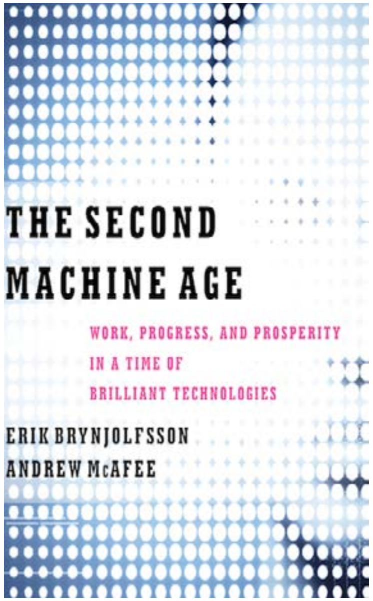
Measuring Scholarly Impact and **Beyond**

Ying Ding
Indiana University
dingying@indiana.edu
http://info.slis.indiana.edu/~dingying/index.html

Scholarly Communication: Challenges and Opportunities in Digital Age

Ying Ding
Indiana University
dingying@indiana.edu
http://info.slis.indiana.edu/~dingying/index.html

Digital Age



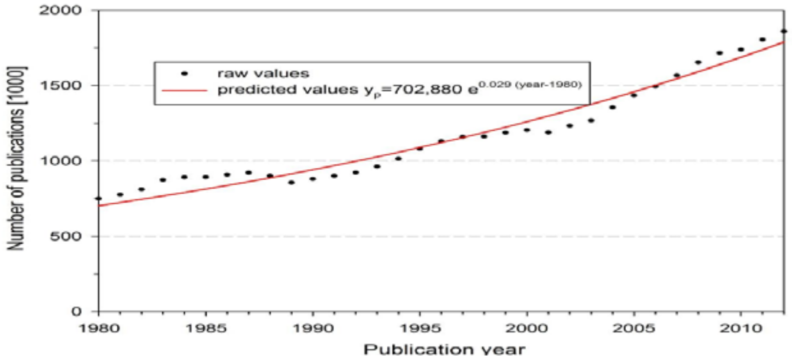
THE SECOND MACHINE AGE

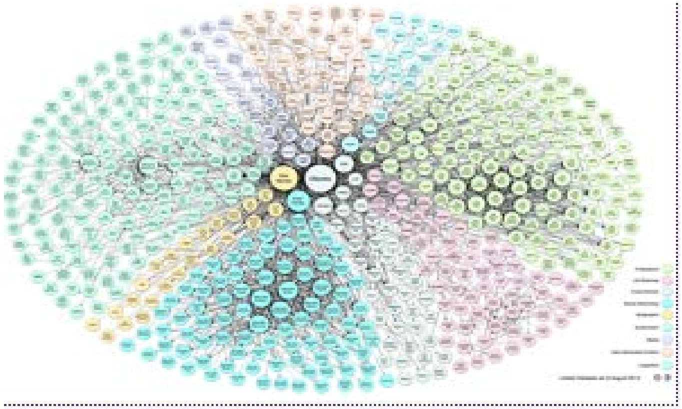
WORK, PROGRESS, AND PROSPERITY IN A TIME OF BRILLIANT TECHNOLOGIES

ERIK BRYNJOLFSSON
ANDREW McAFEE

Machine World → Human World

Digital Challenges





Human World → Machine World

Confusing: Machine or Human World

You and Me are currently living in such a confusing time

Digital Advantages

- Easy to access data
- Easy to share data (e.g., zero cost to make a copy)
- Powerful computing technologies
- Innovative minds and many eyeballs
- Motivated human capitals

- Digital Transformation:
  - Digital DNA (250M photos uploaded to Facebook daily, >5B have mobile phones)
  - Digital Bonding (we spend more time with our smartphones than with our partner, 119 minute/day)

Rich Data and Huge Opportunities

- Smart Phone → Smart Home → Smart Robot → Smart City→ Smart … → Smart doctor

- Scholarly communication → Digital Scholarly Communication → Literature-Driven Discovery →Data-Driven Discovery→ Scientific Discovery

Data2Knowledge

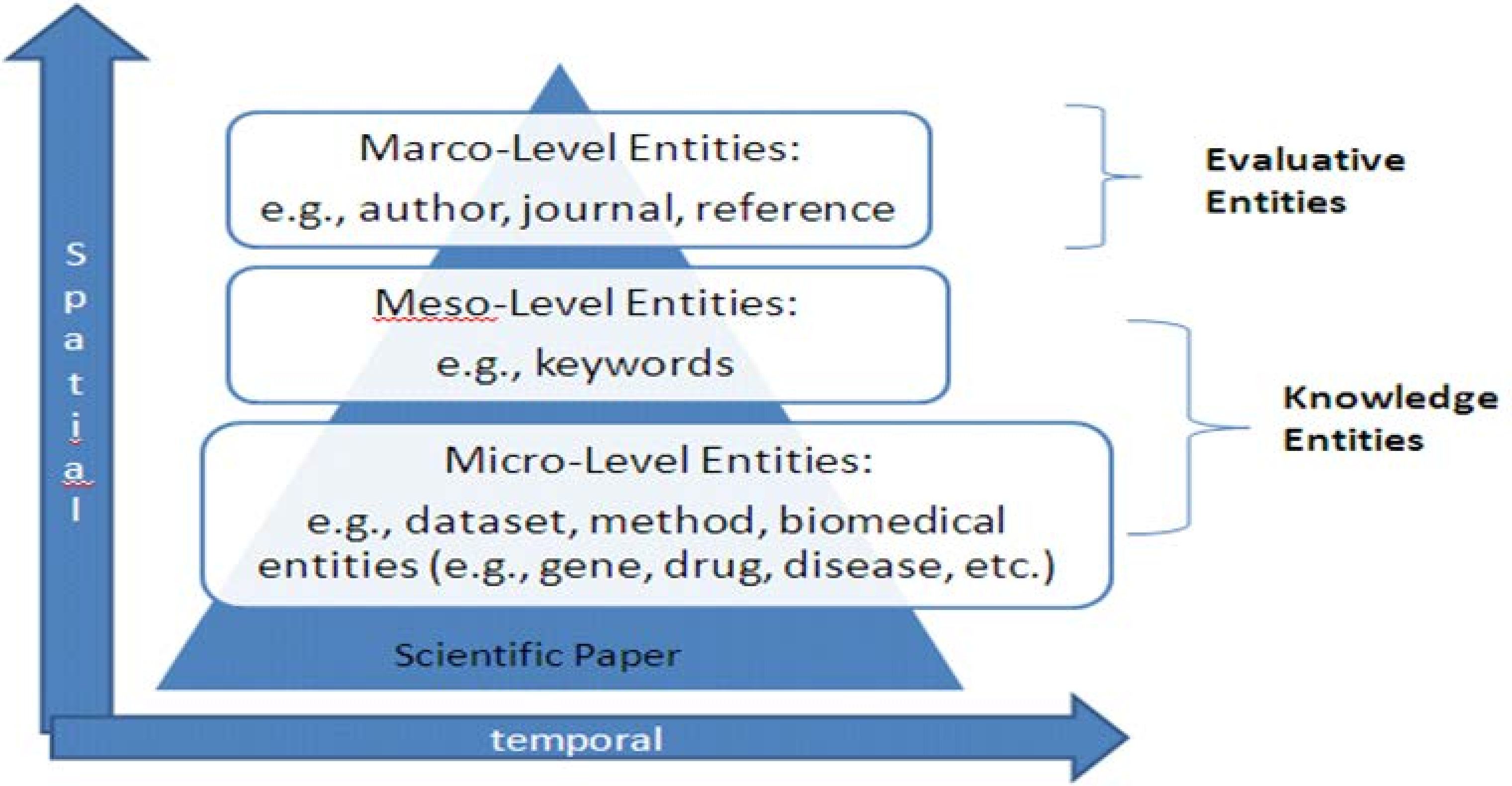
Next Generation of Scholarly Communication

- Newly developed methods allow in-depth analysis of scholarly communication
  - Topic modeling (e.g., Latent Dirichlet Allocation)
  - Information Extraction (e.g., OpenIE)
  - Social Network Analysis (e.g., Community Detection)
- Big data demonstrates the power of connected data to enable knowledge discovery
  - Structured data
  - Unstructured data
  - Social media data
- Digital age incubates transformative innovations
  - Working with domain experts (e.g., biologist, sociologist, historian)
  - Computational discovery in science, social science, and humanities

Ding, Y., Rousseau, R., & Wolfram, D. (Eds.) (2014). *Measuring scholarly impact: Methods and practice*. Springer.

EntityMetrics

Entitymetrics is defined as using entities (i.e., evaluative entities or knowledge entities) in the measurement of impact, knowledge usage, and knowledge transfer, to facilitate knowledge discovery.



Marco-Level Entities:
e.g., author, journal, reference

Meso-Level Entities:
e.g., keywords

Micro-Level Entities:
e.g., dataset, method, biomedical entities (e.g., gene, drug, disease, etc.)

