Original Articles

Analysis of Clinical Flow Cytometric Immunophenotyping Data by Clustering on Statistical Manifolds: Treating Flow Cytometry Data as High-Dimensional Objects

William G. Finn,[1]* Kevin M. Carter,[2] Raviv Raich,[3] Lloyd M. Stoolman,[1] and Alfred O. Hero[2]
[1]Department of Pathology, University of Michigan, Ann Arbor, Michigan 48109
[2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109
[3]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331

Background: Clinical flow cytometry typically involves the sequential interpretation of two-dimensional histograms, usually culled from six or more cellular characteristics, following initial selection (gating) of cell populations based on a different subset of these characteristics. We examined the feasibility of instead treating gated *n*-parameter clinical flow cytometry data as objects embedded in *n*-dimensional space using principles of information geometry via a recently described method known as Fisher Information Non-parametric Embedding (FINE).

Methods: After initial selection of relevant cell populations through an iterative gating strategy, we converted four color (six-parameter) clinical flow cytometry datasets into six-dimensional probability density functions, and calculated differences among these distributions using the Kullback-Leibler divergence (a measurement of relative distributional entropy shown to be an appropriate approximation of Fisher information distance in certain types of statistical manifolds). Neighborhood maps based on Kullback-Leibler divergences were projected onto two dimensional displays for comparison.

Results: These methods resulted in the effective unsupervised clustering of cases of acute lymphoblastic leukemia from cases of expansion of physiologic B-cell precursors (hematogones) within a set of 54 patient samples.

Conclusions: The treatment of flow cytometry datasets as objects embedded in high-dimensional space (as opposed to sequential two-dimensional analyses) harbors the potential for use as a decision-support tool in clinical practice or as a means for context-based archiving and searching of clinical flow cytometry data based on high-dimensional distribution patterns contained within stored list mode data. Additional studies will be needed to further test the effectiveness of this approach in clinical practice.

Key terms: flow cytometry; statistical manifold; information geometry; immunophenotyping; immunophenotype clustering

How to cite this article: Finn WG, Carter KM, Raich R, Stoolman LM, Hero AO. Analysis of clinical flow cytometric immunophenotyping data by clustering on statistical manifolds: Treating flow cytometry data as high-dimensional objects. Cytometry Part B 2009; 76B: 1–7.

Clinical flow cytometric analysis usually involves the interpretation of individual two-dimensional scatter plots culled from sets of simultaneous analysis of up to eight measurements (two light scatter measurements and up to six fluorescence channels or "colors") for routine clinical grade analyzers. However, the multidimensional power of flow cytometry may be instead more effec-

Grant sponsor: National Science Foundation; Grant number: CCR-0325571.

*Correspondence to: William G. Finn, MD, Department of Pathology, University of Michigan, 1301 Catherine Road, Room M5242, Ann Arbor, MI 48109-0602. E-mail: wgfinn@umich.edu

Received 13 February 2008; Accepted 27 May 2008
Published online 18 July 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/cyto.b.20435

tively realized by systems that treat single muticolor analyses as individual high-dimensional datasets (1–8).

The analysis of high-dimensional datasets has become more common in the age of applied genomics and proteomics. However, the fact that all measured characteristics of a given analysis can be traced to each individual cell gives the dimensionality of flow cytometry a uniquely spatial characteristic not shared by other proteomic platforms (7,9). Each individual tube analyzed in a routine *n*-parameter flow cytometry study can be represented conceptually as a single object embedded in *n*-dimensional space and formed in aggregate by thousands of analyzed cells, each of which displays a unique *n*-dimensional signature. Just as an ordinary object is better described by its shape and overall appearance than by the measuring of its individual dimensions, one could consider the possibility that flow cytometry data could be better represented by the general shape of a cell population over all of the dimensions analyzed (5). Since we live in three-dimensional space, direct visualization of a four color (six dimensional) flow cytometry dataset as a six-dimensional object is not feasible. However, rather than utilizing the interpretation of sequential two-dimensional projections of this six-dimensional object (as is the current norm), analytical methods can be devised for the comparison of separate datasets embedded as unique objects in six-dimensional space.

