

GRAPHS FOR REPRESENTING STATISTICS INDEXED BY NUCLEOTIDE OR AMINO ACID SEQUENCES

Daniel B. Carr and Myong-Hee Sung

Key words: Letter sequences, graphics, layouts.

COMPSTAT 2004 section: Graphics.

Abstract: This paper develops coordinates and layouts for graphs that represent statistics indexed by repetitive letter sequences. The need for such graphics arises in a variety of applications. The examples in this paper concern sequences of nucleotides, such as AGTGGC, and sequences of amino acids.

1 Introduction

In contrast to maps that represent statistics indexed by geospatial coordinates, the development of graphics methodology for statistics indexed by repetitive letter sequences has been modest. One interesting exception is the sequence logo display [13] that can show a sequence of categorical frequencies. Statistical graphics methods for categorical data [7], [8] are relevant for relatively simple multivariate combinations but so far have seen little use in nucleotide and amino acid indexing examples.

Journal articles typically show short tables with one column giving the sequence of letters and one more column providing statistics. The rows are often sorted by one of the statistical columns. Both the one-dimensional linear ordering and the restriction to a modest number of rows reduce the opportunity to see patterns that may lead to new understanding. One-dimensional linear orderings produced by clustering, the first principal component, minimal spanning tree traversal, space filling curves or other methods do not exploit the human ability to see multivariate patterns based on 2-D and 3-D connectedness and proximity. Connectedness and proximity are among the most powerful of human perceptual grouping principles [18]. Thus this paper seeks to develop 2-D and 3-D coordinates for representing letter-indexed statistics.

The graphical design objectives include providing an overview along with interactive focusing and re-expression methods. For long sequences, combinatorics grow exponentially with sequence length and quickly lead to an overwhelming number of statistics. Overviews require substantial statistical summarization. The modest research here concerns developing representations for short sequences.

One approach not investigated here is the use of pixel oriented visualization [9]. It is possible to encode univariate statistics on all nucleotide se-

quences of length ten ($4^{10} = 1024 \times 1024$) in a pixel plot on a 1280×1024 monitor. Large high resolution prints will handle somewhat longer length sequences. Interactive pan and zoom methods can support layouts for longer sequences but showing all the values at once is problematic. The color of individual pixels is hard to identify with increased monitor resolution. The use of multiple monitors cannot keep up with the exponentially growing combinations.

Layout details are an issue. A convenient layout for a pixel plot may use lexicographic order for the first half (last half) of the sequence along the x -axis (y -axis). A following sectional on fractal coordinates provide another approach to layouts. In both cases indexing regularity help to keep the analyst oriented with interpreting the plots. However, indexing that is convenient for human memory may be poor at bringing out meaningful patterns. Maps often work well for showing geospatially-indexed statistics because geospatial attributes often have locally similar values. This applies to covariates as well as to the primary variables of interests. Proximity that reflects scientific relationships can be crucial to seeing meaningful patterns.

The layouts in this paper have limitations because they are primarily based on indexing regularity. However, the layouts provide some opportunities to rearrange letter order or axis placement either for perceptual simplification (such as reducing line crossings) or for incorporating physical/chemical properties (such as hydrophobicity) of the sequence constituents. Interestingly these two objectives can lead to the same display. Axes ordering problems are in general NP complete [1]. While many people prefer 2-D layouts, 3-D layouts not only allow better preservation of interpoint distances of higher dimensional points, they also provide more opportunities arranging axes. Thus the layout options are not as restrictive as might be assumed at first glance.

This paper develops three approaches to constructing coordinates while mentioning some alternatives along the way. Three different data sets motivate the development of the coordinates. Section 2 describes self-similar coordinates at different scales. Section 3 concerns self-similar coordinates at the same scale with focus on 3-D extension of parallel coordinates. The application shows cell statistics from a 4-D table. Section 4 illustrates the use of simple additive vector coordinates for showing all quadruples of amino acids (ignoring order). This approach can be useful despite some substantial overplotting problems. The section also hints at other 2-D layouts that avoid overplotting. Comments appear along the way about software and interactive tools used for rendering.