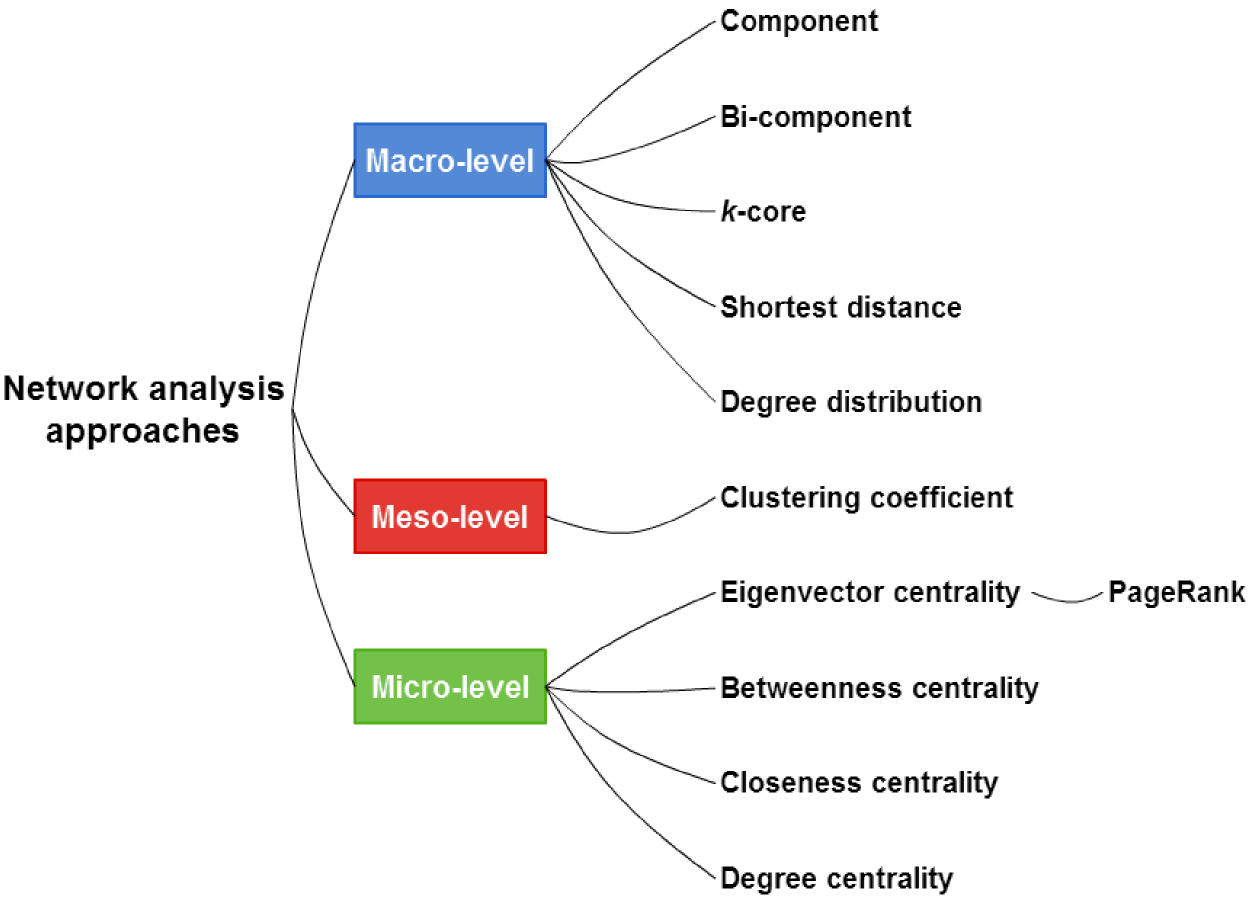
Scientific Paper

Evaluative Entities

Knowledge Entities

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. PLoS One, 8(8): 1-14.

EntityMetrics



PubMed Entities

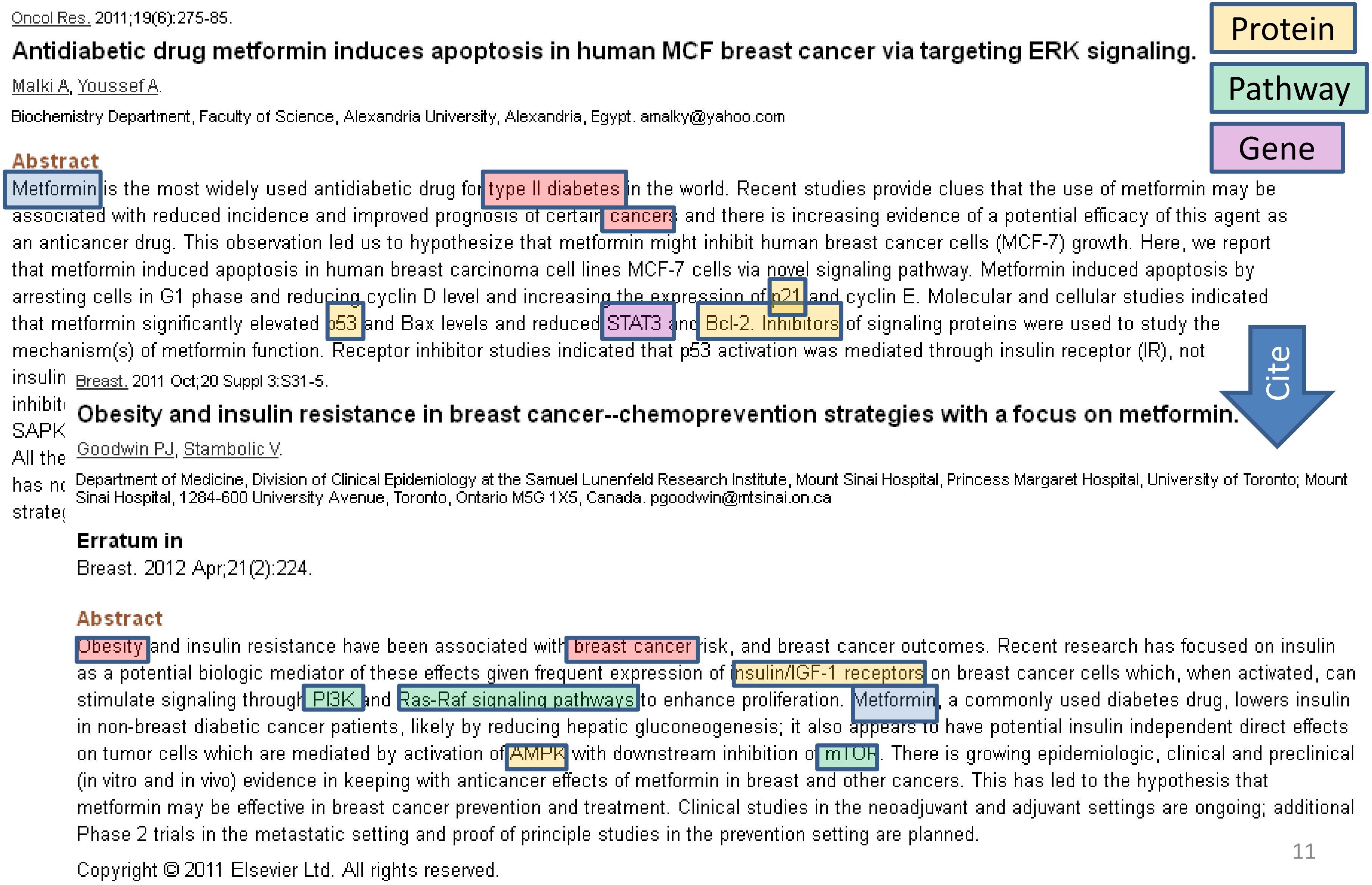
Drug

Disease

Protein

Pathway

Gene



Oncol Res. 2011;19(6):275-85.

Antidiabetic drug metformin induces apoptosis in human MCF breast cancer via targeting ERK signaling.

Malki A, Youssef A.

Biochemistry Department, Faculty of Science, Alexandria University, Alexandria, Egypt. amalky@yahoo.com

Abstract

Metformin is the most widely used antidiabetic drug for type II diabetes in the world. Recent studies provide clues that the use of metformin may be associated with reduced incidence and improved prognosis of certain cancers and there is increasing evidence of a potential efficacy of this agent as an anticancer drug. This observation led us to hypothesize that metformin might inhibit human breast cancer cells (MCF-7) growth. Here, we report that metformin induced apoptosis in human breast carcinoma cell lines MCF-7 cells via novel signaling pathway. Metformin induced apoptosis by arresting cells in G1 phase and reducing cyclin D level and increasing the expression of p21 and cyclin E. Molecular and cellular studies indicated that metformin significantly elevated p53 and Bax levels and reduced STAT3 and Bcl-2. Inhibitors of signaling proteins were used to study the mechanism(s) of metformin function. Receptor inhibitor studies indicated that p53 activation was mediated through insulin receptor (IR), not

insulin

inhibit

SAPK

All the

has n

strate

Cite

Breast. 2011 Oct;20 Suppl 3:S31-5.

