Published in final edited form as:
Cytometry A. 2008 September ; 73(9): 868–874. doi:10.1002/cyto.a.20610.

PolyChromatic Plots: Graphical Display of Multidimensional Data

Mario Roederer[1] and **M. Anthony Moody**[2]

[1] Vaccine Research Center, NIH, Bethesda, MD

[2] Department of Pediatrics, Duke University School of Medicine, Durham, NC

Abstract

Background—Limitations of graphical displays as well as human perception make the presentation and analysis of multidimensional data challenging. Graphical display of information on paper or by current projectors is perforce limited to two dimensions; the encoding of information from other dimensions must be overloaded into the two physical dimensions. A number of alternative means of encoding this information have been implemented, such as offsetting data points at an angle (e.g., three-dimensional projections onto a two-dimensional surface) or generating derived parameters that are combinations of other variables (e.g., principal components). Here we explore the use of color to encode additional dimensions of data.

Methods—PolyChromatic Plots are standard dot plots, where the color of each event is defined by the values of one, two, or three of the measurements for that event. The measurements for these parameters are mapped onto an intensity value for each primary color (red, green, or blue) based on different functions. In addition, differential weighting of the priority with which overlapping events are displayed can be defined by these same measurements.

Results—PolyChromatic Plots can encode up to five independent dimensions of data in a single display. By altering the color mapping function and the priority function, very different displays that highlight or de-emphasize populations of events can be generated. As for standard black-and-white dot plots, frequency information can be significantly biased by this display; care must be taken to ensure appropriate interpretation of the displays.

Conclusion—PolyChromatic Plots are a powerful display type that enables rapid data exploration. By virtue of encoding as many as five dimensions of data independently, an enormous amount of information can be gleaned from the displays. In many ways, the display performs somewhat like an unsupervised cluster algorithm, by highlighting events of similar distributions in multivariate space.

Keywords

Flow Cytometry; Data Analysis; Graphical Displays; Multivariate Data Analysis

Introduction

The ability to analyze complex data sets is limited by the amount of information that can be presented in any single graph type. Typically, flow cytometric data is analyzed by viewing multiple pairs of parameters simultaneously (e.g., using dot plots, contour plots, or density plots). Subsets are identified by hierarchical gating; distributions are compared for these

Address correspondence to: Mario Roederer, Vaccine Research Center, NIH, 40 Convent Dr., Room 5509, Bethesda, MD 20892-3015, Phone: 301-594-8491, Email: Roederer@nih.gov.

subsets by arraying graphs side-by-side. While effective, this mode of data presentation becomes limiting as the number of measured parameters increases. For a two-color flow experiment a single bivariate plot can suffice while for an 18-color experiment 153 bivariate plots are needed to cover all possible fluorochrome combinations.

A number of techniques have been used to increase the information content of graphs (1). For example, “3D” rendering adds a dimension; however, printed 3D graphics are still two-dimensional and distort the represented object. Care must be taken to avoid hiding data points behind other objects in the plot and to avoid introducing excessive or paradoxical distortion. Nonetheless, in some contexts these graphs can aid in the analysis of data sets. Unfortunately, compression of higher-order data onto two dimensional graphs is subject to even greater distortion.

Principal components analysis is another technique for evaluating higher-order data. This approach uses least-squares regression to determine a new set of data axes that account for the majority of the observed variations (2). Principal components analysis does not allow for the pre-selection of axes of most interest and for some data sets the calculated axes may not be scientifically meaningful. Furthermore, if there is not a dominant axis (i.e., the data are truly orthogonal) the dimensionality of the data cannot be reduced. While such analytical tools are useful they may not be intuitively obvious.

Here we report on a graph type, which we term “PolyChromatic Plots”, where color is used to convey information about other parameters in the data. These graphs can convey information from as many as five independent measurements simultaneously. PolyChromatic Plots have a number of advantages and disadvantages over standard graphical displays; they may provide a powerful adjunct in the exploration of complex data sets.