The analysis of high-dimensional datasets often involves characterizing the *manifold* within which the data are assumed to be embedded. In layman's terms, the mathematical concept of a manifold could be defined as a smooth space or surface (of any dimensionality) that is nearly "flat" on small scales, and within which geometrical objects may be embedded. Examples could include a sphere, a torus, Euclidean space in general, and indeed our three-dimensional universe. The field of *manifold learning* involves the discovery of lower dimensional manifolds for objects embedded in higher dimensional space and is often applied to dimensionality reduction of high-dimensional datasets (10).

It is often assumed that high-dimensional datasets can be appropriately represented on Euclidean manifolds (manifolds comprised of points or coordinates embedded within Euclidean space). However, there are many problems in which the data cannot be appropriately represented by a Euclidean manifold, and the model parameters are unspecified and must be learned through the data. In such cases, it may be helpful to assume that the data lie in a manifold composed not of individual spatial coordinates, but of probability density functions. The term *statistical manifold* has been used to describe such manifolds composed of probability density functions rather than spatial coordinates (11,12).

The emerging field of *information geometry* involves the analysis of probability distributions as geometric structures within non-Euclidean space and can be applied to the study of statistical manifolds (13). The distance between points or objects on a statistical manifold can be measured by a distance function known as the *Fisher information metric* (12). However, calculating the Fisher information metric requires knowledge of the underlying parameterization of the assumed manifold, knowledge that is generally not available or feasible in the analysis of flow cytometry datasets.

Recently, Carter et al. described a nonparametric approach to clustering and classification on statistical manifolds using a similarity measurement known as the *Kullback-Leibler divergence* (commonly referred to as the relative entropy of a probability distribution) as an estimate of the Fisher information distance for statistical manifolds for which parameterization is unknown, and for which individual data points lie in reasonably close proximity (as would generally apply to immunophenotypic analysis of distinct cell populations by multiparameter flow cytometry) (12,14). As a given manifold is more densely sampled, the Kullback-Leibler divergence converges to the Fisher information distance. This approach has been termed *Fisher Information Nonparametric Embedding* (FINE) (12).

In this study, we attempted to apply these principles to the interpretation of flow cytometry datasets as high-dimensional objects generated by probability density functions embedded on a statistical manifold (as opposed to sequential groups of individual light scatter characteristics or surface antigens). As an initial test of this approach, we chose to compare the immunophenotypic patterns of leukemic B-precursor lymphoblasts against the immunophenotypic patterns of physiologic B-cell precursors (hematogones), since distinction between these often similar cell types is an important and sometimes challenging task that often confronts practicing hematopathologists on the day-to-day diagnostic service (15).

MATERIALS AND METHODS

Case Selection

The use of previously analyzed clinical flow cytometry data for cluster analysis was approved by our Institutional Review Board. The files of the clinical flow cytometry laboratory at the University of Michigan were searched for cases coded as B-precursor acute lymphoblastic leukemia (ALL) based on complete diagnostic assessment including morphologic assessment of marrow, flow cytometric immunophenotyping, and cytogenetic analysis where indicated per World Health Organization diagnostic criteria (16). From this list, cases were selected that had sufficient available list mode data and sufficient cells for analysis, searching back from the most recent cases available. Thirty-one cases of ALL were retrieved for analysis, spanning an approximately 18-month period. For comparison, the flow cytometry database was manually screened for the presence of cases with hematogone hyperplasia, and from this screen 23 cases were retrieved showing prominent hematogone populations, again based on a combination of morphologic assessment, clinical correlation, and flow cytometric immunophenotyping based on previously published descriptions of hematogone immunophenotypes (15).

