Eurographics Conference on Visualization (EuroVis) 2017
B. Kozlíková, T. Schreck, and T. Wischgoll
(Guest Editors)

Short Paper

Molecular Visualization of Computational Biology Data: A Survey of Surveys

N. Alharbi[1], M. Alharbi[1], X. Martinez[2], M. Krone[3], A. Rose[4], M. Baaden[5], R.S. Laramee[1], M. Chavent[6]

[1] Department of Computer Science, Swansea University, UK
[2] Department of Biochemistry, University of Oxford, UK
[3] Visualization Research Center, University of Stuttgart, Germany
[4] University of California, San Diego, USA
[5] Laboratoire de Biochimie Théorique, UPR 9080 CNRS, France
[6] IPBS, Toulouse, France

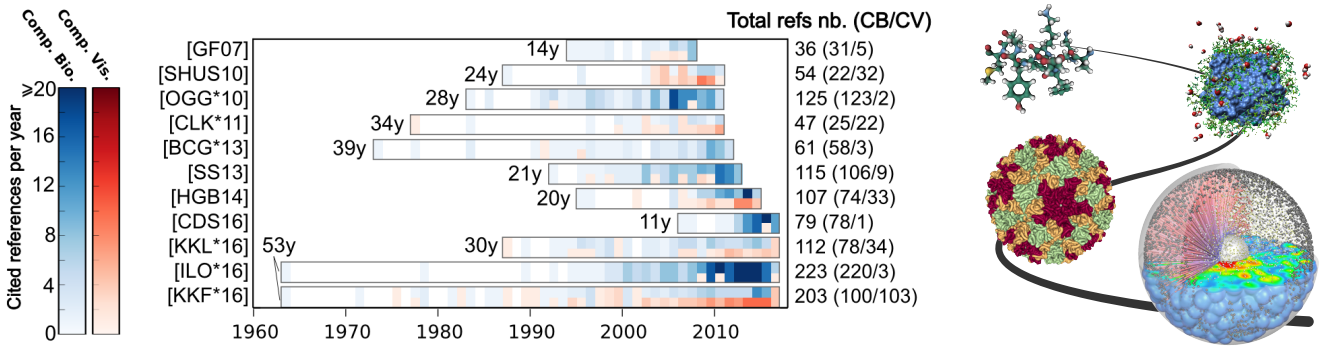

Figure 1: *List of surveys presented in this article indicating their time span, number of cited references per year, total number of references, and the ratio of papers coming from the Computational Biology (CB) and Computer Visualization (CV) fields respectively. If a paper refers to both types (CB and CV) of references for a same year, the cell is divided in two rows of different color. The collage to the right illustrates the scales covered by the visualizations in these surveys, ranging from small molecules over protein complexes to whole cells (screenshots made with UnityMol [DCP*14], MegaMol [GKM*15, KBE09], NGL Viewer [RH15, RBV*16] (norovirus example), and CellVis [FKE13]).*

Abstract

Visualizations for computational biology have been developing for over 50 years. With recent advances in both computational biology and computer graphics techniques, these fields have witnessed rapid technological advances in the last decade. Thus, coping with the large number of scientific articles from both fields is a challenging task. Furthermore, there remains a gap between the two communities of visualization and computational biology, resulting in additional challenges to bridge the divide. A team of computational biology and visualization scientists attempts to address these challenges by presenting unified state-of-the-art reviews from both communities. We apply a variety of data-driven analysis to highlight links or differences between studies from both communities. This approach facilitates the identification of present and future challenges in visualizing and analysing computational biology data. It offers a distinctive step forward in managing the literature on visualization of molecular dynamics and related simulation approaches.

Categories and Subject Descriptors (according to ACM CCS): Human-centered computing → Visualization → Visualization application domains → Scientific visualization

1. Introduction and Motivation

In computational biology—comprising bioinformatics, molecular modeling, and structural biology—visualization is an important means to gain insight into molecular structures and their dynamics. Due to its demanding nature, visualizing molecular data has always been tightly linked to computer hardware development [Lev66]. Originally, papers describing advances in molecular visualization were welcomed by the whole scientific community and published in journals with a broad audience such as *Science* [LFKC81]. More recently, scientific fields have become more specialized, resulting in focused scientific communities publishing in dedicated journals.

