LA-UR -90-1568

LA-UR--90-1568

DE90 012098

TITLE: A CLIPS EXPERT SYSTEM FOR CLINICAL FLOW CYTOMETRY DATA ANALYSIS

AUTHOR(S) Gary C. Salzman, Ph.D., R. E. Duque, M.D., R. C. Braylan, M.D.,
C. C. Stewart, Ph.D.

SUBMITTED TO 1st Annual User's Group Meeting
Houston, TX

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Los Alamos
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

A CLIPS Expert System for Clinical Flow Cytometry Data Analysis

[1]G.C. Salzman, Ph.D., [2]R.E. Duque, M.D., [3]R.C. Braylan, M.D., [4]C.C. Stewart, Ph.D.

[1]Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545

[2]Norwood Clinic, 1528 N. 26th St., Birmingham, AL 35234

[3]Department of Pathology, University of Florida College of Medicine, Gainesville, FL 32610

[4]Roswell Park Cancer Center, Laboratory of Flow Cytometry, 666 Elm Street, Buffalo, NY 14263

Abstract

An expert system is being developed using CLIPS to assist clinicians in the analysis of multivariate flow cytometry data from cancer patients. Cluster analysis is used to find subpopulations representing various cell types in multiple datasets each consisting of four to five measurements on each of 5000 cells. CLIPS facts are derived from results of the clustering. CLIPS rules are based on the expertise of Drs. Stewart, Duque, and Braylan. The rules incorporate certainty factors based on case histories.

Introduction

Flow Cytometry [1-3] has become an accepted technique in the clinical laboratory for rapidly classifying cell types in blood, bone marrow, and some solid tumor samples. Cells are labeled with fluorescent cell-type-specific markers and then passed one-at-a-time through a focused laser beam. In commercial flow cytometers typically used in clinical laboratories three colors of fluorescence can be detected from a cell as well as light scattered in the forward direction and at right angles to the laser beam. Multiple, fluorescent-labeled monoclonal antibodies are used to tag the cells, which are then analyzed one at a time at rates of several thousand cells a second. Samples can be processed through the flow cytometer at rates of more than one a minute. Clinicians are being overwhelmed by the large amount of data that must be analyzed to provide the information needed to assist in disease diagnosis.

Immunophenotyping

Immunophenotyping is the science of using antibodies to identify cells. The outer membrane of a cell contains many structurally specific molecules called surface antigens. These antigens have specific sites called epitopes to which the antibodies bind. Monoclonal antibodies are molecules derived from cells all having a common parent. These antibodies can be tagged with fluorescent dyes that are excited at the same 488 nm argon laser wavelength, but that emit their fluorescence at different wavelengths. Each fluorescence detector has a filter to accept the fluorescence from only one of the markers. Each cell can be tagged with from one to three different monoclonal antibodies. The tags bound to the cell surface enable identification of the cell type.

Cluster Analysis

The flow cytometer [4] converts the four to five signals from each cell into digital values and stores them in a correlated fashion called list mode so that the relationships among the four to five variates for each cell are preserved. The data can be displayed as a group of from six to ten bivariate dot plots. K-means cluster analysis [5-6] is carried out on each dataset to find the means of each subpopulation of cells in the sample. Heuristics, developed to assign numerical thresholds to the words negative, dim, and bright regarding fluorescence values and to the words low, medium, and high for scattered light values, are used to translate the mean values of the clusters for each variate into symbolic facts for use by the CLIPS rules. The rules are a direct translation of the reasoning process used by the clinical laboratory personnel in processing the data by hand. The rule syntax is clear enough that the rules can be examined by these people and changed to reflect new knowledge.

