

Visual analytics for immunologists

Data compression and fractal distributions

Elena N. Naumova

Department of Public Health and Family Medicine; Tufts University School of Medicine; Boston, MA USA

Key words: immunology, immune response, T-cell repertoire, memory, influenza, visual analytics, visualization, data compression, fractal, distributions

Submitted: 06/15/10

Revised: 06/30/10

Accepted: 06/30/10

Previously published online:
www.landesbioscience.com/journals/selfnonsel/article/12876

Correspondence to:
 Elena N. Naumova;
 Email: elena.naumova@tufts.edu

Visual analytics is the science of analytical reasoning that facilitates research through the use of interactive visual interfaces. New techniques of visual analytics are designed to aid the understanding of complex systems versus traditional blind-context rules to explore massive volumes of interrelated data. Nowhere else is visualization more important in analysis than in the emerging fields of life sciences, where amounts of collected data grow increasingly in exponential rates.

The complexity of the immune system in immunology makes visual analytics especially important for understanding how this system works. In this context, our effort should be focused on avoiding accurate but potentially misleading use of visual interfaces. The proposed approach of data compression and visualization that reveal structural and functional features of immune responses enhances systemic and comprehensive description and provides the platform for hypothesis generation. Further, this approach can evolve into a powerful visual-analytical tool for prospective and real-time monitoring and can provide an intuitive and interpretable illustration of vital dynamics that govern immune responses in an individual and populations.

The undertaken explorations demonstrate the critical role of novel techniques of visual analytics in stimulating research in immunology and other life sciences and in leading us to understanding of complex biological systems and processes.

Introduction

There is no greatness where there is not simplicity, goodness, and truth.

-Leo Tolstoy, novelist and philosopher (1828-1910)

A new field of visual analytics, emerging from research in scientific visualization and information visualization, is promising to be exceptionally useful in immunology, where large volumes of complex data must be processed and communicated. Visual analytics provides a platform for a theory-based approach to the selection of graphical tools emphasizing analytical reasoning.¹ With the growing need for effective communication across various disciplines, excellence in performance of diverse analytical tasks, e.g., detecting unusual clusters or patterns, discovering potential linkages, and monitoring change, is a highly desirable feature in modern immunology. In fact, the acceleration of computational power offers an effective medium for data processing, synthesis and communication.

As with any developments in computer-based systems, the proper use of new tools requires proper training. The widespread availability of “do-it-yourself” computerized gizmos has stimulated an enormous production of graphs, charts, plots and maps in both public media and research literature. Practically all fields of science utilize data visualization tools, unfortunately, at times, with little guarantee of

high quality in the visualizations produced. In constructing a graphical display, the ethical challenges, the role of communication and significance of methodology should not be underestimated. As it has been pointed out in our early work, a graphical display must rest on and adhere to a set of operational rules and research methodologies.² We believe that as a product of cognitive process, a visual representation of information implemented correctly should contain new knowledge, convey meaning to the viewers and avoid any misleading interpretations. Although there have been impressive attempts to develop useful guidelines for constructing sound visual displays and proper interpretation,³ there is still a room for improvement.

In this communication we describe how the methodology of visual analytics can be implemented in a routine setting and in the context of a study of a human immune response to influenza virus. Our exploration of memory T-cell repertoires starts by setting a hypothesis, then moves to the nuances of data collection and analytical procedures, progresses to repertoire's distributional representation and finally, ends with inferences on mechanisms as we arrive to a proposed fractal structure of TCR. We utilize a set of illustrative examples, similar to our early paper,² supplemented with additional information.^{4,5} We provide step-by-step instructions for compression of large volumes of data and offer guidance for proper deciphering of each graph. We illustrate how visual analytics can promote a scientific breakthrough when a "great idea" along with "great data" in a "great display" evoke a "great discovery." To aid with the analytical process we will follow these easy-to-remember lines:

Ordering, ranking, tossing, arranging,
Summing, dividing,
 subtracting, equating,
Transforming, transposing,
 sometimes integrating,
Rounding, dithering, approximating,
Compacting, inverting
 and simply forgetting,
But never ignoring,
 refusing, regretting.

We conclude this paper with a list of pitfalls and challenges relating to visual displays and suggest potential solutions. We hope this work demonstrates how new approaches to visualization may improve our understanding of accumulated experimental results, and helps the scientific community in strengthening their confidence in grasping a broad range of scientific visuals and in creating their own ones.

Visual Analytics for Depicting Fractality of Immune Response

In our early communication we concluded that a graph must rest on and adhere to proper definitions, a set of operational rules and research methodologies. As a language of communication, visuals are designed to convey complex contextual concepts. We devised a simple rule of graph construction: a graph without a well-understood and well-defined statistical context or logical path, without good visual properties and without cognizance of the audience is hardly worth drawing.

Therefore, let us start with the first stage of data arranging by identifying the logical path of forming a hypothesis and outlining working definitions.

Step 1: Forming a "great idea". The first step reflects the process of formulating a research hypothesis and working definitions, on which scientists will rely in forming a visual display. A sound hypothesis is the lynchpin of good science regardless of how a hypothesis is formed: by building theoretical concepts or from observed data. The soundness of the scientific premise is required to arrive at the truth or falsity of the premises. The description below shows how immunologists state their research hypothesis about an anticipated mechanism of the polyclonal immune response, as has been defined earlier.^{5,6}

Illustrative example. TCR determines the diversity and efficacy of the immune response to viral challenge and represent one of the most intriguing current problems in immunology. A T-cell repertoire (TCR), composed of two protein chains, is predefined by the use of a particular V and J region combined with the pseudo-random DNA sequence of

the complementary determining region 3 (CDR3). The manner, in which naïve T cells give rise to the immune memory repertoire, how the repertoire is sustained, and how it changes in time, represents an example of a complex system. Assuming that high avidity T cells will be selectively expanded following successive challenges with antigenic peptide presented by the major compatibility complex (MHC) molecules, it is predicted that a memory repertoire would contain multiple copies of T cells expressing limited diversity of T-cell receptors. **Figure 1** below reflects an example of the crystal structure of the TCR-peptide-MHC complex; the CDR3 regions are the segments of the TCR, in closest contact with antigenic peptide (shown in green and magenta). The composition of clonotypes along with their distribution copies defines the main repertoire properties (**Fig. 1**).

