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

The inner life of the cell

http://video.google.com/videoplay?docid=-2351549868099343381&hl=en#
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)
Large amount of data and links for each compound

Ibuprofen - Compound Summary (CID 3672)

A nonsteroidal anti-inflammatory agent with analgesic properties used in the therapy of rheumatism and arthritis.

BioMedical Annotation: (Total: 1)

Ibuprofen

Medication Information

Concentrated Ibuprofen Oral Suspension USP [Tris Pharma, Inc.]
  Warnings

Advil [Lil’ Drug Store Products, Inc.]

ADVIL (CONCENTRATED IBUPROFEN) [Pfizer Consumer Healthcare]

Advil Liqui-Gels (ibuprofen) [Pfizer Consumer Healthcare]

ADVIL MIGRAINE (ibuprofen) [Pfizer Consumer Healthcare]

Good Neighbor Pharmacy Childrens Ibuprofen (CONCENTRATED IBUPROFEN) [Amerisource Bergen]

Leader Childrens Ibuprofen (CONCENTRATED IBUPROFEN) [Cardinal Health]

Leader Childrens Ibuprofen (CONCENTRATED IBUPROFEN) [Cardinal Health]

Links

PubMed (3)
NLM Toxicology Link

Chemical Structure Search
BioActivity Summary:
This Compound with Similar Compounds

Related Compounds:
  Same, Connectivity: 12 Links
  Same, Stereocchemistry: 7 Links
  Same, Isotopes: 3 Links

Similar Compounds: 2712 Links
Similar Conformers: 446 Links
View Conformers

Substances:
  All: 522 Links
Proteins & Genes

Chem2Bio2RDF: The FaceBook of Drug Discovery

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

Journal home > Archive > Commentary > Full Text

Commentary

doi:10.1038/nchembio0807-447

Systems chemical biology

Tudor I Oprea¹, Alexander Tropsha², Jean-Loup Faulon³ & Mark D Rintoul³

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, CB # 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

User interface and applications

Trust

Proof

Unifying logic

Querying: SPARQL

Ontologies: OWL

Rules: RIF/SWRL

Taxonomies: RDFS

Cryptography

Data interchange: RDF

Syntax: XML

Identifiers: URI

Character set: UNICODE

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

Drug

name

company

Lipitor

Pfizer

<RDF>
    <Description about="http://chem2bio2rdf.org/drug/DB01076">  
        <name>Lipitor</name>  
        <company>Pfizer</company>  
    </Description>
</RDF>
Chem2Bio2RDF context

Dereferencable URI

PlotViz: Visualization

Cytoscape Plugin

Linked Path Generation and Ranking

Sparql Endpoints

Third party tools
Chem2Bio2RDF Relationships

Converting data into RDF

External Sources

XML
CSV
TXT
DB

Download

Local copy

Scripts

Relational DB

D2R Mapping

D2R server

Ontology

Dumping

Virtuoso Triple Store

Publishing

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>12345</td>
</tr>
<tr>
<td>name</td>
<td>John</td>
</tr>
<tr>
<td>age</td>
<td>30</td>
</tr>
<tr>
<td>email</td>
<td><a href="mailto:john@example.com">john@example.com</a></td>
</tr>
</tbody>
</table>
Finding multi-target inhibitors of MAPK pathway with a SPARQL query

<table>
<thead>
<tr>
<th>Compound</th>
<th>Pubchem Bio assay: 133649</th>
<th>Pubchem Bio assay: 3388</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>84</td>
<td>91</td>
</tr>
<tr>
<td>Outcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target</td>
<td>Gi: 66932916</td>
<td>Gi: 4758204</td>
</tr>
<tr>
<td>Pathway</td>
<td>Uniprot: P28482</td>
<td>Uniprot: P28562</td>
</tr>
<tr>
<td>Pathway name</td>
<td>kegg_pathway: hsa04010</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MAPK signaling pathway
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

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

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

Doxazosin
PTGS1
PTGS2
hepatic necrosis
VEGF signaling pathway
hepatitis
arachidonic acid metabolism

