Databases and Ontologies

DIVE: A Data Intensive Visualization Engine

Dennis Bromley¹,3, Steven J. Rysavy¹,3, Robert Su², Rudesh D. Toofanny², Tom Schmidlin², and Valerie Daggett¹,2,4

¹Division of Biomedical and Health Informatics, University of Washington Medical School, Seattle, WA, USA
²Department of Bioengineering, University of Washington, Seattle, WA, USA
³These authors contributed equally to this work.

ABSTRACT
Summary: Modern scientific investigation is generating increasingly larger data sets, yet analyzing these data with current tools is challenging. DIVE is a software framework intended to facilitate big data analysis and reduce the time to scientific insight. Here we present features of the framework and demonstrate DIVE’s application to the Dynaemomics project, looking specifically at two proteins.

Availability and Implementation: Binaries and documentation are available at http://www.dynameomics.org/DIVE/DIVESetup.exe

Contact: daggett@uw.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

The advent of massive networked computing resources has enabled virtually unlimited data collection, storage and analysis from low-cost genome sequencing, high-precision molecular dynamics simulations, and high-definition imaging data for radiology, to name just a few examples. This explosion of “big data” is changing traditional scientific methods; instead of relying on experiments to output relatively small, targeted data sets, data mining techniques are being used to analyze data stores with the intent of learning from the data patterns themselves. Unfortunately, data analysis and integration in large data storage environments is challenging even for experienced scientists. Furthermore, most existing domain-specific tools designed for complex, heterogeneous datasets are not equipped to visually analyze big data.

DIVE is a software framework designed for exploring large, heterogeneous, high-dimensional data sets using a visual analytics approach (SI Fig. S1). Visual analytics is a big data exploration methodology emphasizing the iterative process between human intuition, computational analyses, and visualization. DIVE’s visual analytics approach integrates with traditional methods, creating an environment that supports data exploration and discovery.

2 SYSTEM AND IMPLEMENTATION

DIVE provides a rich, ontologically expressive data representation and a flexible, modular streaming-data architecture, or pipeline (SI Fig. S2). It is accessible through an application programming interface (API), command line interface or graphical user interface. Applications built on the DIVE framework inherit features such as a serialization infrastructure, ubiquitous scripting, integrated multithreading and parallelization, object-oriented data manipulation, and multiple modules for data analysis and visualization. DIVE can also interoperate with existing analysis tools to supplement its capabilities, such as the Visualization Toolkit (Schroeder et al., 1996), Cytoscape (Shannon et al., 2003) and Bing maps (http://bing.com) by either exporting data into known formats or by integrating with published software libraries. Furthermore, DIVE can import compiled software libraries and automatically build native ontological data representations, reducing the need to write DIVE-specific software. From a data perspective, DIVE supports the joining of multiple heterogeneous data sources, creating an object-oriented database capable of showing inter-domain relationships. And while DIVE currently focuses on bioinformatics, DIVE itself is data-agnostic; data from any domain may enter the DIVE pipeline.

A core feature of DIVE’s framework is the flexible, graph-based data representation. DIVE data are stored as nodes in a strongly-typed, ontological network defined by the data. These data can be a simple set of numbers or a complex object hierarchy with inheritance and well-defined relationships. Data flow through the system explicitly as a set of data points passed down the DIVE pipeline or implicitly as information transferred and transformed through the data relationships (SI Fig. S3e). A thorough description of the novel technical contributions of DIVE is provided elsewhere (Rysavy, S.J.; Bromley, D; Daggett, V.; DIVE: A Graph-based Visual Analytics Framework for Big Data, submitted for publication, 2013).

3 RESULTS

The impetus for DIVE was data mining the Dynaemomics data set (Van der Kamp et al., 2010). Dynaemomics is a large, data-intensive project that contains atomistic molecular dynamics (MD) simulations of the native state and unfolding pathways of representatives of essentially all protein folds (Van der Kamp et al., 2010). These protein simulations and associated biophysical analyses are stored in a mixed data warehouse (Simms and Daggett, 2012) and file system environment distributed over multiple servers containing hundreds of terabytes of data and over 10⁷ times as many structures as the Protein Data Bank (PDB) (Bernstein et al., 1977), representing the largest collection of protein structures and protein simulations in the world.

