Big Data in Drug Discovery

David J. Wild
Assistant Professor & Director, Cheminformatics Program
Indiana University School of Informatics and Computing
djwild@indiana.edu - http://djwild.info
Epochs in drug discovery

**Empirical – up until 1960’s**
- 754 First pharmacy opened in Baghdad
- Late 1800’s – major pharmaceutical companies, mass production
- 1900-1960 – major discoveries (insulin, penicillin, the pill ...)

**Rational – 1960’s to 1990’s**
- Designing molecules to target protein active sites – “lock and key”
- Computational Drug Discovery
- Biggest success HIV (RT, protease inhibitors)

**Big Experiment – 1990’s to 2000’s**
- High throughput screening
- Microarray Assays
- Gene Sequencing and Human Genome Project

**Big Data – 2010’s onwards**
- Informatics-driven drug discovery
- Accepting the body is amazingly complex and we don’t understand it well
- Everything is connected
The metabolic pathways of a single cell
Big Data in the public domain

- There is now an incredibly rich resource of public information relating compounds, targets, genes, pathways, and diseases. Just for starters there is in the public domain information on:
  - 69 million compounds and 449,392 bioassays (PubChem)
  - 4,763 drugs (DrugBank)
  - 9 million protein sequences (SwissProt) and 58,000 3D structures (PDB)
  - 14 million human nucleotide sequences (EMBL)
  - 19 million life science publications - 800,000 new each year (PubMed)
  - Multitude of other sets (drugs, toxicogenomics, chemogenomics, SAR, ...)

- Even more important are the relationships between these entities. For example a chemical compound can be linked to a gene or a protein target in a multitude of ways:
  - Biological assay with percent inhibition, IC50, etc
  - Crystal structure of ligand/protein complex
  - Co-occurrence in a paper abstract
  - Computational experiment (docking, predictive model)
  - Statistical relationship
  - System association (e.g. involved in same pathways cellular processes)
PubChem growth since 2005

PubChem Substance Size 2005-2010

PubChem Bioassays 2005-2010

Addition of ChemSpider

Addition of ChEMBL

Large amount of data and links for each compound
Proteins & Genes

You are a big pile of data too!
Large-scale predictive modeling adds even more data

Range of ROCV values from different classes of BioAssay data set.

Range of ROCV values from three different classes of BioAssay data set for original models and models built with additional inactive compounds (“improved”).

Informatics-based drug discovery

Predicting new molecular targets for known drugs. Nature 462, 175-181 (12 November 2009)
“Systems chemical biology” and chemogenomics

Commentary

doi:10.1038/ncchembio0807-447

**Systems chemical biology**

Tudor I Oprea, Alexander Tropsha, Jean-Loup Faulon & Mark D Rintoul

1. Tudor I. Oprea is in the Division of Biocomputing, MSC11 6145, University of New Mexico School of Medicine, 2703 Frontier NE, Albuquerque, New Mexico 87131, USA. e-mail: toprea@salud.unm.edu
2. Alexander Tropsha is in the Laboratory for Molecular Modeling, C3 # 7360 Beard Hall, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA.
3. Jean-Loup Faulon and Mark D. Rintoul are at Sandia National Laboratories, PO Box 5800, Albuquerque, New Mexico 87185, USA.

The increasing availability of data related to genes, proteins and their modulation by small molecules has provided a vast amount of biological information leading to the emergence of systems biology and the broad use of simulation tools for data analysis. However, there is a critical need to develop cheminformatics tools that can integrate chemical knowledge with these biological databases and simulation approaches, with the goal of creating systems chemical biology.
Recent enabling technologies for SCB / Chemogenomics

Cloud computing allows processing and data mining on a vast scale

Integrative cheminformatics & bioinformatics connects compounds, targets genes, pathways, diseases and side effects

Semantic technologies and complex systems tools allow seamless integration and human-scale data mining

