Browsing Large Scale Cheminformatics Data with Dimension Reduction

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

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.

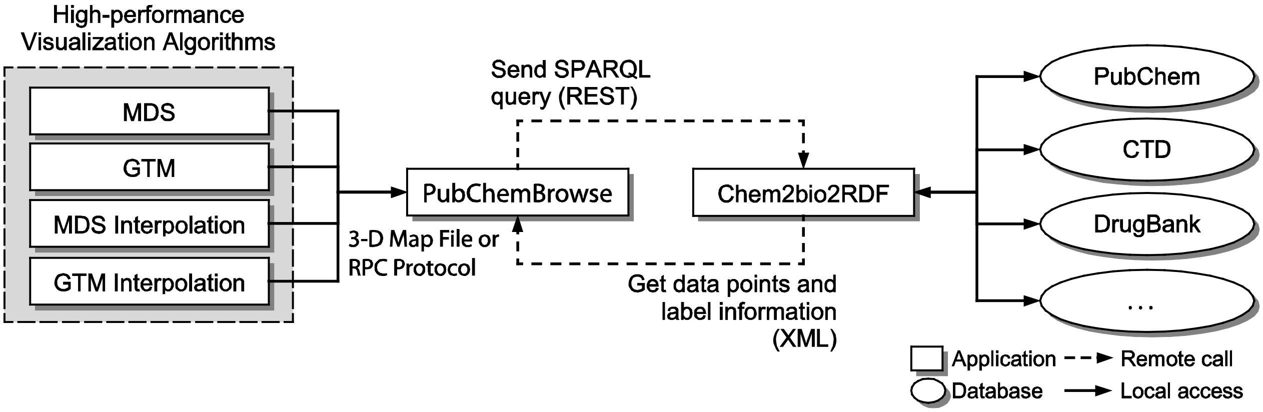


Figure 1. System architecture for PubChemBrowse

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 biological 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.

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 chemical gene information, the DrugBank database for drug target association 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 PubChem data 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 and 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. Although both GTM and SOM share the same object to find a map in low-dimensional user-defined space out of the data in high-dimensional space, GTM, however, finds a mapping based probability density model, which SOM lacks of. On the other hand, MDS tries to construct a mapping in target dimension with respect to the pairwise proximity information, mostly dissimilarity or distance. 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 between two points is approximated to the corresponding original proximity value. In other words, MDS is a non-linear optimization problem with respect to the mapping in the target dimension and the original proximity information. The STRESS [12] value is a well-known objective function of MDS, and Eq (1) represents STRESS function:

$$\sigma(X) = \sum_{i<j\leq N} w_{ij}\left(d_{ij}(X) - \delta_{ij}\right)^2 \quad (1)$$

Among many MDS solution, we use SMACOF algorithm which is based on Expectation Maximization (EM)-like iterative majorization method. For details of the SMACOF algorithm, please refer to [13].

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 to maximize the log-likelihood of Gaussian mixtures (2), as an objective function, by using Expectation-Maximization (EM) method.

$$\mathcal{L} = \underset{\{y_k\},\beta}{\operatorname{argmax}} \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left(-\frac{\beta}{2}\|x_n - y_k\|^2\right) \right\} \tag{2}$$