In order to test this hypothesis, experimentally and analytically, working definitions for a clonotype's repertoire, and relevant distributions should be in place before an analysis starts. The excellence of a scientific hypothesis depends upon the excellence of working definitions. If the terms are improperly defined, then the entire endeavor—including the statistical analysis and developed visuals—falters. In our case, if a definition of "clonotype" is not accurate, the more complex statement, containing the term, will not produce knowledge. In the strictest sense, the proper definition should include an understanding of the reasoned cause.

Illustrative example. As a first step, we define "a clonotype", the smallest observable unit used for modeling and analysis and three clonotype-related entities: clonotype repertoire, distribution and rank-frequency summary, which are the products of aggregating clonotype-specific information at various levels. The CDR3 sequence of a T cell defines its clonotype. In the context of recognition of an antigen fragment, the term clonotype can refer to the amino acid sequence of the CDR3. However, here we use the more restricted definition of clonotype as the nucleotide sequence of the CDR3. On a molecular level, almost all T cells express only one copy of the β -chain and for this reason

we will be generating β -chain CDR3 sequences (reviewed in ref. 7).

In order to devise a hypothesis, the studied elements of the system have to be described and quantified in term of their properties. For simplicity, we postulate that most properties of clonotypes are relatively invariant to experimental conditions and can be directly measured, as can some of the properties of the repertoire. We postulate that each of the three entities can be characterized by its “diversity”; whereas “diversity” is defined as a property of a TCR repertoire that reflects repertoire variability and can be measured at different levels; including CDR3 amino acid sequences and clonotype distribution. For instance, diversity of a clonotype distribution is associated with variety in rearrangement and the multiplicity of selective re-assortment.

Step 2: Collecting “great data”. A researcher must select and apply appropriate and accurate methodology for obtaining data relevant to a formulated hypothesis. This step determines the data quality that is essential for an analyst: biostatistician or bioinformatician, in forming various assumptions (such as statistical assumptions and assumptions of biological plausibility) and selecting procedures for analyses. An exquisitely executed statistical analysis performed at this step is necessary to generate valid research conclusions. From the principles of formal logic we know that the solid research requires validity of methods achieved by proper inferences drawn from the premises.

Illustrative example. The techniques for clonotypes detection and counting are beyond the scope of this paper and the details of such techniques can be found elsewhere.^{5,6} In brief, the memory T-cell repertoires were evaluated by using cultures generated from the peripheral blood mononuclear cells of a healthy mid-aged adult with the strong immune response to the influenza A virus peptide M1₅₈₋₆₆. The blood samples were collected at two time points: in 1994 and 2004. The unit of measure of the T cell repertoire is a clonotype, represented by the unique DNA sequence that encodes the complementary determining region 3 (CDR3) of TCR β -chain and is quintessential for peptide