Health informatics (PHRs and EHRs) allows integration of the molecular and patient models (QP)

Analysis
Visualization, projection, data mining, hypothesis generation, network tools

Integration
RDF, XML, Triple Stores Ontologies, SPARQL, Graph algorithms

Access
Web Services, RPC Information extraction
ChemBioGrid.org: Web service infrastructure for cheminformatics


The Semantic Web – meaning & relationships

- Resource (subject)
- Property (predicate)
- Value (object)

- Drug
  - name: Lipitor
  - company: Pfizer

<RDF>
  <Description about="http://chem2bio2rdf.org/drug/DB01076"/>
  <name>Lipitor"author>
  <company>Pfizer"company>
</Description>
</RDF>

User interface and applications
- Trust
- Proof
- Unifying logic

Querying: SPARQL
- Ontologies: OWL
- Rules: RIF/SWRL

Cryptography
- Taxonomies: RDFS
- Data interchange: RDF
- Syntax: XML

Identifiers: URI
- Character set: UNICODE

Chem2Bio2RDF – RDF integration & SPARQL querying


Over 80 million triples!

Chem2Bio2RDF Datasets


Chem2Bio2RDF Relationships

Finding multi-target inhibitors of MAPK pathway with a SPARQL query
Finding compounds with similar polypharmacology using SPARQL
Projecting queries into chemical space

- GTM / MDS projection and embedding of all PubChem using clouds
- Plotting and embedding unknown compounds with SCB property labels
- Dynamic querying and projection into chemical space
Projecting queries into chemical space


“Doppler Radar Plot” – Kinase Specificity

ACM Symposium for High Performance Distributed Computing Jun 21-25, 2010, Chicago IL
“Doppler Radar Plot” – Kinase Specificity
Chem2Bio2RDF Dashboard: finding paths

- Origin: sider
- Terminus: kegg_pathways

Linked Paths:
1. sider --> chemogenomics_hub --> bindingdb_ligand --> bindingdb_prot
2. sider --> chemogenomics_hub --> ctld --> gene --> gene2uniprot
3. sider --> chemogenomics_hub --> drugbank_dru --> drugbank_target
4. sider --> chemogenomics_hub --> matador --> unprot_hub --> kegg
5. sider --> chemogenomics_hub --> pubchem_bioassy --> gi --> gi2uniprot
6. sider --> chemogenomics_hub --> qsar --> gene --> gene2uniprot
7. sider --> chemogenomics_hub --> ttd_dru --> ttd_target --> unprot

Generate Links

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>sider_call</td>
<td>sider_cid</td>
</tr>
<tr>
<td>DBID</td>
<td>drugbank_drug_CID</td>
</tr>
<tr>
<td>DBID</td>
<td>drugbank_interaction_geneSymbol</td>
</tr>
<tr>
<td>DBID</td>
<td>gene2uniprot_geneSymbol</td>
</tr>
<tr>
<td>DBID</td>
<td>gene2uniprot_unprot</td>
</tr>
<tr>
<td>DBID</td>
<td>pubchem_pathway_id</td>
</tr>
</tbody>
</table>
Pathfinder

http://ella.slis.indiana.edu/~yuysun/flex/pathfinder.html
Dynamic exploration with clouds and Cytoscape

Virtuoso runs Chem2Bio2RDF queries on the cloud

Cytoscape plugins give access to Chem2Bio2RDF, LPG and chemical structure visualization

Dynamic exploration in Cytoscape
**Hydrocortisone – Dexamethasone links**

- **Fig.** Use Case 1. Network diagram of the paths obtained between Hydrocortisone and Dexamethasone using ChemBioScape. Drugbank interaction contains information about every drug’s target. In this case, DB00741 and DB01234 share common targets through several different Drugbank interaction ID's.
**Fig.** Use case 2. Tolcapone and Entacapone are connected to each other through drugbank interaction 2348 and 1962. Also, the two drugs appear in PubMed articles 8119326 and 8223912 via their CID (Compound ID).
Isoniazid and Ethionamide – replicate paper results

Doxorubicin is an anthracyclin antibiotic that inhibits DNA synthesis and disrupts DNA replication. It is used in the treatment of various types of cancer, including breast and lung cancer.

