Coupling Visualization and Data Analysis for Knowledge Discovery from Multi-dimensional Scientific Data

Oliver Rübel\textsuperscript{ab}, Sean Ahern\textsuperscript{c}, E. Wes Bethel\textsuperscript{a}, Mark D. Biggin\textsuperscript{d}, Hank Childs\textsuperscript{a}, Estelle Cormier-Michel\textsuperscript{e}, Angela DePace\textsuperscript{d}, Michael B. Eisen\textsuperscript{g}, Charless C. Fowlkes\textsuperscript{h}, Cameron G. R. Geddes\textsuperscript{i}, Hans Hagen\textsuperscript{b}, Bernd Hamann\textsuperscript{abj}, Min-Yu Huang\textsuperscript{j}, Cris L. Luengo Hendriks\textsuperscript{k}, Soile V. E. Keränen\textsuperscript{d}, David W. Knowles\textsuperscript{l}, Jitendra Malik\textsuperscript{m}, Jeremy Meredith\textsuperscript{c}, Peter Messmer\textsuperscript{e}, Prabhat\textsuperscript{a}, Daniela Ushizima\textsuperscript{a}, Gunther H. Weber\textsuperscript{a}, and Kesheng Wu\textsuperscript{a}

\textsuperscript{a}Computational Research Division, Lawrence Berkeley National Laboratory (LBNL), One Cyclotron Road, Berkeley, CA, 94720, USA.
\textsuperscript{b}International Research Training Group 1131, University of Kaiserslautern, Germany.
\textsuperscript{c}Oak Ridge National Laboratory (ORNL), P.O. Box 2008, Oak Ridge, TN 37831, USA.
\textsuperscript{d}Genomics Division, LBNL, One Cyclotron Road, Berkeley, CA, 94720, USA.
\textsuperscript{e}Tech-X Corporation, 5621 Arapahoe Ave., Suite A, Boulder, CO 80303.
\textsuperscript{f}Harvard Medical School, 200 Longwood Ave., Boston, MA, 02115, USA.
\textsuperscript{g}Howard Hughes Medical Institute and Department of Molecular and Cell Biology, University of California, Berkeley, Stanley Hall 304B, Berkeley, CA 94720, USA.
\textsuperscript{h}Donald Bren School of Information and Computer Science, University of California, Irvine, CA 92697, USA.
\textsuperscript{i}LOASIS program of LBNL, One Cyclotron Road, Berkeley, CA 94720, USA.
\textsuperscript{j}Institute for Data Analysis and Visualization (IDAV), Department of Computer Science, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA.
\textsuperscript{k}Centre for Image Analysis, Swedish University of Agricultural Sciences, Box 337, SE-751 05 Uppsala, Sweden.
\textsuperscript{l}Life Sciences Division, LBNL, One Cyclotron Road, Berkeley, CA, 94720, USA.
\textsuperscript{m}Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA.

Knowledge discovery from large and complex scientific data is a challenging task. With the ability to measure and simulate more processes at increasingly finer spatial and temporal scales, the growing number of data dimensions and data objects presents tremendous challenges for effective data analysis and data exploration methods and tools. The unique combination and close integration of methods from scientific visualization, information visualization, automated data analysis, and other enabling technologies —such as efficient data management— supports knowledge discovery from multi-dimensional scientific data. This paper surveys two distinct applications in developmental biology and accelerator physics, illustrating the effectiveness of the described approach.
1. INTRODUCTION

Knowledge discovery from large and complex collections of today’s scientific datasets is a challenging task. Due to advances in data acquisition and scientific computing, today’s datasets are becoming increasingly complex. With the ability to measure and simulate more processes at finer scales, the number of data dimensions and data objects has grown significantly in today’s scientific datasets, while the phenomena researchers are able to investigate become increasingly complex. Researchers are overwhelmed with data and standard tools are often insufficient to enable efficient data analysis and, hence, discovery of information and knowledge from the data.

We address these challenges via a combination of scientific visualization, information visualization, automated data analysis, and other enabling technologies. The tight coupling of different analysis methods and tools supports knowledge discovery from complex, multi-dimensional scientific data. To illustrate the effectiveness of this approach, we survey the processes and tools used to analyze: i) 3D gene expression data, and ii) laser wakefield particle acceleration data, demonstrating the applicability of the described basic concept to a large range of applications.