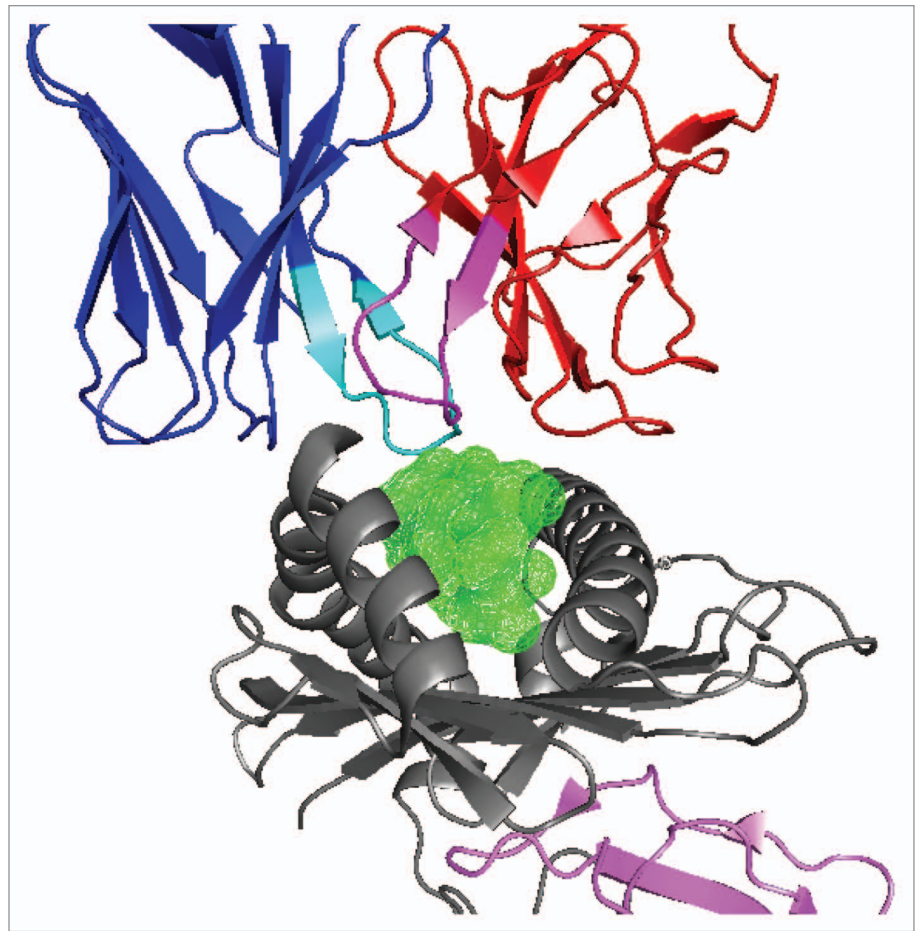


Figure 1. Illustration of the T-cell receptor with clonotype defining region with respect to working definitions, hypothesis and main concepts of data collection. The α -T-cell receptor heterodimer (in blue) is generated by a rearrangement process that results in a random section of genetic information inserted in the position that will encode the part of the β -chain (magenta) that contacts the antigen-derived peptide (green spheres). This random piece of genetic material can be identified and all T cells with the same random piece of DNA counted. They are all assumed to be related and the number reflects the expansion of that T cell. This random genetic segment (magenta) defines the clonotype, the primary unit of investigation. The image of the 3-D structure of the clonotypical T-cell receptor (clone JM22)-M1(58-66)-HLA-A2 (PDB1OGA) was created using MacPyMol (DeLano Scientific, LCC).

recognition or avidity. The clonotypes were identified by the DNA polymerase chain reaction of the TCR β -chain CDR3, subcloning and sequencing.

Step 3: Naming names and forming orders—“ordering, ranking, tossing, arranging” The next stage of data collection involves the process of “naming names and forming orders” following the first line of the rhyme “ordering, ranking, tossing, arranging.” To arrange the data that describes repertoires, each entity or clone/clonotype should be named or assigned to a unique identifier. To order the clonotypes for an observed repertoire we have developed and utilized a

clonotypes nomenclature, a system for assigning a unique name according to a set of specific properties.⁸ The proposed nomenclature has a number of important features: 1) is exhaustive, thus ensures each clonotype has a unique identifier; 2) represents an open system, thus allowing an expansion of identifiers; and 3) is compartmental, thus permits manipulation, sorting, selection, and arranging of identifiers. The decision to form a nomenclature for detected clonotypes stems from the desire to describe not only the composition of observed clonotypes but also to determine the abundance of each representative in the sample and further relate

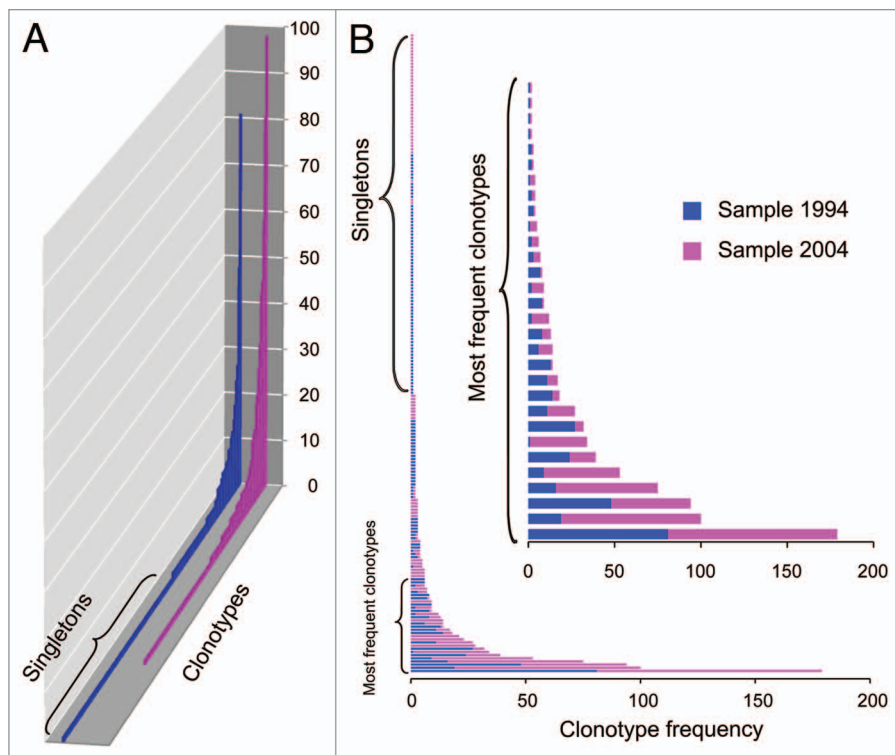


Figure 2. Relative frequencies of clonotypes plotted in descending order: a three-dimensional view of clonotypic frequencies (A) and a stacked bar-graph (B) for samples collected at two time points (1994 and 2004). An inset illustrates most dominant clonotypes.

the description to the main properties of a repertoire, such as diversity, complexity and stability.

To determine a clonotype's abundance, we count the number of times a unique clonotype appears in a sample, e.g., how many times a clonotype was sequenced and identified. Formally speaking, the conceptual basis of a clonotype distribution incorporates the number of times any particular clonotype is observed. Specifically, a repertoire is a set of unique CDR3 sequences: $C = \{C_1, C_2, \dots, C_i, \dots, C_N\}$ where C_i is a unique clonotype and N is a number of unique clonotypes with some pre-selected properties. A null set would indicate no measurable response. A clonotype distribution described as a repertoire, $C = \{C_1, C_2, C_3, \dots, C_N\}$, with each clonotype being observed a number of times, $V = \{V_{C_1}, V_{C_2}, V_{C_3}, \dots, V_{C_N}\}$, so $M = \sum V$, where M is a number of sequences. Since the only clonotypes with the sufficient number of copies can be detected in experimental conditions, the minimum observable frequency is one copy. For the analysis of repertoires presented in our illustrative example we identified 494

bacterial colonies containing TCR inserts accounting for by 135 unique clonotypes at the first time point and 673 sequences with 97 unique clonotypes at the second time point.

Step 4: Preparing a "great display." Visualization is the process of constructing an informative view of large volumes of multidimensional data appropriately grounded in a statistical context and is critical in understanding subject-specific properties, structures and relations and in communicating the information, query or statement to both professional and lay audiences.