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**

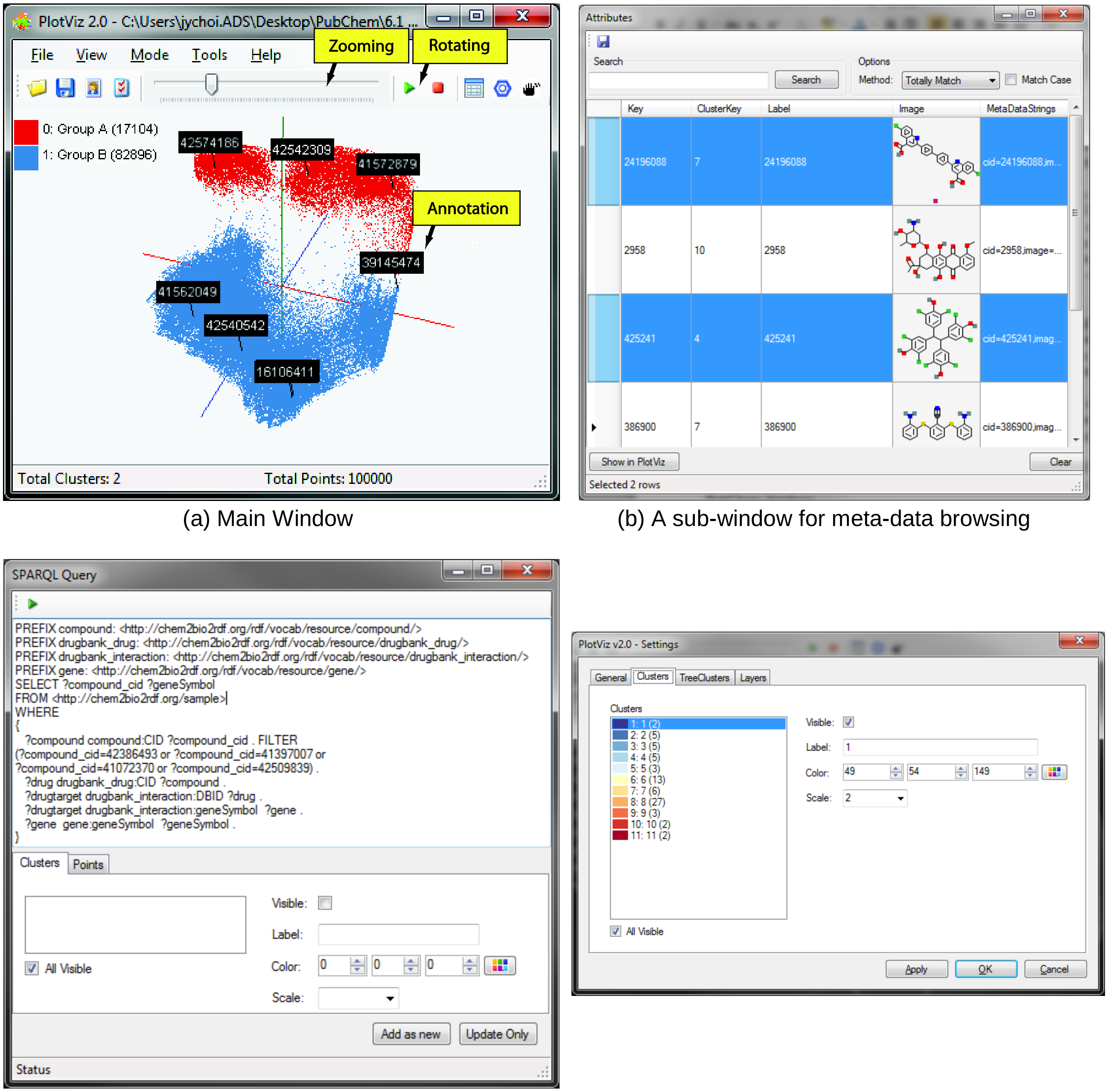
Chem2Bio2RDF is an integrated repository of chemogenomic and systems chemical biology data by aggregating over 20 publicly accessible datasets in Resource Description Framework (RDF) and enable users to access them by using SPARQL which is a standard query language for RDF data and is a part of semantic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

Figure 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.



Total Clusters: 2   Total Points: 100000

(a) Main Window

Method: Totally Match   ☐ Match Case

Key	ClusterKey	Label	Image	MetaDataStrings
24196088	7	24196088		cid=24196088,im...
2958	10	2958		cid=2958,image=...
425241	4	425241		cid=425241,imag...
386900	7	386900		cid=386900,imag...

Selected 2 rows

(b) A sub-window for meta-data browsing

```
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 .
}
```

☑ All Visible

Visible: ☐

Label:

Color: 0 0 0

Scale:

(c) A sub-window for SPARQL

Clusters

1: 1 (2)
2: 2 (5)
3: 3 (5)
4: 4 (5)
5: 5 (3)
6: 6 (13)
7: 7 (6)
8: 8 (27)
9: 9 (3)
10: 10 (2)
11: 11 (2)

☑ All Visible

Visible: ☑

Label: 1

Color: 49 54 149

Scale: 2

(d) A sub-window for cluster management

Figure 2. Screenshots of PubChemBrowse.

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 environments. Mostly those algorithms run in a batch mode but interpolation algorithms can run in an online mode too.