This fragmentation can lead to paradoxical situations where visualization challenges may be published in one type of journal while the solutions may appear in another. With this first survey of surveys (SoS), we reunite the communities by describing both questions posed by the computational biology community and answers provided by (or new challenges for) the visualization community. This work provides, for newcomers and experienced researchers, a unique and concise perspective presenting state-of-the-art literature in molecular visualization.

Survey Scope: Our team of authors consists of experts in both scientific visualization and computational biology. We have selected 11 survey papers spanning both fields. We focused on literature reviews addressing the rapidly expanding fields of structural biology and molecular modelling with a focus on spatio-temporal simulation data. For readers interested in a broader view of visualizing biological data, we refer to O'Donoghue *et al* [OGG*10]. The literature reviews cover selected, related topics: visualization of molecular structures [GF07] [KKF*16] and software dedicated to this task [OGF*10], advances based on Graphics Processing Units (GPUs) [CLK*11] [SHUS10], detection and analysis of cavities in proteins [BCG*13] [KKL*16], time-dependent biological data [SS13], and new challenges in molecular modelling leading to new visualization questions [CDS16] [ILO*16]. As a useful introduction to the links between molecular simulation and visualization, we discuss the review by Hirst *et al.* [HGB14]. As our literature selection covers a large time span, we focus on the last fifty years from the mid-sixties to 2016 (see Figure 1).

2. Survey of Surveys

In this section we describe each review and group them by main common themes such that closely related surveys are together. Details about references cited and literature time span are depicted in Figure 1.

Introduction to Molecular Visualization and Simulation
Hirst *et al.* propose an overview of the recent literature on molecular simulation and visualization [HGB14]. They highlight the increasing importance of Human-Computer Interaction (HCI) and virtual reality in the molecular visualization context. This survey introduces a series of articles dedicated to molecular visualization [far14]. This review contains 107 citations covering 20 years of research with about two thirds of the citations referring to computational biology work and one third to computer science papers.

Visualization of Molecular Structures
O'Donoghue *et al.* review visualization methods and tools that enable the community of structural biologists to gain insight into macromolecular structures [OGF*10]. This report covers an extensive list of web-based and stand-alone tools and discusses the advantages and disadvantages of the most common molecular structure acquisition techniques. The review covers 28 years of scientific literature containing 125 references, almost exclusively related to works published in the biological and experimental communities.

Goddard *et al.* discuss developments and challenges in visualization of molecular structure to better understand molecular systems such as Depth Perception, Level of Detail (LoD), 2D and abstract representations [GF07]. This review focuses on 14 years, citing 36 papers, of which 5 are from computer science.

The recent state-of-the-art report by Kozlíková *et al.* proposes an extensive review of visualizing biological data covering a wide range of spatial scale from atoms to cells [KKF*16]. The authors pay particular attention to molecular surface rendering with an interesting chronological perspective on visualization of the *solvent excluded surface*. Numerous challenges evoked by Goddard *et al.* [GF07] are addressed in this review such as LoD or the effective representation of dynamical data. This review covers more than fifty years of scientific research referring to 203 articles. These references are well balanced between computational biology and computer science literature.

Detection and Visualization of Cavities
While the previous selection of surveys discusses how it is possible to render a structure, here we present two reviews highlighting detection, visualization, and analysis of molecular cavities. These cavities are often important for the proper function of a molecule. This task is especially difficult as it needs to visualize voids which have to be well defined and detected.

Brezovsky *et al.* review programs available to identify, visualize, and analyse protein voids [BCG*13]. As the shape of the void may have an impact on the technique used to detect it, the authors compare different tools to assess which one is the best for a dedicated type of space. The review spans 39 years of literature, presenting a majority of articles published in computational biology journals.