Obesity and insulin resistance in breast cancer--chemoprevention strategies with a focus on metformin.

Goodwin PJ, Stambolic V.

Department of Medicine, Division of Clinical Epidemiology at the Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Princess Margaret Hospital, University of Toronto; Mount Sinai Hospital, 1284-600 University Avenue, Toronto, Ontario M5G 1X5, Canada. pgoodwin@mtsinai.on.ca

Erratum in

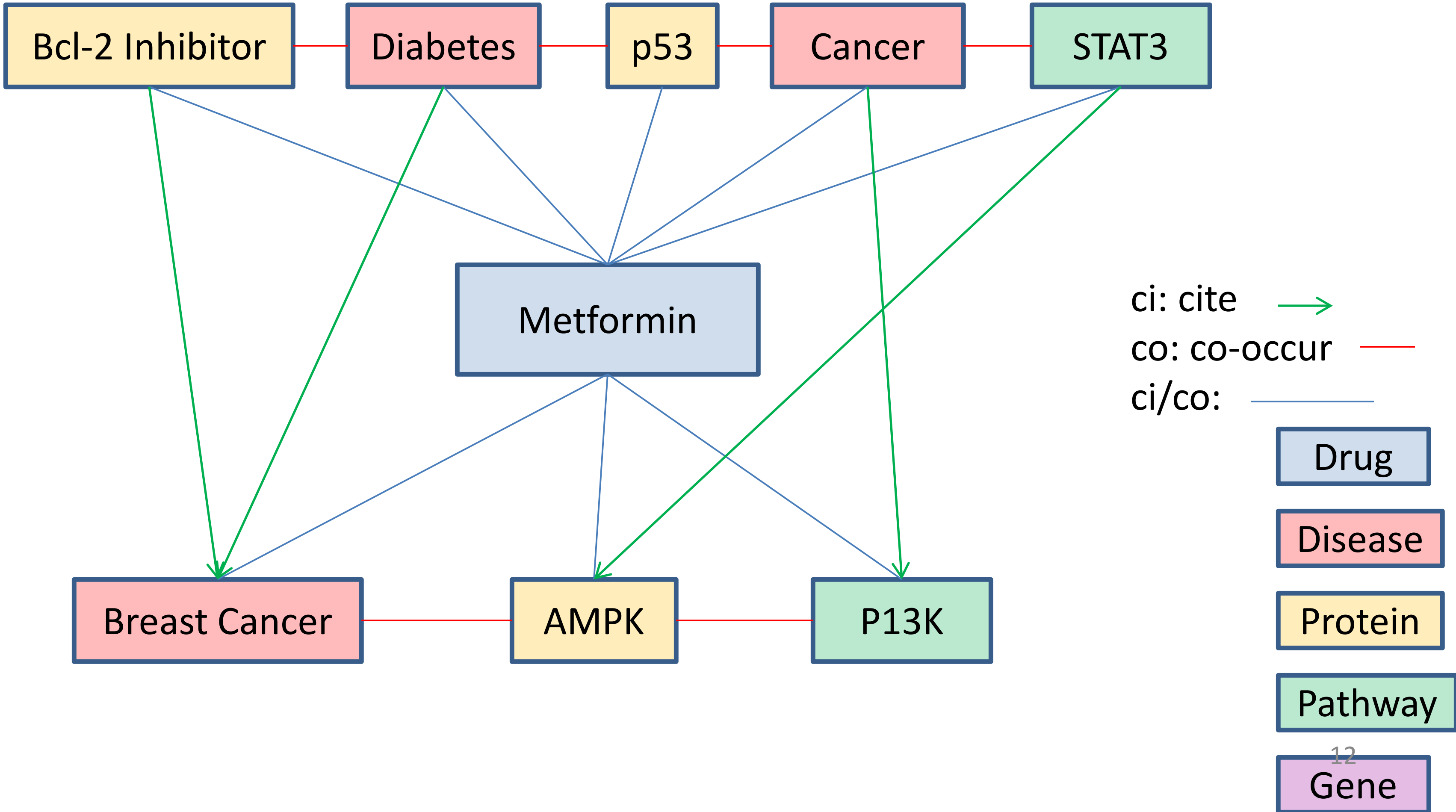
Breast. 2012 Apr;21(2):224.

Abstract

Obesity and insulin resistance have been associated with breast cancer risk, and breast cancer outcomes. Recent research has focused on insulin as a potential biologic mediator of these effects given frequent expression of insulin/IGF-1 receptors on breast cancer cells which, when activated, can stimulate signaling through PI3K and Ras-Raf signaling pathways to enhance proliferation. Metformin, a commonly used diabetes drug, lowers insulin in non-breast diabetic cancer patients, likely by reducing hepatic gluconeogenesis; it also appears to have potential insulin independent direct effects on tumor cells which are mediated by activation of AMPK with downstream inhibition of mTOR. There is growing epidemiologic, clinical and preclinical (in vitro and in vivo) evidence in keeping with anticancer effects of metformin in breast and other cancers. This has led to the hypothesis that metformin may be effective in breast cancer prevention and treatment. Clinical studies in the neoadjuvant and adjuvant settings are ongoing; additional Phase 2 trials in the metastatic setting and proof of principle studies in the prevention setting are planned.

Entity Graph

- Heterogeneous Entity Graph



Metformin related entity-entity citation network

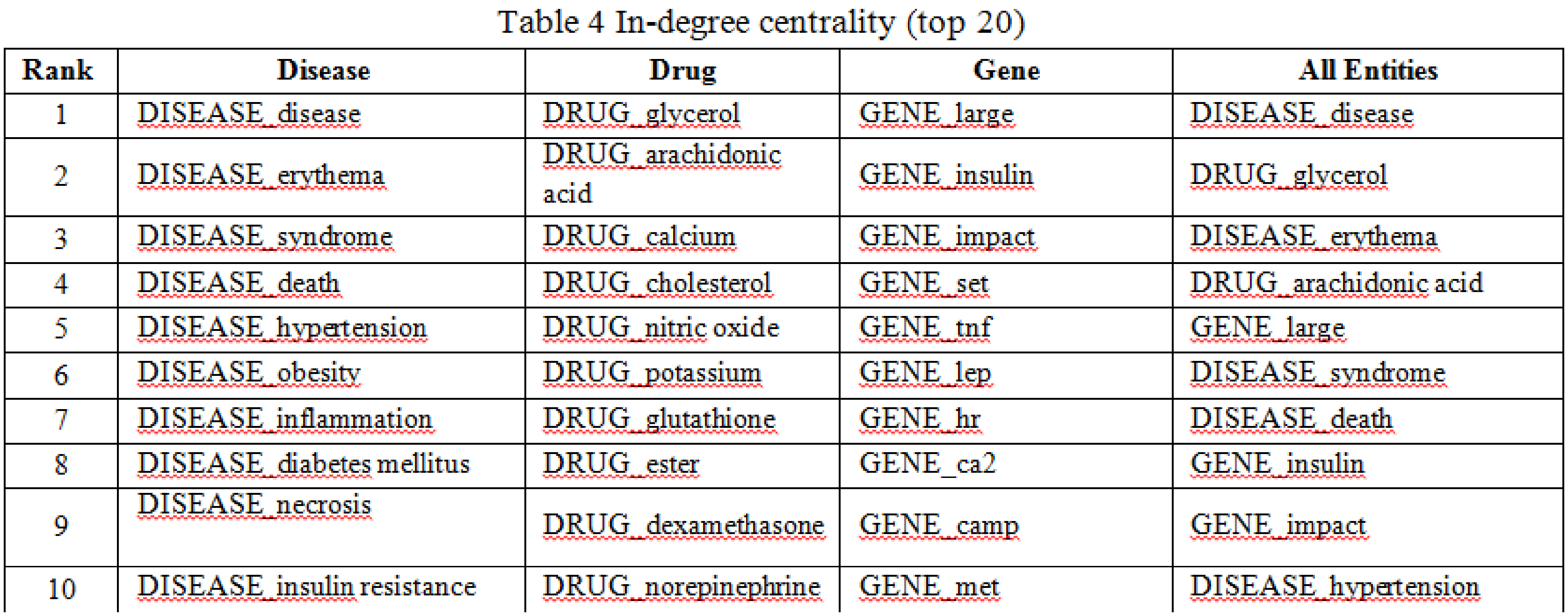
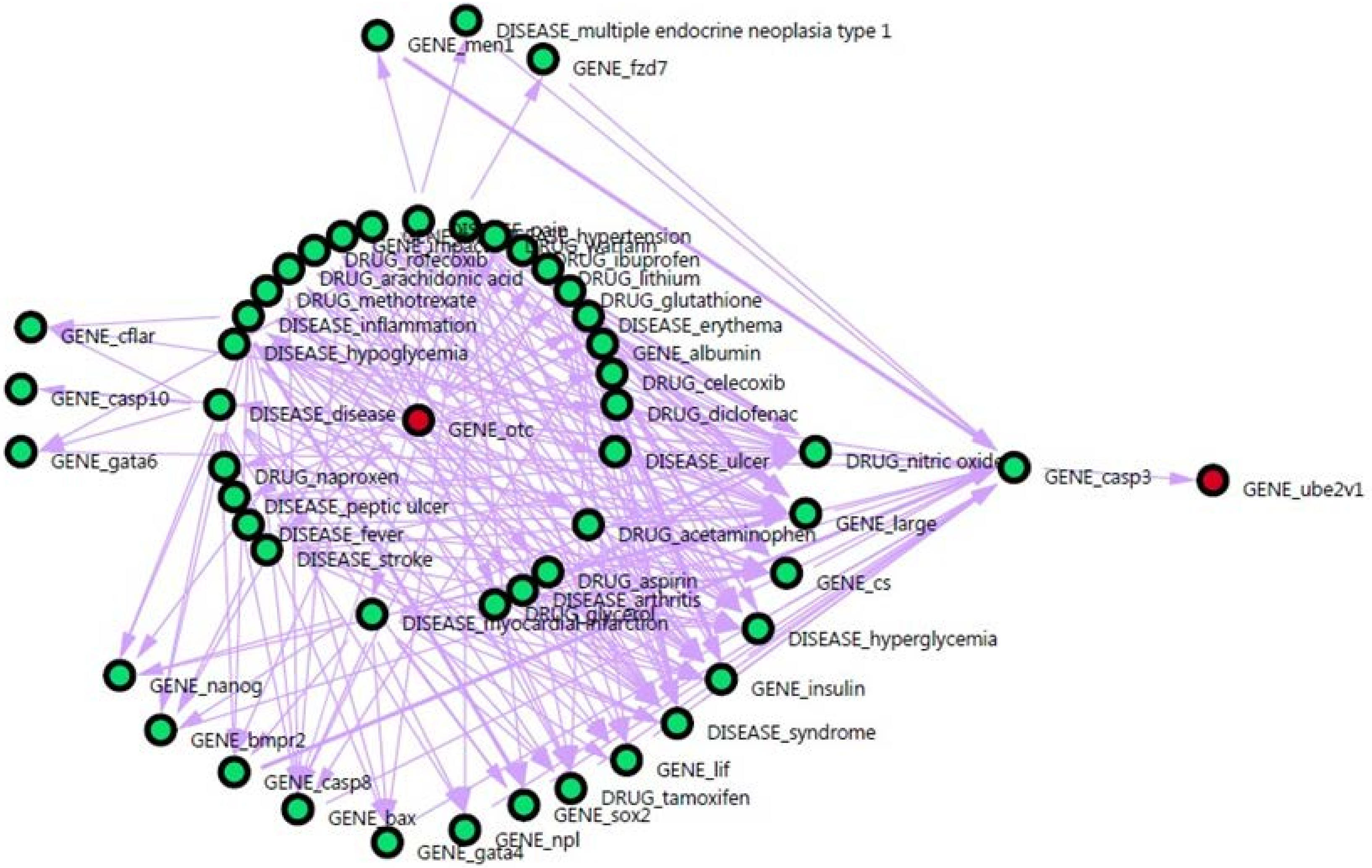