A research hypothesis dictates the selection of a visual aid and the selection of an analytical method, both of which require a special form of data preparation. In our example, to determine clonotypes' abundance and repertoire evenness, as the first step of TCR analysis we start with the simplest form of visualization for clonotypes abundance—a frequency plot. The next step is data arranged by plotting clonotype frequencies in a descending order. When the absolute frequencies or counts, are plotted in descending order, the shape of the clonotype's distribution is revealed:

a monotonic decline from the highest to lowest frequencies with a long tail consisting of singletons, or the clonotypes presented by only one copy (Fig. 2A). Here we note that only a few clonotypes contain multiple copies (shown in the inset panel of Fig. 2B). This step represents the first stage of raw data compression. To emphasize the relative contribution of specific clonotypes, the graph can be re-plotted with relative frequencies instead.

A direct comparison of clonotypic distributions based on the identifiers must be performed with caution. Furthermore, a distribution presented in this form, i.e., with a nominal scale, imposes serious limitations to the range of statistical analyses that can be performed.

Step 5: Making data analyzable—"summing, dividing, subtracting, equating" The next step is to convert a simple form of data organization, e.g. based on counts of elements, to an advanced form that permits elementary arithmetical operations. To enhance potential inferences and develop comprehensive description of a clonotype distribution, we assigned each clonotype a rank based on the absolute counts of copies; therefore Rank 1 consists of clonotypes observed only once, Rank 2 contains clonotypes observed twice, etc. This rank-frequency approach provides an efficient summary of data and allows further conceptualization of repertoire diversity. The technique of converting data to rank-frequencies summaries is well described in statistics and ecology, specifically, in species abundance analyses, and has been widely utilized to predict the probability of occurrence of new species.^{9,10} Furthermore, useful information can be extracted by targeted analysis of distributional components. In our example, these would be the fractions or proportions of singletons in different samples. Implicit in such ranking is the high-ranking clonotypes that have expanded more than others indicate a better fitness for the antigen.

When relative frequencies for each rank are plotted in an increasing rank order, it reveals a power-law-like form in the rank-frequency relationship. Figure 3 illustrates the rank-frequency summary for the clonotype distribution shown in Figure 2. This form represents the next

stage of data compression, which allows for a wide range of analytical procedures. The graph of the relation between the log-transformed rank and the log-transformed frequency is shown in the inset of **Figure 3**. The use of the whole range of log-transformed data, from its minimum to its maximum, helps to compare samples across times. The fitting of the power-law curve was satisfactory in its performance for the first part of the curve (with relatively low frequencies) but the fit for the right tail, representing clonotypes of high frequencies, required special treatment.

Grouping by rank not only reveals the power-law-like shape in the rank-frequency summaries, it also suggests a possibility for a fractal structure of the TCR distribution and even the potential mechanisms of arriving to this form. It is possible that in a rank-frequency summary, a “rank” might mimic a behavior of a real physical parameter or a latent process, reflect affinity/avidity properties, and/or the clonotype’s ability to proliferate or to bind.

Step 6: Curve fitting and model building—“transforming, transposing, integrating” To quantify rank-frequency summaries with rapid decay we applied a power-law equation, $y = a/x^b$, where x is the rank and y is the rank frequency. For the simplest of situations, plotting a log/log transformation of the data: $\log y = \log a - b \log x$, should yield a straight line, where the parameter a indicates the frequency of observing single copy clonotypes and parameter b describes the shape of the curve by indicating how rapidly the curve decays. In the first exploratory step of describing the rank-frequency summaries we employed a simple linear fitting procedure based on a log-transformation, thus plotting a log/log transformation of the data: $\log y = \log a - b \log x$, which should yield a straight line, as shown in the insets of **Figure 3**.

Observing how well the line fit in each segment, especially at extremes, is useful for understanding clonotype distributions: a poor fit for clonotypes with low frequency indicates that prediction of singletons can be over-estimated or under-estimated. A simple power-law equation might poorly approximate the sharp decay in the rank-frequency distribution,

providing greater deviations at lower rather than higher frequencies. Similarly, for distributions with a large fraction of clonotypes of high frequency the linear fit may be unsatisfactory. In complex situations, a repertoire distribution is best described by treating it as a composition of two components (**Fig. 4**). The separation of the curve into two portions at a critical point, x_c^* , ensures a good fit of the first portion to a power law. The first portion includes the ranks represented by the high number of clonotypes, i.e., the extensive single copy tails from **Figure 2**.

There are a number of ways to estimate the parameter for each segment. In

our example, this is performed by using a “broken-stick” regression approach. The estimates of a slope and an intercept for the two portions of log-log rank-frequency plots can be also obtained iteratively using a Monte Carlo approach. For our case, the estimation starts with an initial value of $a_0 = y_1$, $b_0 = 1 - a_0$, $x_{c0} = \exp((\log N + \log a_0)/b_0)$, where N is a number of unique clonotypes. The estimation progresses by allowing a random walk for an a -parameter, say of 0.0001. In each iteration, j , we assign new values for a b -parameter: one—for the first portion of the curve $b_j = \Sigma(\log(a_j/y_i)/\log x_i)/M$, M —is number of ranks, when $x_i \leq x_j$ and $b_j = \log(a_j N)/$

