Browsing Large Scale Cheminformatics Data with Dimension Reduction

Jong Youl Choi, Seung-Hee Bae, Judy Qiu, Geoffrey Fox
School of Informatics and Computing
Pervasive Technology Institute
Indiana University
Bloomington IN, 47408, USA
{jychoi,sebae,xqiu,gcf}@indiana.edu

Bin Chen, David Wild
School of Informatics and Computing
Indiana University
Bloomington IN, 47408, USA
{binchen,djwild}@indiana.edu

ABSTRACT

Visualization of large-scale high dimensional data tool is highly valuable for scientific discovery in many fields. We present PubChemBrowse, a customized visualization tool for cheminformatics research. It provides 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 browser. PubChemBrowse is built around in-house high performance parallel MDS (Multi-Dimensional Scaling) and GTM (Generative Topographic Mapping) services and supports fast interaction with an external property database. These properties can be overlaid on 3D mapped compound space or queried for individual points. We prototype use with Chem2Bio2RDF system using SPARQL query language to access over 20 publicly accessible bioinformatics databases. We describe our design and implementation of the integrated PubChemBrowse application and outline its use in drug discovery. The same core technologies can be used to develop similar high dimensional browsers in other scientific areas.

General Terms

Applications, Visualization

Keywords

Visualization, MDS, GTM, Interpolation, Semantic Web

1. INTRODUCTION

The scale of scientific data generated by new instruments or experiments along with substantial public accessibility make the issue of tools a challenging problem. Volumes of PubChem data need to be visualized towards the end of capture, curation, and analysis pipeline to support interactive studies. To browse 60 million intrinsically high-dimensional NIH PubChem data, we have developed a 3D data point visualization tool, named PubChemBrowse, to display chemical structures with complex properties such as gene and disease relationships established through querying Chem2Bio2RDF system [1] for drug discovery.

We are applying parallel Multidimensional Scaling (MDS) and Generative Topographic Mapping (GTM) algorithms to structural descriptors of around 60 million chemical structures in the PubChem dataset, to provide three-dimensional graphical plot representations of the structural diversity of nearly all of the chemical compounds that are currently known. By labeling these plots with properties of the compound, including simple properties like molecular weight, and complex properties such as gene and disease relationships established through querying our Chem2Bio2RDF system, we can investigate the overall properties of regions of chemical space inhabited by PubChem compounds, as well as embedding new compounds into this framework.

In section 2, we give a brief overview of related work and our high performance parallel dimension reduction technologies including Multidimensional Scaling (MDS) and Generative Topographic Mapping (GTM) algorithms and relational biochemical repository framework – Chem2Bio2RDF and its SPARQL query interface. Section 3 presents details of our design and implementation of an integrated system PubChemBrowse that plots 3D PubChem compounds as well as embedding new compounds retrieved in real time with various labeling capabilities. A case study is given in section 4 with a data set of up to 930,000 data points followed by a summary of our work and future improvements.

2. DATA VISUALIZATION AND REMOTE DATA ACCESS

Large scale data visualization is an active research area in many fields of science: The Sloan Digital Sky Survey (SDSS) project [2] for astronomy, UCSC [3] and Ensemble [4] for genomics, and Google Earth for geology. Although designed with different technologies, they share one common concept: users need only a lightweight client but can access huge data that are curated and processed through intensive computation. We have used a multi-stage pipeline model to explore natural data parallelism for large-scale scientific problems. This has been demonstrated in our biology gene sequencing application from data acquisition, analysis, reduction and aggregation for visualization and implemented with both classic HPC and Cloud technologies [5]. We are taking a similar approach for the cheminformatics research, assisting drug discovery from large dataset.

ECMLS '10 Chicago, Illinois USA

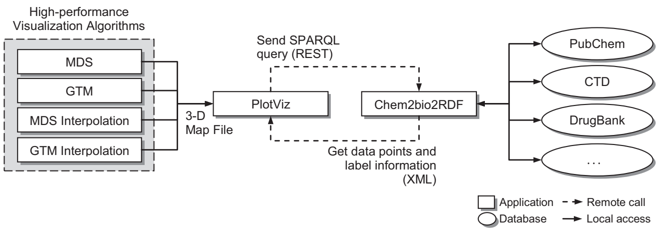


Figure 1: System architecture for PubChemBrowse.

