on a high speed network is substantial.

Solve the problem of remote access to your data and your place of work. The goal is to have full access to your electronic resources (data, programs, documents) from wherever you are. Certainly you want to be able to access these resources from home. You will also want full access when you are on the road.

Getting on the Internet is essential. Email is cheap and easy and saves tremendous time and effort compared to the phone. Consider email lists for notifying groups of people and for sending memos. Create automated web archives of material sent by email so people can review and catch up.

People

Involve your system administrators in your productivity efforts. It is very tempting for system administrators to confine their operations to "the system" and leave the application level work to the users. Explain that the users are part of the system and it all exists to get work done. Of course there are system tasks that must be done for security reasons, maintenance reasons and for reasons of system stability. But good system administrators should be very concerned that the system is useful to the users and willing and able to help with productivity issues.

Reduce training costs by standardizing on software that is easy to use and has the features your department needs — especially features related to automation and integration with other tools.

Learn more about the tools you already have. Most large modern complex software systems have many features that go unused. Skim through the reference manual for a tool you use everyday and consider features that you do not yet use. Some features may simplify your use of the software.

Insist on a stable environment with a high-speed network and standard software tools that interoperate. Consider how these tools can best be used to reduce your effort in performing your tasks.

Michael Conlon
Department of Statistics
Box 100212 HSC
University of Florida
Gainesville, FL 32610–0212
Email: mconlon@stab.ufl.edu
Home page:
http://www.clas.ufl.edu/~mconlon

TOPICS IN SCIENTIFIC VISUALIZATION

Scanning a 4-D Domain for Local Minima: A Protein Folding Application

by Daniel B. Carr

With Contributions From Peter J. Munson and Geetha Vasudevan* (* Analytical Biostatistics Section, LSB, DCRT, National Institutes of Health)*

1. Introduction

Methods from statistical graphics apply to a wide range of applications. At Interface '95 I conjectured to Peter Munson that while protein folding (the collapse of a protein chain into a specific compact structure) occurs in three dimensions, insights might be obtained by considering constraints and using the methodology of higher-dimensional graphics. Peter immediately had a problem for me. He and Geetha Vasudevan had computed theoretical energies of short protein chain segments, described by a 7 x 7 x 7 x 7 lattice of dihedral angles. They knew the point of the minimum energy on the lattice. Peter asked if I could provide a visualization that would shed more insight into the energy surface. In Section 2 Peter and Geetha provide more details about the data. In Section 3 I indicate my design considerations in developing a first display of this data. Peter and Geetha discuss some implications of the display in Section 4. Finally Section 5 indicates some extensions. I have been thinking about extensions because Peter and Geetha have tougher problems at hand.

2. The Computed Protein Folding Data

This data set represents a theoretical potential energy as a function of the shape of a small segment of a protein molecule. The protein molecule is a long chain of residues which collapse into a very specific shape. Predicting this shape is known as the protein folding problem. We have modeled only four links of the protein's polypeptide chain which form a "reverse turn", basically a U-turn in the naturally occurring protein backbone. There are four main types of turns, designated I, I', II, and II' (Schulz 1979). The shape of such turns can be described by four torsion angles in the links of the chain. In an attempt to describe the complex energy landscape surrounding each of these turn types, we used a well-established potential energy function (CHARMm - see

Brooks, 1983). We expect naturally occurring proteins to take conformations which are close to their energy-minimal shape. Type II' turns have canonical values for the torsion angles as follows: phi1=60, psi1=-120, phi2=-80 and psi2=0 degrees. We calculated all the energies in a 4-dimensional, 60 degree window around these canonical values, and hoped to find a well-defined energy-well containing the unique energy-minimal conformation somewhere within. We also expected to see some very high energy conformations corresponding to "impossible" twists of the protein backbone.

The visualization of such energy landscapes is commonly done in two dimensions, phi and psi, (known as the Ramachandran plot). Here the challenge is to visualize the energy surface when there are 4 dimensions.

