EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (2015)
W. Aigner, P. Rosenthal, and C. Scheidegger (Editors)

On the Reproducibility of our Biomolecular Visualization

Katrin Scharnowski[1] and Michael Krone[1] and Guido Reina[1] and Thomas Ertl[1]

[1]Visualization Research Center, University of Stuttgart, Germany

Abstract

*We reflect on the reproducibility of our work presented at EuroVis 2014 [SKR*14], which applies deformable models to compare molecular surfaces. We discuss both negative and positive aspects of our work in terms of reproducibility and put the aspects in a wider, more general context, in particular for the more critical points.*

Categories and Subject Descriptors (according to ACM CCS): I.4.7 [Image Processing and Computer Vision]: Feature Measurement—Size and Shape,

1. Introduction

Scientific research is an incremental process where current research is often based on previous work. Consequently, reproducibility is crucial to this process. By reproducibility, we mean not only the ability to reproduce the presented results, but also a clear description of the algorithm and the implementation that facilitates applying the presented concepts to different problems or other input data. Nonetheless, while the visualization community has matured over the last decades, providing robustness and transparency in terms of reproducibility is still often neglected. In general, reproducibility is often aggravated by limited budget (time and funding). However, the individual nature of the presented work can also be critical.

At EuroVis 2014, we presented a method for comparative visualization for arbitrary attributes of two input molecular surfaces [SKR*14]. We used a deformable model approach for non-rigid partial shape matching. A source surface is triangulated and deformed to match the target using a velocity field that is based on a diffusion process. Consequently, a vertex-wise correspondence between the source surface and the target surface is established, which allowed us to obtain local and global dissimilarity measurements for arbitrary surface attributes. In the following, we want to discuss our work with respect to reproducibility. We consider both negative and positive characteristics regarding different aspects like the employed data sets, the implementation, technical soundness, and expressiveness of the description. We conclude our discussion by transferring individual aspects of our work to more general aspects of the problem.

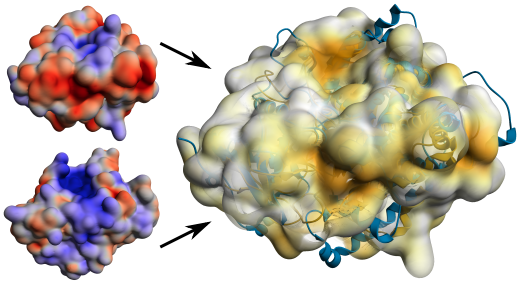

Figure 1: *Comparative visualization of molecular surfaces showing differences in the electrostatic surface potential (modified image from [SKR*14]).*

2. Theoretical Background & Soundness

A detailed explanation of the theoretical background is crucial, especially in a technical paper. Here, it is important to enable the reader to understand the motivation of the choices for the technical implementation, including a clear problem statement and, if possible, a mathematical formulation of the goal. Furthermore, a clear description of limitations of the method should be provided. Thus, readers can anticipate bad results for certain data and will not doubt the correctness of their implementation. These aspects improve the soundness of the paper and, consequently, support reproducibility.

Our paper contains the technical details and the equations necessary to understand our method as well as limitations we noticed during the evaluation of our method. Since our work is based on previously published methods, we also clearly state how our work differs from these. We also provided a lot

DOI: 10.2312/eurorv3.20151142

of illustrations that help to understand the individual steps of our algorithm.

3. Source Code & Implementation

In scientific visualization, research prototypes are often implemented using frameworks like *ParaView* [HA04] that provide basic functionality for data processing, filtering, or rendering. Without access to the respective framework, or the source code in general, it can be time-consuming to reimplement the presented work. This also requires a thorough description of the implementation.

We used *MegaMol* [GKM*15], which is an open source rapid prototyping framework for particle-based visualizations and also provides functionality for field-based data. Our technique is integrated in a plugin for MegaMol and is available online. Consequently, readers can download and compile the respective code. While the build process is well documented, the lack of information regarding the configuration and execution of MegaMol in order to use a specific functionality, like our method, is a major obstacle for potential users. Our code can be executed on a commodity desktop PC equipped with an Nvidia GPU (since our implementation uses CUDA). This enhances reproducibility; however, it is clear that this can not be achieved for all work (consider e.g. high performance visualization requiring a compute cluster).

The implementation of our method is highly parallelized. This includes an optimized CUDA implementation of a modified Marching Tetrahedra algorithm. Although the respective section contains an extensive description, there might still be details that are not entirely clear to a reader, such as the data structures. Ideally, implementation details complement the source code and further enhance the reproducibility. On the other hand, a more detailed description of the implementation requires additional page space, which was not feasible for us, since it would have involved sacrificing details in other parts of the paper. We, nevertheless, think that an experienced programmer who is familiar with CUDA and visualization algorithms should be able to reimplement our algorithm.

For a comparison with other methods, it is crucial that these methods are reproducible as well. We chose two global descriptors for molecular shape for our comparison, namely the RMSD [RR73], which is a widely known method, and comparison based on Zernike descriptors [SLL*08], which can be computed through a web service as mentioned in [LERV*09].

In conclusion, we think the availability of source code is instrumental for reproducing the results and allows other researchers to compare their own research with them. This is often required by reviewers but not necessarily meaningful if the authors had to reimplement previous work for performance comparisons.