2 Self-similar coordinates, dimensionality, and fractals

The regularity of self-similar coordinates speeds learning and helps analysts devote short-term memory to other issues. A natural approach to developing multivariate coordinates encodes letters as integers and then uses Cartesian

coordinate product sets with a coordinate for each position in the sequence. With A=1, C=2, G=3, and T=4 the sequence ATCG is located at (1, 4, 2, 3). The similar treatment of each position in the sequence and the same ordering of nucleotides for each axis motivates the description as a self-similar coordinate system. With coordinates in hand, multivariate glyphs can encode multivariate statistics associated with the sequence. The most immediate problem with this approach is that straightforward graphical representation of points is only available through three dimensions.

2.1 Rendering approaches and difficulties

There are many approaches to rendering multivariate data with more than three coordinates. As further mentioned in Section 4, nested coordinate plots provide one approach. Another approach encodes some coordinates as glyph features. For example with four coordinates the ray angle of a stereo-ray glyph can encode the value for the fourth coordinate [3]. Ray length and color can encode more coordinates. Good perceptual accuracy of extraction for the angle encoding and modest use of “ink” make glyph a good choice for revealing hyperplanes in 4-D data and other tasks. However in the current context all 3-D coordinate glyphs lose self similarity when rendering more than three coordinates.

Before developing 3-D coordinates to represent sequences longer than three, brief comments about limits and merits of 3-D graphics are appropriate. Many people prefer 2-D graphics to 3-D graphics. Common arguments for 2-D rendering are that people only see surfaces, occlusion is a problem in 3-D and that motion and/or binocular parallax depth cues are impossible or inconvenient to convey on a printed page. The position here is that many humans are endowed with the cognitive ability to see 3-D images based on motion and binocular parallax. They should be allowed to utilize this capability whenever it helps in dealing with difficult scientific challenges. Three dimensions provide a richer environment for conveying relationships and produce less distortion than 2D and 1D plots when scaling multivariate data into lower dimensions.

2.2 Weighted vector addition and fractals

Vector addition provides an enticing starting point for developing 3-D coordinates. For nucleotides, associate each letter with a vector from the origin to the vertices of a tetrahedron. Let A=(1,1,1), C=(1,-1,-1), G=(-1,1,-1), and T=(-1,-1,1). Then using of vector addition for each letter in a sequence produces a 3-D coordinate for representing the sequence. However, vector addition is commutative so all permutations of the same set of letters yield the same point. When the goal is to represent all hexamers (six letter sequences) the result should be $4^6 = 4096$ distinct plotting points but rather 84 points corresponding with the multinomial terms in multinomial $(A + C + G + T)^6$.

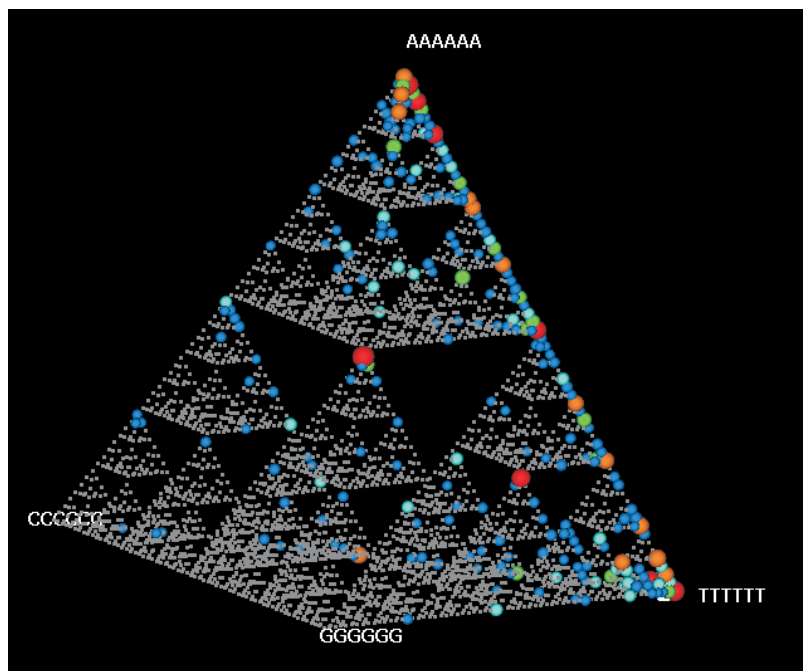


Figure 1: Fractal coordinates for nucleotide sequences. Sphere size (and color) show counts. Small rectangles show the plotting locations of low count spheres.