3. Graphical Design Considerations in Representing the Energy Data

My problem (Dan) was which visualization approach to select. The multivariate arsenal of statistical graphics tools continues to grow. For example cone plots (Dawkins 1995) are a recent addition. Given my historical bias toward ray glyphs in relative low dimensions, I still had the problem of which method to apply first: a 5-D display (e.g. stereo + ray angle + length as in Carr et al., 1986), a one-factor conditioned plot sequence of 4-D displays (stereo + ray), or a two-factor conditioned plot sequence of 3-D plots (rays or stereo). My first choice was the two-factor conditioned plot as illustrated in Figure 1.

Figure 1 conditions on the 7 x 7 levels of Phi1 and Psi1 to produce 49 small plots. In terms of recent history, the layout for Figure 1 (with margins added) dates back to Tukey and Tukey (1983) who called it an X3, X4 plot windowed by X1, X2. Cleveland (1993 and earlier) uses the word coplot to label the collection of conditioned plots. Tufte (1983) refers to the collection as small multiples. Here I use the word coplot and refer to individual conditioned plots as panels. Whatever the label, coplots have proved effective in breaking visual problems down into visually manageable pieces. A coplot seemed a good first choice.

> ***Here the challenge is to visualize the energy surface when there are 4 dimensions.***

Figure 1 differs from the Tukey and Tukey plots because individual plots (or panels) represent three variables: Phi2, Psi2, and energy . Given that the two angles are represented with x and y position (as shown in the legend) the question is how to represent energy? The numerous choices include ray angle, stereo depth, framed rectangles (Cleveland 1985) , circle area, colored contours, perspective views of fitted surfaces and colored dots. For a monochrome static view I chose ray angle to represent energy. Carr, Olsen and White (1992) discuss some merits of this choice.

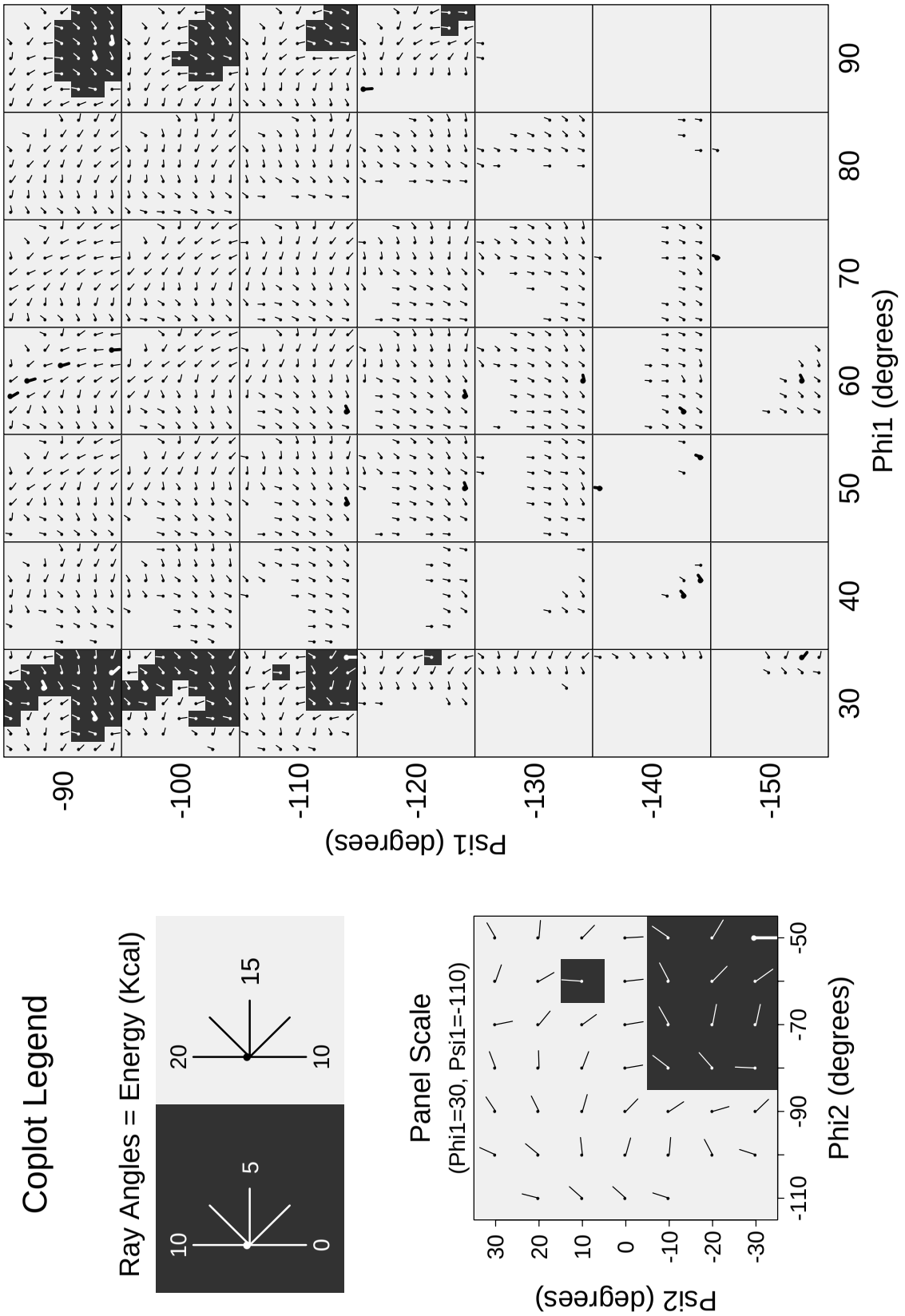
> ***Coplots have proved effective in breaking visual problems down into visually manageable pieces.***