Analysis of 3D gene expression data is linked to the more general problem of understanding the control of embryo development, which is a fundamental question in biology. A cell’s unique fate is determined by specific combinations of developmental regulatory factors. These factors form part of complex genetic regulatory networks, which ultimately coordinate the expression of all genes. In order to study these complex systems, the BDTNP\(^1\) has developed so called PointCloud data, a novel type of spatial and temporal gene expression data. Single PointClouds are obtained via segmentation of two-photon microscopy images of *Drosophila* embryos and provide a quantitative representation of spatial gene expression levels of the *Drosophila* blastoderm at cellular resolution [1]. Multiple PointClouds representing a variety of genes at multiple developmental time intervals are registered into a single Atlas PointCloud describing the expression of about one hundred genes at multiple points in time [2]. Analysis of 3D gene expression data is challenging in particular due to the large number of data dimensions (genes) and the complex interactions between them.

Laser wakefield particle accelerators (LWFAs) [3] utilize an electron plasma wave to accelerate charged particles (e.g., electrons) to high energy levels over very short distances [4,5]. Analysis, understanding, and control of the complex physical processes of plasma-based particle acceleration requires understanding of how particle beams are formed and accelerated. These processes are best understood by tracing the particles that form a beam over time and studying their temporal evolution [6–9]. In laboratory experiments, however, it is impossible to record the complete evolution of a beam and much less to trace single particles within a plasma. Researchers from the LOASIS\(^2\) project perform simulation of LWFA experiments using VORPAL [10], in order to better understand the fundamental physics of plasma-based acceleration and the processes involved in experiments, as well as to improve experiments [11]. The datasets produced by LWFA simulations are (i) extremely large, (ii) of varying spatial and temporal resolution, (iii)

\(^{1}\)Berkeley Drosophila Transcription Network Project (BDTP): http://bdtnp.lbl.gov/Fly-Net/

heterogeneous, and (iv) high-dimensional, making analysis and knowledge discovery from complex LWFA simulation data a challenging task.

Section 2 introduces our general approach for knowledge discovery from multi-dimensional scientific data. We demonstrate the broad applicability of the described methodology in a survey of the analysis processes and tools used to analyze 3D gene expression (Section 3) and laser wakefield particle acceleration data (Section 4).

2. GENERAL METHODOLOGY

While the challenges in developmental biology and accelerator physics research are quite different, the same basic analytic methodology can be used for knowledge discovery from such complex data. The basic approach is based on the unique combination and close integration of: (i) enabling technologies, (ii) visualization, and (iii) data analysis (Figure 1). Enabling technologies are fundamental methods, needed for data analysis, that are not necessarily part of the analysis itself, e.g., methods for data retrieval, access, and management.

Visualization transforms data into readily comprehensible images and is an indispensable part of the scientific discovery process [12]. In particular in the context of multi-dimensional data, a single display is often not sufficient to reveal all aspects of the data. Scientific visualizations support detailed analysis of physical data characteristics, while information visualizations provide means for exploration of variable space and identification of relationships between different data dimensions. We use multiple views — each highlighting different aspects of the data — linked via the concept of data selection (brushing) [13–15]. Selected data subsets can be highlighted in any view enabling detailed analysis and knowledge discovery.

While interactive data exploration based on linked multiple views is effective, it also has limitations. Manual data exploration can be time-consuming — hindering the analysis of large data collections — and visual detection of all fine and subtle data features is often impossible. Automated data analysis methods promise to overcome these limitations of
Figure 2. The system for knowledge discovery from 3D gene expression data. Color indicates the areas the different components of the system belong to, i.e., enabling technologies (light blue), scientific visualization (lilac), information visualization (orange), and data analysis (green).

visual data analysis by assisting in the most complex and time-consuming steps of the analysis pipeline, e.g., through automated feature detection [16].