Table 4 In-degree centrality (top 20)

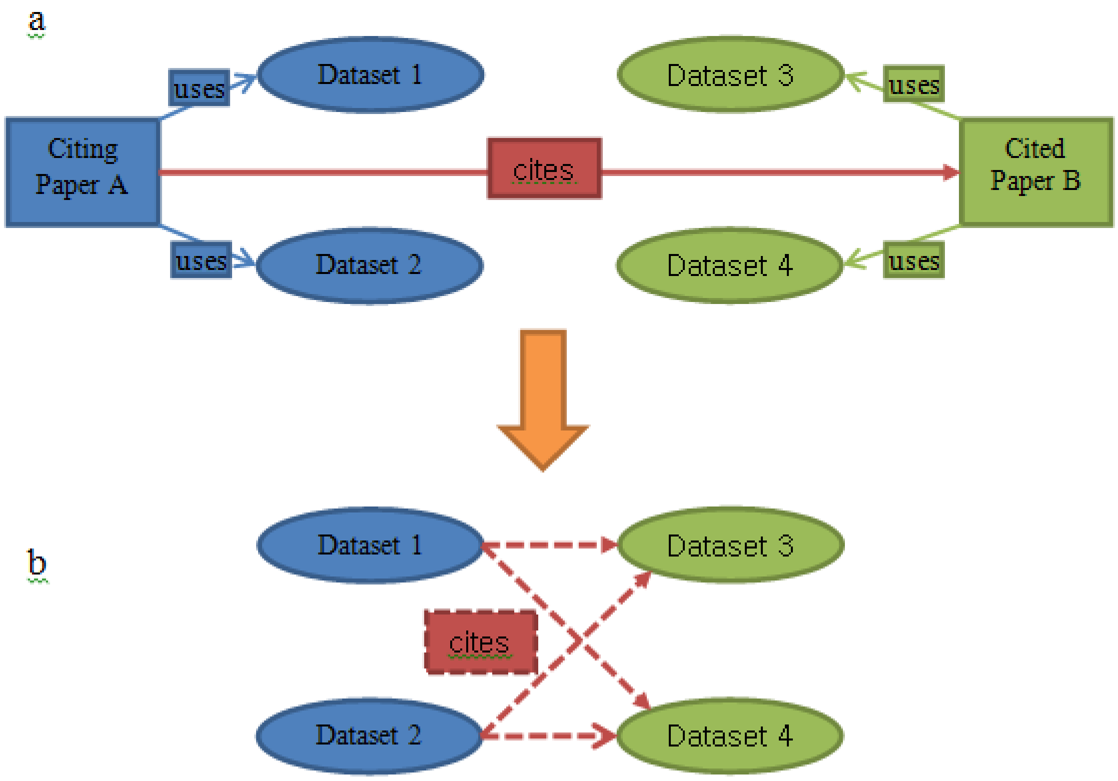
Rank	Disease	Drug	Gene	All Entities
1	DISEASE_disease	DRUG_glycerol	GENE_large	DISEASE_disease
2	DISEASE_erythema	DRUG_arachidonic acid	GENE_insulin	DRUG_glycerol
3	DISEASE_syndrome	DRUG_calcium	GENE_impact	DISEASE_erythema
4	DISEASE_death	DRUG_cholesterol	GENE_set	DRUG_arachidonic acid
5	DISEASE_hypertension	DRUG_nitric oxide	GENE_tnf	GENE_large
6	DISEASE_obesity	DRUG_potassium	GENE_lep	DISEASE_syndrome
7	DISEASE_inflammation	DRUG_glutathione	GENE_hr	DISEASE_death
8	DISEASE_diabetes mellitus	DRUG_ester	GENE_ca2	GENE_insulin
9	DISEASE_necrosis	DRUG_dexamethasone	GENE_camp	GENE_impact
10	DISEASE_insulin resistance	DRUG_norepinephrine	GENE_met	DISEASE_hypertension

Data: 4,770 articles retrieved from PubMed Central with 134,844 references, and 1,969 bio-entities (i.e., 880 genes, 376 drugs, and 713 diseases)

