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

Bin Chen, David Wild
School of Informatics and Computing
Indiana University

SALSA
http://salsahpc.indiana.edu

INDIANA UNIVERSITY
PERVERSIVE TECHNOLOGY INSTITUTE
Drug Discovery

A pipeline process with various stages
- Many screening processes to filter out large number of chemical compounds
- Empirical science

Nature Reviews Drug Discovery 1, 515–528 (1 July 2002)
Data Mining for Drug Discovery

- Modern drug discovery
  - Not an empirical science anymore
  - Data intensive science
  - Use of *in silico* screening methods

- Numerous open databases
  - NIH founded PubChem
  - DrugBank, Comparative Toxicogenomics Database (CTD), …

(Cresset’s FieldAlign, Nature, 2007)
Motivation

▸ To browse large and high-dimensional data
  ➤ Data visualization by dimension reduction
  ➤ High-performance dimension reduction algorithms

▸ To utilize many open (value-added) data
  ➤ Combine data from different sources in one place
  ➤ A uniform interface

▸ A light-weight easy-to-use visualization tool
  ➤ A desktop client with an user-friendly UI
  ➤ Easy to use high-performance computing resources
PubChemBrowse System

- Visualization Algorithms
- Parallel dimension reduction algorithms

PubChemBrowse

- Light-weight client

3-D Map File

SPARQL query

Meta data

Chem2Bio2RDF

- Aggregated public databases
- DrugBank
- CTD
- QSAR
- PubChem
Visualization by Dimension Reduction

- Simplify data
- Preserve the original data’s information as much as possible in lower dimension
- Explore enormous data in 3D
Visualization Algorithms

▸ Compute- and memory-intensive algorithms
  – High-performance is not for free
  – Commodity hardware is not capable of processing large data

▸ In-house high-performance visualization algorithms
  – Parallel GTM (Generative Topographic Mapping)
  – Parallel MDS (Multi-dimensional Scaling)
  – Further performance improvement by interpolation extensions to GTM and MDS
## GTM vs. MDS

<table>
<thead>
<tr>
<th>Purpose</th>
<th>GTM</th>
<th>MDS (SMACOF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-linear dimension reduction</td>
<td>• Find an optimal configuration in a lower-dimension</td>
<td></td>
</tr>
<tr>
<td>• Iterative optimization method</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input</th>
<th>Vector-based data</th>
<th>Non-vector (Pairwise similarity matrix)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Objective Function</th>
<th>Maximize Log-Likelihood</th>
<th>Minimize STRESS or SSTRESS</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Complexity</th>
<th>O(KN) (K &lt;&lt; N)</th>
<th>O(N²)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Optimization Method</th>
<th>EM</th>
<th>Iterative Majorization (EM-like)</th>
</tr>
</thead>
</table>

- **GTM**
  - **Purpose**: Non-linear dimension reduction, find an optimal configuration in a lower-dimension, iterative optimization method
  - **Input**: Vector-based data
  - **Objective Function**: Maximize Log-Likelihood
  - **Complexity**: O(KN) (K << N)
  - **Optimization Method**: EM

- **MDS (SMACOF)**
  - **Purpose**: Minimize STRESS or SSTRESS
  - **Input**: Non-vector (Pairwise similarity matrix)
  - **Objective Function**: Minimize STRESS or SSTRESS
  - **Complexity**: O(N²)
  - **Optimization Method**: Iterative Majorization (EM-like)
Parallel GTM

- Finding K clusters for N data points
  - Relationship is a bipartite graph (bi-graph)
  - Represented by K-by-N matrix (K << N)
- Decomposition for P-by-Q compute grid
  - Reduce memory requirement by 1/PQ

Example:
A 8-byte double precision matrix for N=100K and K=8K requires 6.4GB
Parallel MDS

- Decomposition for P-by-Q compute grid
  - Reduce memory requirement by 1/PQ

Example:
A 8-byte double precision matrix for N=100K requires 80GB
Interpolation extension to GTM/MDS

- Full data processing by GTM or MDS is computing- and memory-intensive

- Two step procedure
  - *Training*: training by M samples out of N data
  - *Interpolation*: remaining (N-M) out-of-samples are approximated without training
PubChemBrowse

- Light-weight desktop client
- Interactive user interface
- Display 3D embedding and meta data
Chem2Bio2RDF

- Value-added database of databases
  - Aggregate over 20 public databases (PubChem, CTD, DrugBank, …)
  - Stored in RDF (Resource Description Framework)
  - Support SPARQL query language

- SPARQL query
  - A W3C standard query language for RDF

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?email
WHERE {
  ?person a foaf:Person.
  ?person foaf:name ?name.
}
```
Query Interface
CTD data for gene-disease

**PubChem data with CTD visualization by using MDS (left) and GTM (right)**

About 930,000 chemical compounds are visualized as a point in 3D space, annotated by the related genes in Comparative Toxicogenomics Database (CTD)
Chemical compounds shown in literatures, visualized by MDS (left) and GTM (right)
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.
Solvent screening

Visualizing 215 solvents
215 solvents (colored and labeled) are embedded with 100,000 chemical compounds (colored in grey) in PubChem database
Conclusion

▸ Modern drug discovery
  – Data intensive process
  – High-throughput *in silico* screening methods

▸ PubChemBrowse
  – A light-weight desktop client
  – Parallel high-performance visualization algorithms
  – Access multiple databases via Chem2Bio2RDF by using an uniform interface, SPARQL query
Thank you

Question?

Email me at jychoi@cs.indiana.edu