Both ray-glyphs (dot at the base) and arrows (arrow head at the tip) can represent angle. Arrow heads have line terminators and introduce additional visual angles. I use rays because reducing the number of line terminators simplifies the plot appearance, and eliminating extra angles makes it easier to focus attention on the information encoded as angles. Without explanation or experience, ray-glyphs are ambiguous as to direction. In my informal checks, some people have preferred arrows because their direction is not ambiguous. Careful cognitive testing may be required to shed more insight into the relative merits of the two representations for angle. Those who want to see the arrows' variation can obtain the plot by `ftp` as indicated below.

> ***Careful cognitive testing may be required to shed more insight into the relative merits of the two representations for angle.***

The dependent variable is energy. The units, Kilocalories/mole, can be negative and are to be interpreted relative to the minimum value. A simple Boltzmann formula converts the difference from the minimum into a probability that the system will appear with specified dihedral angles given by (phi1, psi1, phi2, psi2). I subtracted the minimum value before producing Figure 1. This simplifies the scale labels in the legend. The translated energy values cover a large range, [0–1097], relative to the region of primary interest which is only a few units from zero. To provide resolution in the region of interest, Figure 1 masks values above 20 and uses a nested scale. As the legend shows, white rays on a dark gray background encode values in the interval [0–10] and black rays on a light gray background encode values in the interval [10–20]. For Figure 1, I chose 20 as the upper limit to show a substantial portion of the data. My first picture for Peter and Geetha narrowed attention to values below 10. Using gray backgrounds rather than black and white backgrounds reduces the contrast and makes the figure easier to study. The two dark-gray regions in the Figure 1 make it immediately obvious that there are at least two local minima. This was not expected as described in Section 4.

Figure 1. A Coplot of an Energy Surface.