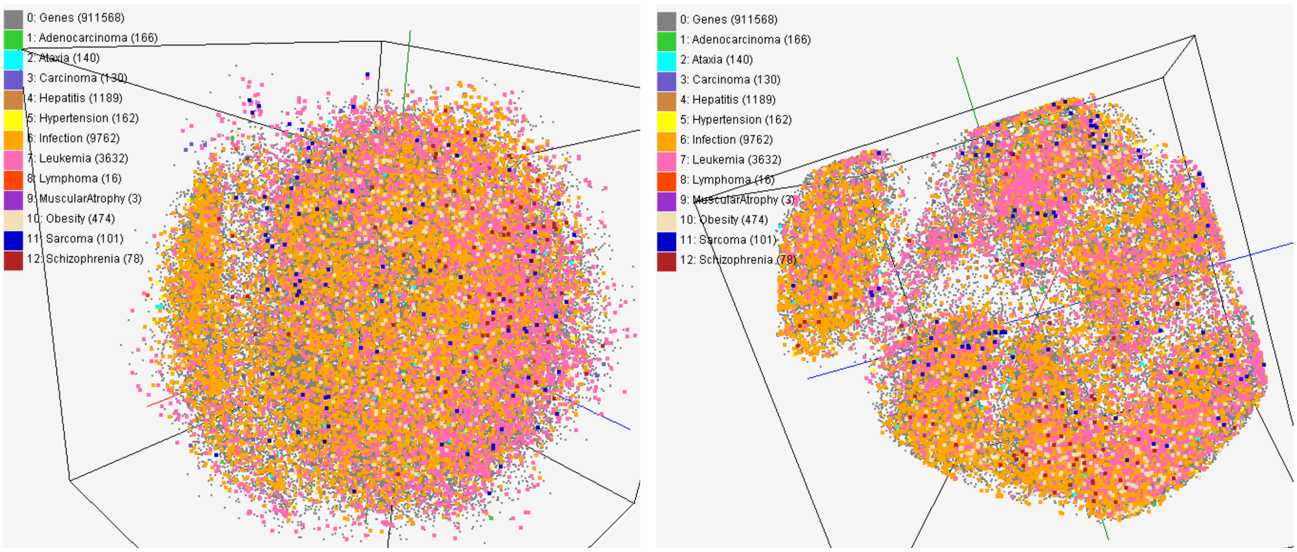
Currently PubChemBrowse can visualize the outputs from our in-house parallel implementations of 4 main dimension reduction algorithms: MDS, GTM, MDS interpolation, and GTM interpolation. All of them can run in a batch mode by utilizing high-throughput multi-core clustering infrastructure.

3D data browser : As shown in Figure 2 (a) and (b), 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 Figure 2 (c), 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. **Applications**

We have been used our PubChemBrowse in various cheminformatics 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.