Materials and Methods

Example data set

Human PBMC were stimulated with staphylococcus enterotoxin B (SEB) superantigen for six hours. Cytokine-expressing T cells were identified using standard intracellular cytokine staining protocols; cells were stained with antibodies detecting CD3 (conjugated to Cy7APC), CD4 (FITC), CD8 (Quantum Dot (QD) 585), CD27 (Cy5PE), CD45RA (QD605), CCR5 (Cy7PE), CCR7 (Alexa (Ax) 680), IL2 (Ax594), IFNγ (PE), TNF (APC), CD14 (Pacific Blue) as well as the viability dye ViViD. Approximately one million PBMC were analyzed on a modified 18-color LSR II.

Results

We hypothesized that in a bivariate plot of flow cytometric data additional information could be displayed by using measured parameters to define the color of an event. For example, in a dot plot, color intensity of an event could be related to the measured value of a third parameter (such as fluorescence intensity); thus, events which have similar values in that third parameter would display a similar color intensity and could be distinguished in the dot plot from other events.

Thus, we implemented a “color mapping” algorithm, in which the red, green, or blue component of an (RGB) color assigned to an event were defined by one, two, or three measurement parameters for that event. In general, the color mapping is a monotonic function on the measurement parameter, although other maps (e.g., step functions, non-linear scalings) are possible. The choice of RGB color space was not arbitrary; “perceptual” color mappings such as the hue-saturation-value (HSV) system have lower resolution in certain regions of the color space (3). In addition, many investigators are already familiar with heat maps (using red-green intensity) and so such combinations are likely to be more easily interpreted. In addition, the human eye has cone receptors for red, green, and blue wavelengths and computer and television displays use the RGB system to display color.

Consider a 48-bit color model, in which the color of any pixel is given by three different 16 bit values (one each for red, green, and blue). Thus, each red, blue, or green color value ranges from 0 to 2^{16}-1 (65,535); the color mapping function therefore transforms a measured parameter into a value ranging from 0 to 65,535. A simple transformation would linearly re-scale the measurement range of a parameter from its minimum and maximum values to this range. However, the range of measured parameters often occupies only a subset of the potential range of measurements (e.g., the fluorescence signal might range only into the second decade), thus assigning many color gradations to ranges of fluorescence with no measured events.

The simplest scaling we use is a linear transformation of displayed measurement space, ranging from the 1st to the 99th percentile of displayed values:

$$C=65,535 * (F - F_1)/(F_{99} - F_1)$$

where C is the R, G, or B color value; F is the display-transformed measured fluorescence of the parameter assigned to that color; F_1 is the 1st percentile of all measurements for that parameter, and F_{99} is the 99th percentile for that parameter. (Note: F is expressed in graph units: if the parameter is logarithmically scaled, then F is related to the logarithm of the fluorescence; if the parameter is linearly scaled, then F is proportional to the fluorescence; if the parameter is scaled by a biexponential, then F also is scaled by the same biexponential). This results in a color gradation that starts at black (C = 0) for the events with the least fluorescence, ranges to full color intensity (C = 65,535) for the events with the greatest fluorescence, and having a straight-line relationship in between. See the purple line Figure 1A and 1B that illustrates the relationship between the color intensity (shown on the right of the histogram) and the measured fluorescence intensity (shown on the abscissa of the histogram); we term this form of color mapping “Uniform”.

An alternative color mapping is obtained by using a linear transformation of the percentile value:

$$C=65,535 * (P - 0.01)/(0.99 - 0.01)$$

where P is the percentile of the measurement within the entire distribution (i.e., for an event with the median fluorescence, P = 0.5). Unlike uniform mapping, this function changes the color linearly within the cumulative distribution of events; i.e., in regions of the measurement space with many events (local peaks), the mapped color intensity rapidly changes as fluorescence intensity increases. This color mapping function therefore tends to make cells within a subset appear more heterogeneous, since the colors assigned to the lower end of fluorescence of the subset can be very different than those assigned to the upper end of the same subset. In Figure 1A and 1B, this mapping function is shown by the orange line (which is equivalent to the cumulative distribution function, CDF, between the 1st and 99th percentile of fluorescence); we term this form of color mapping “Percentile”.