It is also worth mentioning that [6] has introduced data visualization for drug discovery and a visualization tool. We present the same type of tool for drug discovery but we integrate our system with more cutting-edge technologies: high-performance visualization algorithms based on HPC and Cloud environments, large data access by using semantic web technologies, and a lightweight 3D visualization client.

In the field of drug discovery, which aims at mining cause-and-effect relationships between genes and diseases from large volume of data sources, one needs to access various kinds of databases – such as PubChem for chemical compounds and structures, the Comparative Toxicogenomics Database (CTD) for gene information, the DrugBank database, to name a few – and visualize them in 2D or 3D space to explore findings and verify results.

More specifically, such process typically involves three main tasks: i) visualizing multi-dimensional chemical compounds in the PubChem database in 3D space, ii) finding relationships from various public databases, such as gene-compounds and/or disease-compounds, and assigning labels to chemical compounds, and iii) displaying labels by using different colors or symbols in the previous 3D visualization and exploring the results. The last two steps can be repeated until we have meaningful discovery by overlaying new information to the previous results.

Authors have been researching on developing high performance visualization algorithms, such as parallel MDS and GTM and their interpolation extensions [7, 8], to visualize large PubChem dataset in 3D space by using our in-house 3D data point visualization tool and now we extend its functionality to access external data sources in a dynamic way.

An issue in dealing with various types of databases in drug discovery is that databases are often too big to store in local storage and hard to maintain consistency with the original dataset since their frequent updates. Another issue can arise due to non-uniform data structures. Typically those databases are not compatible to each other so that elaborated work needs to be done to use them together. To overcome such problems, Chem2Bio2RDF system has been developed to aggregate various public databases for chemical or biological information and provide an uniform interface which enables users to access the multiple databases by using SPARQL as a standard query language in semantic web technology. Our new tool, named PubChemBrowse, can interact with Chem2Bio2RDF system by using SPARQL query to access various databases and large number of datasets in an online and uniform way.

2.1 Data visualization algorithms

Among many dimension reduction algorithms, we focus on using Multidimensional Scaling (MDS) [9, 10] and Generative Topographic Mapping (GTM) [11] due to their popularity and theoretical strength. More details are as follow.

MDS : Multidimensional scaling (MDS) is a general term of the techniques to configure low dimensional mappings of the given high-dimensional data with respect to the pairwise proximity information, while the pairwise Euclidean distance within the target dimension of each pair is approximated to the corresponding original proximity value. To solve MDS problem, we use SMACOF algorithm. For details of the SMACOF algorithm, please refer to [12].

GTM : GTM is an unsupervised learning algorithm for modeling the probability density of data and finding a non-linear mapping of high-dimensional data in a low-dimension space. GTM is also known as a principled alternative to Self-Organizing Map (SOM) which does not have any density model, GTM defines an explicit probability density model based on Gaussian distribution [11] and seeks the best set of parameters associated with Gaussian mixtures by using Expectation-Maximization (EM) method.

MDS and GTM Interpolation : Both are an extension to the original MDS and GTM algorithm, designed to process much larger data points with sampling approaches. With minor trade-off of approximation, interpolation approach for MDS and GTM can visualize millions of data points with modest amount of computations and memory requirement. In [7], up to 2 million PubChem data points has been visualized by using parallelized MDS and GTM interpolation algorithms.

2.2 Remote data access

Chem2Bio2RDF is an integrated repository of chemogenomic and systems chemical biology data by aggregating over 20 publicly accessible datasets in Resource Description Framework (RDF) format and enable users to access them by using SPARQL which is a standard query language for RDF data and is a part of semantic web technology.