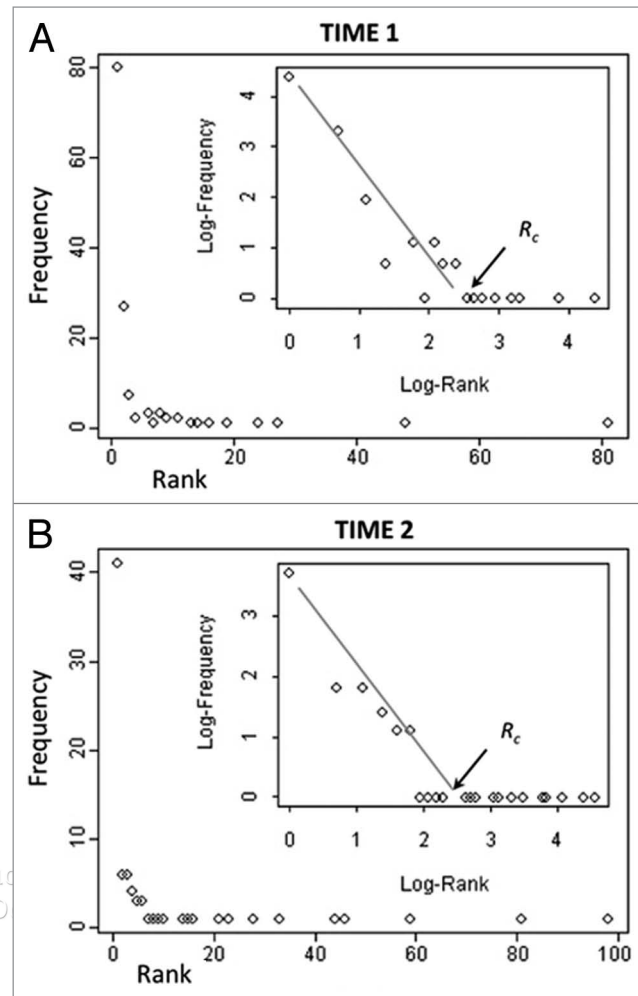


Figure 3. Visualization of experimental data after the second step of data compression: rank-frequency summaries of clonotypes distributions for samples collected at two time points (1994 and 2004). To describe properties of the clonotype distribution we assigned each clonotype a rank based on the absolute counts of copies (Rank 1 consists of clonotypes observed as single copies, Rank 2 those observed twice, etc.). By plotting relative frequencies in increasing rank order a power-law-like rank-frequency relationship is revealed. In the inset of each plot, the first steps of computational assessment are depicted: the predicted values are obtained by fitting the linear regression model applied to log-transformed data (shown as a green solid line).

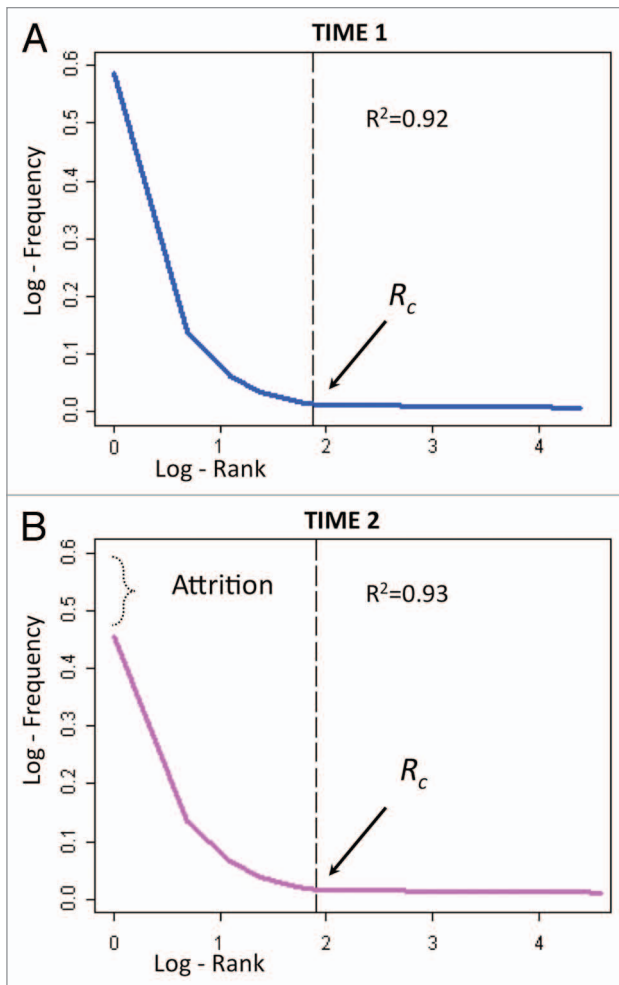


Figure 4. Computational assessment of rank-frequency summaries at two time points. The predicted values were obtained by fitting the linear regression model to log-transformed data, parameters and quality of model's fit (R^2 values) are also shown in the graph. R^2 values reflect percent variability explained and indicate a very good, over 90%, fit. Similarly to the insets in Figure 3, R_c indicate the critical inflection points, essential for a good fit. To ease the direct comparisons of predicted curves, the vertical axes use the units of relative frequency with identical ranges. The fitted curves demonstrate that flu-specific $\nu\beta 19$ repertoire underwent attrition in its low-frequency component.

$\log(x)$, for the second portion, when $x_i > x_{c_j}$. In each iteration, for a given rank, i , the difference (residual) between the observed frequency, y_i and expected frequency, y_i^* was calculated. The expected frequencies can be calculated as $y_i^* = a/x_i^{b_j}$ by minimizing sum of the squared residuals, $R = \Sigma(y_i - y_i^*)^2$. The estimates of a and b parameters, at which sum of squares reaches its' minimum, are used for estimating and plotting predicted curves shown in Figure 4. This visual display allows us to show that an observed repertoire had lost its low frequency component.

The separation of the curve into two portions substantially improves fitting and

suggests a possible biphasic mechanism for repertoire generation. This analysis triggers a number of new questions related to location and meaning of critical points. For example, a critical point may indicate a fraction of high frequency clonotypes reactive to the most acute infection. A hypothesis built on such assumptions can be tested in experimental conditions.

The conversion of repertoires into a set of regression coefficients helps to describe the shape of the curve by indicating how rapidly the curve decays. This step represents the next stage of data compression, so a long list of detected clonotypes, observed in large numbers of clonotype's copies,

is replaced by two or three numbers with a defined range of values. By including diversity measures as additional characteristics the analysis can be further enriched. At this stage, the conversion of repertoires into a set of quantifiable units provides an overall summary for a studied population, and allows for comparison of repertoire characteristics over time⁵ and/or over a spectrum of various experimental conditions.¹¹ Furthermore, comparisons between and among the groups of parameters can be performed by using parametric and non-parametric tests and guide efficient designs in expensive longitudinal studies.