Data Retrieval

Raw flow cytometry data for this study were generated by analysis on a Beckman-Coulter FC-500 flow cytometer using Beckman-Coulter CXP acquisition software (Beckman-Coulter, Hialeah, FL) and stored as list mode data in standard fcs format. Within our routine acute leukemia flow cytometry panel, we include a single four-color (six-parameter) tube including CD45 (ECD conjugate), CD10 (phycoerythrin-cyanin 5 conjugate), CD19 (phycoerythrin conjugate), and CD38 (fluorescein isothiocyanate conjugate) (all antibody reagents obtained from Beckman-Coulter/Immunotech, Hialeah, FL), designed for the isolation of hematogones and aberrant lymphoblast populations, based on known differential patterns of these markers in these cell types. Although it may take additional markers to render fine distinctions in practice, this tube was selected for analysis since the methods being tested in this study require single high-dimensional datasets acquired in a single analysis, and since this marker combination is highly useful in distinguishing these cells subsets in most cases.

List mode data were prepared for analysis as follows. First, the cell population of interest (either hematogones or lymphoblasts, depending on the case) was selected by manually examining the datasets using an iterative gating strategy to evaluate for the presence of distinct cell clusters based on the most effective discriminator for that particular case. In most cases, the initial evaluation was of a CD10 versus CD19 histogram or of a CD10 versus CD38 histogram (depending on the separation of cell clusters), due to the tendency for lymphoblasts and hematogones to coexpress these markers. From here, data were projected onto CD45 versus side angle light scatter histograms to exclude higher side scatter events that could potentially represent nonspecific binding of antibodies to nonlymphoid cells. Data were then reprojected onto CD10 versus CD19 and CD19 versus CD38 histograms to assure the appearance of well-distributed clustered data without evidence of artificial "shelves" or cut-off thresholds for any given marker, and without evidence of inclusion of extraneous cell clusters that would represent nonlymphoblast (or nonhematogone) cell populations. Care was taken during this approach to target the selection of the cell subpopulation of interest (either hematogones or leukemic lymphoblasts) based on differential cell clusters on histograms, without being artificially restrictive as to the exclusion of cells beyond a prescribed level of light scatter or marker expression.

Once the gated data for the cells of interest were isolated via this iterative approach, the data were converted from standard flow cytometry list mode format to tab-delimited text using WinMDI software, version 2.8 (Scripps Research Institute, La Jolla, CA).

Data Analysis

The gated data files were analyzed using the three-step FINE process, described in more detail by Carter et al. (12,14).

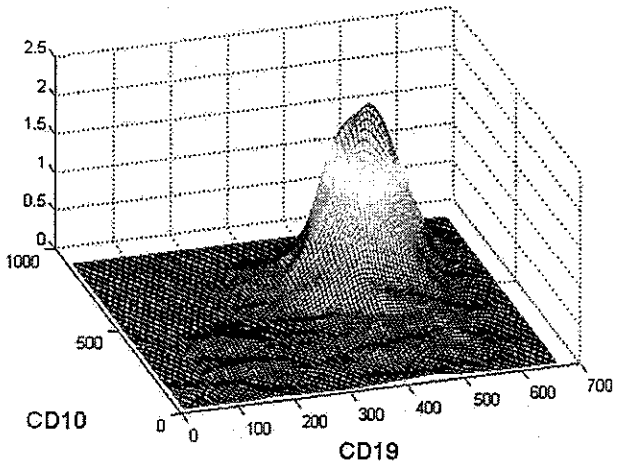

FIG. 1. Illustration of flow cytometry list mode data after conversion by a kernel density estimate. This smoothed the data by converting individual data points into Gaussian distributions, which were then summed and normalized to form an overall distribution of the same shape and density variation of the initial dataset. The conversion was performed over all six dimensions for each dataset, but the figure depicts a two-dimensional projection of the kernel density estimate (in this case CD19 vs. CD10).