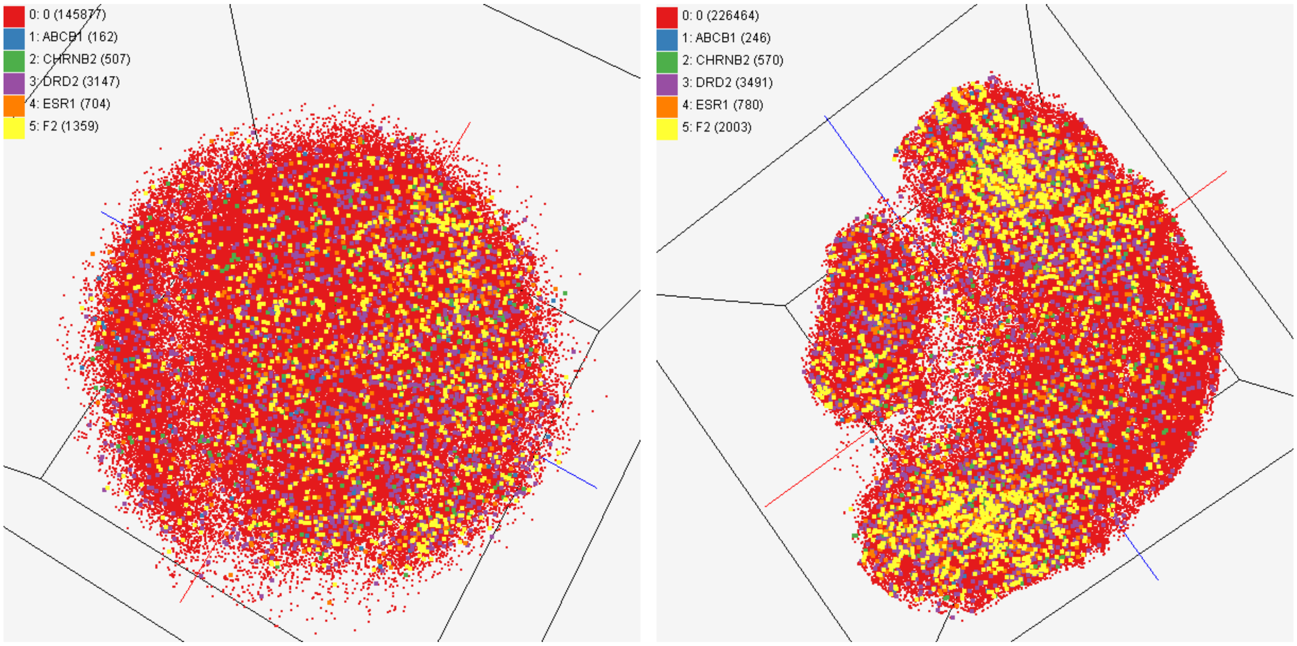


(a) MDS (b) GTM

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).

CTD data for gene-disease : In Figure 3, we have visualized about 930,000 biologically interesting 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 chemical 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.



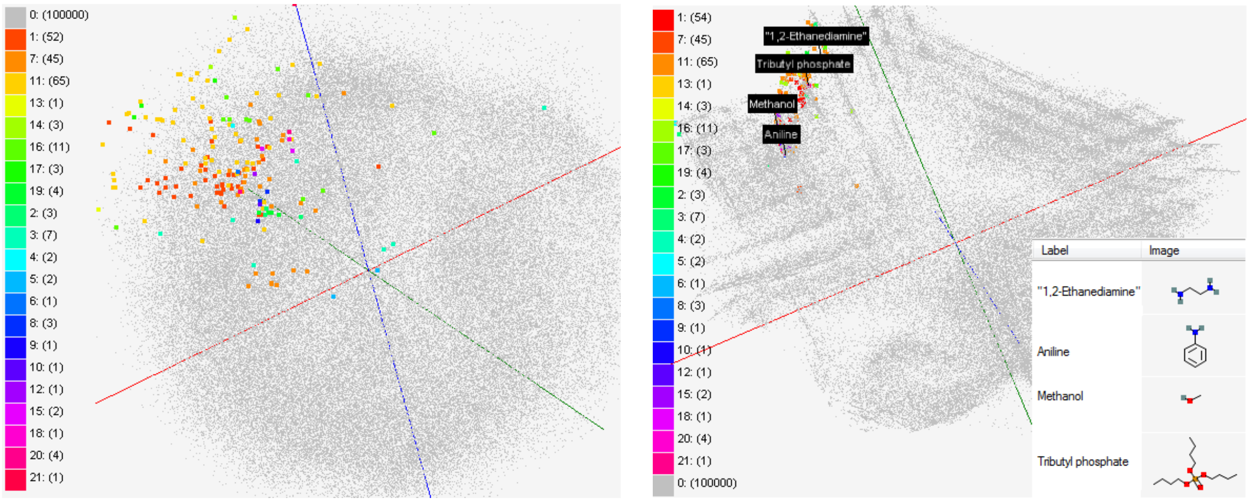
(a) MDS

(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).

Chem2Bio2RDF : In Figure 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 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.



(a) MDS (b) GTM

Figure 5. MDS (a) and GTM (b) visualization of 215 solvents (colored and/or labeled) with 100k PubChem dataset (colored in grey) to navigate chemical space.

Solvent screening : In Figure 5, we have visualized 215 solvents used in a pharmaceutical pre-screening process [14] along with 100 thousand chemical compounds. The result shows that our tool can clearly separate solvents from other chemicals based on the structural characteristics and users can navigate the large chemical space with visualization.