Metformin related entity-entity citation network

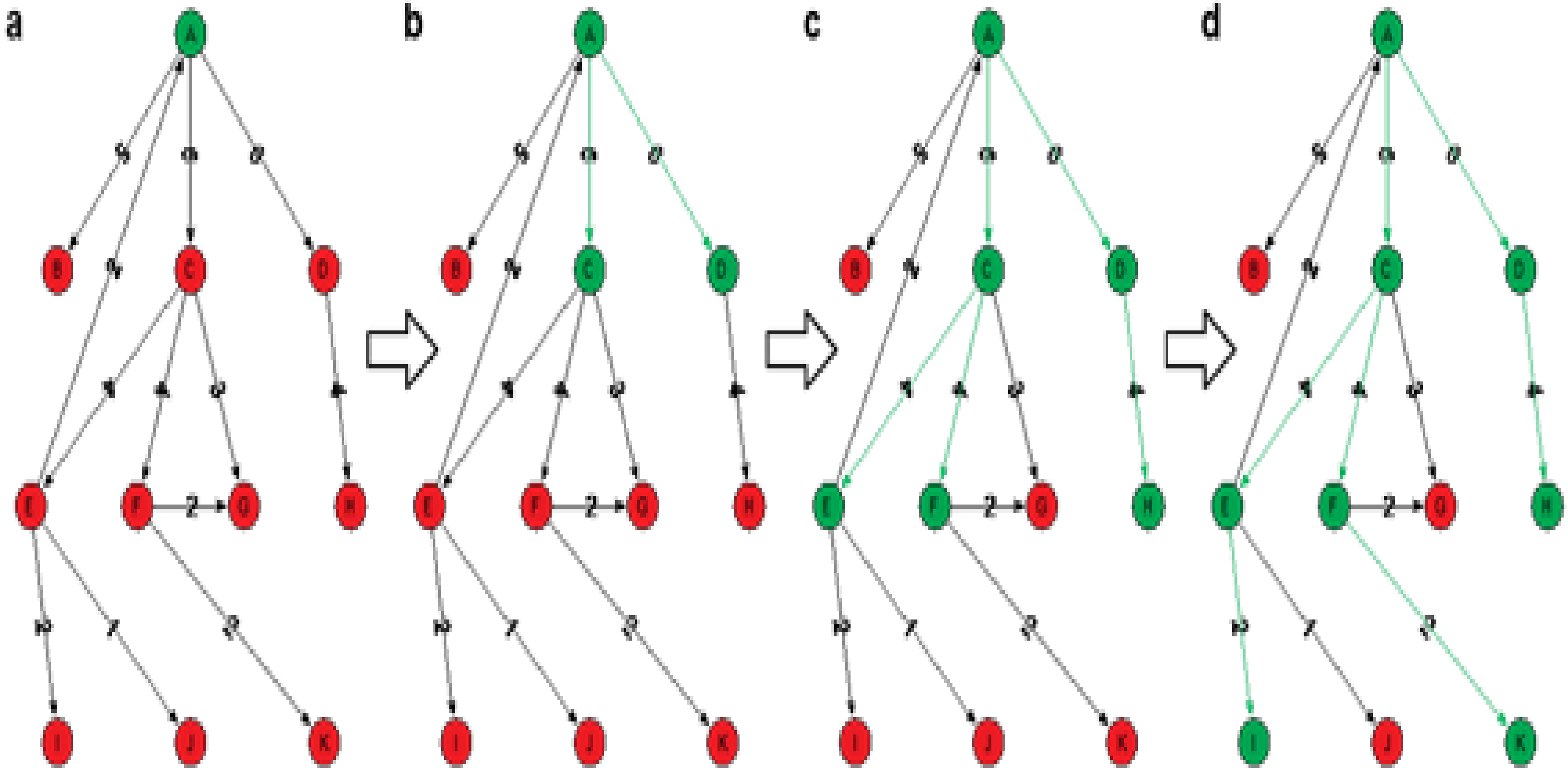


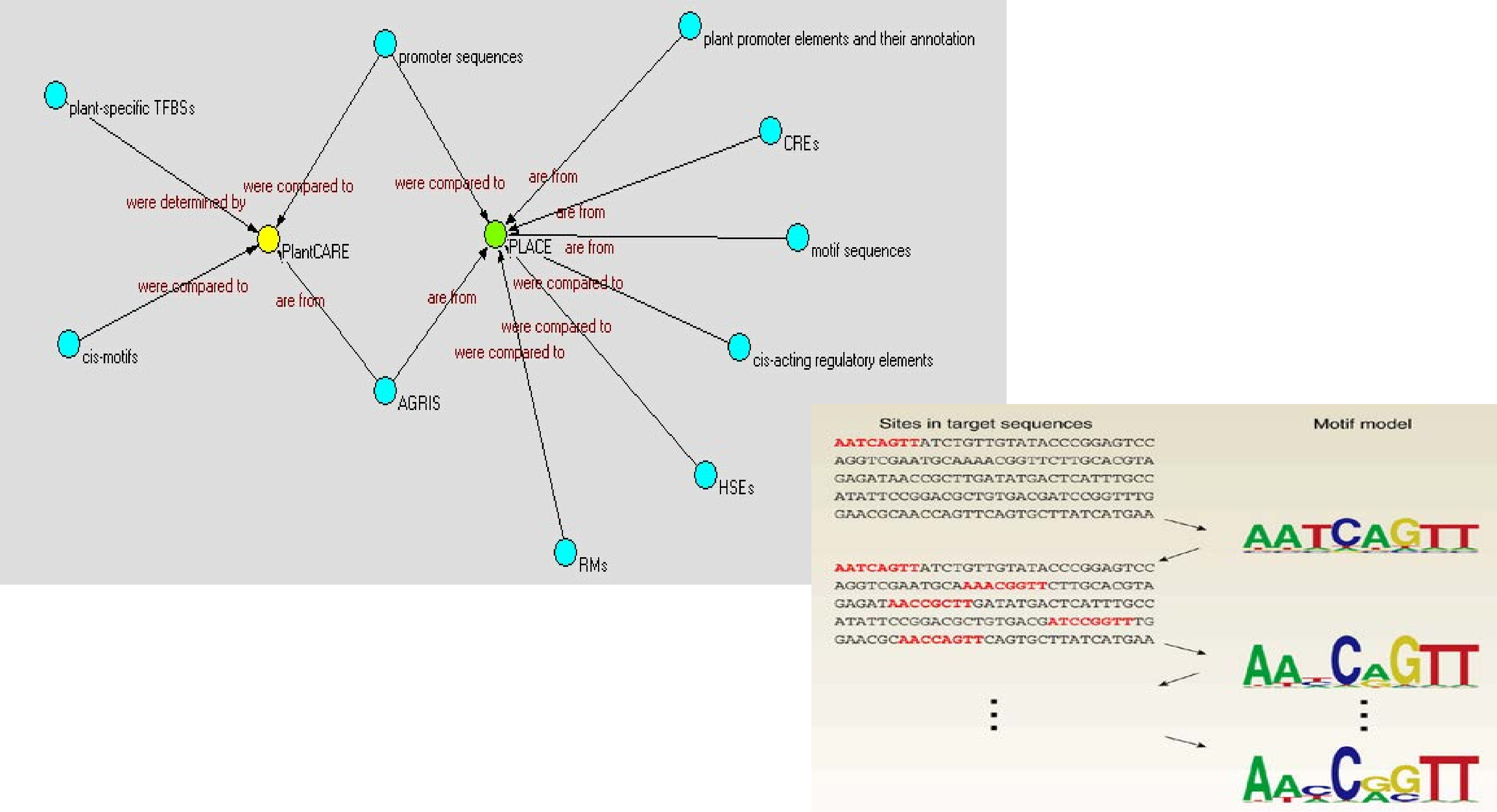
Dataset



Yu, Q., **Ding, Y.**, Song, M., Song, S., Liu, J., & Zhang, B. (forthcoming). Tracing database usage: Detecting main paths in database link networks. *Journal of Informetrics*.

Main Path




plant promoter elements and their annotation
promoter sequences
plant-specific TFBSs
CREs
were compared to
were compared to
are from
were determined by
are from
PlantCARE
PLACE
are from
motif sequences
were compared to
were compared to
are from
are from
were compared to
were compared to
cis-motifs
cis-acting regulatory elements
AGRIS
HSEs
RMs
Sites in target sequences
Motif model
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA
AATCAGTT
AATCAGTTATCTGTTGTATACCCGGAGTCC
AGGTCGAATGCAAAACGGTTCTTGCACGTA
GAGATAACCGCTTGATATGACTCATTTGCC
ATATTCCGGACGCTGTGACGATCCGGTTTG
GAACGCAACCAGTTCAGTGCTTATCATGAA
AA CAGTT
AA CGGTT

PLACE is a database of motifs found in plant cis-acting regulatory DNA elements
PlantCARE is a database of plant cis- acting regulatory elements, enhancers and repressors. The motifs are collected in this database as well,

Entity Citation Network vs. Entity Co-Occurrence Network

- Gene Gene Co-Occurrence Network (GG) vs. Gene Cite Gene Network (GCG)
  - The GCG network shares many genes with the GG network and as a result is a competitive complement to the GG network
  - Using gene relationships based on citation relation extends the assumption of gene interaction being limited to the same article and opens up a new opportunity to analyze gene interaction from a wider spectrum of datasets.
  - 1,149 gene pairs from GCG were found in GG. A total of 164 pairs out of 1,149 were not found in GG before 2005, but were found in GCG before 2005. In particular, the PARK2 and PINK1 gene pair ranks fifth by co-occurrence frequency in the GG network, implying the gene pair has highly been studied since 2005