Weighted vector addition provides an approach that can produce unique points for plotting. Consider power of two weights $2^{(6-i)}/63$ where i is the position along the sequence and the weights sum to 1. The sequence ACGTTC then maps into the point $(.555, .270, .206)$.

Rendering and rotating reveals the coordinates creating a Sierpinski Gasket similar to one shown in Mandelbrot [11]. The self-similarity provides means of decoding the indexing of a point based on its location. A point in the large tetrahedron toward the C attractor has C as the first letter. If a point within this C tetrahedron is as close as possible to the T attractor, then all the remaining letters are T. A tetrahedron zooming widget can reveal the sequence of conditioning letters. Similarly, zooming can be controlled by entering the first letters of the sequence.

A delight occurs when rotating the gasket. In some orthographic projections the appearance is a square lattice. Construction using a pair of coordinates indicated earlier makes this clear. However, this is not intuitive from looking at the gasket in other views. The 2-D layout is also self-similar and extends immediately to the 10-mer by 10-mer pixel display mentioned earlier.

The motivation for Figure 1 was an early effort to find transcription regulation docking sites for the Stanford yeast genes [5]. The study clustered genes based on their expression levels. This produced groups of seemingly co-regulated genes. For the genes in a group, the 300 (nucleotide)-letter regions upstream of the protein-coding regions of the genes were scanned with a sliding window of length six. This produced the basic statistics on the occurrence frequencies of the different hexamers encountered. The sphere glyphs in the plot encode counts using size and color. A glance reveals that most of the higher count hexamers appear along the AAAAAA to TTTTTT edge. Relatively little was known about transcription regulation when the Stanford yeast data was first made available and the statistics for Figure 1 was produced. Today many transcription regulation sites of various lengths have been identified and the regions as far as 800 nucleotides upstream are relevant for some genes. The plots could be improved by obtaining better data and by highlighting the hexamers known to be associated with transcription regulation.

Different kinds of software can produce figures similar to Figure 1. With a little work most standard statistical software can produce projected static views. Software that provides rotation, filtering and brushing, such as Xgobi and CrystalVision, provide better visualization environments. Efforts to produce multilayer 3-D visualization methodology similar to GIS software led to the development of software called GLISTEN (geometric letter-indexed statistical table encoding). GLISTEN supports point and path layers that are used in the graphs below.

Efforts to extend the fractal layout to amino acids were not very successful. One generalization used the 20 face centers of the icosahedron as attractors and adapted the weights so the clouds of points associated with each of the 20 attractors would be separated. Only three letter sequences were shown to restrict the view to $20^3 = 8000$ points. The high density of points for the smallest scale icosahedra and the occlusion, partly due to more points, made this layout less desirable.

2.3 Connecting coordinates for longer sequences

Paths that connect points can represent longer letter sequences. For example, a path through three points in Figure 1 can represent nucleotide sequences 18 letters long. The use of paths is advantageous since occlusion problems do not grow too quickly. Experimenting with translucent triangles and tetrahedra for showing triples and quadruples were not successful due to inability to see through much more than two layers.

Paths can encode statistics using path thickness and color. The path direction also needs to be encoded unless the sequence is explicit or intended to be reversible. There are limitations to this approach. If the data contains a large number of sequences, overplotting precludes providing an overview. Filtering widgets can then help to cope with very large databases. A second

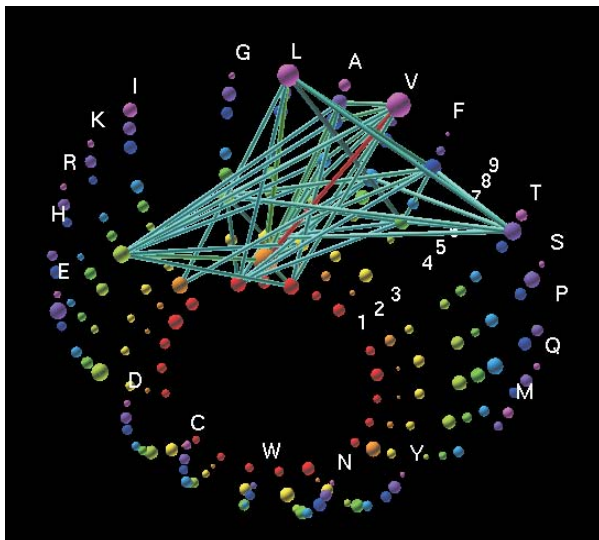


Figure 2: Capless hemisphere coordinates. Sphere size shows counts from 1-D table. Path thickness, color, and filtering enable focus on high counts from 2-D tables.