Complementary to Brezovsky *et al.*, Krone *et al.* detail the technical background of the algorithms [KKL*16]. Their report also covers visualization methods for cavities. The authors present the definition and the classification of cavities. They classify the methods according to the underlying algorithms or the type of cavity definition. This study constitutes a very comprehensive review, spanning 30 years and citing 112 papers. The ratio of computer science to computational biology related papers is about one third.

GPU Computing
With the developments of programmable graphics cards in the early 2000's, development of new algorithms that harness this relatively new computing power are evolving rapidly.

Chavent *et al.* focus on studies that redesign traditional algorithms to exploit Graphics Processing Units (GPUs) [CLK*11]. This survey covers techniques that display small molecules up to macromolecular assemblies, and discusses visual effects to enhance molecular structure perception. It covers 34 years of research and cites 47 papers almost equally balanced between computer science and computational biology.

Even though it is not completely focused on visualization, we mention a closely related review from Stone *et al.* discussing the development of GPU-computing to accelerate molecular simulations [SHUS10]. This work covers 24 years of research and refers to 54 papers predominantly from computer science. Note that some of the previously cited reviews also discuss GPU computing (e.g. [GF07], [HGB14], [KKL*16] and [KKF*16]).

Visualizing Time-dependent Biological Data
Improved rendering efficiency now enables visualization of dynamical systems. Several reviews discuss this topic. O'Donoghue *et al.*

present different tools to render molecular motions [OGF*10]. Kozlíková *et al.* dedicate a full section to the visualization of molecular dynamics data [KKF*16].

In addition, we include the review by Secrier *et al.* which discusses the visualization of biological processes at different time scales [SS13]. This survey reviews time-dependent biology visualization tools by categorizing them into seven groups based on their time scale: molecular level (nano- to micro-seconds), gene level (micro-seconds/hours), network level (micro-seconds/days), cellular level (hours/days), level of an organism (days/weeks), population level (billions of years) and evolutionary scales (multiple levels). This review covers 21 years and cites 115 references with 9 computer science papers.

Challenges in Computational Biology

Computational biology is evolving very quickly, thus, new challenges appear regularly. Here, we highlight two recent reviews that outline challenges in computational biology. For computer scientists, these reports can inspire future research directions. For computational biologists, these reports cover the latest state-of-the-art.

Chavent *et al.* discuss the advances in molecular simulations of membrane proteins with a focus on protein-lipid interactions and modelling complex membranes at different scales [CDS16]. At the nanoscale resolution, simulations are used to predict and investigate fine lipid-protein interactions. Beyond the nanoscale, it is necessary to model very large and crowded systems requiring significant computing power. Reaching time-scales probed in experiments will require the development of new types of models. This review covers very recent work (the last 11 years), almost exclusively from the computational biology field.

Im *et al.* explore the modelling of biological systems at different scales [ILO*16]. They discuss how to move from one scale to another while simultaneously maintaining a high resolution to develop meaningful models. The next big challenge is to reach the cell scale and combine models with experimental data. This survey covers a long time span (up to 53 years) and is constituted by 223 references, mostly from the computational biology field.

3. From Text to Information: Meta-analysis of the Reviews

We perform a meta-analysis of all eleven surveys based on reference origins (CB or CV), shared references, and extracted keywords. These analyses yield new comparisons and insights not available from simply reading each paper separately.

Methods: To construct Figures 1 and 2, we extract the references from the Scopus database [sco] and analyze them using in-house Python scripts. For Figure 1, the references are curated by us to define which category a reference belongs to. Briefly, if the reference was published in an ACM, IEEE or related conference and journal it is categorized as a computer visualization paper, otherwise it was tagged as a *"computational biology"* paper. This category is kept very simple due to the paper format. We also investigated the concordance of important words across the surveys using the Natural Language Toolkit [Bir06] and Python scripts (see supplementary material for more details). Figure 3 shows a parallel coordinates plot that highlights the most represented words for each survey and a word cloud generated using the script by Müller [Mue].