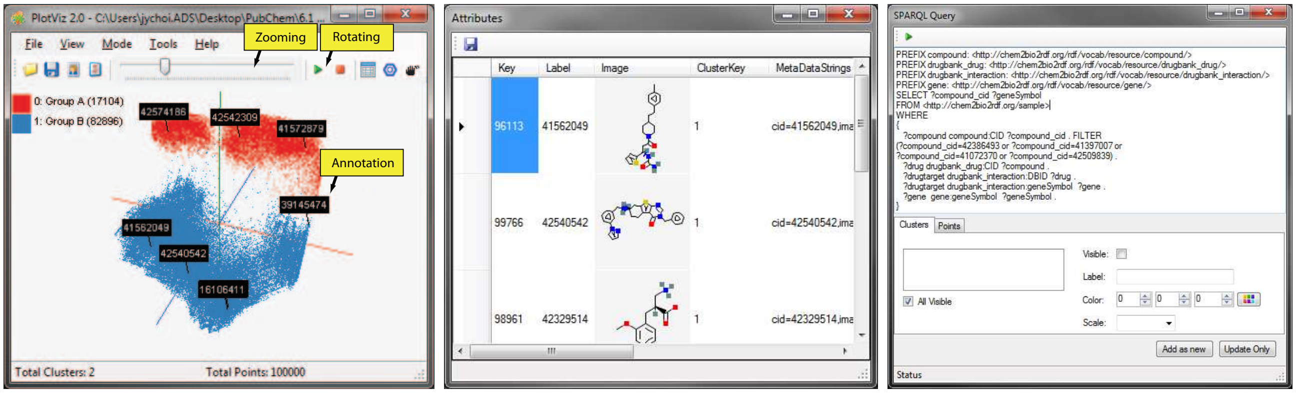
We can take an advantage of Chem2Bio2RDF system for accessing multiple data sources in an online manner by sending SPARQL query and parsing results in RDF format. In this way, users can visualize data with more information-rich context by mashing up with other data sources with ease.

3. THE PUBCHEMBROWSE SYSTEM

We have developed PubChemBrowse tool to visualize 3D data as an output of high-performance visualization algorithms as well as interact with external data sources via Chem2Bio2RDF system to provide rich and on-line information by utilizing semantic web interfaces.

Fig. 1 illustrates our PubChemBrowse architecture. It consists of mainly 3 components: i) high-performance visualization algorithms, such as parallel MDS and GTM, which generate 3D maps for large and high-dimensional data by utilizing parallel clustering infrastructure [7, 8], ii) an user-friendly 3D browsing interface to rotate, pan, and zoom in-/out data space and display annotations or meta-data, and iii) SPARQL query interface to access the remote data repository Chem2Bio2RDF for updating and adding new data points in an on-line manner. More details of each component are as follow.

High-performance visualization algorithms : Currently PubChemBrowse can visualize the outputs from our in-house parallel implementations of 4 main dimension reduction algorithms: MDS, GTM, MDS interpolation, and GTM interpolation. All of them can run in high-throughput clustering infrastructure by maximally utilizing multi-core



Key	Label	Image	ClusterKey	MetaDataStrings
96113	41562049		1	cid=41562049,ima
99766	42540542		1	cid=42540542,ima
98961	42329514		1	cid=42329514,ima

```
PREFIX compound: <http://chem2bio2rdf.org/rdf/vocab/resource/compound/>
PREFIX drugbank_drug: <http://chem2bio2rdf.org/rdf/vocab/resource/drugbank_drug/>
PREFIX drugbank_interaction: <http://chem2bio2rdf.org/rdf/vocab/resource/drugbank_interaction/>
PREFIX gene: <http://chem2bio2rdf.org/rdf/vocab/resource/gene/>
SELECT ?compound_cid ?geneSymbol
FROM <http://chem2bio2rdf.org/sample>
WHERE
{
  ?compound compound:CID ?compound_cid . FILTER
(?compound_cid=42386493 or ?compound_cid=41397007 or
?compound_cid=41072370 or ?compound_cid=42509839) .
  ?drug drugbank_drug:CID ?compound .
  ?drugtarget drugbank_interaction:DBID ?drug .
  ?drugtarget drugbank_interaction:geneSymbol  ?gene .
  ?gene  gene:geneSymbol  ?geneSymbol .
}
```

Visible: □
Color: 0 0 0
☑ All Visible

(a) A main window

(b) A sub-windows for meata-data browsing

(c) A sub-windows for SPARQL

Figure 2: Screenshots of PubChemBrowse.

environments.

3D data browser : As shown in Fig. 2a and Fig. 2b, one can visualize 3D data points and browse them in various ways: rotating, zooming, and viewing meta-data including chemical structures available from PubChem database.

SPARQL query : As shown in Fig. 2c, our system has an interface for users to compose SPARQL query and send to Chem2Bio2RDF system remotely by using REST protocol, which is a standard web service protocol. Then, the Chem2Bio2RDF system will fetch the query and return back results in XML format. Our system can parse the XML information and update or mash up the results in user's working plot.