Often, however, the opposite functionality is desired: i.e., we would like cells within the same subset (or that share similar fluorescence characteristics) to be displayed with similar colors. For this, we define the color mapping function:

$$C=65,535 * (X - 0.01)/(0.99 - 0.01)$$
$$F_i=[\text{Log}(H_{max}+1) - \text{Log}(H_i)]^3$$

where X is the percentile of the event's measurement value within the distribution of the function F; F is computed from the histogram H of the parameter's fluorescence within the range of the 1st to the 99th percentile of fluorescence; H_i is the value of this histogram at channel i and H_{max} is the maximum value of all H_i. This color mapping changes very slowly in regions of measurement space with many events (local peaks) and thus behaves as the opposite of the Percentile function. The result is that events which are closely related in terms of fluorescence characteristics are similarly colored. In Figure 1A and 1B, this mapping function is shown by the green line; we term this form of color mapping "Clustered".

To explore the use of these functions, we used a multicolor immunofluorescence dataset (gating shown in Supplemental Figure 1). Figure 1 illustrates the PolyChromatic Plot color mapping functions applied to this data set. The colored lines show the relationship between the fluorescence distribution (the black histogram) and the selection of the intensity of the output color for an event of any given fluorescence intensity.

The advantage of PolyChromatic Plots is that color heterogeneity can be used to discriminate subsets in the data. The three color mapping functions we explored generate quite different views. "Uniform" mapping changes the color gradually as the displayed fluorescence measurement increases. A potential disadvantage of this is that if cells are highly clustered, most of the color change will occur in "empty" regions, resulting in less heterogeneity in the final display. "Percentile" mapping, in which color changes according to the distribution of measured fluorescences, generates the greatest color heterogeneity amongst the displayed cells. This mapping has the potential to highlight differences within a tightly clustered population, but this may be a disadvantage in that clustered cells will be drawn with different shades of color and may create the appearance of a difference that is not scientifically significant. The "clustered" function tends to keep color relatively constant for cells within a subset, and changes colors most rapidly in regions of the measurement space with few events – thus providing the greatest discrimination between subsets.

Figure 1 shows the successive assignment of one, two, or three measurement parameters to control the blue, red, and green colors of events in the displays. In a three-color system with three different color mapping algorithms, there are 27 possible displays that can be generated for any bivariate dot plot. These may look quite different; thus, it will likely be important to look at all possible combinations and select the one that best conveys the conclusions being drawn about the data.

Frequency information

The primary disadvantage of the polychromatic plot is the loss of frequency information (as is true for standard black-and-white dot plots), since events with the same measurement values for the two displayed axes will obscure each other (only one event will appear in any given location/pixel). Pseudo-color plots use color to convey frequency information (e.g., in Supplemental Figure 1, blue indicates low-frequency; red indicates high frequency). Since polychromatic plots use color to indicate measurement parameters, frequency information is largely lost (except as provided by the density of the dots, which can be misleading when more than about 10,000 events are displayed in a bivariate graph).

More importantly, however, rare populations may be obscured by more frequent subsets occupying the same region of the plot. Since plots are generally drawn in the order that the events occur, if two subsets occupy the same region of space, the more frequent population will “cover” the rarer subset. Paint-Agate™ software introduced the ability to draw populations such that the rarest subsets are drawn last (“on top”); this emphasized rare subsets and made them more easily viewed.

For PolyChromatic Plots, we take a different approach with similar results. For any given display, a separate priority weighting can be assigned to each of the three colors. For each event, an overall drawing priority is computed:

$$P=C_R * P_R + C_B * P_B + C_G * P_G$$

where C_X is the color value assigned for a given event to the color X (red, green, or blue) and P_X is the priority assigned to color X. For any given output pixel in the graphical display, only the event with the highest priority score is drawn. If the P_X values are zero (neutral), then all events have the same priority and the display functions as a standard dot plot. If $P_X > 0$, then events with that color will be over-emphasized (they will be drawn “on top” of other events); if $P_X < 0$, then events with that color will tend to be de-emphasized (as other events would have higher priority).

Figure 2 illustrates the effects of changing color prioritization using cytokine expression data. To emphasize the phenotypes of cytokine-expressing cells, the drawing priority can be increased for those events. As shown in Figure 2B, increasing the priority of events with high cytokine signals significantly emphasizes those relatively rare cytokine-expressing cells making those populations much more apparent. Caution must be taken, however, to not over-interpret the appearance of the plot as implying a high frequency of that population.