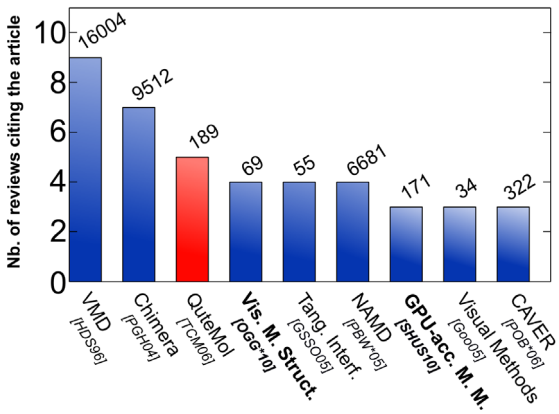

Figure 2: *The most common references shared by our 11 selected surveys. The two papers cited in bold print are included in our selection. We only displayed papers shared by at least 3 surveys. On top of each bar is the number of citations for each paper. Blue: computational biology papers; Red: scientific visualization paper.*

References as a Function of Time

Figure 1 shows that the selected reviews focus mainly on the last 25 years, even though some highlighted works were published before 1980. There is an imbalance between references from the CV and CB fields. The latter is clearly more represented. There is of course an intrinsic bias, as selected reviews are more from computational biology (9: [GF07], [OGF*10], [SHUS10], [CLK*11], [BCG*13], [SS13], [HGB14], [CDS16], [ILO*16]) than pure data visualization (2: [KKF*16] and [KKL*16]). Nevertheless, at least three of them ([SHUS10], [CLK*11], [HGB14]) are focusing on molecular graphics or algorithms development, which counter-balances the ratio to 6:5. Furthermore, even the reviews published in the scientific visualization field cite numerous computational biology papers. To explain this imbalance, we hypothesize that the technical orientation of CV papers and the dissemination through very dedicated conferences may prevent some researchers of being aware of these studies. Recent initiatives such as the VizBi [viz] and BioVis [bio] conference series may help to highlight work from computer visualization researchers. Another reason may be that, even if some CV papers are published in journals, some papers are only published as conference proceedings and may not be referenced in scientific article databases such as PubMed [pub] commonly used by CB researchers. This situation may cause large parts of CV research to be almost invisible to the CB community. Some CB journals also publish methods dedicated to molecular visualization and analysis such as *Journal of Molecular Graphics and Modeling*, *Journal of Computational Chemistry*, *PloS Computational Biology* etc. This topical intersection may create some competition with journals dedicated to computer science.

Shared References

These surveys share several references (see radial representation in supplementary material). Figure 2 shows that the most shared references are associated with software (VMD [HDS96], Chimera [PGH*04], NAMD [PBW*05], and CAVER [POB*06]). Only one reference comes from the CV field: Tarini *et al.* presented an Ambient Occlusion method applied to molecular visu-

