2009 GMOD Meeting

Dhileep Sivam & Isabelle Phan

Seattle Biomedical Research Institute
Seattle Biomedical Research Institute (SBRI)

- Founded in 1976
- About 250 full-time staff
- Focus on infectious disease
- 13 Labs
- Strong ties to the University of Washington
- Bioinformatics Core
How we first came to use Chado

LmjF V4.0  LmjF V5.2  LinJ V2.0  LinJ V3.0  LinJ V4.0

LmjF Probe Set  LinJ Probe Set

Mapping  Mapping  Mapping

↓  ↓  ↓

Result Set  Result Set  Result Set
Microarray Project

Nimblegen Data → Parsers → Chado

Analysis Tools
- Normalization
- Scaling
- Feature-level aggregation
- Remapping
- Visualization
Use Case: SSGCID

*Seattle Structural Genomics Center for Infectious Disease*

**Project Aim**

- 3D Protein Structure
- NIAID Emerging and re-emerging priority pathogens
- Structures will serve as a starting point for drug development
- Multi-center

**Diagram:**

- Bioinformatic Screening
- Gene Cloning & Expression
- Protein Crystallization
- Structure Determination
- Vaccine Targets!
SSGCID

Bioinformatic Screening

Gene Cloning & Expression

Protein Crystallization

Structure Determination

Vaccine Targets!

SSGCID structure determination pipeline

<table>
<thead>
<tr>
<th>Target selected</th>
<th>Clone 1</th>
<th>Clone 2</th>
<th>Clone 3</th>
<th>Clone 4</th>
<th>Clone 5</th>
<th>Clone 6</th>
<th>Clone 7</th>
<th>Clone 8</th>
<th>Clone 9</th>
<th>Clone 10</th>
<th>Clone 11</th>
<th>Clone 12</th>
<th>Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloned</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Expressed</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Soluble</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Purified</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RNAi/SiRNA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Diffraction/NMR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Stable</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Usable data</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Structure solved</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Abbreviations:
- NAD: NADP
- BGC: Biosynthetic Gene Cluster
- SiRNA: Small Interfering RNA
- MPR: Medium Proportionality Ratio
- CIP: Closest Intracellular Pathway
Things that have come up...

- Complexity of querying BLAST results
- Complexity of querying microarray data
- “Grouping of Genes”
- Gene Models

Materialized Views

DBXrefs

Simplest Possible Model
Sequence data management at SBRI

Proteomics
Microarray
*Structural genomics*

Automated analysis pipeline

Warehouse

curation

Data access
Chado + GUS: why do we need both?

• Chado
  – Collaboration with IGS
  – Annotation tools: Manatee (apollo), Ergatis
    • Internal data production

• Gus
  – Collaboration with UPenn
  – Web front end
    • External data access
Proteomics
Microarray
Structural genomics

Manatee
Manual annotation

Chado

Ergatis
Analysis pipeline

Sequence data management at SBRI

GUS

GUS WDK
Chado2GUS: Lost in translation

- **Chado**
  - Denormalized schema
    - Polymorphism
  - Mysql (IGS Chado)

- **GUS**
  - Normalized schema
    - Subclassing
  - Postgres port from Oracle
Picking the best of two worlds

- **Chado**
  - Biological data model
  - Flexibility

- **GUS**
  - Software engineering
  - Flexibility
The future?

• **SQL-free data production**
  – Instead of custom wrappers over raw SQL:
    • ORMs: Chado Hibernate, ActiveRecords
    • Unified object model

• **RDBMS-free data mining**
  – Instead of GUS predefined query + set combination
    • Biomart + Galaxy
    • RDF + triple store + sparql (object store + Lucene)