Journal of Molecular Graphics and Modelling 67 (2016) 44–53

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Visualizing ensembles in structural biology

Ryan L. Melvin, Freddie R. Salsbury Jr.*

Department of Physics, Wake Forest University, NC, United States

ARTICLE INFO

Article history:
Received 1 March 2016
Received in revised form 26 April 2016
Accepted 2 May 2016
Available online 4 May 2016

ABSTRACT

Displaying a single representative conformation of a biopolymer rather than an ensemble of states mistakenly conveys a static nature rather than the actual dynamic personality of biopolymers. However, there are few apparent options due to the fixed nature of print media. Here we suggest a standardized methodology for visually indicating the distribution width, standard deviation and uncertainty of ensembles of states with little loss of the visual simplicity of displaying a single representative conformation. Of particular note is that the visualization method employed clearly distinguishes between isotropic and anisotropic motion of polymer subunits. We also apply this method to ligand binding, suggesting a way to indicate the expected error in many high throughput docking programs when visualizing the structural spread of the output. We provide several examples in the context of nucleic acids and proteins with particular insights gained via this method. Such examples include investigating a therapeutic polymer of FdUMP (5-fluoro-2-deoxyuridine-5-O-monophosphate) – a topoisomerase-1 (Top1), apoptosis-inducing poison – and nucleotide-binding proteins responsible for ATP hydrolysis from *Bacillus subtilis*. We also discuss how these methods can be extended to any macromolecular data set with an underlying distribution, including experimental data such as NMR structures.

1. Introduction

Discussing the static images of proteins found in journals, Henzler-Wildman and Kern comment, "Physicists, however, will object to a static picture: they see proteins as soft materials that sample a large ensemble of conformations. . .." [1] Indeed, we do object. Since the crystallography work by Austin et al. in 1975 on myoglobin [2–4] there is widespread understanding that proteins exist in statistical ensembles of states rather than static conformers, with similar understandings emerging for biopolymers in general [1,4–7]. The specific objection presented here is that despite decades of such knowledge, one of the most basic practices when dealing with distributions of data has yet to become standard for visualizing macromolecules – error bars. Displaying single structures without some indication as to the variance in the distribution those structures represent at best leaves it to the reader to imagine an ensemble and at worst deceives a reader into believing that there is a single state in which the molecule exists [1]. Here we argue for the need to indicate underlying distributions of states when showing macromolecular structures and present statistically rigorous methodologies – involving expected error and standard deviation (see Section 4) – for doing so.

Admittedly, the leap from error bars around a point on a scatter plot or a bar on a histogram to visualizing the uncertainty in a conformation is not obviously straightforward. Therefore, we present several examples and the underlying methodologies. Our first example is of representative structures from Molecular Dynamics (MD) simulations, a useful and increasingly common tool in drug discovery [8,9], generating ever larger ensembles of states as available computational power increases [10,11]. Here we show structures from all-atom MD simulations of a therapeutic oligomer of FdUMP (5-fluoro-2-deoxyuridine-5-O-monophosphate) [12–15]. The large number of structures generated in MD trajectories often requires partitioning conformation space into a handful of macrostates. Typically a representative from each macrostate – say a mean or median structure – is then displayed. However, displaying only the representative with no indication of uncertainty obscures the width of the underlying distribution. We propose a method based on combining clustering – quality threshold clustering [16] for the purposes of this manuscript – combined with statistical analysis as described in Section 4.

As an example of this method's application to structure prediction and protein-protein interaction (PPI), we show predicted interactions between SufC and SufD, Fig. 4, from *Bacillus subtilis* [17–24] responsible for ATP hydrolysis and thought to be analogous to similar proteins in *Escherichia coli* [25]. In addition to the most likely complex, we use the expected error of the PPI calculations to choose additional structures to display – Fig. 4b–d. Our next two

* Corresponding author.
E-mail address: salsbufr@wfu.edu (F.R. Salsbury Jr.).