limitation is that a single coordinate system with each point representing a sequence of p letters does not accommodate sequence lengths that are not a multiple of p . A third issue that especially applies to fractal coordinates is that the apparent distance between paths is heavily influenced by the subset of coordinates receiving heavy weight. Fourth, there can be ambiguity when multiple paths go through the same point. Still, such displays can often turn up meaningful patterns that are otherwise missed.

3 Parallel coordinates escape the plane

Parallel coordinates (PC) plots also provide self-similar representations. Analysts are increasingly using these plots to show multivariate data and to provide interactive input in a multivariate context. A limitation of parallel coordinates is the lack of a natural way to connect non-adjacent axes. Figure 2 shows a capless hemisphere coordinate system that partially addresses the problem. The coordinate system encodes the 20 natural amino acids as longitude and nine positions along a sequence as latitude. The gap create perceptual groups that facilitate focusing on subsets. The path connecting L2 and L9 and the path between L2 and its neighbor L3 do not overlap due to the hemisphere curvature. The 3-D setting with curvature means the axes are not longer parallel, but there is a simple mapping to parallel coordinates.

The data motivating Figure 2 comes from a database of peptides [2], or

in this case amino acid sequences of length 9 known to bind to important immune molecules called HLA. This binding reaction is crucial in initiating the recognition by the human body of peptides from 'foreign' sources such as viruses or cancer. When a T-cell finds the peptide-HLA combination on the target cell surface, it activates coordinated processes for the purpose of clearing the infected cells. The immune system is an incredibly complex 'search-and-destroy' system. Autoimmune diseases are examples of false positives, where the immunity is mistakenly directed toward normal tissue. An example of an immune system false negative is the inability to detect and clear certain infections. Bioinformatic prediction of peptides from pathogenic proteomes such as HIV has emerged as a valuable tool in vaccine development and cancer immunotherapy [16].

The data used in the example concerns peptides binding to the HLA A-2 molecule (a specific form among many different genetic versions of this HLA). Most of the binding peptides listed were 9-mers. Typical statistics would be just the counts of amino acid for each of the nine positions. Such can be represented by sequence logo displays or as sequence of bar charts. The sphere size and color (when not shown in gray level) in Figure 2 convey this information just as effectively. The paths in Figure 2 encode counts from the nine choose two (36) two-way (20×20) tables. A filtering widget removed all but the highest count cells. The counts are encoded by the line color (when not shown in gray level) based on a color ramp. The layout also requires some sorting considerations. Putting the hydrophobic amino acids adjacent to each other reduces line crossing.

Since paths can have more than one line segment, the capless hemisphere coordinate framework can also represent statistics from higher dimensional tables. The three-segment paths in Figure 3 show high count (frequency) cells from the nine choose four (126) four-way ($20 \times 20 \times 20 \times 20$) tables (see also [10]). The lowest count path shown goes through L2, A7, A8, and V9 as might be expected from the 1-D margin counts. There are large counts for L2 and V9. The rest of the high frequency paths shown go through G4. This is not apparent from the 1-D margins counts.

4 Additive vector coordinates and overplotting

When the ordering of letters in a sequence is not important, the additive vector coordinate approach mentioned in Section 2 is more appropriate. A 3-D tessellation application provides such data [15], [17]. The data arise from tessellating the space of proteins based on the location of their backbone Carbon alpha atoms. The Delauney tessellation yields tetrahedra indexed by the associated amino acid residues at the vertices. The ordering of the amino acids for a tetrahedron is not considered important. There are $(20 - 1 + 4)$ choose 4 or 8855 distinct tetrahedra (see [6] for a discussion of classical occupancy problems.)