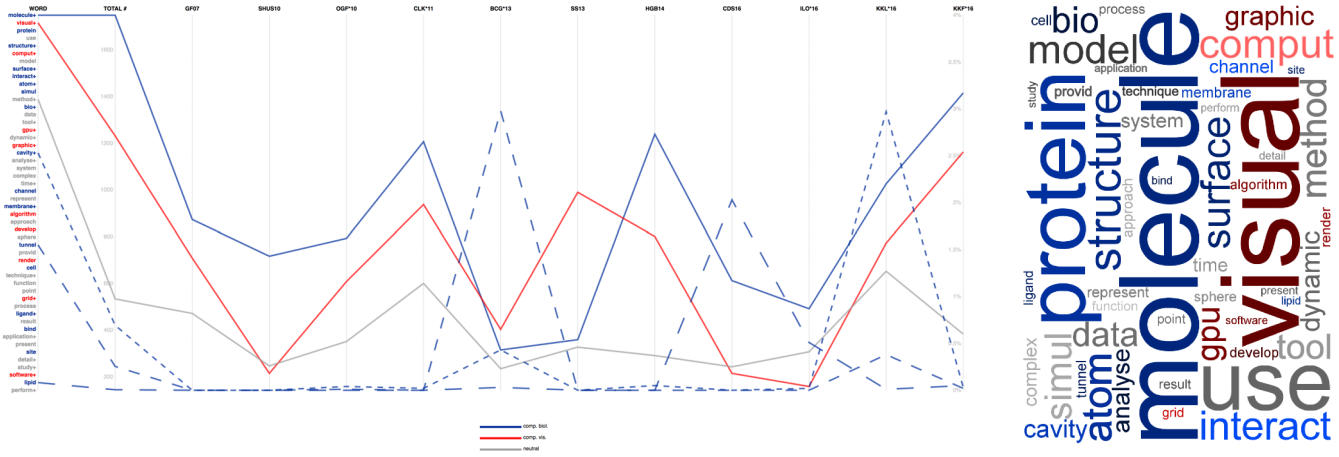

Figure 3: *Result of the text analysis. Left: Parallel coordinate plot displaying the collective concordance of the most frequent words in each survey. Right: Word cloud based on the collective concordance ranking. We categorized the keywords based on our expertise in both fields (category shown by color; blue: computational biology keywords; red: scientific visualization keywords; grey: neutral keywords).*

alization [TCM06]. This paper furthermore describes a software application, the molecular viewer QuteMol. Thus, making computer graphics programs available, even just as visualization prototype, is a key step to highlight CV researchers' work. Another good example is the fast *QuickSurf* molecular surface visualization by Krone et al. [KSES12], which was published at a major visualization conference but was also made available in the popular molecular visualization tool VMD [HDS96]. This makes the method widely known in both fields, as can be seen in the number of citations as well as the usage and feedback by CB researchers. Three references shared by the selected reviews are survey papers: [OGF*10], [SHUS10], [Goo05] with two of them discussed in section 2. The last paper mentioned discusses the combination of molecular visualization and 3D printing [GSSO05]. The number of shared references in the selected survey is in very good agreement with the overall number of citations for each paper. We observe one clear outlier: the NAMD program for Molecular Dynamics simulations [PBW*05] which is important for creating dynamic models but is out of the scope of these surveys.

Text Analysis

We performed a text analysis using the parallel coordinates plot depicted in Figure 3. An interactive version of the plot is available as supplementary material to allow interested readers to further investigate the data we collected for our survey of surveys. The interactive parallel coordinates plot is a useful way of exploring themes throughout the surveys. The user can exploit mouse motion to observe trends in the collecton of text over time. For example, if we hover the mouse over "cavity" we can see that it is a popular topic in the surveys, i.e [BCG*13] and [KKL*16]. Another example is with the term "lipid" which reoccurs often in [CDS16] and [ILO*16] but is never mentioned previously, with the exception of [BCG*13], but only twice in the references. This may indicate an emerging important visualization topic. In contrast to the interactive plot that can show correlations or concordances between the individual surveys, the word cloud presented in Figure 3 gives a static overview of the most important keywords. This figure highlights biological topics (such as protein, cell, ligand, membrane, molecule, lipid) or a part of it (channel, cavities, atom, structure*, tunnel) that can be interpreted as important application fields for CV reserachers. Some words are related to 3D objects (points, grid, surface, sphere) describing the essential graphical primitives used to render molecular objects. Some are potentially related to biological processes (binding, interact*) which are important to analyse and visualize.

4. Solved Problems and Future Challenges

Visualizing molecular structures and models is one of the first analysis steps every computational biologist takes to assess their results. A broad spectrum of tools are available to visualize objects ranging from protein structure to cavities both as static items or dynamical data sets. Recent advances in GPU computing improve the efficiency and the quality of the rendering. Nevertheless, molecular visualization remains challenging due to the increasing amount of simulation data [KKF*16]. First, dealing with models that can expand on different scales both in terms of structure [GF07, OGF*10, CLK*11, ILO*16] and time [SS13, KKF*16, OGF*10] is not yet solved. This type of visualization needs to be coupled with other methods to grasp the full complexity of molecular systems. Thus, there is a need for real time 3D annotation [CLK*11] and filtering [KKL*16]. These visualization advances may be combined with HCI and VR [HGB14, OGF*10] to help the user *immerse* in the system. Automating rendering and analysis [BCG*13] and storing the result for further analyses [CDS16] will be equally important. Finally, a huge gap still exists between CV and CB posing the challenge to turn innovations developed by computer visualization researchers into useful tools for computational biologists [GF07] and making the respective CV publications visible to the CB field. As elucidated by the *QuickSurf* example in section 3, making novel visualization methods available in existing open-source tools is a solution and a rewarding way to foster exchange between the two communities, even if it may require additional implementation effort. We think that this target will require further collaboration between the communities. Our survey of surveys is an important step in this direction.