Since color priority is assigned on a continuous scale, there are an infinite number of distinct displays. However, most of the differences in displays can be explored by comparing low (<0), neutral (0), or high (>0) priority for each of the three colors (for a total of 27 combinations of priorities). Such an analysis is shown in Supplemental Figure 2.

Discussion

Studies of visual perception and the use of color for enhancement of data analysis have a long history (4). Many early studies on the use of color to enhance visual sensitivity were focused on improved recognition of military targets, and the data showed that in certain cases the use of color was helpful while in others it was not (5). One consistent finding, however, was that adding channels of stimulation improved the ability of people to discriminate among different populations within data sets (6,7). Based upon this, many novel schemes have been created, including the use of faces to encode up to 18 dimensions of data (8). Color encoding has been shown to improve cluster recognition in complex data sets but also shows limitations especially in certain areas of perceptual color space (3).

In the development of PolyChromatic Plots we used the RGB color space. Colors generated using this model vary more uniformly across all regions to display variations in the data. The hue-saturation-value (HSV) space is a “perceptual” color map, but it has disadvantages compared to RGB space. Since the hue value is mapped onto a color circle, values near the bottom (1st percentile) and top (99th percentile) would have nearly the same color appearance. This could be avoided by restricting hue to a subset of possible colors but with a loss of resolution. At low saturation the resolution of hue becomes reduced and at low value (brightness) both saturation and hue lose resolution. Other perceptual color mappings have both of these latter problems. The cyan-magenta-yellow-black (CMYK) system maps color in a manner similar to the RGB system and offers no obvious advantage. Since there are many examples of red-green color maps in the literature currently, the RGB system will likely appear more "natural" at first glance.

PolyChromatic Plots are a novel display for flow cytometry data. Their primary purpose is to assist in the exploration of multivariate data by revealing heterogeneity (or homogeneity) of underlying subsets. PolyChromatic Plots are different than other kinds of displays in that they use shades of color to convey the values of some measurement parameters (not necessarily those being used to define the positions of the dots in the graph itself). This approach has a number of advantages. It can aid in data exploration by highlighting clusters of events that would not be evident except by multiple rounds of gating and comparing different gated subsets. Principally, it provides a fast way to explore multiple dimensions simultaneously. PolyChromatic Plots can display information from as many as five independent measures (two displayed as Cartesian coordinates along the ordinate and abscissa; and three displayed via the RGB color model).

PolyChromatic Plots shift the time utilization during data analysis. More time is necessary during data exploration to determine the optimal combination of variables that generate informative PolyChromatic Plots – leading to a longer session during the initial analysis of prototypical samples. However, this added time is more than offset by the savings during the subsequent comparison phase – where many samples are compared. Since PolyChromatic Plots can convey in a single image information from as many as five parameters (which may require 3 or more "standard" graphs), they may be particularly useful for scanning or interpreting data from large numbers of samples.

PolyChromatic Plots also have a number of disadvantages. The primary one is that frequency information is obscured when large numbers of events are displayed. The ability to prioritize colors will emphasize or de-emphasize certain subsets out of proportion to their representation in the raw data. Hence, care must be taken when evaluating PolyChromatic Plots to understand that the visual appearance of a plot may suggest larger or smaller populations than are actually present. Indeed, these graphs should be used simply to convey patterns of expression; any quantitative information should be provided by standard means.

Given the large number of variables that can generate many different views of exactly the same data (as illustrated, for example, in Figure 2 and Supplementary Figure 2), care must also be taken to use the same settings for different samples being compared. For publication purposes, we recommend that the settings for any given graphic be explicitly shown (for example, in the Figure Legend); if multiple PolyChromatic Plots are shown, it must be assumed that all use the same settings unless explicitly noted by the authors. Since some of the color assignment algorithms depend on the scaling used to display the data, different displays may arise when using log-scaled vs. biexponential-scaled (9,10) data.

Finally, a disadvantage of PolyChromatic Plots is their inherent use of color to convey information. This puts individuals with color blindness at a disadvantage for evaluating the plots either as investigators or as consumers of lectures or journal articles. The most common form of color-blindness, X-linked red-green defect, affects from 8–10% of males and up to 0.5% of females (11). Other forms of color vision defects are more rare, but all would alter the appearance and the utility of PolyChromatic Plots. We also note that the ability to accurately reproduce color in print or by projection can impact the quality of the display and the analysis.