In the domain of structural biology, Dynaemomics exemplifies the challenges of big data. Here we present DIVE applications involving two proteins where specialized modules built on the DIVE framework are used to accelerate biophysical analysis. The first protein is the transcription factor p53, mutations in which are implicated in cancer. The second protein is human Cu-Zn superoxide dismutase 1 (SOD1), mutations in which are associated with amyotrophic lateral sclerosis (Rakhit and Chakrabartty, 2006).

To whom correspondence should be addressed.
The Y220C mutation of the p53 DNA binding domain is responsible for destabilizing the core (Joerger et al., 2006), leading to approximately 75,000 new cancer cases annually (Boeckler et al., 2008). We have used the DIVE framework to analyze the structural and functional effects of the Y220C mutation through a module called ContactWalker (Bromley et al., 2013), which identifies amino acids’ interatomic contacts disrupted significantly as a result of mutation. The contact pathways between disrupted residues are identified using DIVE’s underlying graph-based data representation.

Figure 1a shows the most disrupted contacts in the vicinity of the Y220C mutation. Specific residues, contacts and simulations were identified for more focused analysis. Interesting interatomic contact data are isolated and then specific MD time points and structures are selected for further investigation. For example, see the contact data mapped onto a structure containing a stabilizing ligand, which docks closely to many of the disrupted residues, suggesting a correlation between the mutation-associated effects and the observed stabilizing effects of the ligand (Fig. 1a).

As another example of the use of DIVE, we have over 300 simulations of 106 disease-associated mutants of SOD1 (Schmidlin et al., 2009). Through extensive studies of A4V mutant SOD1 simulations, Schmidlin et al. (2009) previously noted the instability of two β-strands in the SOD1 Greek Key β-barrel structure. However, that analysis took several years to complete and such manual interrogation of simulations does not scale to allow us to search for general features linked to disease across hundreds of simulations. Using DIVE, we were able to further explore the formation and persistence of the contacts and packing interactions in this region across multiple simulations of mutant proteins. DIVE facilitates isolation of specific contacts, rapid plotting of selected data, easy visualization of the relevant structures, and geographic locations of specific mutations while providing intuitive navigation from one view to another (Fig. 1 and SI Fig. 1).

The top panel of Fig. 1b maps secondary structure for different variants as an example of DIVE’s charting tools. This chart is quickly generated, contains results for over 300 SOD1 mutant simulations, is customizable, and links to the protein structure property data (in this case the change in the structure over time) with a single mouse click (Fig. 1b). These data are in turn linked to protein structure modules, allowing interactive visualization of over 60,000 structures from each of the 300 simulations, all streamed from the SQL data warehouse (Fig. 1b). With DIVE, we simplified the transition between high-level protein views and atomic level details, facilitating rapid analysis of large amounts of data. DIVE can also show the context of the detailed results on other levels, such as worldwide disease incidence (SI Fig. S1).

DIVE’s utility is not limited to protein simulations. To demonstrate its versatility, usability and data-agnostic nature, we applied it to additional domains. Brief details of these applications are provided in Supplementary Information. One example shows an interaction with the Gene Ontology (Ashburner et al., 2000) and another example explores professional baseball statistics.

4 CONCLUSIONS

Overall, DIVE provides an interactive data-exploration framework that expands upon conventional analysis paradigms and self-contained tools. We provided analytic examples in the protein simulation domain, but the DIVE framework is not limited to this field. DIVE can adapt to existing data representations, consume non-DIVE software libraries, and import data from an array of sources. As research becomes more data-driven and reliant on data mining and visualization, big data visual analytics solutions should provide a new perspective for scientific investigation.

REFERENCES


Bromley,D. et al. (2013) Structural Consequences of Mutations to the α-Tocopherol Transfer Protein Associated with the Neurodegenerative Disease Ataxia with Vitamin E Deficiency. Biochemistry, 52, 4264–4273.