4. Data Sets

One important aspect for reproducibility is the data sets used to produce the results. We applied our visualization method to publicly available data sets from the *RSCB Protein Data Bank (PDB)* [BWF*00] and provided the respective PDB IDs. The electrostatic potential can for example be computed using the *PMEPot*-plugin provided with *Visual Molecular Dynamics (VMD)* [HDS96, AS05] or the *Adaptive Poisson-Boltzmann Solver (APBS)* [BSJ*01]. Here, it is necessary to assign feasible partial charges to the atoms, which can be derived from force fields of open-source simulation packages like *GROMACS* [HKvdSL08] or *Amber* [SFCW13]. Alternatively, there is a web service that assigns partial charges to PDB data files [DCL*07]. While an experienced user can easily reproduce these data, our paper lacks a detailed description of how to obtain the electrostatic fields.

We also used data provided by co-authors from the field of biochemistry. Although these data are not publicly available, this is only a minor issue, since we also evaluated our method using publicly available data. In addition, we used simple synthetic data to show certain aspects of the algorithm, which can easily be recreated by the reader. We think that it is possible to produce results comparable to the ones presented in our paper.

5. Parameter Sensitivity

Another critical aspect for the reproducibility is sensitivity of the method to parameters. We included the values of all important parameters for the test data. However, even if the results can be reproduced for the exact data sets used in the paper, readers might also want to use the method with other data. Here, it can be difficult to find an appropriate set of parameters. We tried to remedy this issue by explaining the effect of all parameters.

6. Conclusion

One of the main problems in visualization concerning reproducibility is the lack of space for detailed explanation of the implementation. During the review process, reviewers need to be able to evaluate the presented work, which can to some extent be achieved by providing source code. However, for a blind review process, this is problematic since it might reveal the authors' identities (when using a frame work of limited currency like MegaMol). Alternatively, one could provide only parts of the code or implementation details. Another aspect, particularly with respect to molecular visualization, is that there are no commonly used data sets for testing and comparison, as it is for example the case in volume rendering. The respective data should be publicly available and be recognized by the community. In conclusion, reflecting on the reproducibility aspects of our own work reinforced our awareness of the relevance of reproducible research and will certainly influence our future publication practices.

References

[AS05] AKSIMENTIEV A., SCHULTEN K.: Imaging α-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map. *Biophysical Journal 88* (2005 Jun 2005), 3745–61. doi:10.1529/biophysj.104.058727. 2

[BSJ*01] BAKER N. A., SEPT D., JOSEPH S., HOLST M. J., MCCAMMON J. A.: Electrostatics of Nanosystems: Application to Microtubules and the Ribosome. *Proceedings of the National Academy of Sciences 98* (2001), 10037–10041. doi:10.1073/pnas.181342398. 2

[BWF*00] BERMAN H., WESTBROOK J., FENG Z., GILLILAND G., BHAT T., WEISSIG H., SHINDYALOV I., BOURNE P.: The Protein Data Bank , 2000. www.rcsb.org. 2

[DCL*07] DOLINSKY T. J., CZODROWSKI P., LI H., NIELSEN J. E., JENSEN J. H., KLEBE G., BAKER N. A.: PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research 35*, suppl 2 (July 2007), W522–W525. doi:10.1093/nar/gkm276. 2

[GKM*15] GROTTEL S., KRONE M., MÜLLER C., REINA G., ERTL T.: MegaMol – A Prototyping Framework for Particle-based Visualization. *IEEE Transactions on Visualization and Computer Graphics 21*, 2 (2015), 201–214. doi:10.1109/TVCG.2014.2350479. 2

[HA04] HENDERSON A., AHRENS J.: *The Paraview Guide : A Parallel Visualization Application*. Kitware, Inc., New York, 2004. 2

[HDS96] HUMPHREY W., DALKE A., SCHULTEN K.: VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics 14* (1996), 33–38. doi:10.1016/0263-7855(96)00018-5. 2

[HKvdSL08] HESS B., KUTZNER C., VAN DER SPOEL D., LINDAHL E.: GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation 4*, 3 (2008), 435–447. doi:10.1021/ct700301q. 2

[LERV*09] LA D., ESQUIVEL-RODÍRGUEZ J., VENKATRAMAN V., LI B., SAEL L., UENG S., AHRENDT S., KIHARA D.: 3D-SURFER: software for high-throughput protein surface comparison and analysis. *Bioinformatics 25*, 21 (2009), 2843–2844. doi:10.1093/bioinformatics/btp542. 2

[RR73] RAO S. T., ROSSMANN M. G.: Comparison of super-secondary structures in proteins. *Journal of Molecular Biology 76*, 2 (1973), 241–256. 2

[SFCW13] SALOMON-FERRER R., CASE D. A., WALKER R. C.: An overview of the amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science 3*, 2 (2013), 198–210. URL: http://ambermd.org/, doi:10.1002/wcms.1121. 2

[SKR*14] SCHARNOWSKI K., KRONE M., REINA G., KULSCHEWSKI T., PLEISS J., ERTL T.: Comparative Visualization of Molecular Surfaces Using Deformable Models. *Computer Graphics Forum 33*, 3 (2014), 191–200. doi:10.1111/cgf.12375. 1

[SLL*08] SAEL L., LA D., LI B., RUSTAMOV R., KIHARA D.: Rapid Comparison of Properties on Protein Surface. *Proteins Struct. Funct. Bioinforma. 73*, 1 (2008), 1–10. doi:10.1002/prot.22141. 2