Bivariate Plots

Figure 1 shows the six bivariate dot plots from four variate flow cytometry measurements on the blood of a patient with acute leukemia. The 1.2 label in the lower left hand bivariate plot indicates that the x-axis is variate 1, which is FSC, the intensity of forward scattered light and that the y-axis is variate 2, which is SSC, the intensity of light scattered by a cell at 90° to the laser beam axis. Variate 3 is FL1, the emission from a fluorescent-labeled monoclonal antibody. Variate 4 is FL2, which is propidium iodide (PI) fluorescence. PI stains the DNA of dead cells, which can then be excluded from further analysis. The bivariate plots are arranged so that variate 1 is the x-axis for column 1, variate 2 is the x-axis for column 2, and variate 3 is the x-axis for column 3. The cluster analysis program has labeled the data for each cell with a letter corresponding to its cluster association. The clustering algorithm has been instructed to find three clusters in these data. Each ellipse is centered on a cluster and is two standard deviations wide in the x-direction and two standard deviations wide in the y-direction. The important bivariate plot is the one in the upper right hand corner in which x is 3 and y is 4. Clusters B and C represent cells that have taken up PI and so are dead. Cluster A would be called "negative" for variates 3 (monoclonal antibody) and 4 (PI). The axes for the fluorescence measurements (FL1, FL2 and FL3) span four decades on a logarithmic scale. Figure 2 shows the bivariate plots for the same patient using a different monoclonal antibody for variate 3 (FL1). Here cluster A in bivariate 3.4 is "dim" for variate 3 and "negative" for variate 4. Figure 3 shows the bivariate plots for the same patient for another monoclonal antibody for variate 3. Cluster A in bivariate 3.4 (upper right hand corner) is "bright" for variate 3 and "negative" for variate 4.

Results and Discussion

After the cluster analysis has been run on the samples for a patient, a C function translates the means and standard deviations into CLIPS facts, which can then be pattern matched against the conditions in the rules. The first prototype [7] used a rigid decision tree on five variate data in which multiple monoclonal antibodies were used to label the cells. Only nine of eleven acute leukemia cases were correctly assigned. Misclassifications occurred because of the boundaries between "negative" and "dim" and "dim" and "bright" were fixed. Variability in staining intensity and laser power were sufficient to move the means of clusters so that the incorrect choice was sometimes made at a decision node.

A second prototype is now being developed that incorporates uncertainty through the use of certainty factors and measures of belief [8-10]. Facts and rules are selected for evaluation in the order that gives the greatest improvement in certainty for one of the possible outcomes. This approach has been successfully used recently in the diagnosis of colonic lesions [11]. The certainty factors are being assigned initially from *a priori* probabilities calculated from a database containing a large number of case histories. The results for this second prototype will be reported at the meeting.

References

1. Howard M. Shapiro. Practical Flow Cytometry, 2nd Ed., Alan Liss, NY, 1988.
2. M.A. Van Dilla, P.N. Dean, O.D. Laerum, M.R. Melamed (eds). Flow Cytometry: Instrumentation and Data Analysis. Academic Press, NY, 1985.
3. Andrew Yen. Flow Cytometry: Advanced Research and Clinical Applications. Vols I and II. CRC

Press, Boca Raton, FL, 1989.

4. FACSCAN, Becton-Dickinson, Inc., San Jose, CA.
5. H. Spath. Cluster dissection and analysis, theory, FORTRAN programs and examples. Halsted Press, John Wiley and Sons, NY (1985).
6. J.A. Hartigan, Cluster algorithms. John Wiley and Sons, NY, 1975.
7. G.C. Salzman, C.C. Stewart, R.E. Duque, "Expert Systems for Flow Cytometry Data Analysis: A Preliminary Report," New Technologies in Cytometry and Molecular Biology, Gary C. Salzman, Editor, Proc. SPIE 1206, (in press) 1990.
8. J. Giarratano and G. Riley, Expert Systems, Principles and Programming. PWS-KENT Publishing Co., Boston, 1989.
9. E.H. Shortliffe and B.G. Buchanan, "A model of inexact reasoning in medicine," Mathematical Biosciences 23, 1975.
10. E. Rich, Artificial Intelligence, McGraw-Hill, NY, 1983.
11. J.E. Weber, P.H. Bartels, W. Griswold, W. Kuhn, S.H. Paplanus, A.R. Graham. Colonic Lesion Expert System. Performance Evaluation. Analytical and Quantitative Cytology 10, 150-159, 1988.