Like coplots, nested scales allow readers to focus attention while limiting the mental burden through the use of identical structures for the scales. Typically color distinguishes the different scales. The monochrome Figure 1 shows two scales. This could be extended using shades of gray. Full color plots provide more options for distinguishing scales. The use of ordered colors (saturation and lightness) helps in mentally gluing the pieces together. The use of distinctive hues brings preattentive vision into play and promotes rapid evaluation within individual scales. A single plot cannot optimize both for overview and detail.

For a single plot, understanding the patterns across the nested scale transitions involves extra work. In Figure 1 white rays pointing up have energies similar to black rays pointing down. White rays pointing down have energies much less than black rays point up. Keeping track of this distinction is an extra mental burden. Rather than leave all the work to the reader in the search for local minima, it helps to algorithmically flag candidate values. Figure 1 shows the candidate local minima as enlarged ray-glyphs. Surprisingly there are 28 such local minima in the plot.

Like coplots, nested scales allow readers to focus attention while limiting the mental burden through the use of identical structures for the scales.

The local minimum values found depend on the definition of a local neighborhood. The square, cube, and hypercube lattices are awkward because "neighboring" points fall into two distinct classes, non-diagonals and diagonals. Non-diagonal neighbors are much closer than diagonals in the following sense. If one looks at the near-neighbor regions (hypercubes) about the lattice points, the hypercubes about non-diagonal points will share "cube faces" with the hypercube about the center point. The non-diagonal hypercubes barely touch the hypercube about the center point. Given a choice, I prefer to use the body-centered hypercube lattice in 4-D because the near neighbor regions (24-cells) for both non-diagonals and diagonals share "octahedron faces" with the 24-cell about the center point. A local minimum is established by comparison against the 24 neighboring points. For the current hypercube lattice, I choose to use non-diagonal neighbors to define the local neighborhood. Each point has 2 x 4 = 8 neighbors except for points on the edge of the domain that have fewer neighbors. A consequence of this definition is that diagonal troughs of local minima can appear. The four points in the (60,-90) panel are part of such a trough. The five points in the (50, 60) x (-130, -120, -110) set of panels form another diagonal trough. The two points in the top right panel are diagonally connected.

The coplot layout is important because it attempts to keep points in multivariate space close to each other in the plot. The layout represents a compromise since closeness in the plot is not equivalent to closeness in the 4-D space. The non-diagonal neighbors with a panel are much closer together than non-diagonal neighbors that have the same position in adjacent panels. At first glance, one might think that the point just to the left of the bottom right point in the (70, -100) panel is a local minimum. However, the values get lower as one goes up a panel and then left a panel. Reversing the roles of (phi1 psi1) and (phi2 psi2) would facilitate the study of energy as a function of phi1 or psi1 given the other values.

The coplot layout is important because it attempts to keep points in multivariate space close to each other in the plot.

Scanning the plot for more local minima is easier with interactive tools. Consider coloring the local regions dark gray or light gray depending on the whether or not the energy is above or below a particular cutoff value. When this cutoff value is under slider control, the user can increase the energy cutoff and immediately see when dark gray appears in visually disjoint regions. Slight slider oscillation will blink the new regions. For complicated surfaces many disjoint regions may appear and spotting new ones gets progressively harder. Switching to a new distinctive color can help. Interactive visualization can convey information about the energy well depth and shape.

Lattice plots like Figure 1 have some interpretational dangers. The dependent variable may change radically between lattice points. The plot does not reveal the exact location of the apparent local minima. Further, the plot provides no indication of a local minimum if an energy well is completely contained in a region between lattice points. Unless the lattice point separation is known to be smaller that the scale of all energy wells, the possibility of missing local minima remains.