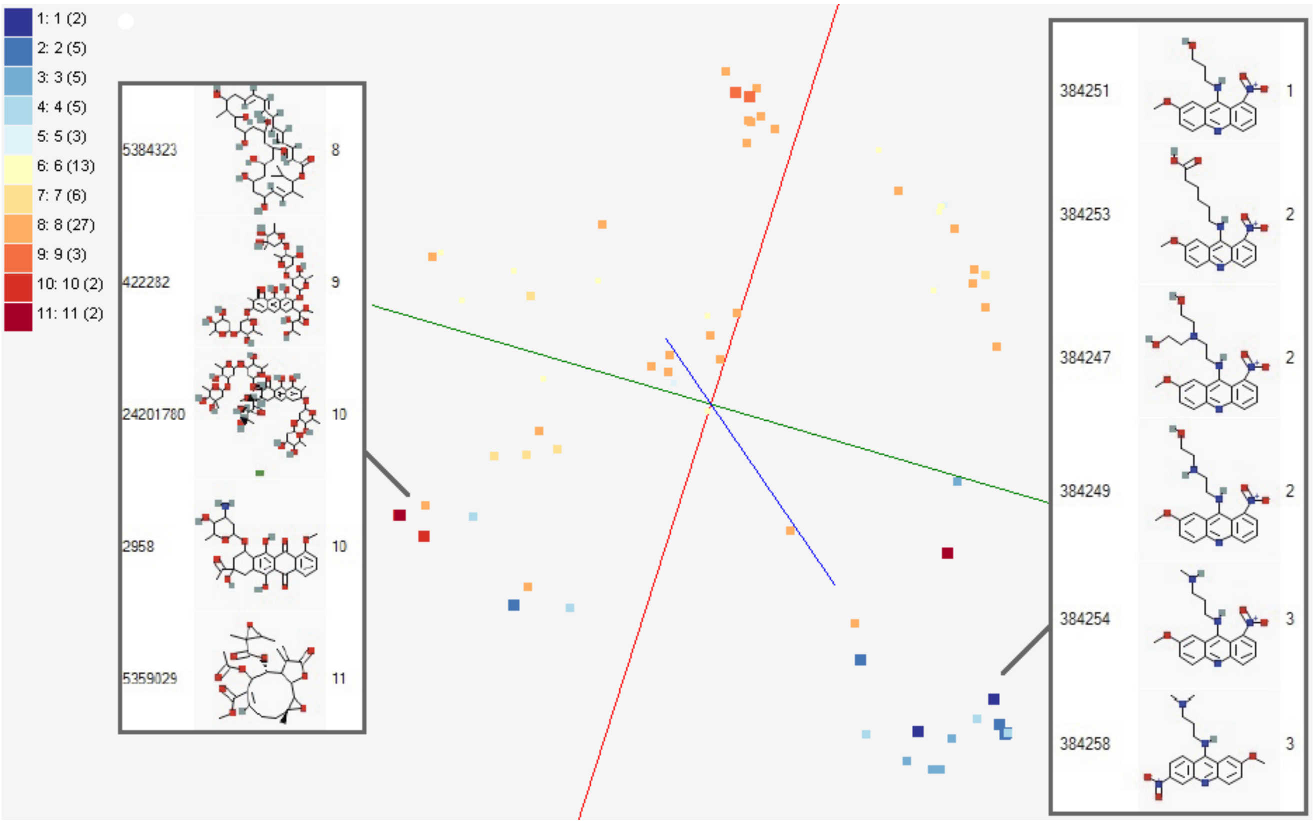


Figure 6. GTM visualization of chemical comActivity Cliffs

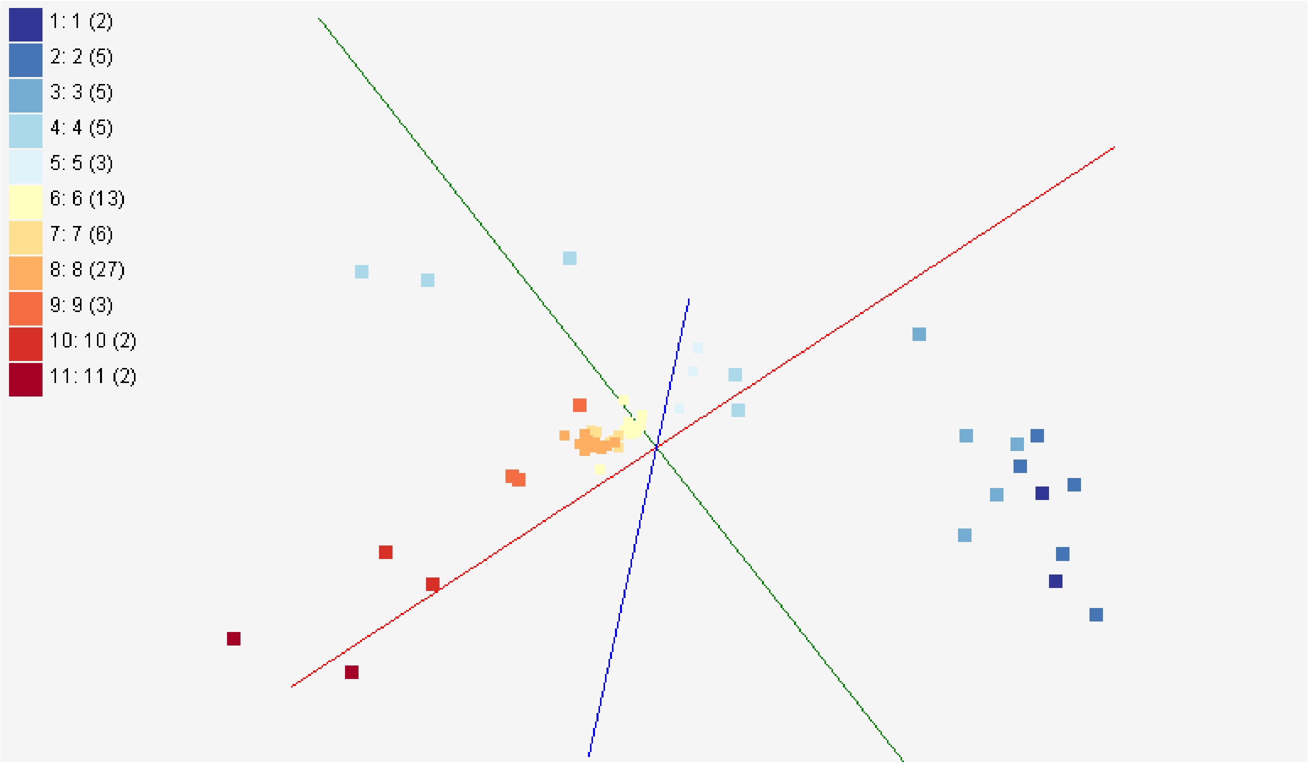


Figure 7. MDS mapping for Activity Cliffs, based on SALI pairwise matrix

Activity Cliffs : This is a research project to study about activity cliffs with SALI index [15] in PubChem Bioassays for applications in polypharmacology and chemogenomics. In this study, visualization of bioassay activities and SALI values in 3D chemical space is necessary. Figure 6 shows an example of visualizing activities of 73 chemical compounds in 3D chemical space representing structural similarity, processed by GTM. Figure 7 also describes the same example of visualizing activities of Figure 6 based on pairwise SALI indices [15] of those points instead of chemical feature vectors in 3D SALI space. Note that SALI space mapping can be only generated by MDS since MDS can generate a mapping in target dimension with pairwise proximity information.

Other Life Science Applications: Although our PubChemBrowse system is mainly developed to provide special functions for cheminformatics applications, we can process and visualize various other life science data by using our PubChemBrowse system, as well. To show those extensions, we present two different life science applications. First, the visualization of health data to study child-obesity is represented in the PubChemBrowse system. The health data we processed is the children's health property data, such as Body Mass Index (BMI), blood pressures, and so on, as well as environmental factors geographically linked in the dataset, like greenness of neighborhoods, incomes of households, etc. We generate pairwise dissimilarity matrix of 10 thousands of points based on 98-dimensional normalized feature vector, and a 3D mapping of that data constructed by MDS as shown in Figure 8. Another example shown here is a biological sequence data. We produce 3D mapping of 30 thousands metagenomics sequence data based on pairwise sequence alignment scores by a well-known sequence alignment algorithm called Smith-Watermann [16] algorithm. Figure 9 depicts the 3D mapping results of 30 thousands sequence data with respect to the metagenomics study, configured by MDS algorithm.