### Compound Knowledge Space

<table>
<thead>
<tr>
<th>Compound</th>
<th>Probability</th>
<th>Cell Line(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doxorubicin</td>
<td>1.000</td>
<td>NCI-H69, NCI-H727</td>
</tr>
<tr>
<td>Colorectal</td>
<td>0.69</td>
<td>NCI-H69, NCI-H727</td>
</tr>
<tr>
<td>Breast</td>
<td>0.63</td>
<td>NCI-H69, NCI-H727</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.59</td>
<td>NCI-H69, NCI-H727</td>
</tr>
</tbody>
</table>

### Toxic Hazards

- **High Risk (Class III)** according to Cramer rules.

### Conclusion

Doxorubicin is a potent and effective drug, but its use is often limited by dose-limiting toxicities. Understanding the compound knowledge space can help in optimizing treatment protocols and minimizing side effects.
Chronic myeloid leukemia (CML) with P190BCR-ABL - analysis of characteristics, outcomes and prognostic significance.


Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, United States.

The most common BCR-ABL transcripts in CML are e13a2(b2a2) and e14a2(b3a2). Other transcripts like e1a2 are rare and their outcome with TKI therapy is undefined. We analyzed 1292 CML patients and identified 14 with only e1a2 transcripts, 9 in chronic (CP), 1 accelerated (AP), and 4 blast phase (BP). Of the CP 4 achieved CHR, 2 CCR, 2 PCR, and 1 did not respond to imatinib. Five patients progressed to myeloid BP (3), lymphoid BP (1), or AP (1). The AP patient received various TKIs sequentially and achieved only CHR. BP patients received Hyper-CVAD-Imatinib/dasatinib or idarubicin+Ara-C; 2 did not respond, 1 had CCR, and 1 short-lasting CMR. Overall, cytogenetic responses lasted 3–18 months, only 2 achieved MMR on TKI. P190(BCR-ABL) CML is rare and is associated with an inferior outcome to therapy with TKI. These patients need to be identified as high-risk patients.

PMID: 19631057 [PubMed - as supplied by publisher]
WENDI v2.0 - Automated reasoning with RDF

- Simple OWL ontology for relationships
- Large RDF network expands out from Query
- RDF inference engines applied & results filtered / prioritized
Semantic text mining of journal articles

Table 1. List of Interaction Keywords (Verbs Only) Used for Identifying Interactions

<table>
<thead>
<tr>
<th>list of interaction keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>accelerate</td>
</tr>
<tr>
<td>acetylate</td>
</tr>
<tr>
<td>activate</td>
</tr>
<tr>
<td>affect</td>
</tr>
<tr>
<td>associate</td>
</tr>
<tr>
<td>bind</td>
</tr>
<tr>
<td>block</td>
</tr>
<tr>
<td>carboxylate</td>
</tr>
<tr>
<td>catalyse/catalyse</td>
</tr>
<tr>
<td>control</td>
</tr>
<tr>
<td>convert</td>
</tr>
<tr>
<td>deacetylate</td>
</tr>
<tr>
<td>decline</td>
</tr>
<tr>
<td>decrease</td>
</tr>
<tr>
<td>eliminate</td>
</tr>
</tbody>
</table>

Table 5. Evaluation of the NER Component, Interaction Extraction Component, and the Whole System