http://dx.doi.org/10.1016/j.jmgm.2016.05.001
1093-3263/

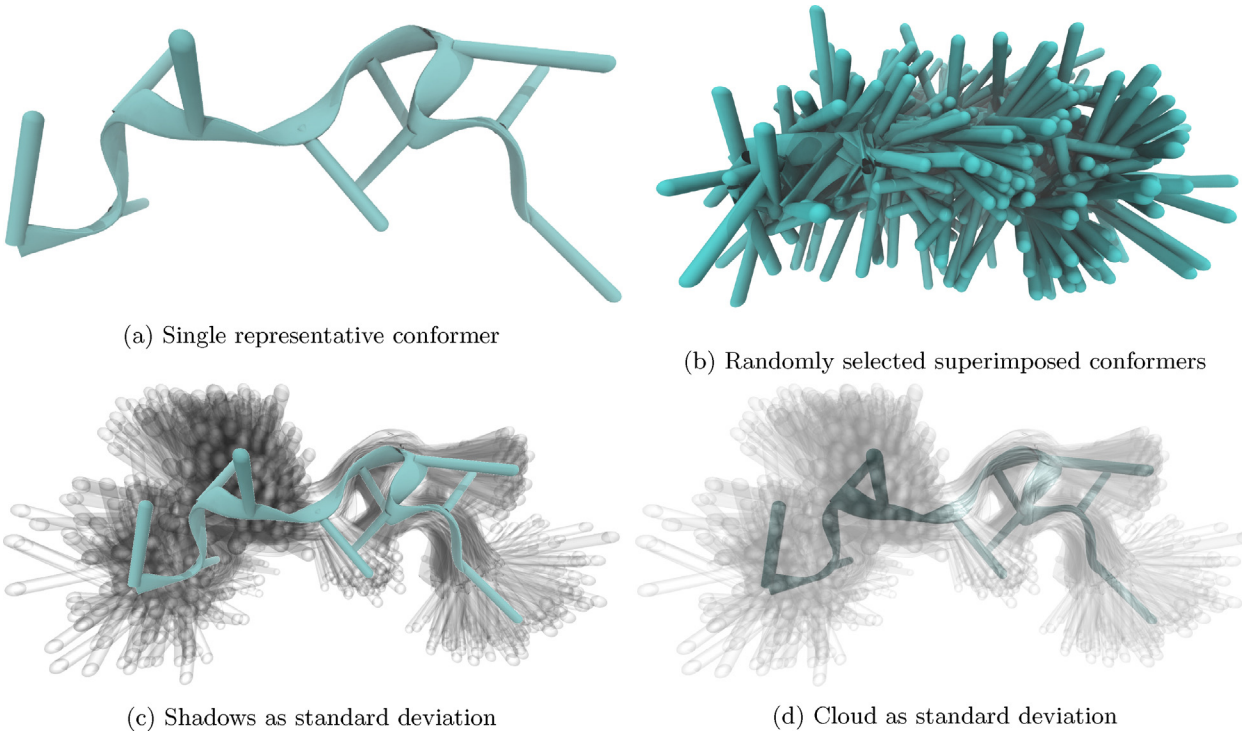

Fig. 1. Side-by-side comparison of conformations of F10 in 150 mM $CaCl_2$ with these four visualization techniques reveals (compare a and b) the deceptiveness of using only a single conformer, (c) the width of the distribution indicated with little loss of visual simplicity, and (d) the ability to see standard deviation in three dimensions.

examples, Figs. 5 and 6, extend to high-throughput, small ligand docking software that predicts *in vivo* interactions. Such docking software carries with it an expected error, which we translate to conformation space (see Section 4) and display along with the highest ranked docking pose – Figs. 5b–d and 6b–d. We also discuss how these methods can be extended to any macromolecular data set with an underlying distribution, including experimental data from NMR and X-ray crystallography studies, concluding with a final example of such an application, Fig. 7. To aide others in using these methods, we have made example Python [26] scripts publicly available online [27], supporting information Fig. 1.

One common way to show the distribution of states underlying a representative structure is to visualize a superposition of states [28] that have been aligned by minimizing a distance metric between the structures – a method that has been applied to MD, NMR and X-ray crystallography data sets, of which we cite a few examples [28–32]. Within these examples we see both sets of all solid structures superimposed, similar to panel b in all figures herein, where every structure is given equal visual weight. However, this visualization method is often used without a statistically rigorous method for selecting frames. To emphasize the difference between such selection methods and ours, in Figs. 1–3, we choose the additional frames in panel b randomly. In the remaining figures of the paper, we use the statistical measures described in Section 4.

For displaying additional conformers, we also see the use of transparency, in which a single representative structure is given the most visual weight with the additional conformations seen as a shadow, panel c in all figures, or diffuse cloud, panel d in all figures. For each example system, we include a side-by-side comparison of a traditional structural ensemble diagrams (panels a and b, except Fig. 5) and our suggested styles (panels c and d), drawn with the same representation and viewpoint so that readers may judge for themselves both the value in displaying a clearly defined metric of uncertainty and the usefulness of our preferred visualization styles. No matter the display method, we clearly indicate the statistical metric for choosing these conformations, as the primary point of this work is to show the value in indicating uncertainty in visualization of macromolecule conformers.

Another method is to adjust the width of a residue – or other substituents – based on the width of its distribution of positions [28,33]. These visualizations have the advantage of maintaining the visual simplicity of a single structure and provide information on each substituent's degree of contribution to overall structural uncertainty. There are also more abstract methods for visualizing these structural varieties, such as that used by Best and Hege to plot sets of interchanging structures or even entire trajectories on planar maps [34]. Such a method brings with it the analysis methods for planar graphs in general [35–38], allowing for a great deal of information to be extracted from the transitions among conformations [34,39,40]. While far less abstract than mapping to a planar graph, the methodology we suggest has a higher level of statistical rigor than simple overlays and is intuitively more straightforward than more abstract methods.

In addition to a representative structure, we choose which substituents to display based on a statistical measure relevant to the data. We display a median – or otherwise representative – structure as solid and each frame in some subset of the distribution as a shadow or cloud – e.g., those falling within one standard deviation of the median, panels c and d in Figs. 1–3, or those within some expected error of the representative, Figs. 4–6. In the shadow representation, overlapping shadows add their opacity to one another, distinguishing between isotropic and anisotropic motion based on the relative darkness of areas in the shadow. In the cloud representations, this distinction is diminished in favor of seeing the variance in three dimensions. Our methods differ from others primarily in the use of a statistical measure – other than random selection, even sampling or the entire distribution width – to choose the superimposed frames. To demonstrate the value of indicating the statistical measure used, we select additional frames using an uncertainty-based cutoff and state the measure used.