4. Implications of Figure 1

Figure 1 reveals several important facts about our problem. First, the global energy minimum does not lie right at the center, but is just on the edge of the domain. In fact, it is on the boundary in three out of four dimensions, suggesting that the real minimum lies outside the window and more energy calculations are needed with a shifted domain. These energy calculations may require several hours on a workstation, so we

can't expect to modify the domain interactively. We also notice that there is a second energy well (phi1=90) which is completely disconnected from the first well (phi1=30). Such multiple minima are a common feature in molecular modeling. If the energy barriers separating multiple minima are high enough, molecules can be completely trapped in local minima even though a nearby, lower-energy states exist. Likewise, numerical energy-minimization algorithms can become trapped in these local minima which makes use of simple gradient descent or Newton-Gauss type minimizers unreliable. We did not expect to see significant, multiple energy minima in this very simplified system (full proteins have hundreds of times the complexity of our modeled fragment), and indeed, until we visualized our data set, we had no idea that they existed here. One should not forget that the mathematical model employed here (CHARMm potential energy) may not be a good description of the actual forces influencing real molecules in its natural, watery environment. More complex, and time-consuming, but more realistic calculations would include the effects of water on the intramolecular forces and energies.

5. Extensions

Peter and Geetha indicate that there are many extensions of the protein folding problem to consider. I have started to think about them. For example they may compute at higher resolution. If need be, I will answer with a three feet wide plot of up to 9 feet in length. While static views cannot provide progressive disclosure, good old human pan and zoom is not bad. When I first spoke to Peter I had naively anticipated that 5-D plots would do the trick. For starters he tells me about domain dimensions in steps of 2 from 2-D to 30-D. Figure 1 suggests that 4-D domains are quite manageable. At first consideration I think that 6-D domains are within the limits of divide and conquer comprehension. I suspect that understanding a function on a 6-D domain in any kind of overview sense will take a lot of mental energy even with the best of visualization methods. Right now thinking about 8-D domains is too hard.

The statistical graphics community has much to offer in developing higher-dimensional visualizations for protein folding applications and a host of other applications. The opportunity seems so great that I felt compelled to write about it even though I have only been working on this problem for two days. If there are existing methods in the protein folding field that I don't know about, there is still a good chance that our community can produce something better. Figure 1 represents my first effort. The variations cited in Section 2 and those that occur to readers still need to be evaluated. I anticipate that my notion of preferred graphics will evolve as I try different variations and as Peter and Geetha guide my efforts. I hope others take up the challenge. Those that want to try their hand at visualizing the results of protein folding computation experiments can contact Peter at munson@helix.nih.gov. Readers can obtain my data sets and Splus script files by anonymous `ftp` to `galaxy.gmu.edu`. The files will be under `/pub/submissions/protein`.

Acknowledgments

Graphics research related to this article was supported by EPA under cooperative agreement No. CR8280820-01-0. The article has not been subject to the review of the EPA and thus does not necessarily reflect the view of the agency and no official endorsement should be inferred.

References

Brooks, B. et al, 1983. *Journal of Computational Chemistry*, Vol 4, No. 187.

Carr, D. B., A. R. Olsen, and D. White. 1992. "Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data." *Cartography and Geographic Information Systems*, Vol. 19, No. 4, pp. 228-236, 271.

Carr, D. B., W. L. Nicholson, R. J. Littlefield, and D. L. Hall. 1986. "Interactive Color Display Methods for Multivariate Data." *Statistical Image Processing and Graphics*, eds. E. J. Wegman and D. J. DePriest,, New York NY: Marcel Decker, pp 215-250.

Cleveland, W. S. 1985. *The Elements of Graphing Data*, Monterey CA: Wadsworth Advanced Books and Software.

Cleveland, W. S. 1993. *Visualizing Data*, Summit NJ: Hobart Press.

Dawkins, B. P. 1995. "Investigating the Geometry of a P-Dimensional Data Sets," *Journal of the American Statistical Association*, Vol 90, No. 429, pp. 350-359.

Tufte, E. R. 1983. *The Visual Display of Quantitative Information.* Cheshire CN: Graphics Press.

Schulz, G. E. and Schirmer, R. H. 1979. *Principles of Protein Structure*, New York: Springer-Verlag.

Daniel B. Carr
George Mason University
`dcarr@voxel.galaxy.gmu.edu`