location of peaks relative to elution volume rather than elution time.

Moreover, in the study of drug metabolism, the comparison of chromatograms with controls and (or) with normals, the normalization of peak areas relative to an internal standard, and the identification of compounds present research objectives that have yet to be treated.

The major drawback in processing chromatographic patterns off-line on another computer consists in the transfer of the spectral data. In our implementation we are faced with the intermediary

---

**Fig. 4. Sample printout for single peak**

\[ \text{KPS} = \text{peak start}; \text{KPE} = \text{peak end}; \text{NOB} = \text{number of component peaks}; \text{LFP} = \text{left flex point}; \text{MAX} = \text{maximum}; \text{RFP} = \text{right flex point}; \text{BRP} = \text{baseline reference point}; \text{BI} = \text{baseline elevation at left BRP}; \text{B2} = \text{baseline elevation at right BRP} \]
Fig. 5. Gaussian resolution of envelope of three fused peaks

The numeral 1 outlines the observed complex; the numeral 2 outlines the calculated complex.

of the punched paper tape in conveying data from the Hewlett-Packard to the IBM computer, a procedure that is both awkward and delays results. Direct communication between the computers and (or) implementation of COCOA on the minicomputer by means of overlays are possibilities in improving data transfer. Our use of COCOA has shown that an iterative learning process is involved in determining a judicious selection of input parameters that "optimizes" the analysis. Two to six runnings of the program are commonly made to match the pattern recognition of the programs with that of the chromatographer. This remote, quasi-interactive procedure stands intermediate between the limited, one-shot, on-line processing method and the dynamically interactive processing approach.

We thank Mr. Larry L. Simms for his valuable assistance in the development of the software.

References