**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 (anthracyclin antibiotic)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel</td>
<td>Renal</td>
<td>Renal</td>
<td>Non-Small Lung</td>
<td>Non-Small Lung</td>
<td>Renal</td>
<td>Leukemia</td>
<td>Colon</td>
<td>Non-Small Lung</td>
<td>Colon</td>
<td>Colon</td>
<td>Leukemia</td>
<td>Non-Small Lung</td>
<td>Colon</td>
<td>Ovarian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>probability</td>
<td>.69</td>
<td>.72</td>
<td>.73</td>
<td>.74</td>
<td>.65</td>
<td>.72</td>
<td>.72</td>
<td>.73</td>
<td>.70</td>
<td>.69</td>
<td>.66</td>
<td>.69</td>
<td>.66</td>
<td>.67</td>
<td>.66</td>
<td>.69</td>
</tr>
</tbody>
</table>


CID 1691

1.000 AID 115 NCI human tumor cell line growth inhibition assay. Data for the SR Leukemia cell line

CID 871219

0.964 AID 1 NCI human tumor cell line growth inhibition assay. Data for the NCI-H23 Non-Small Lung cell line

CID 15940175

0.964 AID 1064 A Validation Screen to Identify Compounds that Suppress the Growth of Human Colon Tumor Cells Lacking Oncogenic Beta Catenin Expression Using a Diverse Compound Set

Doxorubicin

Doxorubicin is an antineoplastic in the anthracycline class. General properties of drugs in this class include: interaction with DNA in a variety of different ways including intercalation (squeezing between the base pairs), DNA strand breakage and inhibition with the enzyme topoisomerase II. Most of these compounds have been isolated from natural sources and antibiotics. However, they lack the specificity of the antimitobal antibiotics and thus produce significant toxicity. The anthracyclines are among the most important antitumor drugs available. Doxorubicin is widely used for the treatment of several solid tumors while daunorubicin and idarubicin are used exclusively for the treatment of leukemia. Daunorubicin may also inhibit polymerase activity, affect regulation of gene expression, and produce free radical damage to DNA. Doxorubicin possesses an antitumor effect against a wide spectrum of tumors, either grafted or spontaneous. The anthracyclines are cell cycle nonspecific.

Daunorubicin

Daunorubicin is an antineoplastic in the anthracycline class. General properties of drugs in this class include: interaction with DNA in a variety of different ways including intercalation (squeezing between the base pairs), DNA strand breakage and inhibition with the enzyme topoisomerase II. Most of these compounds have been isolated from natural sources and antibiotics. However, they lack the specificity of the antimitobal antibiotics and thus produce significant toxicity. The anthracyclines are among the most important antitumor drugs available. Doxorubicin is widely used for the treatment of several solid tumors while daunorubicin and idarubicin are used exclusively for the treatment of leukemia. Daunorubicin may also inhibit polymerase activity, affect regulation of gene expression, and produce free radical damage to DNA. Daunorubicin possesses an antitumor effect against a wide spectrum of tumors, either grafted or spontaneous. The anthracyclines are cell cycle nonspecific.

Idarubicin

Idarubicin is an antineoplastic in the anthracycline class. General properties of drugs in this class include: interaction with DNA in a variety of different ways including intercalation (squeezing between the base pairs), DNA strand breakage and inhibition with the enzyme topoisomerase II. Most of these compounds have been isolated from natural sources and antibiotics. However, they lack the specificity of the antimitobal antibiotics and thus produce significant toxicity. The anthracyclines are among the most important antitumor drugs available. Doxorubicin is widely used for the treatment of several solid tumors while daunorubicin and idarubicin are used exclusively for the treatment of leukemia. Idarubicin may also inhibit polymerase activity, affect regulation of gene expression, and produce free radical damage to DNA. Idarubicin

Toxic Hazard: High (Class III) according to Cramer rules.
WENDI v1.0 - insights from the literature

**Basic 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 e1a2(b2a2) and e1a2(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 CCyR, 2 PCyR, 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 CCyR, 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: 19531667 [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

![Diagram showing relationships between entities and RDF properties](image-url)
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/catalyze</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></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!

Table 1. Statistics of the bio-terms extraction.

<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>5,856,084</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>

Wang HJ 2010
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.