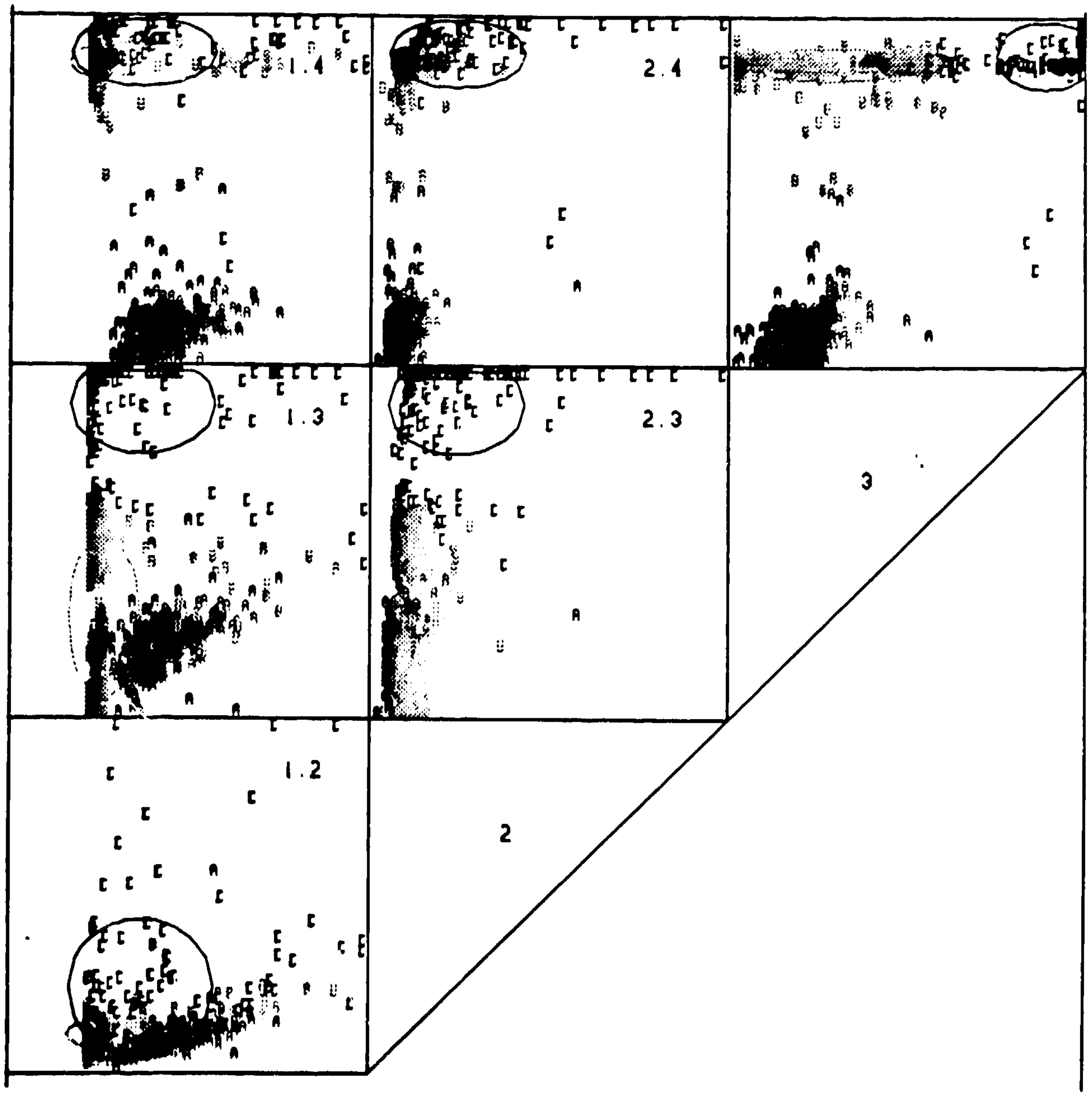

Figure 1. Bivariate dot plot of of four variate data flow cytometry data from a leukemia patient. The two numbers in the upper right hand corner of each box (X.Y) are the x-axis variate number followed by the y-axis variate number. Cluster analysis has divided the data into three clusters. Each data point is labeled with a letter designating its cluster membership. Variate one is forward scatter (FSC), variate 2 is side scatter (SSC), variate three is fluorescence one (FL1), a monoclonal antibody, and variate four is fluorescence two (FL2), which is the dye propidium iodide (PI) that stains the nuclei of dead cells. The population A in bivariate 3.4 in the upper right hand corner is "negative" for the monoclonal antibody of interest (FL2), and "negative" for PI.