Step 7: Pioneering a “great discovery”
 What if the phenomena of a power-law-like behavior observed in a single experiment, as described above, is common for an immune response to an antigen? To explore such responses in a wide range of experimental conditions we have to design the tools to study power-law structures. A power-law structure can be expressed as a self-similar iterative process. Power-laws and self-similarity are expected properties of a self-similar fractal system. A simple test to determine whether a system is fractal is to describe it by a recursive self-similar rule. A textbook example of a simple fractal, the Koch curve, starts with a single line in which the middle third is substituted by the sides of an equilateral triangle of the length of the original middle third. The rule is then repeated: substitute the middle third of each line with the sides of an equilateral triangle, etc.

In our case, we can take advantage of the power law exponent b to generate a recursive rule for a visual presentation of the clonotype distribution of the component of the repertoire for which b has been derived. As demonstrated earlier, the fitted curve of the clonotype's rank-frequency summary forms a decreasing sequence: $1/1^b, 1/2^b, \dots, 1/n^b$. When $b > 1$, as is our case, the curve diverges and can be mapped using a polygonal spiral. The resulting curve can be thought of as the sweep-out areas (squares) corresponding to clonotypes in the first rank, the second rank, etc. In fact, this is similar to the representation of the “golden spiral” encountered in many life sciences, such as biology, where growth follows an optimum pattern of accretion of related units.

We display a rank–frequency summary using a polygonal spiral, where the radii of the spiral reflect relative proportions of distinct clonotypes observed at a particular rank and form a sequence, where the ratio of two consecutive radii is $g = y_i/y_{i+1}$, where i is the rank. We let g be proportional to $(i + 1)b/b^*$ and b^* be equal to 1.67. We calculate the Cartesian coordinates for the points on a spiral using the distance to the origin, r_i , the angle λ_i and the reduction parameter, $1/b^{*i}$: $x_{ij} = (r_i \cos(\lambda_{ij}))/b^{*i}$ and $y_{ij} = (r_i \sin(\lambda_{ij}))/b^{*i}$, where i is the rank of j -th bead, $j = 1, 2, \dots, k_i$. The coordinates of the origin shift depend on the rank and the branch of a spiral.

This iterative approach for polygonal spiral mapping generates a representational model of the entire repertoire, shown in **Figure 5A**. In this representation, each segment of the spiral corresponds to the rank, and the number of beads within the segment corresponds to the number of clonotypes in that rank. As the curve branches out or trifurcates, at the end of the first segment of the spiral reflecting actual clonotype frequencies, the new two spirals continue as mirror images. The first segment of each of these two branches maps the clonotypes that appear twice or clonotypes of the second rank. The third spiral continues into the upper right quadrant and describes the remaining levels in a similar manner starting with rank three. The same process continues for the remaining portion of the distribution and then repeats again. These steps describe the self-similar recursive rule for generating a map of the repertoire. If one subtracts the first segment and the first two mirror spirals, the remaining portion of the model is an identical image of the entire model, providing a graphic demonstration of repertoire self-similarity. We define the mathematical properties of this fractal structure (Naumova, in preparation), which we called a Mondrian set after paintings of Piet Mondrian.¹²

This visual representation of a clonotypic structure suggests that a repertoire can be described as a fractal system (**Fig. 5B**). Furthermore, a temporal link helps to identify changes occurred over time: while the fraction of singletons decreases, the overall shape remained (**Fig. 6**).

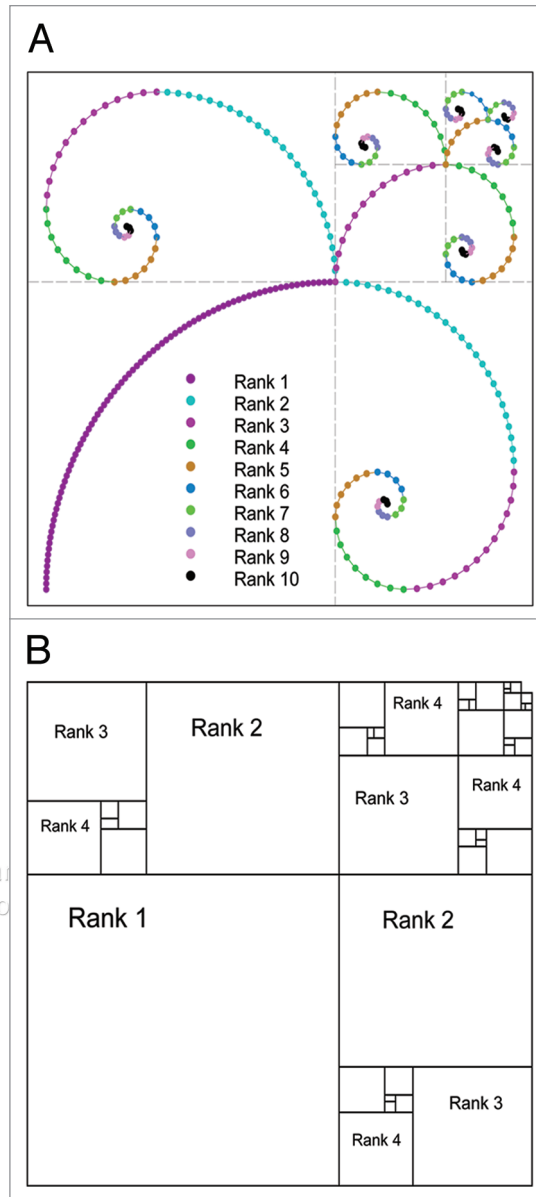


Figure 5. Visual analytics illustrate the fractal nature of T-cell repertoire responding to influenza in a form of a spiral (A) and as a fractal Mondrian set (B). The color-coded spiral depicts clonotypes starting with singletons as the first branch and progressing up to 10 ranks. A Mondrian set mimics the spiral representation (for the ease of resolution depicts only the first 8 ranks). I named this fractal structure a Mondrian set after painting *Composition II in Red, Blue and Yellow*, 1930 by Piet Mondrian, a Dutch painter [Pieter Cornelis “Piet” Mondrian, after 1912 Mondrian (1872–1944)].

