BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 4 2014, pages 593–595
doi:10.1093/bioinformatics/btt721

Databases and ontologies Advance Access publication December 13, 2013

DIVE: a data intensive visualization engine

Dennis Bromley[1,†], Steven J. Rysavy[1,†], Robert Su[2], Rudesh D. Toofanny[2], Tom Schmidlin[2] and Valerie Daggett[1,2,*]

[1]Division of Biomedical and Health Informatics, University of Washington Medical School and [2]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Modern scientific investigation is generating increasingly larger datasets, 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 Dynameomics 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.

Received on August 21, 2013; revised on November 13, 2013; accepted on December 8, 2013

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 datasets, 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 datasets using a visual analytics approach (Supplementary 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 (Supplementary Fig. S2). It is accessible through an application programming interface, 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 although 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 (Supplementary Fig. S3e). A thorough description of the novel technical contributions of DIVE is provided elsewhere (Rysavy,S.J *et al.*, 2014).

3 RESULTS

The impetus for DIVE was data mining the Dynameomics dataset (Van der Kamp *et al.*, 2010). Dynameomics 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 >10^4 times as many structures as the Protein Data Bank (Bernstein *et al.*, 1977), representing the largest collection of protein structures and protein simulations in the world.

In the domain of structural biology, Dynameomics 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.

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

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).

The Y220C mutation of the p53 DNA binding domain is responsible for destabilizing the core (Joerger *et al.*, 2006), leading to ~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 >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 Supplementary Fig. S1).

The top panel of Figure 1b maps secondary structure for different variants as an example of DIVE's charting tools. This chart is quickly generated, contains results for >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 >60 000 structures from each of the 300 simulations, all streamed from the Structured Query Language (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 (Supplementary 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.

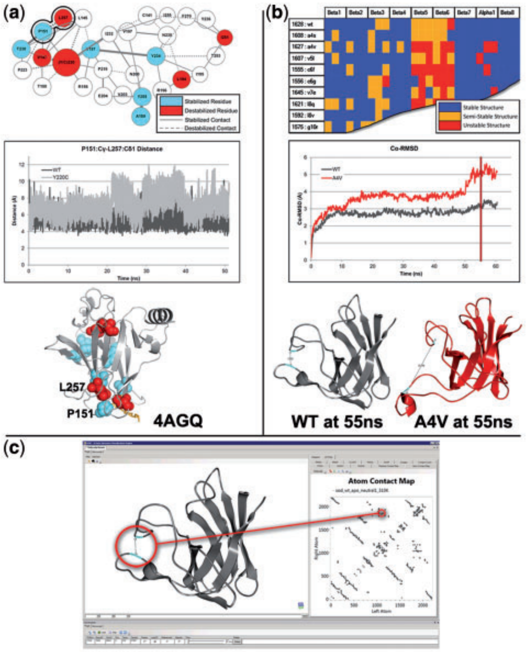

Fig. 1. Interactive visualizations in DIVE. (**a**) The p53 analysis visualizations. Top, ContactWalker summary of contact differences between wild-type and Y220C simulations. The highlighted residues have contacts with ≥50% occupancy change. Middle, distances between P151 and L257, outlined in black in the map above. Bottom, p53 with ligand (stick figure at bottom) (Protein Data Bank code 4AGQ) in proximity to disrupted colored residues. (**b**) SOD1 analysis visualizations. Top, aggregated secondary structural data from mutant simulations. Middle, plot of the Cα root-mean-squared (RMS) deviation of the wild-type and A4V mutant simulations. Bottom, MD structures. (**c**) Protein dashboard application showing a viewer and interactive contact map

4 CONCLUSIONS

Overall, DIVE provides an interactive data-exploration framework that expands on 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.

Funding: This work was supported by Microsoft through the External Research Program at Microsoft Research (www.microsoft.com/science) [to V.D.]; by the National Institutes of Health [GM50789 to V.D.]; and by the National Library of Medicine [5T15LM007442 to D.B. and S.J.R.]. Dynameomics simulations were performed using computer time through the Department of Energy Office of Biological Research as provided by the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy [contract no. DE-AC02-05CH11231].

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for unification of biology. *Nat. Genet.*, **25**, 25–29.

Bernstein,F.C. *et al.* (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Boeckler,F.M. *et al.* (2008) Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. *Proc. Natl Acad. Sci. USA*, **105**, 10360–10365.

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.

Joerger,A.C. *et al.* (2006) Structural basis for understanding oncogenic P53 mutations and designing rescue drugs. *Proc. Natl Acad. Sci. USA*, **103**, 15056–15061.

Rakhit,R. and Chakrabartty,A. (2006) Structure, folding, and misfolding of Cu,Zn superoxide dismutase in amyotrophic lateral sclerosis. *Biochem. Biophys. Acta*, **1762**, 1025–1037.

Rysavy,S.J., Bromley and Daggett,V. (2014) DIVE: A graph-based visual analytics framework for big data. IEEE Computer Graphics and Applications, accepted for publication.

Schmidlin,T. *et al.* (2009) Structural changes to monomeric CuZn superoxide dismutase caused by the familial amyotrophic lateral sclerosis-associated mutation A4V. *Biophys. J.*, **97**, 1709–1718.

Schroeder,W. *et al.* (1996) *The Visualization Toolkit: An Object-Oriented Approach to 3-D Graphics.* Prentice Hall, Upper Saddle River, NJ.

Shannon,P. *et al.* (2003) Cytoscape: a SOFTWARE environment for integrated models of Biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Simms,A.M. and Daggett,V. (2012) Protein simulation data in the relational model. *J. Supercomp.*, **62**, 150–173.

Van der Kamp,M.W. *et al.* (2010) Dynameomics: a comprehensive database of protein dynamics. *Structure*, **18**, 423–435.