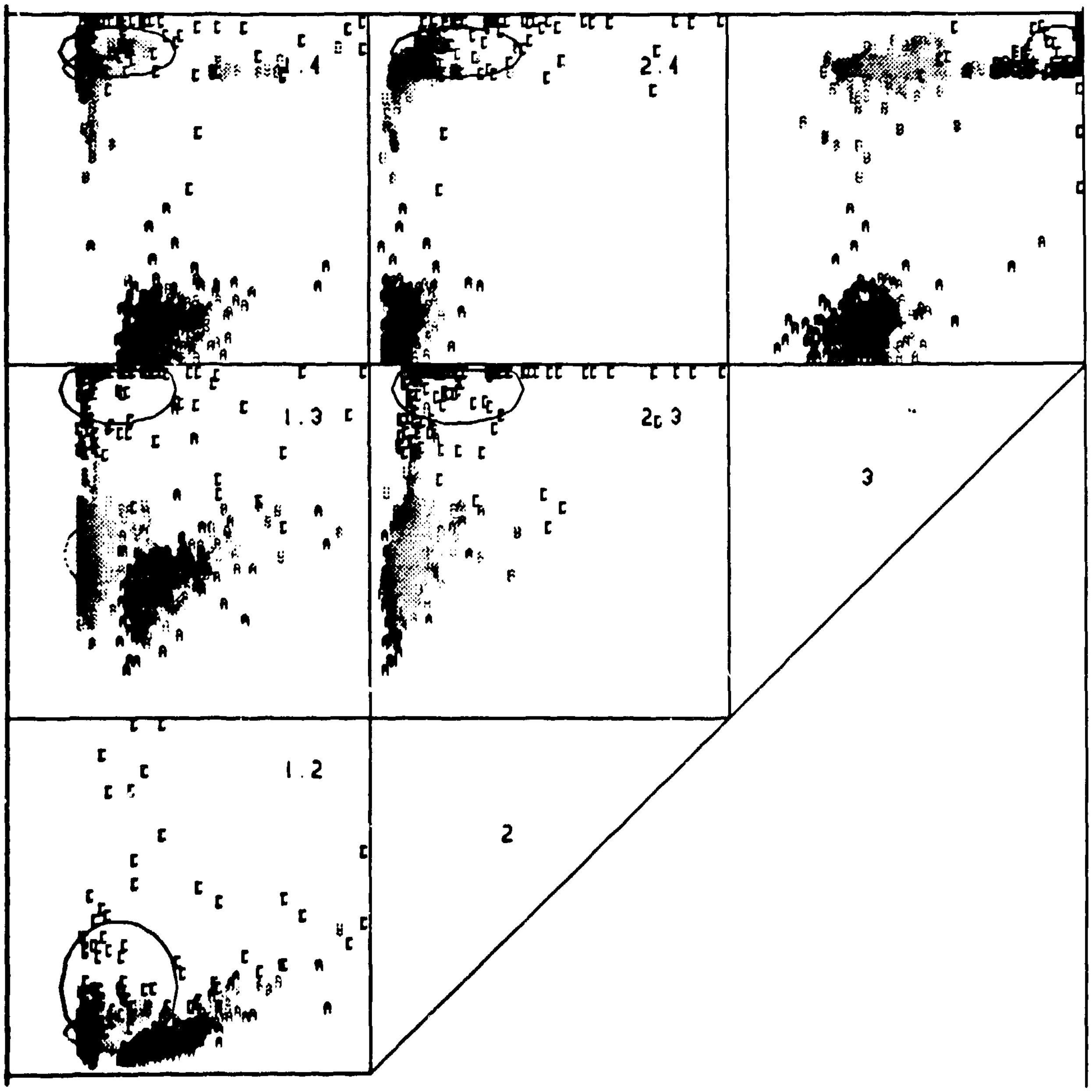

Figure 2. Same as Figure 1 except that a different monoclonal antibody has been used for FL1 (variate 2). Here cluster A in bivariate 3,4 is "dim" for the antibody.