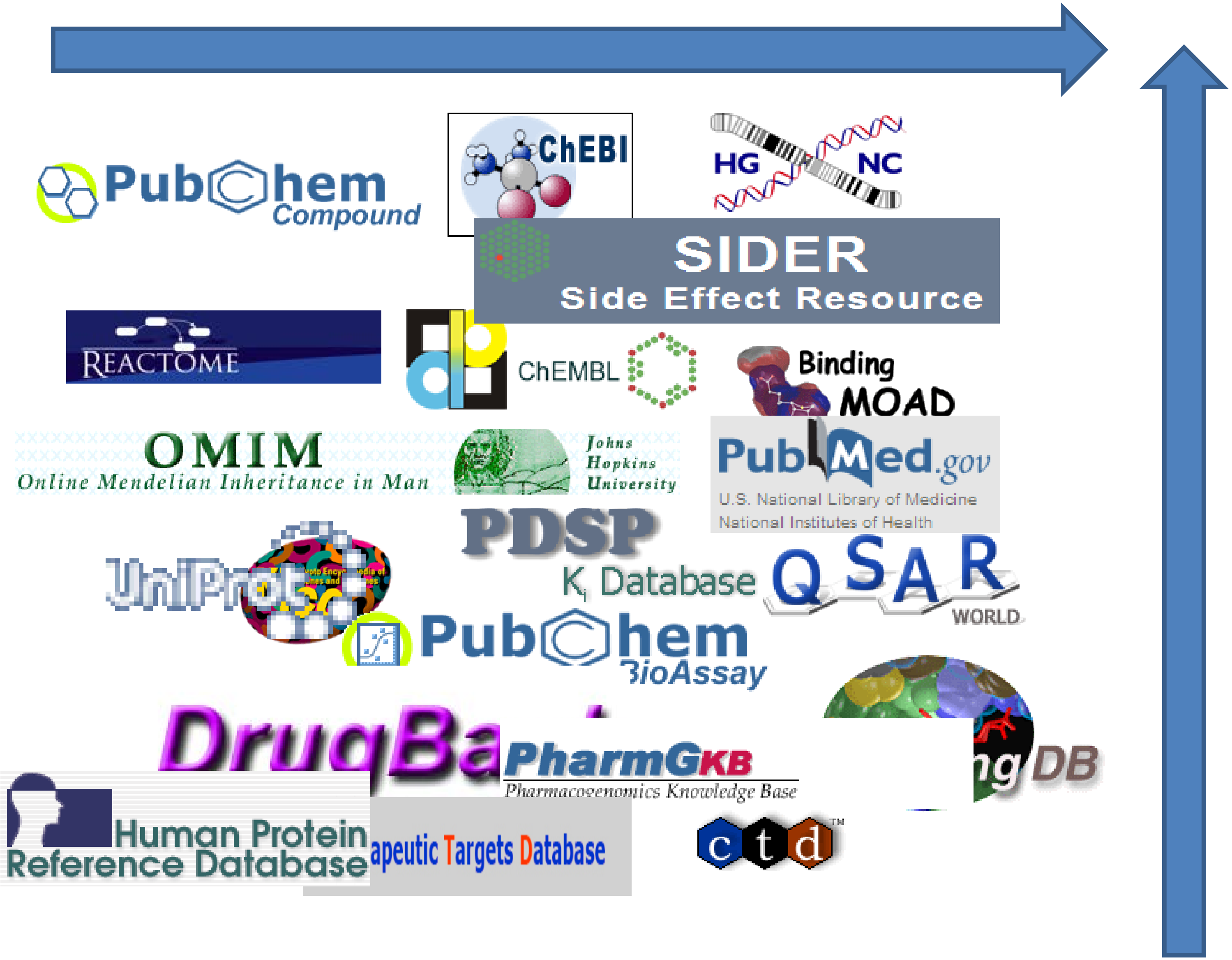
Song, M., Han, N., Kim, Y., Ding, Y., & Chambers, T. (2013). Discovering implicit entity relation with the gene-citation-gene network. *PLoS One*, 8(12), e84639

Big Data in Life Sciences

- 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)
  - **59 million compound bioactivities** (PubChem Bioassay)
  - **4,763 drugs** (DrugBank)
  - **9 million protein sequences** (SwissProt) and 58,000 3D structures (PDB)
  - **14 million human nucleotide sequences** (EMBL)
  - **22 million life sciences publications** - 800,000 new each year (PubMed)
  - Multitude of other sets (drugs, toxicogenomics, chemogenomics, metagenomics …)

- 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)

Wild, D. J., Ding, Y., Sheth, A. P., Harland, L., Gifford, E. M., & Lajiness, M. S. (2012). System chemical biology and the Semantic Web: What they mean for the future of drug discovery research. *Drug Discovery Today* (impact factor=6.422), 17(9-10), 469-474.



Text CSV Table HTML XML

Patient

Disease

Tissue

Cell

Pathway

DNA

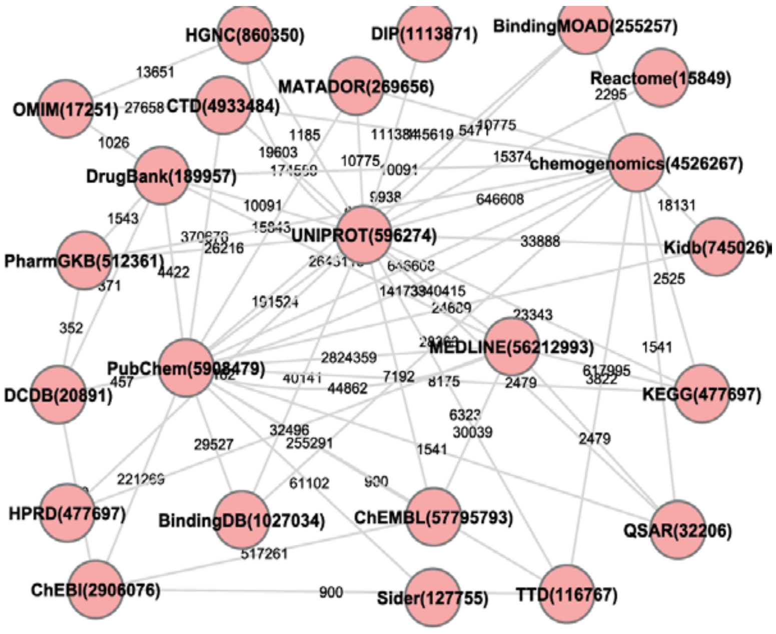
RNA

Protein

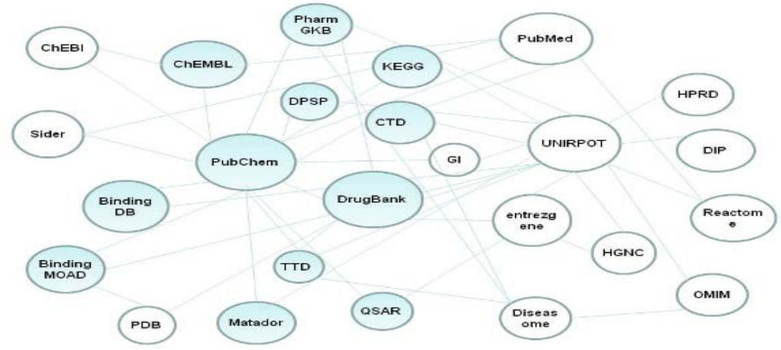
Drug

Chem2Bio2RDF

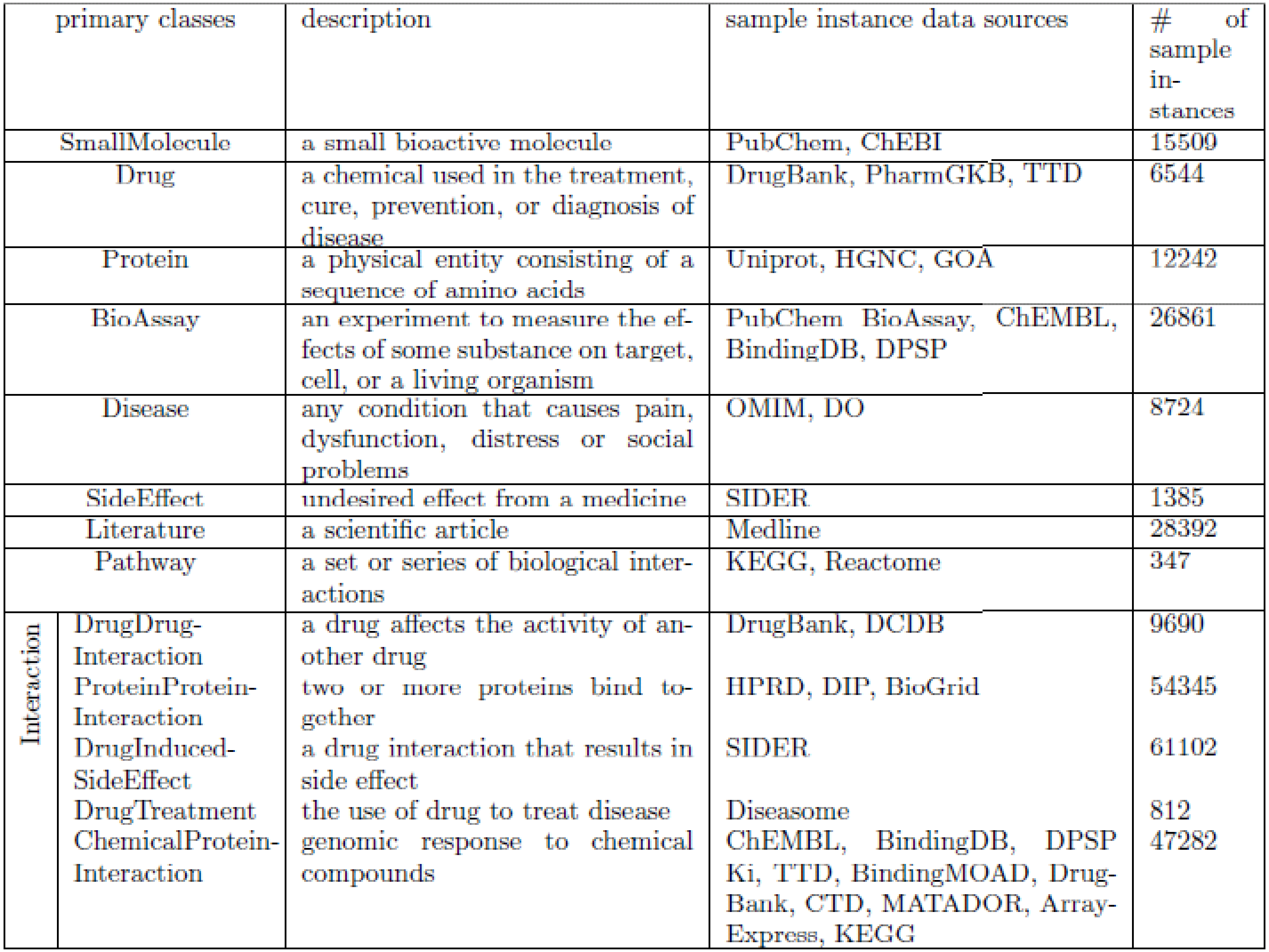
- NCI Human Tumor Cell Lines Data
- PubChem Compound Database
- PubChem Bioassay Database
- PubChem Descriptions of all PubChem bioassays
- Pub3D: A similarity-searchable database of minimized 3D structures for PubChem compounds
- Drugbank
- MRTD: An implementation of the Maximum Recommended Therapeutic Dose set
- Medline: IDs of papers indexed in Medline, with SMILES of chemical structures
- ChEMBL chemogenomics database
- KEGG Ligand pathway database
- Comparative Toxicogenomics Database
- PhenoPred Data
- HuGEpedia: an encyclopedia of human genetic variation in health and disease.