PolyChromatic Plots are not a mathematical clustering technique nor are they quantitative. Unlike vector algebra methods such as support vector machines, linear algebra methods such as principal components analysis or multidimensional scaling, no calculation of the robustness of visibly apparent groupings is performed. Once such groupings are suggested by PolyChromatic Plots these methods can be applied to determine, in a quantitative fashion, the magnitude of the differences between subsets. The advantage of PolyChromatic Plots is that, being a visual display method, they are suited to use in presentations and publications where pictures may be more easily interpreted than tabulated statistics.

Clearly, the most powerful aspect of PolyChromatic Plots is their ability to provide “cluster-like” displays of data. While not a clustering algorithm per se, the displays can show distinct subsets fairly clearly. The ability to discriminate subsets by this graphical technique is directly dependent on the color mapping function. It is in this area that future development will provide the most fruitful advances; conceivably, more sophisticated color mapping algorithms will enhance the contrast between different subsets of cells. In its current form, the use of side-by-side comparisons of multiple PolyChromatic Plots can encourage new comparisons between data sets.

In conclusion, we describe here a graphical display for flow cytometry that encompasses as many as five parameters independently. These displays will aid in the analysis and exploration of high-dimensional data, and thereby render complex data analysis more efficient and presentable.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Drs. Pratip Chattopadhyay and Joanne Yu for the data set used here, Dr. Chattopadhyay for suggesting color prioritization, and Drs. Chattopadhyay and Carl-Magnus Hogerkorp for reading the manuscript. This work was supported in part by the Intramural Research Program of the NIAID, NIH.

References

1. Streit M, Ecker RC, Osterreicher K, Steiner GE, Bischof H, Bangert C, Kopp T, Rogojanu R. 3D Parallel Coordinate Systems—A New Data Visualization Method in the Context of Microscopy-Based Multicolor Tissue Cytometry. Cytometry A. 2006; 69:601–611. [PubMed: 16680710]
2. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 1901; 2(6):559–572.
3. Ware C, Beatty J. Using color dimensions to display data dimensions. Human Factors. 1988; 30(2): 127–42. [PubMed: 3384442]
4. Cohn T, Lasley D. Visual Sensitivity. Annual Review of Psychology. 1986; 37(1):495–521.
5. Christ R. Review and Analysis of Color Coding Research for Visual Displays. Human Factors. 1975; 17(6):542–570.
6. Eriksen C, Hake H. Multidimensional stimulus differences and accuracy of discrimination. Journal of Experimental Psychology. 1955; 50(3):153–160. [PubMed: 13252189]
7. Millar G. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review. 1956; 63(2):81–97. [PubMed: 13310704]
8. Chernoff H. The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association. 1973; 68(342):361–368.

9. Bagwell C. Hyperlog - a flexible log-like transform for negative, zero, and positive valued data. Cytometry A. 2005; 64(1):34–42. [PubMed: 15700280]

10. Parks DR, Roederer M, Moore WA. A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytometry A. 2006; 69(6):541–51. [PubMed: 16604519]

11. Swanson W, Cohen J. Color vision. Ophthalmology Clinicls of North America. 2003; 16(2):179–203.

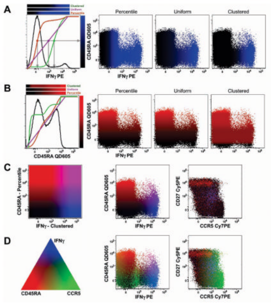

Figure 1.