Briefly, in the first step the gated tab-delimited list mode files were smoothed by converting from sets of individual points into probability density functions using kernel density estimation. Kernel methods are nonparametric techniques used for estimating probability densities of data sets and involve the conversion of discrete data points into the normalized sum of identical densities centered about each data point. In essence, each data point is converted into a probability distribution and the aggregate of these distributions is summed and normalized to form a single smooth distribution. Kernel methods have been used in previous work on the analysis of flow cytometry data (3). For our data, we chose a Gaussian kernel, essentially converting each discrete data point into a Gaussian probability function, the total of which were summed and normalized to form a non-Gaussian distribution corresponding to the overall "shape" of the cloud of individual cellular events measured in each six-dimensional analysis (Fig. 1). The derived distribution for each six-dimensional flow cytometry analysis would be represented as follows:

$$f_i(x) = \frac{1}{(N_i \times h)} \times \sum_{i=1}^{N_i} K\left(\frac{(x - x_i)}{h}\right)$$

where $K(x) = \frac{1}{(\sqrt{2\pi})} e^{\frac{-x^2}{2}}$ is the zero mean unit variance Gaussian kernel, h is the bandwidth or smoothing parameter around each data point, and $f_i(x)$ is the resulting probability density function for the ith patient sample based on the normalized sum of distributions centered on the N_i cells in the sample. The bandwidth parameter is very important to the overall density estimate. Choosing a bandwidth parameter too small will yield a peak filled density, whereas a bandwidth that is too large will generate a density estimate that is too smooth and loses

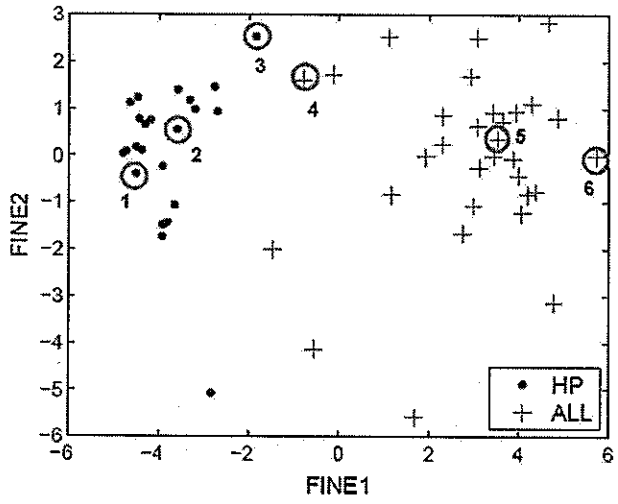

FIG. 2. Two-dimensional embedding of neighborhood map data generated by the comparison of six-dimensional flow cytometry datasets by Fisher information nonparametric embedding (FINE) using the Kullback-Leibler divergence as a distance measurement. Cases of B-precursor acute lymphoblastic leukemia (ALL) were effectively separated from benign hematogone hyperplasia (HP) by this method. The circled points correspond to the density plots illustrated in Figure 3, numbered respectively.

most of the features of the distribution. For this analysis, the parameter h was chosen separately for each analyzed patient sample using the maximal smoothing principle (17) under the assumption that each dimension was independent (and therefore may have different values in the kernel width vector). The result of the kernel density estimation step was conversion of discrete dot-plots into six-dimensional probability density functions.

In the second step, we calculated the relative differences among individual six-dimensional datasets for each case using the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(f_i||f_j) = \int \log\left(\frac{f_i(x)}{f_j(x)}\right) f_i(x)dx$$

to form the following similarity matrix between any given patient samples i and j:

$$D_{ij} = D_{\mathrm{KL}}(f_i||f_j) + D_{\mathrm{KL}}(f_j||f_i).$$

The similarity matrix was constructed to assure symmetry, since the Kullback-Leibler divergence is not symmetric. The result was a high-dimensional neighborhood map depicting the relative difference in information (i.e., similarities) among the 54 total samples analyzed based on distributions defined in six dimensions.

Since the similarity matrix represents a high-dimensional neighborhood map, an additional step of dimensionality reduction is included as the third step in the procedure so that the similarities between cases in the high-dimensional neighborhood map may be visualized on a two-dimensional plot. This dimensionality reduction step was carried out using classical multidimensional scaling (12). Multidimensional scaling is the term used for a group of methods by which high-dimensional dis-