4. APPLICATION STUDY

We have been used our PubChemBrowse in various chem-informatics researches aiming at exploring and discovering complicated compound-gene-disease relationships from large sets of data. In the following we will show a few examples of our results.

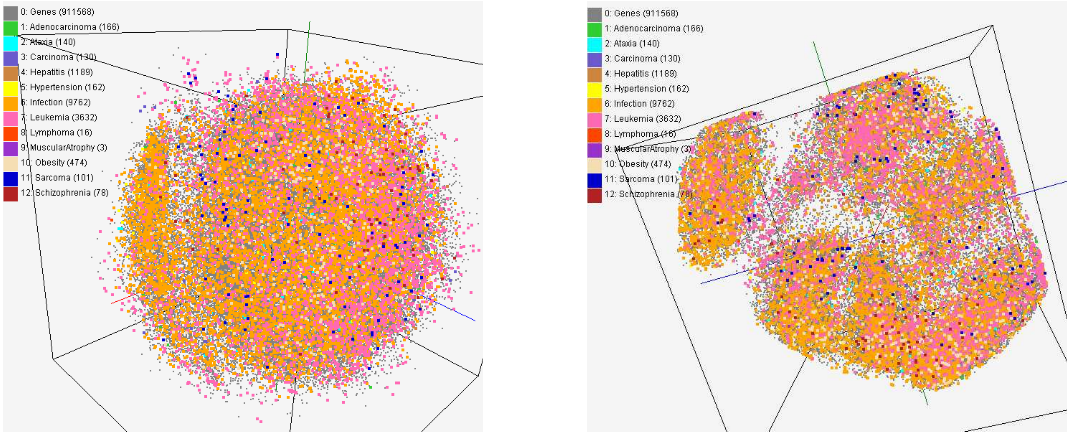
4.1 CTD data for gene-disease

In Fig. 3, we have visualized about 930,000 gene-related chemical compounds having 166 dimensions in PubChem database by using both MDS and GTM algorithms and labeled as different colors to display cause-and-effect associations between genes and diseases based on Comparative Toxicogenomics Database (CTD) dataset so that distances between points represent structural similarity and colors of points are labels based on CTD data.

By using our PubChemBrowse, one can easily identify points of interest by colors or select a group of points distinguished by structural distribution in 3D space. One can also easily browse the data by rotating, zooming, or panning the 3D space to search for more details. Also, updating the labels of points or adding new ones in the figure can be easily done by sending on-line SPARQL query to Chem2Bio2RDF system. With our tool, researchers can easily browse very large set of data with ease.

4.2 Chem2Bio2RDF

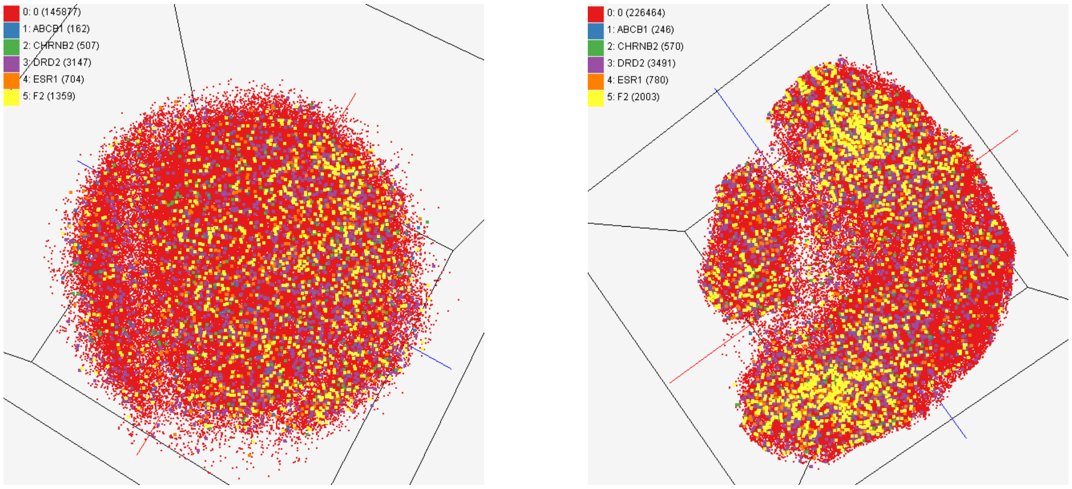
In Fig. 4, we have visualized 234,000 chemical compounds which may be related with a set of 5 genes of interest (ABCB1, CHRNB2, DRD2, ESR1, and F2) based on the dataset collected from major journal literatures which is also stored in Chem2Bio2RDF system. The purpose of this research is