Challenges, Traps, and Pitfalls of Visual Displays

The charms of colorful display can captivate an attentive audience, as well as lead to grossly catastrophic misunderstanding of crucial facts. As precautionary principles for building reliable visuals, we offer the following comments.

Comment 1: Soundness of hypothesis and definitions. At any stage of analytical

assessment it is assumed that a scientist—in the appropriate fields of applications—properly defines the things, events and processes. Thus, in designing a graphical display, a scientist deals with the definitions as final products; rather than testing the definitions themselves, she/he is testing the hypothesis that contains the definitions. As we alluded in our early work,² research is only as sound as the definitions of terms used in stating the hypothesis.

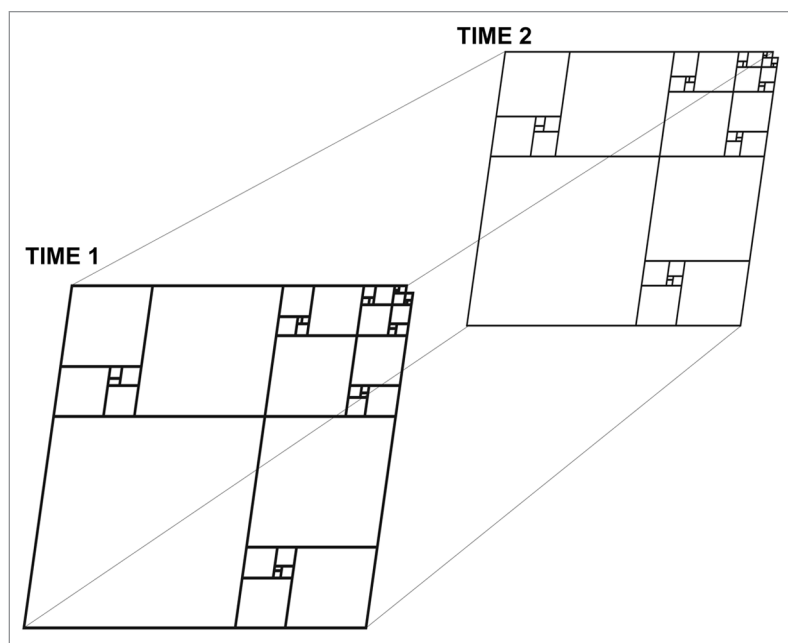


Figure 6. Predicted Rank Frequency relationship obtained by fitting the linear regression model to log-transformed data for two time points and portrayed as a temporal transformation of a Mondrian set.

On occasion, a researcher may deduce, from a variety of sources, that the terms of a hypothesis are not properly defined and should make inquiries in order to continue the research as this initial flaw jeopardizes the entire scientific endeavor.

A reliable visual display can only be made if the process of abstraction from empirical observations was reached based on proper definitions. Further still, graphical presentations based on research using incorrect or haphazard definitions can and will confuse or mislead the viewer, thus, violating a basic ethical premise. It is of utmost importance to consider how a research presentation may lead to wrong impressions and at the extreme, cause harm.

Comment 2: Validity and inferential reasoning. The accuracy of research results depends upon proper definitions of terms—those that contain the “whatness”—and upon the selection of appropriate methods of obtaining data, applying appropriate analytical methods to the data, and using appropriate graphical tools to demonstrate the results. It is assumed that the researcher applies appropriate methods to obtain data, so that, again, the analyst is dealing with data as final products. Needless to say, reliable

statistical analysis and relevant graphs can be made only if the data are correct. It is well known that data visualization, as exploratory data analysis, can be exceptionally helpful in identifying implausible or incorrect data.

On other hand, knowing the “what, how and why” about the collected data as subjects for analysis and visualization, is critical in and for constructing efficient and reliable graphs. Visuals can be created by using a wide variety of graphical tools such as graphs, charts, plots, maps and diagrams. When selecting data visualization techniques, consider the following: the appropriateness, the accuracy and the visual perception of the graph.

Comment 3: Completeness and the role of language. The role of language in a graphic display is dual: a graphical display can have some properties of a language itself and at the same time it serves as shorthand for a language. A graphical presentation and its verbal explanation should be co-extensive; therefore, a presentation standing alone should conform to the explanation given by a researcher or analyst in words.

If and when a researcher needs to qualify a visual display, or must explain to the audience in words what the display

does not say, or if the graphical presentation says more than the research, then the graphical display may be misleading. A proper graph, if and when translated into a language, should clearly reflect the data and research results without losing details or overemphasizing a point. In fact, translating a display into a language can serve as a standard for determining whether a graphical presentation is overstating results, overemphasizing a point, and/or suggesting more than the data supports. To avoid the miscommunication, a set of rules, operations and terms should be formulated and then applied when designing scientific graphical displays, including convention, in illustrating commonly used terms and relations as in ontology applications.

Comment 4: Visual perception and plausibility. The selection of symbols, color schemes and embedded comments is an important part of the data visualization process. Symbols play a special, dual role in a graphical presentation. A graph can be viewed as a symbolic representation of a thing, event or process, and at the same time, symbols are essential attributes of a graph. Overall, the function of a symbol is to supply a relevant and a reliable image; as Lonergan pointed out, “...there is no doubt that, though symbols are chosen by convention, still some choices are highly fruitful while others are not.”¹³ In the process of selecting appropriate symbolic reflection for a graph one must consider the symbols that offer clues, hints and suggestions. It should be acknowledged that selected symbols should not be taken literally. As a fact, we do not claim that the responding clonotypes are folding into a spiral; rather the use of a spiral is the most compact form of data representation.

