Cytometry 5:553–555 (1984)

TECHNICAL NOTE

A Proposal for a Flow Cytometric Data File Standard

Robert F. Murphy and Thomas M. Chused

Center for Fluorescence Research in Biomedical Sciences and Department of Biological Sciences, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213 (R.F.M.), and Laboratory of Microbial Immunity, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20205 (T.M.C.)

Received for publication January 2, 1984; accepted March 20, 1984

The increasing complexity of multiparameter data collection and analysis in flow cytometry and the development of relatively inexpensive arc-lamp-based flow cytometers, which increases the probability that laboratories or institutions may have more than one type of instrument, creates a need for sharable analysis programs and for the transport of flow cytometric data files within an installation or from one institution to another. To address this need, we propose a standard file format to be used for all flow cytometric data. The general principles of this proposal are: (1) The data file will contain a minimum of three segments, TEXT, DATA, and ANALYSIS; (2) The TEXT and ANALYSIS segments consist of KEYWORDS, which are the names of data fields, and their values; (3) All TEXT is encoded in ASCII; (4) KEYWORDS and their values may be of any length; (5) Certain KEYWORDS will be standard, i.e., having specified formats to be recognized by all programs. The structure of the DATA segment will be uniquely defined by the values of KEYWORDS in the TEXT area. It may be in any bit resolution, facilitating compatibility between machines with different word length and/or allowing bit compression of the data. The structured nature of the TEXT area should facilitate management of flow cytometric data using existing data base management systems. The proposed file format has been implemented on VAX, PDP-11, and HP9920 based flow cytometry data acquisition systems.

Key terms: Flow cytometry, data file format, proposed standard, data file transport, data compression, data base management

The recent development of flow cytometers capable of simultaneous three-color measurements emphasizes the increasing complexity of multiparameter data collection and analysis in analytical cytology. Computers must be used to collect and analyze substantial quantities of this type of data. A number of different types of computers running a variety of operating systems are presently used for flow cytometric data collection. Most of the currently used systems are ill-equipped to deal with large volumes of data and do not allow retrieval of specific data files on the basis of associated parameters, such as cell type, antibody preparation, or patient name. In order to facilitate the development of sharable analysis programs, to allow the transport of flow cytometric data files from one installation to another, and to provide a uniform and controlled means for including textual information in data files, we propose a standard file format to be used for all flow cytometric data.

GENERAL PRINCIPLES

The general principles of the standard are as follows:

1. The data file will contain three segments or areas: Text, Data, and Analysis.
2. TEXT consists of KEYWORDS, which are the names of data fields, and their associated values, the contents of the field.
3. All TEXT is encoded in ASCII.
4. KEYWORDS and their values may be of any length.
5. Certain KEYWORDS, which begin with the "$" character, will be standard, i.e., recognized by all programs and with formats as specified below.
6. All space within a file that is not being used will be filled with the space character (ASCII 32.).

Table 1

KEYWORD	Typical contents	Meaning
$FIL	[100,001]112783001.DAT	Name of file to operating system
$SYS	RSX-11M 4.1	Operating system of computer on which file was created
$INST	LMI, NIAID, NIH	Institution where file was created
$CYT	FACS-II	Type of cytometer
$OP	Flo Bighelp	Operator of cytometer
$EXP	Joe Scientist	Name of experimenter
$PROJ	Fluorescence Standards	Name of experimenter's project
$DATE[a]	27-NOV-83	Data file was created; format: dd-mmm-yyyy
$SMNO	23	Sample number
$BTIM[a]	14:25.06	Time of beginning of data collection format: hh:mm.ss or hh:mm:ss
$ETIM[a]	14:27.10	Time at end of data collection
$SRC	Jim Jones	Source of cells (patient's name, cell line, animal strain, etc.)
$CELLS	Peripheral blood	Type of cells
$MODE[a]	U	Uncorrelated single-parameter histograms
	C	One correlated multiparameter histogram
	L	List mode
	S	Sorted list mode
	Z	Correlated multiparameter histogram with zeroes suppressed (e.g., compacted list mode)
$*xn*N		*x* = **P** to indicate a PARAMETER that is collected and is in the file, or *x* = **G** to indicate a GATE parameter that is not collected; *n* = the GATE or PARAMETER number in ASCII digits, both beginning at 1
	FS	Forward scatter
	FSO.1–5	Forward scatter specifying angles
	FL1	Fluorescence 1
	DF1	Delayed fluorescence 1
	FC1	Corrected fluorescence 1
	DC1	Corrected delayed fluorescence 1
	SS	Side scatter
	SS85-105	Side scatter specifying angles
	CV	Coulter volume
	PO	Fluorescence Polarization
	EA	Emission anisotropy
	ET	Energy transfer efficiency
	AE1	Axial extinction 1
	TI	Time
$*xn*S	Anti-IgM	Stain or biological name for *xn* parameter or gate
$*xn*L	488	Laser line or excitation band
$*xn*O	200	Excitation ouput power in mw
$*xn*F	520LP/530LP	Optical filters
$*xn*T	PMT9524A	Tube or detector type
$*xn*V	800	Tube high voltage
$*xn*P	50	Percent of emitted light directed to *xn* detector
$*xn*R	1024	Parameter range (i.e., number of channels)
$G*m*G*n*W	10,20;40,80;70,20	Gate *n* vs. *m* window settings (polygon coordinates)
$P*n*B	16	Number of bits stored for P*n*
$TR	FS,8	Trigger parameter name and setting
$DFC*mn*	5	Dual fluorescence compensation from parameter *m* to parameter *n* in percent (0 indicates no correction)
$TOT	100000	Number of cells in distribution(s) or list
$LOST	2154	Number of cells lost due to CPU busy
$ABRT	1539	Number of cells aborted due to coincidence
$PK*n*	25	Peak channel for parameter *n*
$PKN*n*	12083	Number of cells in peak channel of parameter *n*

[a]The values of these KEYWORDS should be restricted to the format shown. All others may contain ACSII numbers or strings, as appropriate.