Figure 4 provides an additive vector coordinate example with the vec-

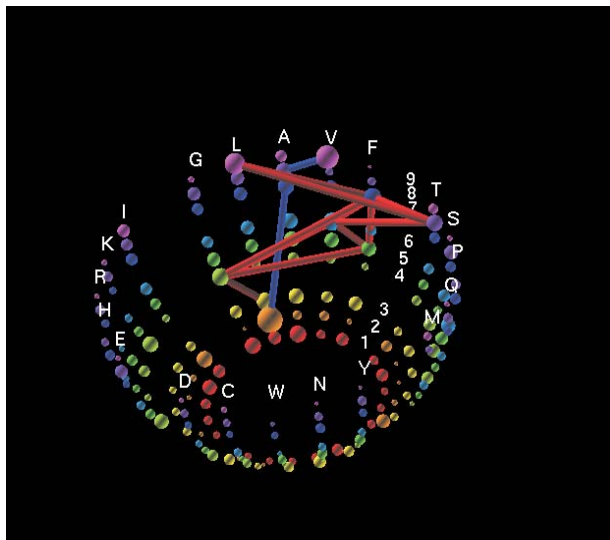


Figure 3: Paths with three segments show frequently occurring quadruples, i.e. high counts from 4-way tables.

tors point to twenty points evenly spaced around a circle. Again point size and color encode the counts, and dynamic filtering has removed low count tetrahedra. High count patterns jump out. One is a circle involving three Cysteins and one each of the amino acids.

While Figure 4 reveals a lot of structure, there are at least three problems worth noting. First, over some 2000 points are overplotted. This is partly related to symmetric construction with equal angles between vectors. Second, zooming reveals many points that too close together to see closer in an overview. Third, for over 4000 points involving four distinct amino acids the connection between plotting location and the indexing is almost impossible to untangle without mouseovers. Figure 4 is mostly useful for points in an outer annulus of the circle.

There are several possibilities for alternative views. It is possible to show a statistic encoded by color in a casement display of $20^4 = 160000$ points [12]. In this example the casement display is a 20×20 layout of 20×20 matrices. However it remains desirable to study plots with a factor of 18 less points. The 8855 points can be placed in a 4-D simplex. (See also pentagonal numbers [4].) Space prohibits showing a layout composed of two-dimensional slices of the simplex. There is also a layout in the plane for all tetrahedra with 2 or more of the same amino acid. While this layout involves duplicates the regularity makes the layout easier to study. Such a layout can provide a starting point for drilling down to conditioned views of the 1-1-1-1 combinations.

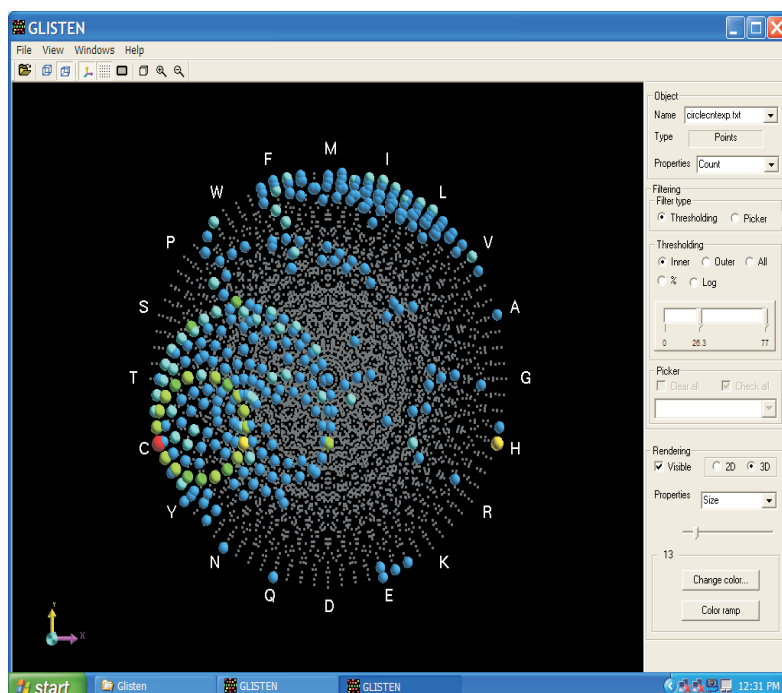


Figure 4: Vector addition coordinates. Sphere size (and color) encode statistics for protein tetrahedra. Small rectangles show plotting locations of low count spheres.