Using color to indicate undisplayed measurement values. (A) The three different color mapping functions are shown for the IFNγ measurement. In the first graphic, the histogram of IFNγ fluorescence for stimulated, CD4$^+$ T cells is shown in black (see Supplemental Figure 1 for gating). Each of the colored lines shows the mapping of IFNγ intensity (abscissa) to a shade of blue (heat map shown on the right side of the graphic). The purple line represents "uniform" mapping, where the color varies smoothly from black to full blue intensity linearly (with respect to the displayed fluorescence scale), bounded by the 1st and 99th percentiles of fluorescence. The orange line represents "percentile" mapping, where the color changes according to the fraction of events below any given intensity. For this, the cumulative display function (CDF) is used; here, the median IFNγ fluorescence is assigned the mid-point color between black and blue. The green line illustrates the "clustered" mapping (see text for derivation). The grey lines identify the relative IFNγ fluorescence intensity that maps to the midpoint of blue color intensity for each of the three different functions. The three heatmaps above the histogram show the complete output color mapping as a function of IFNγ intensity. The three bivariate dot plots illustrate the result using each of the three different color mapping functions. (B) The same illustration as in (A), but using the CD45RA measurement to determine the intensity of the red color assigned to each event. Note how the "clustered" function tends to give events in the same cluster the same color; the red color intensity changes most rapidly in the CD45RA distribution where there are relatively fewer events. (C) The color mapping schemes illustrated in (A) and (B) are simultaneously applied. The first box illustrates the mapping of color within the bivariate distribution of IFNγ and CD45RA; cells expressing both markers will be drawn with a mixture of red and blue, resulting in shades of violet; cells expressing neither will be close to black. The two biviariate graphs show the results of this color mapping. (D) Adding a third parameter, CCR5, to control the green color, results in the full range of colors being used. The triangle illustrates roughly the mixture of colors that are achieved by different amounts of each parameter: for example, events expressing only CD45RA will be red; those with CD45RA and CCR5 will be yellow (i.e., red + green); cells expressing all three markers will be white. Note that it is impossible in two dimensions to display a legend for all possible colors combinations; the triangle is only a rough guide. The two bivariate displays show the full color mapping result; subsets of cells can be readily identified by the color scheme.

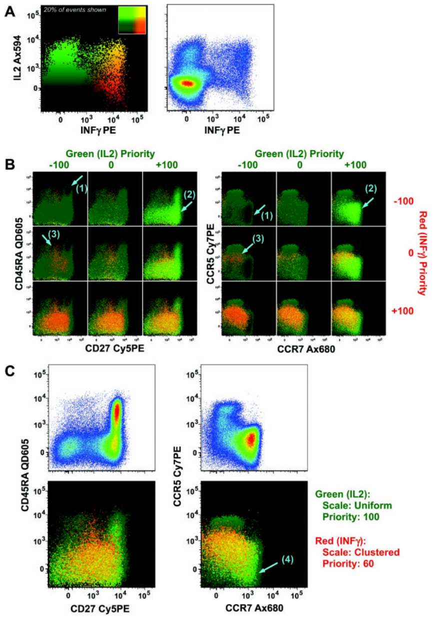

Figure 2.

Color priority to control the highlighting of certain events. (A) A bivariate display of IL2 vs. IFNγ from the same data as shown in Figure 1. Both a standard pseudo-color dot plot and a polychromatic plot are shown. The inset of the polychromatic plot shows the color mapping scheme for the two parameters. (B) The same data graphed in (A) are shown for two different pairs of other parameters. For each pair, the nine combinations of low (−100), neutral (0), and high (+100) prioritization for the two color assignments are shown. The prioritization value controls how events which occupy the same pixel are displayed: the event shown is that with the highest priority. When both cytokines are given a low priority, the non-cytokine producing cells tend to obscure cytokine-producing cells, leading to emphasis of darker-colored events (low IL2, low IFNγ), which are typically naïve ($CD45RA^+CD27^+CCR7^+CCR5^−$), as shown by the arrow marked (1). Selectively increasing the priority for green (IL2) events highlights the cells producing only IL2, which tend to be naïve or central memory ($CD45RA^±CD27^+CCR7^+CCR5^−$), shown by the arrows marked (2). As the priority for IFNγ is elevated (with IL2 low priority) the cells producing only IFNγ become emphasized; these tend to be effector memory ($CD45RA^−CD27^−CCR7^−CCR5^+$), as shown by the arrows marked (3). (C) The final display chosen emphasizes green (IL2) fully and red (IFNγ) partially. This allows for cells producing only IL2 (arrow marked (4)) to be emphasized over other cytokine producing cells, which in turn are emphasized over non-cytokine-producing cells. The upper graphs illustrate the distributions of the surface markers for all gated events using a standard pseudo-color plot.