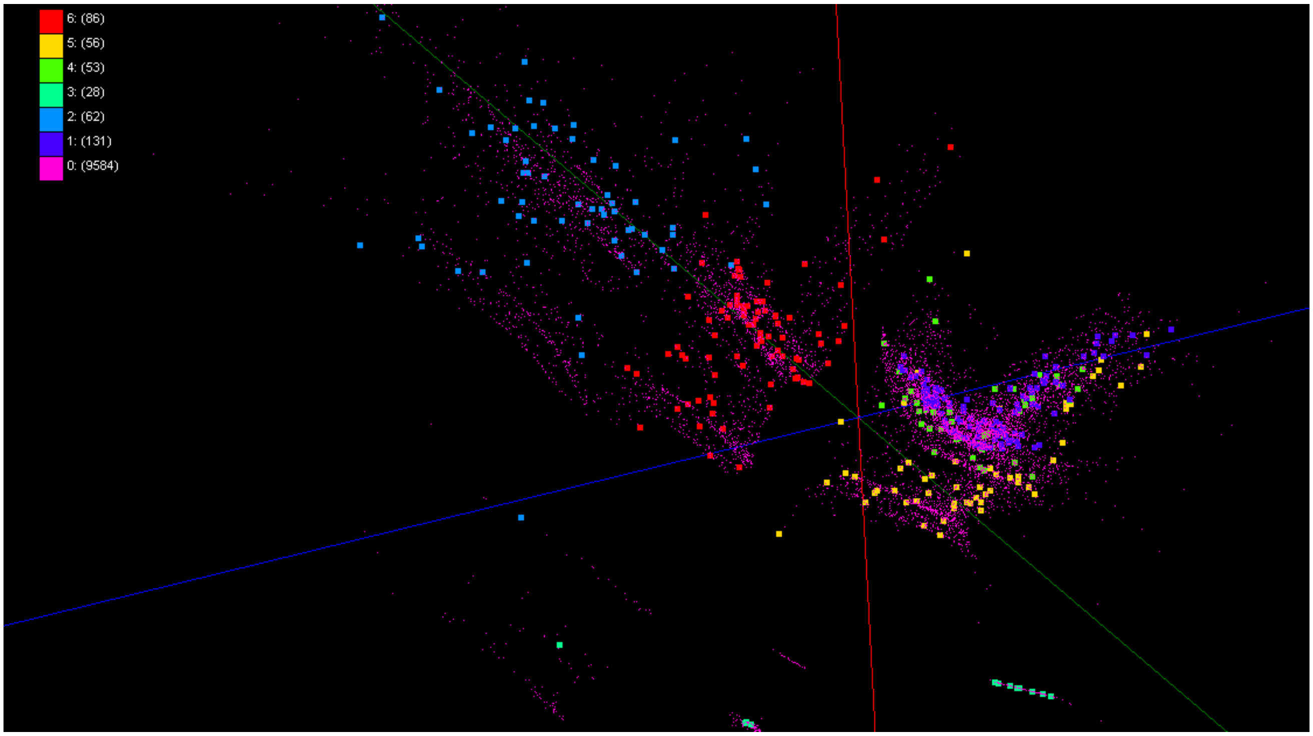


Figure 8. 3D mapping of 10 thousands children obesity data

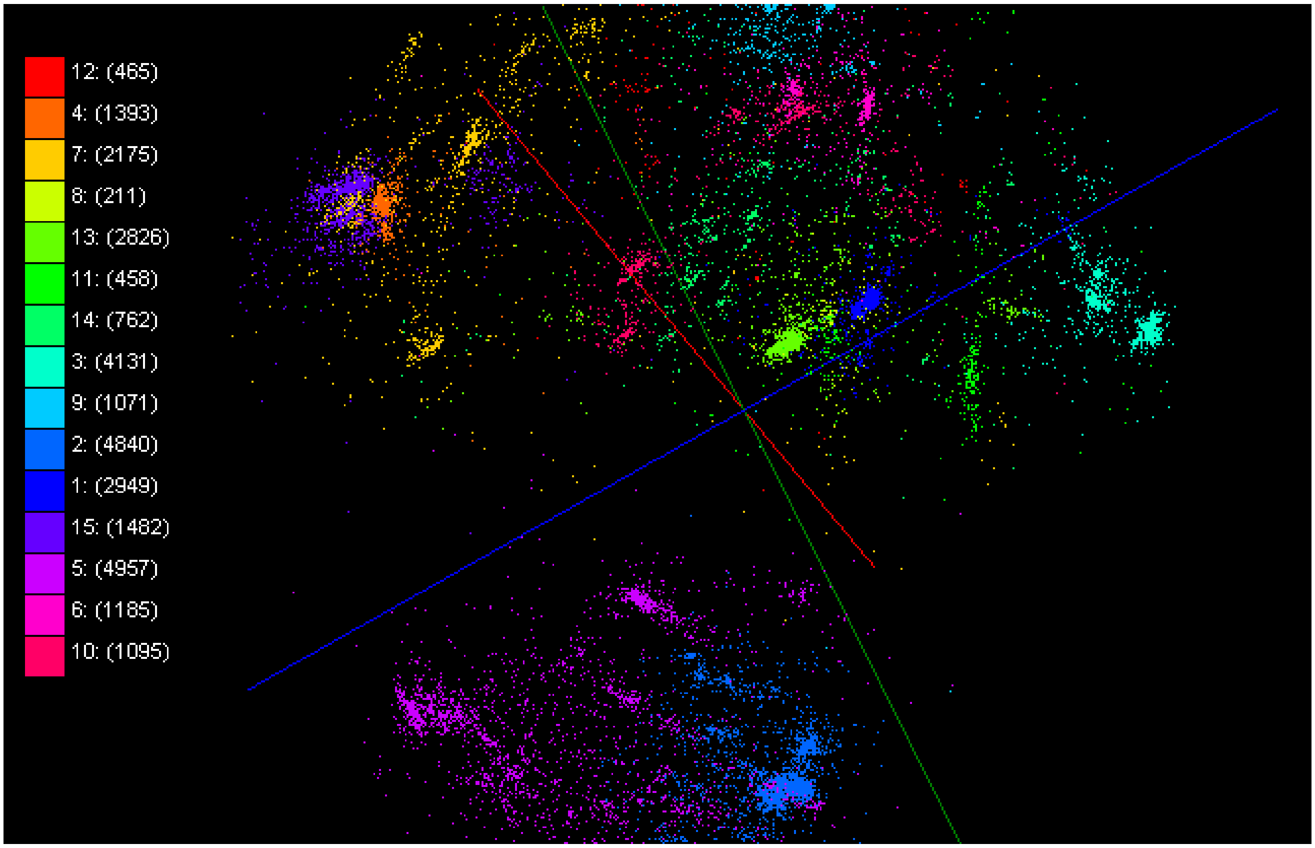


Figure 9. mapping result of 30 thousands metagenomics sequences

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. **Acknowledgement**

We appreciate Microsoft for their technical support. This work was made possible using the computing use grant provided by Amazon Web Services which is titled "Proof of concepts linking FutureGrid users to AWS". This work is partially funded by Microsoft "CRMC" grant and NIH Grant Number RC2HG005806-02. This document was developed with support from the National Science Foundation (NSF) under Grant No. 0910812 to Indiana University for "FutureGrid: An Experimental, High-Performance Grid Test-bed." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

7. **References**

[1] B. Chen*, et al.*, "Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data," *BMC Bioinformatics,* vol. 11, p. 255, 2010.
[2] D. G. York*, et al.*, "The Sloan Digital Sky Survey: Technical summary," *Astronomical Journal,* vol. 120, pp. 1579-1587, Sep 2000.
[3] R. M. Kuhn*, et al.*, "The UCSC Genome Browser Database: update 2009," *Nucleic Acids Research,* vol. 37, pp. D755-D761, Jan 2009.
[4] T. Hubbard*, et al.*, "The Ensembl genome database project," *Nucleic Acids Research,* vol. 30, pp. 38-41, Jan 1 2002.
[5] Judy Qiu*, et al.*, "Data Intensive Computing for Bioinformatics," *Technical Report,* December 29 2009.