In practice, interpretation of automated analysis results is often unintuitive, may lead to false interpretations, and proper definition of analysis parameters is often complicated. By linking automated data analysis and visualization, we overcome the difficulties with both visual and automated data analysis. Automating the detection of data features of interest enables, e.g., development of advanced visualizations that focus on the main data portions of interest, significantly reducing clutter and occlusion of important information. At the same time, visualization eases validation and interpretation of analysis results and definition of input parameters. Ultimately, it is the tight and meaningful integration of all these different methods that enables us to effectively discover new knowledge.

3. APPLICATION I: 3D GENE EXPRESSION DATA

With the availability of 3D PointCloud gene expression data, new ways for analyzing the complex genetic regulatory networks controlling animal development are becoming possible. PointCloud data describes the output of these complex networks quantitatively, enabling analysis of the spatial patterns of gene expression, their temporal variation, and regulation.

PointCloudXplore (PCX) [17] is a visualization and analysis system specifically developed for the analysis of 3D gene expression data (Figure 2). PCX supports analysis of spatial gene expression patterns via dedicated 2D and 3D physical model representations of the embryo blastoderm [18]. Dedicated information visualizations (abstract views) provide means for exploration of gene expression space and identification of relationships
between genes. The concept of cell selection (brushing) allows the user to correlate the information shown in different views [19]. The user can select cells of interest in any view, e.g., via drawing on the embryo surface or via thresholding in parallel coordinates. This mechanism allows features of interest to be defined and highlighted in any view (Figure 3).

While visualization is a powerful approach for knowledge discovery from complex data sets, visual detection of all existing features is very difficult in this case due to the large number and subtlety of features and intricate nature of 3D gene expression data. A typical feature of interest defines, e.g., various groups of cells behaving similarly with respect to the expression of several genes or a single gene over time. In the context of conceptually simpler forms of expression data — such as microarray experiments — data clustering has already shown to be able to reveal details hidden in the data [20]. However, appropriately defining clustering parameters — such as the number of clusters — as well as validation and interpretation of clustering results, is still complicated.

PCX integrates data clustering directly with the visualization. Using a combination of visualization and dedicated algorithms for evaluating the quality of clustering results, the user can intuitively identify appropriate clustering parameters [21]. A cluster defines a selection of cells behaving similarly with respect to the expression of the set of genes used in the clustering process. Similar to user-defined cell selections, PCX can display and highlight automatically computed clusters in any view. The meaningful integration of data clustering with the visualization improves the visualization as well as the clustering process. Data clustering supports automatic detection and highlighting of data features in the visualization, enabling a more focused and accurate analysis process. Visualization provides effective means for accurate definition of clustering input parameters and allows
intuitive validation and interpretation of clustering results.

In particular in the context of novel scientific data — such as PointCloud data —, researchers need to be able to quickly develop new analysis functions. PCX addresses this need by providing an interface to MATLAB (The MathWorks Inc., Natick, MA, USA), allowing researchers to integrate custom analysis capabilities with PCX and providing biologists faster and more convenient access to advanced analysis functions [22]. With its interface to MATLAB, PCX supports fast prototyping and testing of new ideas, facilitating communication between bioinformatics researchers and experimental biologists. The close integration of MATLAB with the visualization improves the visualization by providing simple access to new, advanced analysis capabilities as well as the analysis implemented in MATLAB by providing efficient means for validation and exploration of analysis results.

4. APPLICATION II: LASER WAKEFIELD PARTICLE ACCELERATION DATA

Analysis of and knowledge discovery from large, complex, multi-dimensional laser wakefield particle accelerator (LWFA) simulation data is a challenging task. One main feature researchers are interested in are beams of high-energy particles formed during the course of LWFA simulations. To enable efficient and accurate data analysis, dedicated mechanisms for beam selection and detection are needed.

Figure 4 provides an overview of the system for knowledge discovery from LWFA simulations. The index/query system FastBit [23] serves as main interface to the data enabling
Coupling Visualization and Data Analysis for Knowledge Discovery ...

Figure 5. a) Parallel coordinates of time step $t = 86$ of a particle dataset showing all particles with a momentum in $x$ direction of $px > 1 \times 10^1$ (gray) and a selected particle beam (red). b) Semi-transparent rendering of the paths of the selected particles over time, using time steps 0 through 105. Color indicates $px$ (the momentum in the acceleration direction $x$) and height indicates momentum in $y$ direction ($py$). The particles of the selected beam at time step $t = 86$ are shown in addition (red). The transition in color from blue to red along the particle paths shows that the selected particles are constantly accelerated over time. The cork-screw-like structure of the paths illustrates the oscillating motion of the particles in the wake. c) Close-up view of the region in figure b showing the selected particles at time step $t = 86$.