References

[BCG*13] BREZOVSKY J., CHOVANCOVA E., GORA A., PAVELKA A., BIEDERMANNOVA L., DAMBORSKY J.: Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnology advances 31*, 1 (2013), 38–49. 2, 3, 4

[bio] BioVis. http://biovis.net (last accessed: 31.01.17). 3

[Bir06] BIRD S.: Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (2006), Association for Computational Linguistics, pp. 69–72. 3

[CDS16] CHAVENT M., DUNCAN A. L., SANSOM M. S.: Molecular dynamics simulations of membrane proteins and their interactions: from nanoscale to mesoscale. *Current Opinion in Structural Biology 40* (2016), 8–16. 2, 3, 4

[CLK*11] CHAVENT M., LÉVY B., KRONE M., BIDMON K., NOMINÉ J.-P., ERTL T., BAADEN M.: Gpu-powered tools boost molecular visualization. *Briefings in Bioinformatics* (2011), bbq089. 2, 3, 4

[DCP*14] DOUTRELIGNE S., CRAGNOLINI T., PASQUALI S., DERREUMAUX P., BAADEN M.: UnityMol: Interactive scientific visualization for integrative biology. In *IEEE Symposium on Large Data Analysis and Visualization (LDAV)* (2014), pp. 109–110. http://www.baaden.ibpc.fr/umol/ (last accessed: 14.02.17). 1

[far14] :. *Molecular Simulations and Visualization* (2014), vol. 169 of *Faraday Discussions*. 2

[FKE13] FALK M., KRONE M., ERTL T.: Atomistic Visualization of Mesoscopic Whole-Cell Simulations Using Ray-Casted Instancing. *Computer Graphics Forum 32*, 8 (2013), 195–206. 1

[GF07] GODDARD T. D., FERRIN T. E.: Visualization software for molecular assemblies. *Current opinion in structural biology 17*, 5 (2007), 587–595. 2, 3, 4

[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. http://www.megamol.org (last accessed: 14.02.17). 1

[Goo05] GOODSELL D. S.: Visual methods from atoms to cells. *Structure 13*, 3 (2005), 347–354. 4

[GSSO05] GILLET A., SANNER M., STOFFLER D., OLSON A.: Tangible interfaces for structural molecular biology. *Structure 13*, 3 (2005), 483–491. 4

[HDS96] HUMPHREY W., DALKE A., SCHULTEN K.: VMD: visual molecular dynamics. *Journal of Molecular Graphics 14*, 1 (1996), 33–38. 3, 4

[HGB14] HIRST J. D., GLOWACKI D. R., BAADEN M.: Molecular simulations and visualization: introduction and overview. *Faraday discussions 169* (2014), 9–22. 2, 3, 4

[ILO*16] IM W., LIANG J., OLSON A., ZHOU H.-X., VAJDA S., VAKSER I. A.: Challenges in structural approaches to cell modeling. *Journal of molecular biology* (2016). 2, 3, 4

[KBE09] KRONE M., BIDMON K., ERTL T.: Interactive visualization of molecular surface dynamics. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 1391–1398. 1

[KKF*16] KOZLÍKOVÁ B., KRONE M., FALK M., LINDOW N., BAADEN M., BAUM D., VIOLA I., PARULEK J., HEGE H.-C.: Visualization of biomolecular structures: State of the art revisited. *Computer Graphics Forum* (2016). 2, 3, 4

[KKL*16] KRONE M., KOZLÍKOVÁ B., LINDOW N., BAADEN M., BAUM D., PARULEK J., HEGE H.-C., VIOLA I.: Visual analysis of biomolecular cavities: State of the art. *Computer Graphics Forum 35*, 3 (2016), 527–551. 2, 3, 4