The use of symbols might become customary and dictate inference. An unexplained change of pre-selected symbolic patterns in a graphical presentation may lead to confusion and give a wrong impression. If, in the process of graph construction, the impact of selected symbols—color, scale, size and style—is not taken into account, then a graphic display may lose its efficiency and mislead the viewer.

Comment 5: Ethical challenges. At times, an analyst may face ethical

challenges of scientific integrity. If an analyst believes that the terms are not properly defined, or the data quality is under suspect, then she/he must make a decision to either proceed with the analysis, or—often a more difficult path—set limits on what can, should, and will be done with the data. An extending dimension of a somewhat different kind of ethical challenge is in the assurance of the graphical presentation as reflective of the results and does not weave a false tale.

Provided within scientific disciplines are guidelines for the ethical conduct of research. As an example, the statistical practice code addresses many specific types of obligations, which appear in the “Ethical Guidelines for Statistical Practice” established by the American Statistical Association.¹⁴

Summary

Enhancing our appreciation of complex phenomena and probing deeper into existing theories to further stimulate discoveries is the goal of visual analytics. An increase in the popularity of graphical applications for data visualization has stimulated an enormous production of graphs, charts, plots and diagrams in research literature. Impressive attempts have been made to develop useful guidelines for proper construction and interpretation of sound visual displays;³ however, there still remains a lack of specific recommendations for the systematic use of complex visuals in life-science research.

Although we illustrated our concepts using actual experimental conditions, experimental design and datasets representing T-cell response to influenza, our approach can be adapted to a multitude of research questions and derived experimental data. This approach can be applied not only to studying TCR but also B-cell repertoires and many other complex systems that follow the rules of self-similarity or self-organization. The presented process of data compression is universal for the analysis of diversity and complexity of cell populations in various fields of immunology.

The result of our analytical and philosophical analysis concludes by necessitating further experimentation with visual analytics as devised steps for data compression and effective communication of knowledge. Our simple rules are intuitive, not operational and are designed to provide sensible insights into how a well-constructed graphical display can be created: (1) a graphical display should contain a well-understood statistical context or logical path, for which one is able to give a verbal description; (2) a graphical display should help to explain data or concepts by taking advantage of visual perception; and (3) a graphical display should force the viewer to notice the unexpected, to motivate challenging questions and to clarify statements, results and/or concepts. As any new technological advancement the use of visual analytics should be approached wisely, not to overwhelm and confuse by a mesmerizing explosion of colorful images, but to highlight our understanding and bring us closer to truth.

Acknowledgments

I wish to thank the funding source, the National Institute of Allergy and Infectious Diseases (N01 AI-50032: HHSN266-200500032); Drs. Yuri Naumov and Jack Gorski for providing original data, Dr. Beth Rosenberg for editorial wisdom and Dr. Nina Fefferman for hints into naming the invented fractal structure as a Mondrian set.

I am indebted to Dr. Eileen O’Neil (1949–2010), an outstanding philosopher and wonderful friend, who passed away before this manuscript was completed. Eileen inspired me to bring together mathematics, philosophy, public health and immunology—a constellation of disciplines that rarely work together but have the highest potential for meaningful discoveries. Eileen’s brilliance and vision were impeccable and can not be overestimated—she left us too early, with so many ideas to be challenged, so many paradoxes to be dissected and so many patterns to be revealed.

References

1. Thomas JJ, Cook KA. Illuminating the path: The research and development agenda for visual analytics. IEEE CS Press 2005.
2. Naumova EN, O’Neill E. Graph, word and whatness: musings on the philosophy of curves. Proceedings of the Joint Statistical Meetings Section: Statistical Graphics 2001.
3. Wilkinson L, Wills G. The grammar of graphics. New York: Springer 2005.
4. Naumov YN, Hogan KT, Naumova EN, Pagel JT, Gorski J. A class I MHC-restricted recall response to a viral peptide is highly polyclonal despite stringent CDR3 selection: implications for establishing memory T cell repertoires in “real-world” conditions. *J Immunol* 1998; 160:2842-52.
5. Naumova EN, Gorski J, Naumov YN. Two compensatory pathways maintain long-term stability and diversity in CD8 T cell memory repertoires. *J Immunol* 2009; 183:2851-8.
6. Naumov YN, Naumova EN, Hogan KT, Selin LK, Gorski J. A fractal clonotype distribution in the CD8⁺ memory T cell repertoire could optimize potential for immune responses. *J Immunol* 2003; 170:3994-4001.
7. Naumova EN, Gorski J, Naumov YN. Simulation studies for a multistage dynamic process of immune memory response to influenza: experiment in silico. *Ann Zool Fennici* 2008; 45:369-84.
8. Yassai MB, Naumov YN, Naumova EN, Gorski J. A clonotype nomenclature for T cell receptors. *Immunogenetics* 2009; 61:493-502.
9. Hill BM. The rank-frequency form of Zipf’s law. *JASA* 1974; 69:1017-26.
10. Good IJ. The population frequencies of species and the estimation of population parameters. *Biometrika* 1953; 40:237-64.
11. Naumov YN, Naumova EN, Clute SC, Warkin LB, Kota K, Gorski J, et al. Complex T cell memory repertoires participate in recall responses at extremes of antigenic load. *J Immunol* 2006; 177:2006-14.
12. Bax M. Complete Mondrian. Aldershot, Hampshire U.K; Burlington, VT: Lund Humphries 2001.
13. Lonergan BJF. *Insight; a study of human understanding*. New York: Philosophical Library 1957.
14. American Statistical Association. Ethical guidelines for statistical practice. <http://www.amstat.org>.