SPECIFIC FILE SEGMENTS

The file will begin with the flow cytometry standard version identifier, which will occupy the first 10 bytes (e.g., "FCS 1.0" for files that adhere to the standard described herein). The next 8 bytes will contain the offset, in bytes, from the beginning of the file (taken as O) to the start of the TEXT area. This will be in ASCII, right justified. The next 8 bytes will be the offset, again in bytes from the beginning of the file, to the last byte of the TEXT area. The next four 8-byte numbers will give the offsets to the beginning and end of the DATA area, and the offsets to the beginning and end of the analysis area. If one of the areas is not included in the file, the offsets will be "O" or blank. Additional pairs of locally defined offsets may follow the standard three pairs if desired. Thus the first 58 bytes of the file are records of fixed length that identify the format and point to the beginning and end of the file segments.

The *first* byte of the TEXT area defines the "separator" character (for that segment). The separator is inserted after each KEYWORD and after each KEYWORD VALUE. If the separator appears in a KEYWORD or KEYWORD VALUE it must be "quoted" by being repeated. The remainder of the TEXT area consists of repeats of "KEYWORD, separator, KEYWORD VALUE, separator." Note that although standard keywords begin with "$," it is not necessary to "quote" the "$" to include it within either a KEYWORD or KEYWORD VALUE, since null keywords are not allowed (and therefore "$" only has significance in the first position of a keyword name). The standard keywords and the format of their values are described in Table 1. It should be emphasized that none of the keywords are mandatory, although sufficient keywords to define the data structure are normally required. Use of the standard keywords and data formats will facilitate exchange of data and sharing of analysis programs.

The DATA area is simply a list of the data values whose length in bits is defined by the field "$PnB," where *n* is the parameter number. This allows the data word length to be defined dynamically, facilitating compatability between machines with different data word lengths and/or allowing bit compression of the data. Thus the DATA area for a file with the sequence **$MODE/U/$PIR/64/$P1B/16/$P2R/256/$P2B/16** in its TEXT area (assuming "/" were the separator for that segment) would consist of 64 16-bit words containing a histogram for parameter 1 followed by 256 16-bit words containing a histogram for parameter 2. Likewise **$MODE/L/$P1R/1024/$P2R/1024/$P3R/256/$P4R/256/$P1B/16/$P2B/16/$P3B/16/$P4B/16/$TOT/10000/** would signify 10,000 sets of four 16-bit words of which the lowest 10 bits for the first two 16-bit words and the lowest 8 bits of the last two 16-bit words would contain actual data. Note that combinations such as **$P1B/7/$P2B/9/** are possible, in which case significant computation might be involved to read and write the data, depending on the instruction set of the computer being used.

The ANALYSIS area would typically contain information added to the file after it was collected and stored. Examples are the results of cell cycle analysis or percent positive enumeration. While no specific structure for the ANALYSIS area is being defined in this version of the FCS standard, the default structure of all file segments is that of the TEXT area.

CONCLUSIONS

In addition to permitting transport of flow cytometric data files and analysis programs between institutions, adoption of the proposed standard would also significantly facilitate the use of data base management systems for storage and retrieval of specific flow cytometric data. This would normally be implemented in two steps. Shortly after data acquisition (and often by submission of a batch job), those analyses desired for a given sample, such as enumeration of percent positive and estimation of cell cycle parameters, are performed. The results of these analyses and the keywords and values from the TEXT segment are then incorporated into the data base either by using the data base management system's interface for user-written programs (if it exists), or by creating a command file to run the system's data entry program. The original data file can then be archived, and the data base can be examined as desired using the appropriate query languge.

The standard described in this paper has been implemented on a VAX 11/750 (Digital Equipment Corp., Maynard, MA) running VMS (written in Fortran 77), a PDP 11/34 (Digital Equipment Corp.) running RSX-11M (written in MACRO 11), and on an HP 9920 (Hewlett Packard Corp.). The authors welcome comments or inquiries regarding this proposed flow cytometry data file standard.

NOTE ADDED IN PROOF:

This proposal was discussed in a workshop at Analytical Cytology X held June 3–8, 1984 at the Asilomar Conference Grounds, Pacific Grove, California. It was agreed that the following keywords should be added to those in Table 1: (1) $ASC with the values true (T) or false (F); if true, the DATA segment is encoded in ASCII and $*xn*B is the number of ASCII characters in each integer value. (2) $PAR is the number of acquired parameters in the file. (3) $GATE is the number of gate parameters (if any are present).