Fast computation of conditional histograms, threshold queries, and ID-based queries. The close integration of FastBit and the state-of-the-art visualization system VisIt [24] supports fast visual exploration of very large datasets [8]. VisIt implements an efficient rendering method for parallel coordinates based on 2D histograms, computed directly using FastBit. Histogram-based parallel coordinates serve as main interface for defining multi-dimensional range queries used for selection of particle beams (Figure 5a). Once a subset of particles of interest has been identified, the particle IDs are saved as a named selection. Named selections can be applied to any plot in VisIt, enabling effective linking of multiple physical and abstract data views (Figure 5b, c). The close integration of multiple views together with efficient data management supports interactive data exploration based on the iterative refinement and validation of data queries.

While interactive selection of particle beams is effective, it requires substantial manual input from the user and can be time-consuming. Automating the detection of particle beams supports a more focused and efficient analysis process [9,25]. The beam path analysis algorithm [25] defines an efficient analysis pipeline that supports fast detection of particle beams (Figure 4 right). First, each time step is analyzed independently to detect individual particle bunches. The derived information is merged to define a single description for each main particle beam. Finally, the algorithm computes the different temporal phases of each beam — defining, e.g., the time frame when a beam was formed and accelerated — as well as two distance fields $d_s$ and $d_m$, defining the distance of particles to the beam in physical and momentum space, respectively.
The automated beam detection is linked with the visualization in two ways. First, the beam path analysis automatically creates a set of named selections — one per detected beam — that can be applied to any plot in VisIt, enabling a fully automated beam analysis process. Second, a set of complementary files is created, containing additional information about the particle paths. These files can be visualized directly in the context of the simulated data and enable a more efficient manual data exploration process by providing information about: i) the temporal phases of a beam, ii) an appropriate reference time step for each beam, and iii) the beam distance fields $d_s$ and $d_m$, enabling a faster and more accurate selection of particle beams. Linking the automated beam detection with the visualization improves the visualization by enabling a more streamlined and focused analysis process as well as the automated analysis by providing effective means for investigation and validation of analysis results.

5. CONCLUSIONS

The increasing complexity and size of today’s scientific data poses tremendous challenges for data understanding and knowledge discovery. We have described an integrative approach for knowledge discovery from multi-dimensional scientific data based on the concept of linking visualization and data analysis. We have illustrated the effectiveness of this methodology by describing how it is applied in practice in the analysis of 3D gene expression and laser wakefield particle accelerator data. While different applications require different visual representations and analysis methods, it is ultimately the powerful combination and close integration of multiple different methods that enables effective knowledge discovery. The meaningful integration of visualization, data analysis, and enabling technologies supports a more efficient, detailed, and focused analysis process than possible based solely on either visualization or data analysis methods alone.

6. ACKNOWLEDGMENTS

The work concerning the analysis of 3D gene expression data was supported by the National Institutes of Health through grant GM70444 and by Lawrence Berkeley National Laboratory (LBNL) through the Laboratory Directed Research Development (LDRD) program. Research performed at LBNL was also supported by the Department of Energy under contract DE-AC02-05CH11231. In addition, we gratefully acknowledge the support of an International Research Training Group (IRTG 1131) grant provided by the German Research Foundation (DFG), awarded to the University of Kaiserslautern, Germany.

The work concerning the analysis of LWFA data was supported by the Director, Office of Science, Offices of High Energy Physics and Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through the Scientific Discovery through Advanced Computing (SciDAC) program’s Visualization and Analytics Center for Enabling Technologies (VACET) and the COMPASS project. This research used resources of the National Energy Research Scientific Computing Center (NERSC), which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

We thank the members of IDAV at UC Davis, the BDTNP and LOASIS at LBNL, the IRTG 1131, LBNL’s Visualization Group, and the VORPAL team.
REFERENCES


24. VisIt is available from https://wci.llnl.gov/codes/visit/.