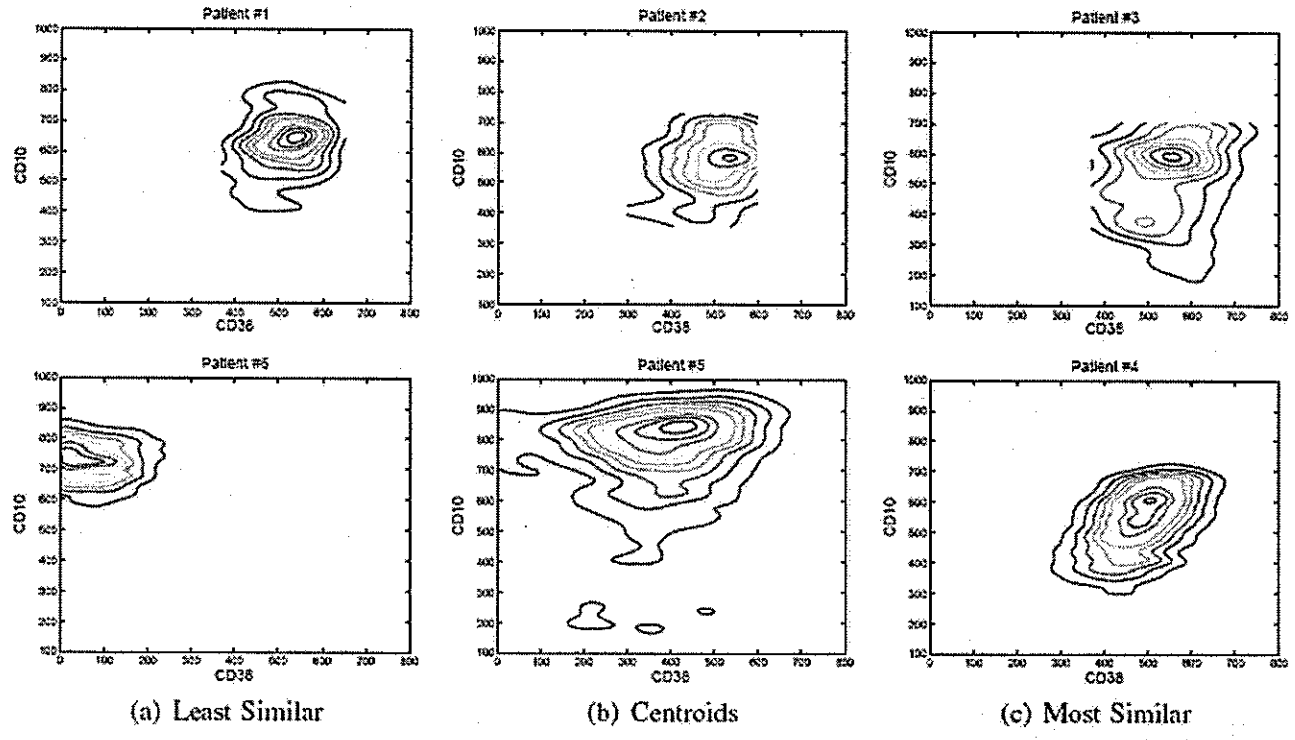

FIG. 3. Contour plots of CD38 versus CD10 expression for several data sets. The top row corresponds to hematogone hyperplasia (HP) cases, and the bottom row represents acute lymphoblastic leukemia (ALL) cases. The selected patients are those most similar between disease classes, the centroids of each disease class, and those with little similarity between disease classes, as highlighted in Figure 2.

tance matrices may be embedded in lower dimensional space. Classical multidimensional scaling (cMDS) is a particular type of multidimensional scaling in which each point on a matrix of dissimilarities is embedded in Euclidean space, by first centering the dissimilarities about the origin, then calculating the eigenvalue decomposition of the centered matrix. This method allows for the low-dimensional graphic representation of data points while revealing any natural separation or clustering of the data (12).

RESULTS

Of patients from whom all samples were obtained, 18 were male and 13 were female, with an average age of 25 years at the time of bone marrow biopsy (range 2–74 years). Of patients from whom hematogone samples were obtained, 14 were male and 9 were female, with an average age of 41 years at the time of bone marrow biopsy (range 9 months to 66 years).

