# Browsing Large Scale Cheminformatics Data with Dimension Reduction

Judy Qiu[1], Jong Youl Choi[1,2], Seung-Hee Bae[1,2], Thilina Gunarathne[1,2], Geoffrey Fox[1,2], Bin Cao[2], David Wild[2]

*[1]Pervasive Technology Institute, [2]School of Informatics and Computing,*
*Indiana University*
*Bloomington IN, U.S.A.*
*{ xqiu, jychoi, sebae, gcf@indiana.edu}*

**SALSA project** (salsahpc.indiana.edu) investigates new programming models of parallel multicore computing and Cloud/Grid computing. It aims at developing and applying parallel and distributed Cyberinfrastructure to support large scale data analysis. We demonstrate this with a project for life sciences and present PubChemBrowse, a customized visualization tool for Cheminformatics research. Visualization of large-scale high dimensional data tool is highly valuable for scientific discovery in many fields. We present a novel 3D data point browser that displays complex properties of massive data on commodity clients. As in GIS browsers for Earth and environment data, chemical compounds with similar properties are nearby in the high dimensional space. PubChemBrowse is built around in-house high performance parallel MDS (Multi-Dimensional Scaling) and GTM (Generative Topographic Mapping) [1] [2] services and supports fast interaction with an external biochemical repository database. We provide robust deterministic annealing and interpolation for adding addition points. The browser will scale up to 60 million points of full NIH PubChem.

We demonstrate the following key features of PubChemBrowse:

i) A lightweight 3D data visualization client to browse large (a few million) and high-dimensional data backed by high-performance cloud technology [3]. Displaying various kinds of meta-data as extra information.

ii) On-line data fetching by connecting a remote external system, Chem2Bio2RDF [4], which is an integrated repository of chemogenomic and systems chemical biology data.

iii) Research results for drug discovery with mining cause-effect relationship between large number of chemical compounds and diseases.

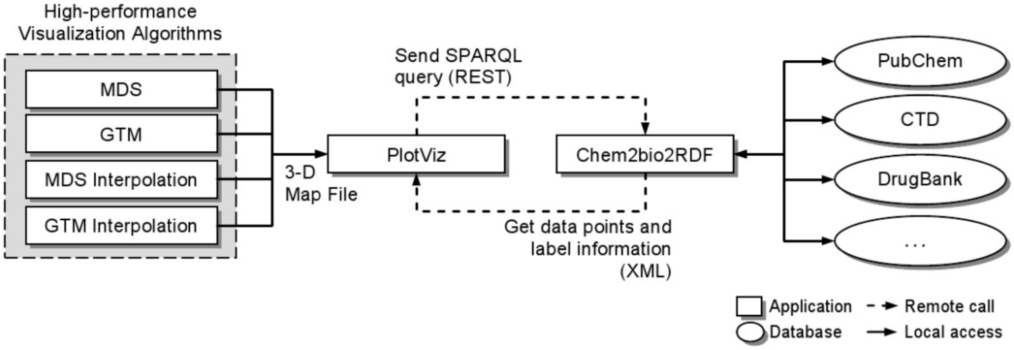
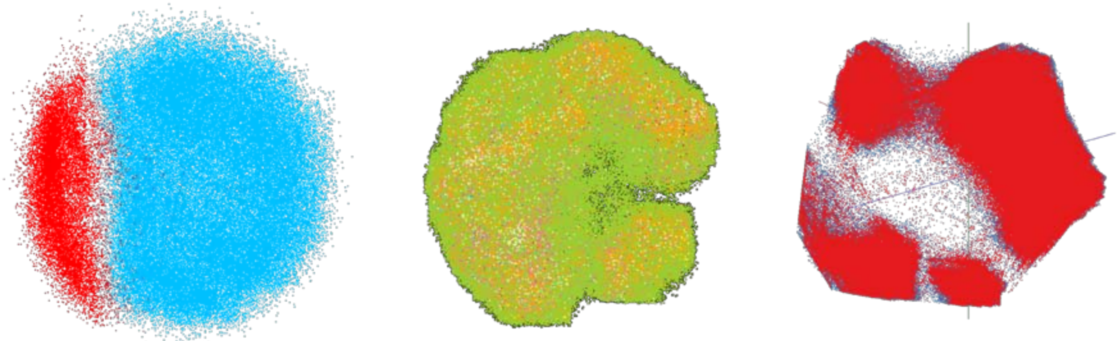


**Figure 1. Architecture of PubChemBrowse**

By using PubChemBrowse, one can easily identify points of interests by colors or select a group of points distinguished by structural distribution in 3D space. Additional functions include browsing the data by rotation, zooming or panning the 3D space to search for details. Dynamic updating the labels of points or adding new data points are supported by sending on-line SPARQL query to Chem2Bio2RDF system. With our tool, researchers can easily browse very large datasets with ease. We've developed parallel MDS and GTM algorithms [1] [2] to visualize large and high-dimensional data. As shown in Figure 2, we processed 0.1 million PubChem data with 166 dimensions and used parallel interpolation algorithms to speed up the process for up to 2M PubChem points.



***MDS for 100k PubChem data***
100k PubChem data having 166 dimensions are visualized in 3D space. Colors represent 2 clusters separated by their structural proximity.

***GTM for 930k genes and diseases***
Genes (green color) and diseases (others) are plotted in 3D space, aiming at finding cause-and-effect relationships.

***GTM with interpolation for 2M PubChem data***
2M PubChem data is plotted in 3D with GTM interpolation approach. Blue points are 100k sampled data and red points are 2M interpolated points.

**Figure 2. Visualization of PubChem database**

## References

[1] Seung-Hee Bae, Jong Youl Choi, Judy Qiu, Geoffrey Fox, **Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation**, to appear in the Proceedings of ACM High Performance and Distributed Computing (**HPDC** 2010) conference, June 20-25, 2010, Chicago, USA.

[2] Jong Youl Choi, Seung-Hee Bae, Xiaohong Qiu and Geoffrey Fox, **High Performance Dimension Reduction and Visualization for Large High-dimensional Data Analysis**, to appear in the Proceedings of the 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (**CCGrid** 2010), May 17-20, 2010, Melbourne, Australia.

[3] Jaliya Ekanayake, Xiaohong Qiu, Thilina Gunarathne, Scott Beason, Geoffrey Fox, **High Performance Parallel Computing with Clouds and Cloud Technologies**, to appear as a book chapter to Cloud Computing and Software Services: Theory and Techniques, CRC Press (Taylor and Francis), ISBN-10: 1439803153.

[4] Semantic Systems Chemical Biology at http://chem2bio2rdf.org/

## About DSC of Indiana University

The Digital Science Center (https://pti.iu.edu/dsc/) focuses on creating an intuitively usable cyberinfrastructure with tremendous capabilities for supporting collaboration and computation. Easy-to-use, human-centered interfaces to cyberinfrastructure created by the Digital Science Center will enable the many thousands of researchers in the public and private sectors to use the capabilities of cyberinfrastructure and accelerate innovation and discovery.

DSC is part of the Pervasive Technology Institute at Indiana University (www.pti.iu.edu). Supported by a $15 million grant from the Lilly Endowment, Inc., PTI is dedicated to the development and delivery of innovative information technologies and technology policies to advance research, education, industry and society. PTI leads a major new NSF funded $15 million project FutureGrid developing a testbed linked to TeraGrid for innovative new approaches for large scale scientific computing (http://uitspress.iu.edu/news/page/normal/11841.html).