[KSES12] KRONE M., STONE J. E., ERTL T., SCHULTEN K.: Fast Visualization of Gaussian Density Surfaces for Molecular Dynamics and Particle System Trajectories. In *EuroVis - Short Papers* (2012), vol. 1, pp. 67–71. 4

[Lev66] LEVINTHAL C.: Molecular model-building by computer. *Scientific American 214*, 6 (1966), 42–52. cited By 131. 1

[LFKC81] LANGRIDGE R., FERRIN T. E., KUNTZ I. D., CONNOLLY M. L.: Real-time color graphics in studies of molecular interactions. *Science 211*, 4483 (1981), 661–666. 1

[Mue] MUELLER A.: word_cloud: A little word cloud generator in python. https://github.com/amueller/word_cloud (last accessed: 31.01.17). 3

[OGF*10] O'DONOGHUE S. I., GOODSELL D. S., FRANGAKIS A. S., JOSSINET F., LASKOWSKI R. A., NILGES M., SAIBIL H. R., SCHAFFERHANS A., WADE R. C., WESTHOF E., ET AL.: Visualization of macromolecular structures. *Nature methods 7* (2010), S42–S55. 2, 3, 4

[OGG*10] O'DONOGHUE S. I., GAVIN A.-C., GEHLENBORG N., GOODSELL D. S., HÉRICHÉ J.-K., NIELSEN C. B., NORTH C., OLSON A. J., PROCTER J. B., SHATTUCK D. W., ET AL.: Visualizing biological data - now and in the future. *Nature methods 7* (2010), S2–S4. 2

[PBW*05] PHILLIPS J. C., BRAUN R., WANG W., GUMBART J., TAJKHORSHID E., VILLA E., CHIPOT C., SKEEL R. D., KALE L., SCHULTEN K.: Scalable molecular dynamics with namd. *Journal of computational chemistry 26*, 16 (2005), 1781–1802. 3, 4

[PGH*04] PETTERSEN E. F., GODDARD T. D., HUANG C. C., COUCH G. S., GREENBLATT D. M., MENG E. C., FERRIN T. E.: Ucsf chimera–a visualization system for exploratory research and analysis. *Journal of computational chemistry 25*, 13 (2004), 1605–1612. 3

[POB*06] PETŘEK M., OTYEPKA M., BANÁŠ P., KOŠINOVÁ P., KOČA J., DAMBORSKÝ J.: Caver: a new tool to explore routes from protein clefts, pockets and cavities. *BMC bioinformatics 7*, 1 (2006), 316. 3

[pub] PubMed. https://www.pubmed.gov (last accessed: 31.01.17). 3

[RBV*16] ROSE A. S., BRADLEY A. R., VALASATAVA Y., DUARTE J. M., PRLIĆ A., ROSE P. W.: Web-based molecular graphics for large complexes. In *Proceedings of the 21st International Conference on Web3D Technology* (2016), Web3D '16, pp. 185–186. 1

[RH15] ROSE A. S., HILDEBRAND P. W.: NGL Viewer: a web application for molecular visualization. *Nucleic Acids Research 43*, W1 (2015), W576. http://proteinformatics.charite.de/ngl (last accessed: 14.02.17). 1

[sco] Scopus. (https://www.scopus.com (last accessed: 09.02.17). 3

[SHUS10] STONE J. E., HARDY D. J., UFIMTSEV I. S., SCHULTEN K.: Gpu-accelerated molecular modeling coming of age. *Journal of Molecular Graphics and Modelling 29*, 2 (2010), 116–125. 2, 3, 4

[SS13] SECRIER M., SCHNEIDER R.: Visualizing time-related data in biology, a review. *Briefings in bioinformatics* (2013), bbt021. 2, 3, 4

[TCM06] TARINI M., CIGNONI P., MONTANI C.: Ambient occlusion and edge cueing to enhance real time molecular visualization. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (2006), 1237–1244. cited By 189. 4

[viz] VIZBI - Visualizing Biological Data. https://vizbi.org (last accessed: 31.01.17). 3