[6] D. M. Maniyar*, et al.*, "Data visualization during the early stages of drug discovery," *Journal of Chemical Information and Modeling,* vol. 46, pp. 1806-1818, Jul 24 2006.
[7] S.-H. Bae*, et al.*, "Dimension Reduction and Visualization of Large High-dimensional Data via Interpolation," presented at the HPDC'10, Chicago, Illinois USA, 2010.
[8] J. Y. Choi*, et al.*, "High Performance Dimension Reduction and Visualization for Large High-dimensional Data Analysis," presented at the The 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid2010), Melbourne, Australia, 2010.

[9] J. B. Kruskal and M. Wish, *Multidimensional Scaling*: Sage Publications Inc., 1978.

[10] I. Borg and P. J. Groenen, *Modern Multidimensional Scaling: Theory and Applications*: Springer, 2005.

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

[12] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. *29*, pp. 1-27, 1964.

[13] J. de Leeuw, "Applications of convex analysis to multidimensional scaling," *Recent Developments in Statistics,* pp. 133-145, 1977.

[14] M. Alleso*, et al.*, "Solvent diversity in polymorph screening," *Journal of Pharmaceutical Sciences,* vol. 97, pp. 2145-2159, Jun 2008.

[15] R. Guha and J. H. Van Drie, "Structure-activity landscape index: Identifying and quantifying activity cliffs," *Journal of Chemical Information and Modeling,* vol. 48, pp. 646-658, Mar 2008.

[16] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology,* vol. *147*, pp. 195-197, 1981.