31m chemical structures
59m bioactivity data points
3m/19m publications
~5,000 drugs

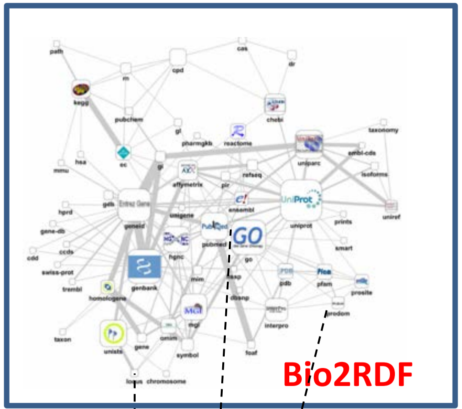


Chen, B., Dong, X., Jiao, Dazhi, Wang, H., Zhu, Q., Ding, Y. and Wild, D. (2010). Chem2Bio2RDF: A semantic framework for linking and mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, 2010, 11, 255.

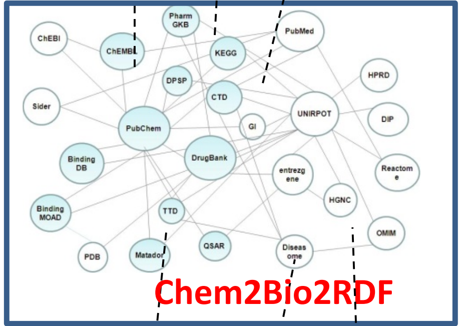


primary classes		description	sample instance data sources	# of sample instances
SmallMolecule		a small bioactive molecule	PubChem, ChEBI	15509
Drug		a chemical used in the treatment, cure, prevention, or diagnosis of disease	DrugBank, PharmGKB, TTD	6544
Protein		a physical entity consisting of a sequence of amino acids	Uniprot, HGNC, GOA	12242
BioAssay		an experiment to measure the effects of some substance on target, cell, or a living organism	PubChem BioAssay, ChEMBL, BindingDB, DPSP	26861
Disease		any condition that causes pain, dysfunction, distress or social problems	OMIM, DO	8724
SideEffect		undesired effect from a medicine	SIDER	1385
Literature		a scientific article	Medline	28392
Pathway		a set or series of biological interactions	KEGG, Reactome	347
Interaction	DrugDrug-Interaction	a drug affects the activity of another drug	DrugBank, DCDB	9690
	ProteinProtein-Interaction	two or more proteins bind together	HPRD, DIP, BioGrid	54345
	DrugInduced-SideEffect	a drug interaction that results in side effect	SIDER	61102
	DrugTreatment	the use of drug to treat disease	Diseasome	812
	ChemicalProtein-Interaction	genomic response to chemical compounds	ChEMBL, BindingDB, DPSP Ki, TTD, BindingMOAD, DrugBank, CTD, MATADOR, ArrayExpress, KEGG	47282

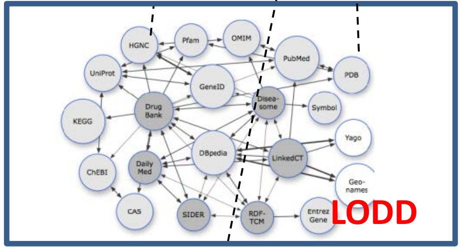
Chen, B., Ding, Y., & Wild, D. J. (2012). Improving integrative searching of systems chemical biology data using semantic annotation. *Journal of Cheminformatics*, 4:6 (doi:10.1186/1758-2946-4-6).



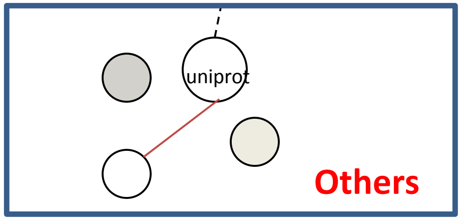
Bio2RDF



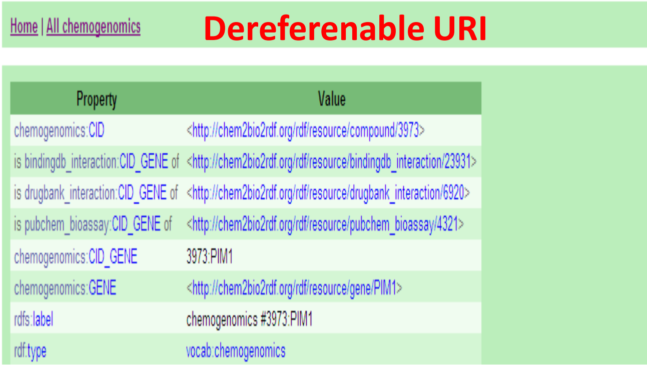
Chem2Bio2RDF



LODD



Others

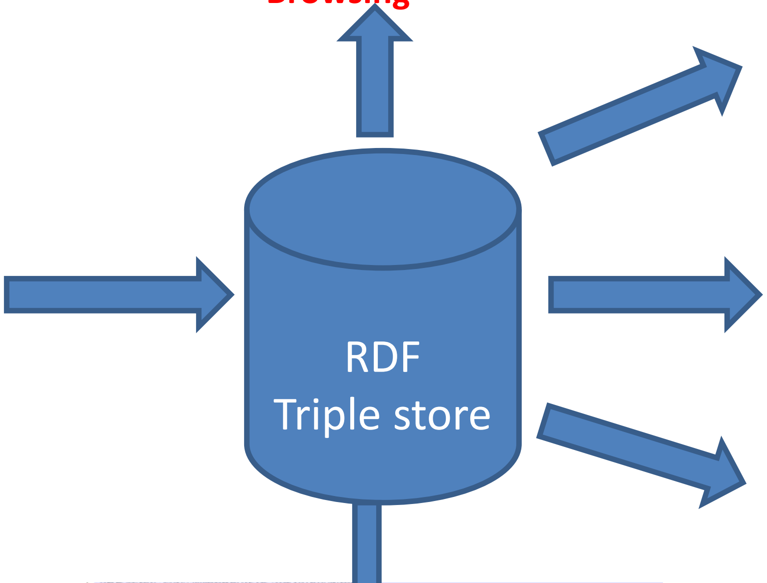


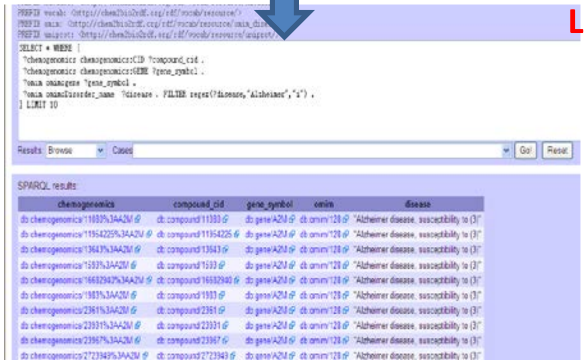
Home | All chemogenomics

Dereferenable URI

Property	Value
chemogenomics:CID	<http://chem2bio2rdf.org/rdf/resource/compound/3973>
is bindingdb_interaction:CID_GENE of	<http://chem2bio2rdf.org/rdf/resource/bindingdb_interaction/23931>
is drugbank_interaction:CID_GENE of	<http://chem2bio2rdf.org/rdf/resource/drugbank_interaction/6920>
is pubchem_bioassay:CID_GENE of	<http://chem2bio2rdf.org/rdf/resource/pubchem_bioassay/4321>
chemogenomics:CID_GENE	3973:PIM1
chemogenomics:GENE	<http://chem2bio2rdf.org/rdf/resource/gene/PIM1>
rdfs:label	chemogenomics #3973:PIM1
rdf:type	vocab:chemogenomics