Two-dimensional maps (generated via multidimensional scaling) depicting projections of the relative Kullback-Leibler divergences of six-dimensional data among cases studied are shown in Figure 2. To illustrate the differences depicted in Figure 2, traditional plots of CD10 versus CD38 (two of the 6 measured dimensions in each analysis) are shown in Figure 3 for paired cases from each cluster (ALL or hematogones) that are relatively similar, relatively dissimilar, and near the center of each cluster.

In general, the algorithm used in this study was effective in the discrimination and clustering of cases of ALL from cases showing hematogone expansions. The hematogone cases were more tightly clustered, likely reflecting the greater immunophenotypic variability of leukemic lymphoblasts relative to the more consistent and uniform immunophenotype typical of hematogones (15).

DISCUSSION

The method outlined in this study represents a novel approach to the analysis of clinical flow cytometry data in which multicolor flow cytometry datasets are treated as virtual objects embedded in high-dimensional space and compared with one another by approximating information distances on statistical manifolds. In the current demonstration of concept, this system was generally effective in the unsupervised distinction of patient samples containing leukemic lymphoblasts from patient samples containing normal B cell precursors (hematogones). Formal data on sensitivity and specificity cannot be derived in this proof-of-principle study since we did not randomly select cases from our normal workflow, and therefore the pretest prevalence of each diagnostic condition, which is required for derivation of such statistics, would not be represented.

In contrast to other proteomic or immunophenotyping methods, flow cytometry allows simultaneous analysis of numerous surface markers traceable in any combination to a specific individual cell. In day-to-day practice, attempts to harness this dimensional power of flow cytometry are usually in the form of sequential two-dimensional analyses linked to additional dimensions of data via previous analytical iterations. Although a useful practical method for the study of flow cytometry datasets, this approach has limited value in unsupervised discovery or in proteomic style analysis.

Analogy may be made to the common endeavor of face recognition. Individuals recognize other individuals by interpreting the overall appearance or shape of one's face, not by evaluating individual facial measurements in a step-by-step selection and analysis process. Similarly, in this study we set out to devise a method whereby the single high-dimensional object formed by the flow cytometric analysis of a given cell population could be evaluated as a whole rather than as sequential parts.

There are numerous potential applications for the type of analysis outlined here. The ability of this statistical manifold learning method to adapt as additional cases are added to the database augments its potential usefulness as a clinical decision support tool. Analysis of any given case could be queried against the known neighborhood maps constructed by previous analyses, and a list of most probable diagnoses could then be generated to assist the hematopathologist or flow cytometrist in rendering a final diagnostic impression. One could envision a role for this type of approach in borderline classification issues (such as lymphoma subtyping based on immunophenotype), issues of minimal residual disease detection, etc.

Aside from its potential diagnostic utility, a system of clustering flow cytometry data within statistical manifolds could potentially be used as a context-based searching and databasing method for case retrieval or research. For example, in our laboratory, we currently list cases in our database according to the final diagnosis assigned to that case following our interpretation of the flow cytometry data. If we wish to search for cases of, for example, ALL, we enter the appropriate text code into the search engine, and it finds cases that we diagnosed as ALL, irrespective of the actual immunophenotypic pattern contained within the list mode files. The approach outlined in this study would potentially allow us to store raw list mode files of selected cell populations and, subjecting them to the manifold learning process, search the database not for cases by diagnostic label, but by the actual similarity of the flow cytometry dataset over the entire group of markers contained within an individual analysis tube. The system would adapt, with information distances across the overall neighborhood map adjusting with each added case. Such context-based searches have been proposed for histologic images (18) and would also be of use in retrieval of flow cytometry data from archives.

Our approach could also have potential value in clustering and classifying disease processes through unsupervised discovery, analogous to approaches used in functional genomic and proteomic applications. Indeed, flow cytometric immunophenotyping is at its essence a proteomic method, albeit on a relatively small scale (19–21). Our approach allows a proteomic-style analysis of the

entire distribution formed by multicolor flow cytometric analysis of cell suspensions. Our study was performed using archived clinical four-color datasets. The power of this approach could be magnified considerably if applied to higher dimensional datasets (10 color and beyond) currently deployed in research settings (22).