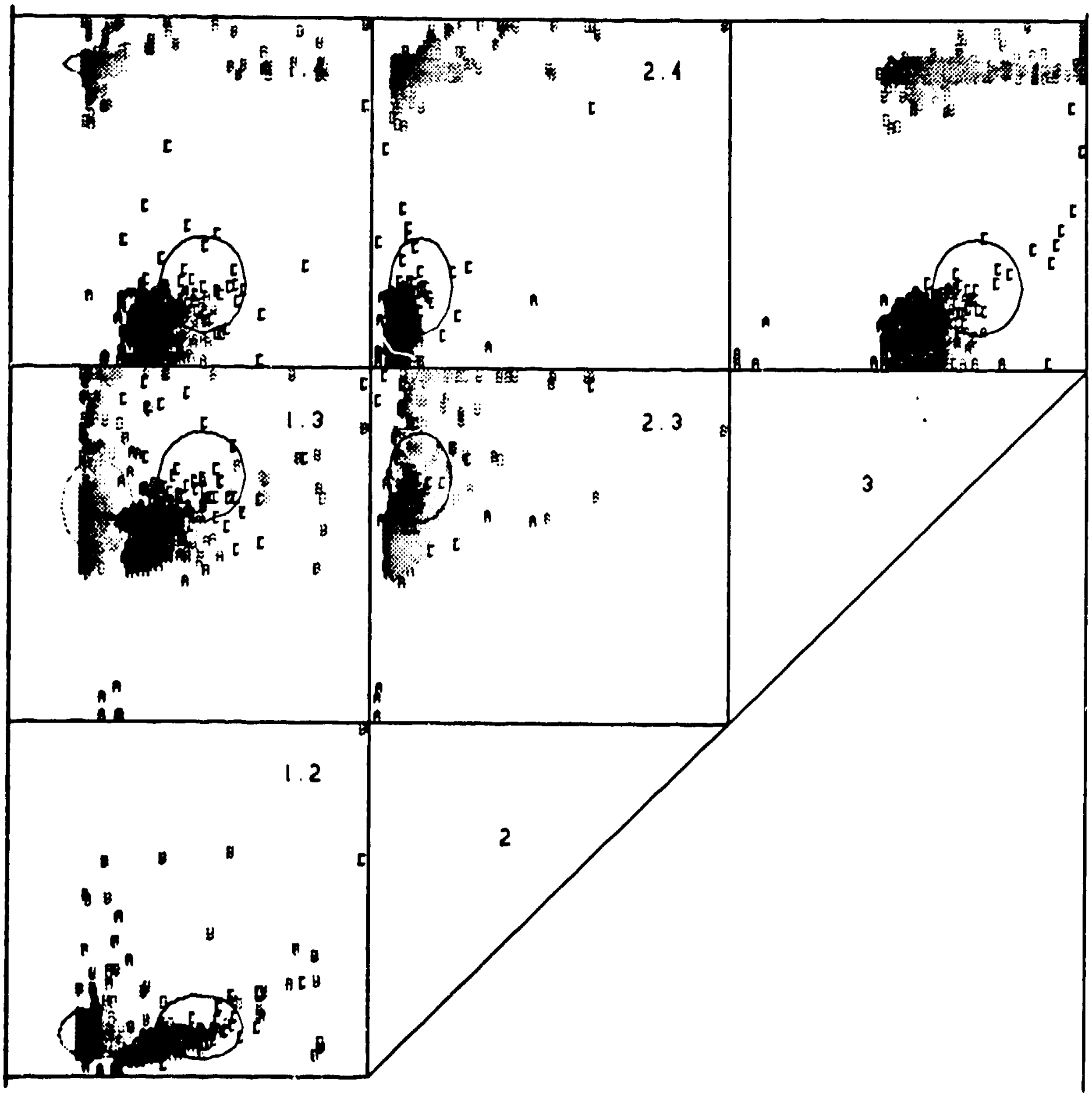

Figure 3. Same as Figure 1 except that another different monoclonal antibody has been used for FL1 (variate 2). Here cluster A in bivariate 3.4 is "bright" for the antibody.