5 Closing remarks

Just as map projections have been devised to serve different purposes, coordinates systems for encoding statistics can be developed to serve different purposes. A worthy goal is to develop coordinates systems with a regularity that minimizes memory burdens and helps analysts keep oriented with respect to the coordinates. A tension arises when one desires to show complex relationships faithfully in some abstract sense while keeping the relationships cognitively accessible. In many cases there are no easy answers and the graphics are a compromise. Still, analysts can make discoveries from imperfect graphs. It is worthwhile to consider graphics that lean toward the cognitive accessibility and work toward incorporating as much scientific structure as possible options. Accessible graphics enable analysts to look and, if they look, they have a chance to see.

References

- [1] Ankerst M., Berchtold S., Keim D.A. (1998). *Similarity clustering of dimensions for an enhanced visualization of multidimensional data*. Proceedings IEEE Symposium on Information Visualization, IEEE Computer Society, Washington, 51–60.
- [2] Brusic V., Rudy G., Harrison L.C. (1998). *MHCPEP, a database of MHC-binding peptides: update 1997*. Nucleic Acids Research **26** (1), 368–371.
- [3] Carr D.B., Nicholson W.L. (1988). *EXPLOR4: a program for exploring four-dimensional data*. Dynamic Graphics for Statistics, W.S. Cleveland and M.E. McGill (eds.), Wadsworth, Belmont, California, 309–329.
- [4] Conway J.H., Guy R.K. (1996). *The book of numbers*. Copernicus Books, Inc. New York.
- [5] DeRisi J.L., Iyer V.R., Brown P.O. (1997). *Exploring the metabolic and genetic control of gene expression on a genomic scale*. Science **278**, 680–686.
- [6] Feller W. (1968). *An introduction to probability theory and its applications*. Third Edition. John Wiley and Sons. New York.
- [7] Friendly M. (1999). *Extending mosaic displays: marginal, conditional, and partial views of categorical data*. Journal of Computational and Graphical Statistics **8** (3), 373–395.
- [8] Hoffman H. (2000). *Exploring categorical data: interactive mosaic plots*. Metrika **51**, 11–26.
- [9] Keim D.A. (1996). *Pixel-oriented visualization techniques for exploring very large databases*. Journal of Computational and Graphical Statistics, 58–77.
- [10] Lee J.P., Carr D., Grinstein G., Kinney J., Saffer J. (2002). *The next frontier for bio- and cheminformatics visualization*. T-M Rhine, Ed. IEEE Computer Graphics and Applications, 6–11.
- [11] Mandelbrot B.B. (1983). *The fractal geometry of nature*. W.H. Freeman and Company.
- [12] Munson P.J., Singh R.K. (1997). *Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignments*. Protein Science **6**, 198–201.
- [13] Schneider T.D., Stephens R.M. (1990). *Sequence logos: a new way to display consensus sequences*. Nucleic Acids Research **18**, 6097–6100.
- [14] Segal M., Cummings R., Hubbard A. (2001). *Relating amino acid sequences to phenotype: analysis of peptide binding data*. Biometrics V57, 632-643
- [15] Singh R.K., Tropsha A., Vaisman I.I. (1996). *Delaunay tessellation of proteins: four body nearest neighbor propensities of amino acid residues*. J. Computational. Biology. **3** (2), 213–221.

- [16] Sung M.-H., Simon R. (2004). *Genome-wide conserved epitope profiles of HIV-1 predicted by biophysical properties of MHC binding peptides* J. Computational Biology **11** (1), 125–145.
- [17] Vaisman I.I., Tropsha A., Zheng W. (1998). *Compositional preferences in quadruplets of nearest neighbor residues in protein structures: statistical geometry analysis*. Proceedings of the IEEE Symposia on Intelligence and Systems, 163–168.
- [18] Ware C. (2000). *Information visualization, perception for design*. Morgan Kaufman Publishers, New York.

Acknowledgement: The work was supported by NSF cooperative grant # 9983461. Duoduo Liao and Yanling Liu implemented versions of GLIS-TEN.

Address: D.B. Carr, George Mason University, Dept. AES MS 4A7, George Mason University, Fairfax, VA 22030, USA

M.-H. Sung, Biometric Research Branch, National Cancer Institute Room 8146 6130 Executive Plaza Rockville, MD 20852

E-mail: dcarr@gmu.edu, sungm@mail.nih.gov