To our knowledge, our study is the first to employ the principles of information geometry and statistical manifold embedding in the comparison of flow cytometry results between different patient samples. However, previous studies have described methods that treat flow cytometry output as single high-dimensional datasets rather than as collections of two-dimensional projections. Roederer et al. described systems based on probability binning of *n*-dimensional data, including the use of an algorithm that identified geographic regions in *n*-dimensional space that contain significantly more or fewer events than other areas (7,23). They termed this statistical comparison of event numbers in high dimensional space "frequency difference gating." Zeng et al. and Zamir et al. described approaches with some conceptual similarity to ours but with different methods (2,6). Zamir et al. evaluated single four-color (six-dimensional) flow cytometry assays by converting each of them into a single matrix with the number of rows equal to the number of cells analyzed, and the number of columns equal to the number of measured flow cytometry characteristics (in this Case 6), each normalized to a mean of zero and standard deviation of 1. The matrices were then subjected to statistical clustering methods for the classification of different cell populations within the sample. Although this method was based on the analysis of a six-dimensional dataset as a single entity, it maintained the identity of each cell as a discrete point in the matrix, without conversion to probability density functions as in our study. Zeng et al. used a kernel density estimation method similar to ours to convert high-dimensional flow cytometry datasets into probability density functions, but then used histogram features extracted from each dimension of the probability density function to guide *k*-means clustering as a means to identify discrete cell populations within a given dataset. Pedreira et al. described a multidimensional classification approach for automated flow cytometry analysis that, like our method, treated flow cytometry datasets as objects embedded in *n*-dimensional space and did not require the application of an assumed distribution onto the flow cytometry dataset, but did not use the specific principles of information geometry outlined in the current study (8).

There are limitations to the treatment of entire flow cytometry datasets as single high-dimensional distributions. For example, patients with immunophenotypically identical abnormal cell populations would likely be clustered separately depending on the nature of the non-neoplastic background cells or on the sheer percentage of abnormal cells in the sample. For this reason, we chose in this study to purify the cells of interest through an iterative list-mode selection process before application of the manifold learning algorithm. One could argue, however, that the analysis of entire datasets (including both normal and abnormal cell types) would be of potential value, since the nature of the host response may be distinct in a given disease process and may be represented by the immunophenotypic pattern of non-neoplastic cells in the sample. Furthermore, the nature of flow cytometry data allows for the virtual selection of numerous different cell types without preanalytical sorting or isolation, and subsequent analysis of these subsets via manifold learning. A caveat, of course, is that any given process of selection for cell populations of interest could influence the subsequent clustering algorithm, and minor differences in cell selection strategies could harbor the potential to inordinately affect the clustering due to potential inconsistencies in initial data selection. The influence of various preanalytical factors (number of colors in the analysis, presence of normal cell populations, cell selection strategies, etc.) on the performance of this statistical manifold clustering approach will have to be evaluated in expanded prospective studies.

In summary, this study was an attempted demonstration of principle for the analysis of clinical flow cytometry data as individual high-dimensional datasets using the principles of information geometry and statistical manifolds. Such an approach may harbor potential for the development of decision support tools and context-based search capability in clinical flow cytometry laboratories, and for the analysis of flow cytometry data as a proteomic discovery tool. Additional studies will be required to formally assess the potential utility of this approach for such specific applications.