<table>
<thead>
<tr>
<th></th>
<th>NER (CYP)</th>
<th>NER (chemical)</th>
<th>interaction extraction</th>
<th>whole system</th>
</tr>
</thead>
<tbody>
<tr>
<td>training set</td>
<td>7112 entities</td>
<td>30957 entities</td>
<td>90 sentences (LOO) 189 interactions</td>
<td>10 sentences 18 interactions</td>
</tr>
<tr>
<td>testing set</td>
<td>867 entities</td>
<td>4088 entities</td>
<td>90 sentences (LOO) 189 interactions</td>
<td>68.4%</td>
</tr>
<tr>
<td>precision</td>
<td>85.9%</td>
<td>89.3%</td>
<td>76.0%</td>
<td>72.2%</td>
</tr>
<tr>
<td>recall</td>
<td>86.6%</td>
<td>89.2%</td>
<td>82.6%</td>
<td>70.2%</td>
</tr>
<tr>
<td>F-score</td>
<td>86.3%</td>
<td>89.3%</td>
<td>79.2%</td>
<td>70.2%</td>
</tr>
</tbody>
</table>

Jiao, D. and Wild, D.J. Extraction of CYP Chemical Interactions from Biomedical Literature Using Natural Language Processing Methods, Journal of Chemical Information and Modeling, 49(2); pp263-269

Chemical & Biological Literature Extraction

Covering 1865-2009
18,502,916 PubMed/Medline literature records!

Wang HJ 2010

<table>
<thead>
<tr>
<th>Bio-Terms</th>
<th># of unique terms</th>
<th># of term-citation pairs</th>
<th># of unique citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound</td>
<td>56,383</td>
<td>11,775,891</td>
<td>3,856,084</td>
</tr>
<tr>
<td>Drug</td>
<td>2,820</td>
<td>5,624,529</td>
<td>3,427,067</td>
</tr>
<tr>
<td>Gene</td>
<td>13,022</td>
<td>5,252,844</td>
<td>3,735,517</td>
</tr>
<tr>
<td>Disease</td>
<td>3,848</td>
<td>12,612,636</td>
<td>7,066,084</td>
</tr>
<tr>
<td>Side Effect</td>
<td>1,363</td>
<td>10,489,676</td>
<td>6,310,741</td>
</tr>
<tr>
<td>Pathway</td>
<td>180</td>
<td>916,754</td>
<td>838,090</td>
</tr>
</tbody>
</table>

Validating topics by experimental relationships

Topic 26: cell, expression, cancer, tumor,…
Related Disease: DNA Damage, Melanoma, Glioblastoma,…
Bio-LDA III

- Entropy
  - In information theory, entropy is a measure of the uncertainty associated with a random variable.
  - Here we can compute the bio-term entropies over topics

- Kullback-Leibler divergence (KL divergence)
  - a non-symmetric measure of the difference between two probability distributions.
  - Here we used the KL divergence as the non-symmetric distance measure for two bio-terms over topics
Combining path finding and Bio-LDA

- Detect semantic association
  - Path finding algorithm
  - Millions of RDF triples from Chem2bio2rdf

- Assess semantic association
  - Bio-LDA model
  - Entropy and KL divergence
  - Additional knowledge base: 50, 100 and 200 topics using the recent 336,899 MEDLINE abstracts, which contains 13,338 identical bio-terms
Summary

- Drug discovery is entering a new era that is arguably centered on informatics analysis of the vast amount of biological and chemical data now being produced, and which looks at the effect of drugs on biological systems as a whole. This new approach underlies the new fields of *systems chemical biology* and *chemogenomics*.

- Analyzing this data and particularly the relationships between compounds, drugs, proteins, genes, diseases, pathways and people promises to provide important understanding of the nature of disease and treatment.

- The Semantic Web provides an effective framework for logically managing the data, and Cloud Computing provides a physical framework for computation and searching.

- Early-stage methods developed at Indiana allow integrated access to this data, path finding between any two points, visualization in chemical space and network tools, and advanced handling of the scholarly literature.

- Critical next steps include ranking and intelligent filtering of paths and relationships to provide aggregate evidence-based approaches, and integration of NGS and patient data.