(a) MDS

(b) GTM

Figure 3: Visualization of disease-gene relationships based on the data in Comparative Toxicogenomics Database (CTD). About 930,000 chemical compounds are visualized as a point in 3D space, computed by MDS (a) and GTM (b).



(a) MDS

(b) GTM

Figure 4: Visualization of compound-gene relationship by using multiple data sources in Chem2Bio2RDF system. About 234,000 points are visualized in 3D space. Each point represents a chemical compound in PubChem database and its configuration is computed by using MDS (a) and GTM (b).

to find chemical compounds related with genes reported in scientific literatures and verify their relationships through visualization.

As such, researchers need to gather information from different data sources and overlay them in one figure. Our system can help this repetitive procedure by supporting online multiple data access through SPARQL queries.

5. CONCLUSION

In this paper we discuss our design and implementation of PubChemBrowse, an integrated 3D data point visualization tool customized for browsing massive cheminformatics data. Our system is generated by our in-house high-performance visualization algorithms and interact with external Chem2Bio2RDF system by using SPARQL query language to access publicly accessible multiple bioinformatics databases. Further, we can embed new points and label various bioinformatics related datasets to assist on-going research results from data mining projects for drug-discovery.

As for the future work, we will integrate clustering service into our system to display entire compounds as a hierarchical structure so that users can easily zoom in and zoom out on a region with different levels of details. We will also continue to develop our system to integrate with more external systems transparently via semantic web interfaces.

6. REFERENCES

[1] B. Chen, D. Wild, Q. Zhu, Y. Ding, X. Dong, M. Sankaranarayanan, H. Wang, and Y Sun. Chem2bio2rdf: A linked open data portal for chemical biology. In *Future of the Web in Collaboratice Science (FWCS) 2010*, 2010.

[2] D. G. York. The sloan digital sky survey: Technical summary. *Astron.J.*, 120:1579–1587, 2000.

[3] RM Kuhn, D. Karolchik, AS Zweig, T. Wang, KE Smith, KR Rosenbloom, B. Rhead, BJ Raney, A. Pohl, M. Pheasant, et al. The UCSC genome browser database: update 2009. *Nucleic acids research*, 37(Database issue):D755, 2009.

[4] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al. The Ensembl genome database project. *Nucleic acids research*, 30(1):38, 2002.

[5] Judy Qiu, Jaliya Ekanayake, Thilina Gunarathne, Jong Youl Choi, Seung-Hee Bae, Yang Ruan, Saliya Ekanayake, Stephen Wu, Scott Beason, Geoffrey Fox, Mina Rho, and Haixu Tang. Data intensive computing for bioinformatics. In *Data Intensive Distributed Computing*. IGI Publishers, 2010.

[6] D.M. Maniyar, I.T. Nabney, B.S. Williams, and A. Sewing. Data visualization during the early stages of drug discovery. *Journal of chemical information and modeling*, 46(4):1806–1818, 2006.

[7] Seung-Hee Bae, Jong Youl Choi, Judy Qiu, and Geoffrey Fox. *Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation*. Pervasive Technology Institute at Indiana University, technical report edition, Jan 2010.

[8] Jong Youl Choi, Seung-Hee Bae, Xiaohong Qiu, and Geoffrey Fox. High performance dimension reduction and visualization for large high-dimensional data analysis. In *Proceedings of CCGRID 2010 (accepted)*, 2010.

[9] Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling*. Sage Publications Inc., Beverly Hills, CA, U.S.A., 1978.

[10] Ingwer Borg and Patrick J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, NY, U.S.A., 2005.

[11] C.M. Bishop, M. Svensén, and C.K.I. Williams. GTM: A principled alternative to the self-organizing map. *Advances in neural information processing systems*, pages 354–360, 1997.

[12] Jan de Leeuw. Applications of convex analysis to multidimensional scaling. *Recent Developments in Statistics*, pages 133–145, 1977.