LITERATURE CITED

1. Valet GK, Hoffkes HG. Automated classification of patients with chronic lymphocytic leukemia and immunocytoma from flow cytometric three-color immunophenotypes. Cytometry 1997;30:275–288.
2. Zamir E, Geiger B, Cohen N, Kam Z, Katz BZ. Resolving and classifying haematopoietic bone-marrow cell populations by multi-dimensional analysis of flow-cytometry data. Br J Haematol 2005;129:420–431.
3. Collins GS, Krzanowski WJ. Nonparametric discriminant analysis of phytoplankton species using data from analytical flow cytometry. Cytometry 2002;48:26–33.
4. Boddy L, Wilkins MF, Morris CW. Pattern recognition in flow cytometry. Cytometry 2001;44:195–209.
5. Toedling J, Rhein P, Ratei R, Karawajew L, Spang R. Automated in-silico detection of cell populations in flow cytometry readouts and its application to leukemia disease monitoring. BMC Bioinformatics 2006;7:282.
6. Zeng QT, Pratt JP, Pak J, Ravnic D, Huss H, Mentzer SJ. Feature-guided clustering of multi-dimensional flow cytometry datasets. J Biomed Inform 2007;40:325–331.
7. Roederer M, Hardy RR. Frequency difference gating: A multivariate method for identifying subsets that differ between samples. Cytometry 2001;45:56–64.
8. Pedreira CE, Costa ES, Arroyo ME, Almeida J, Orfao A. A multidimensional classification approach for the automated analysis of flow cytometry data. IEEE Trans Biomed Eng 2008;55:1155–1162.
9. Perez OD, Nolan GP. Phospho-proteomic immune analysis by flow cytometry: From mechanism to translational medicine at the single-cell level. Immunol Rev 2006;210:208–228.
10. Law M. Manifold Learning (Web Page). 2008. Available at: http://www.cse.msu.edu/~lawhiu/manifold/. Accessed January 25, 2008.
11. Lee S, Abbott AL, Clark N, Araman P. Active contours on statistical manifolds and texture segmentation. In the IEEE International Conference on Image Processing, IEEE; Genoa, Italy: 2005. pp 828–831.

12. Carter KM, Raich R, Hero AO. FINE: Information embedding for document classification. In the Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, NV, IEEE; 2008. pp 1861–1864.
13. Amari S, Nagoaka H. Differential-Geometrical Methods in Statistics. New York: Springer; 1990.
14. Carter KM, Raich R, Hero AO. Learning on statistical manifolds for clustering and visualization. In 45th Allerton Conference on Communication, Control, and Computing, Monticello, Illinois; 2007.
15. McKenna RW, Washington LT, Aquino DB, Picker LJ, Kroft SH. Immunophenotypic analysis of hematogones (B-lymphocyte precursors) in 662 consecutive bone marrow specimens by 4-color flow cytometry. Blood 2001;98:2498–2507.
16. Brunning RD, Borowitz M, Matutes E, Head D, Flandrin G, Swerdlow SH, Bennett JM. Precursor B lymphoblastic leukaemia/lymphoblastic lymphoma. In: Jaffe ES, Harris NL, Stein H, Vardiman JW, editors. World Health Organization Classification of Tumours: Pathology & Genetics: Tumours of the Haematopoietic and Lymphoid Tissues. Lyon: IARC Press; 2001. pp 111–114.
17. Terrell GR. The maximal smoothing principle in density estimation. J Am Stat Assoc 1990;85:470–477.
18. Balis UJ. Implementation of a region of interest-based query using vector quantization, generalized affine class-based vocabularies, and multimodal Chebyshev polynomial normalization to retrieve context-matched imagery from existing digital image repositories (abstract). Arch Pathol Lab Med 2005;129:811.
19. Habib LK, Finn WG. Unsupervised immunophenotypic profiling of chronic lymphocytic leukemia. Cytometry B Clin Cytom 2006;70B:124–135.
20. De Zen L, Bicciato S, te Kronnie G, Basso G. Computational analysis of flow-cytometry antigen expression profiles in childhood acute lymphoblastic leukemia: An MLL/AF4 identification. Leukemia 2003;17: 1557–1565.
21. Maynadie M, Picard F, Husson B, Chatelain B, Cornet Y, Le Roux G, Campos L, Dromelet A, Lepelley P, Jouault H, Imbert M, Rosenwadj M, Verge V, Bissieres P, Raphael M, Bene MC, Feuillard J. Immunophenotypic clustering of myelodysplastic syndromes. Blood 2002;100:2349–2356.
22. De Rosa SC, Brenchley JM, Roederer M. Beyond six colors: A new era in flow cytometry. Nat Med 2003;9:112–117.
23. Roederer M, Moore W, Treister A, Hardy RR, Herzenberg LA. Probability binning comparison: A metric for quantitating multivariate distribution differences. Cytometry 2001;45:47–55.