Browsing



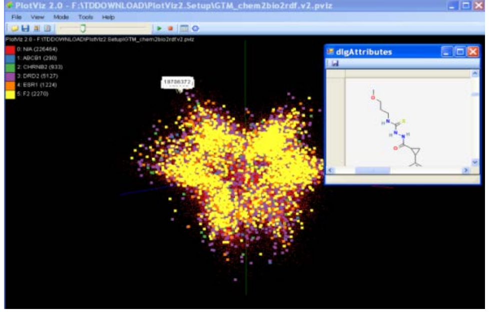


```
PREFIX vocab: <http://chem2bio2rdf.org/rdf/vocab/resource/>
PREFIX omim: <http://chem2bio2rdf.org/rdf/vocab/resource/omim_dise
PREFIX uniprot: <http://chem2bio2rdf.org/rdf/vocab/resource/uniprot/>
SELECT * WHERE {
 ?chemogenomics chemogenomics:CID ?compound_cid .
 ?chemogenomics chemogenomics:GENE ?gene_symbol .
 ?omim omim:gene ?gene_symbol .
 ?omim omim:Disorder_name  ?disease . FILTER regex(?disease,"Alzheimer","i") .
} LIMIT 10
```

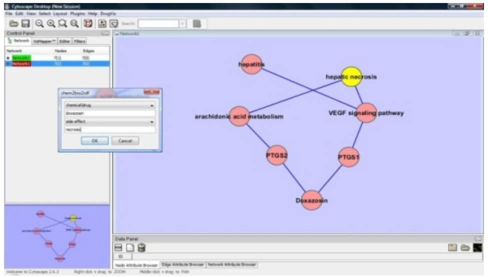
SPARQL results:

chemogenomics	compound_cid	gene_symbol	omim	disease
db:chemogenomics/11083%3AA2M	db:compound/11083	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/11954225%3AA2M	db:compound/11954225	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/13643%3AA2M	db:compound/13643	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/1593%3AA2M	db:compound/1593	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/16682940%3AA2M	db:compound/16682940	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/1983%3AA2M	db:compound/1983	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/2361%3AA2M	db:compound/2361	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/23931%3AA2M	db:compound/23931	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/23967%3AA2M	db:compound/23967	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"
db:chemogenomics/2723949%3AA2M	db:compound/2723949	db:gene/A2M	db:omim/128	"Alzheimer disease, susceptibility to (3)"

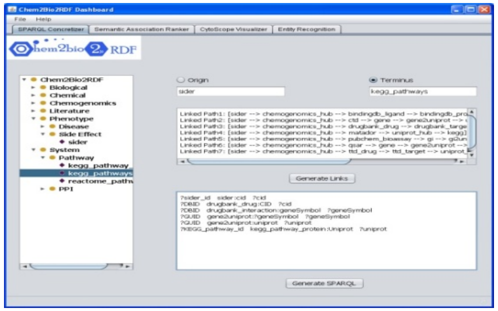
SPARQL ENDPOINTS



PlotViz: Visualization


hepatitis
hepatic necrosis
arachidonic acid metabolism
VEGF signaling pathway
PTGS2
PTGS1
Doxazosin
Welcome to Cytoscape 2.6.3
Right-click + drag to ZOOM
Middle-click + drag to PAN

Cytoscape Plugin



sider

kegg_pathways

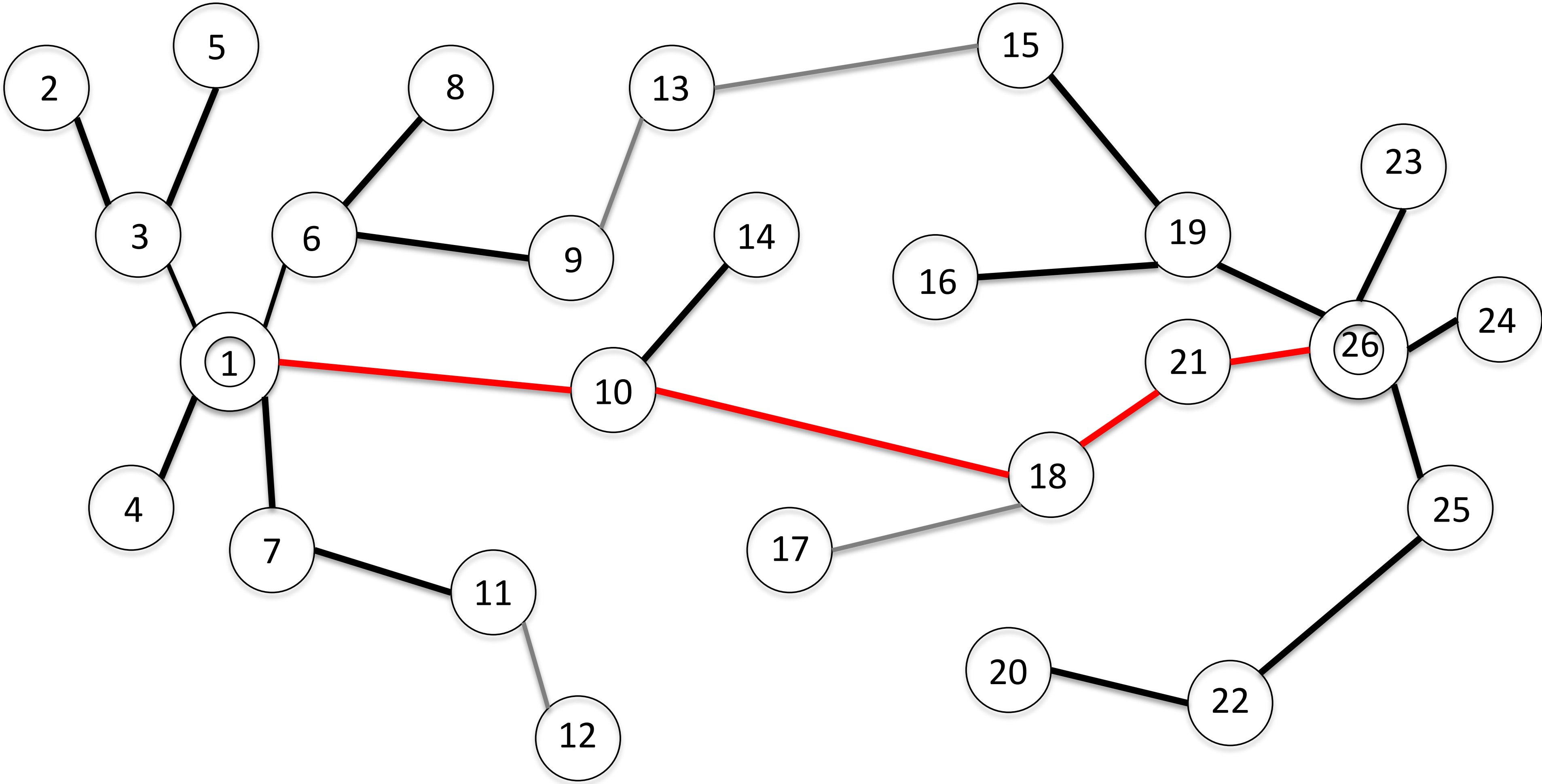
```
Linked Path1: [sider --> chemogenomics_hub --> bindingdb_ligand --> bindingdb_pro
Linked Path2: [sider --> chemogenomics_hub --> ctd --> gene --> gene2uniprot --> 
Linked Path3: [sider --> chemogenomics_hub --> drugbank_drug --> drugbank_targe
Linked Path4: [sider --> chemogenomics_hub --> matador --> uniprot_hub --> kegg]
Linked Path5: [sider --> chemogenomics_hub --> pubchem_bioassay --> gi --> gi2un
Linked Path6: [sider --> chemogenomics_hub --> qsar --> gene --> gene2uniprot -->
Linked Path7: [sider --> chemogenomics_hub --> ttd_drug --> ttd_target --> uniprot_
```

```
?sider_id   sider:cid   ?cid
?DBID   drugbank_drug:CID   ?cid
?DBID   drugbank_interaction:geneSymbol   ?geneSymbol
?GUID   gene2uniprot:?geneSymbol   ?geneSymbol
?GUID   gene2uniprot:uniprot   ?uniprot
?KEGG_pathway_id   kegg_pathway_protein:Uniprot   ?uniprot
```

Linked Path Generation and Ranking

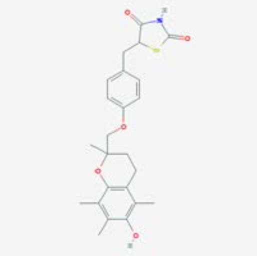
Third party tools

Semantic graph mining: Path Finding Algorithm

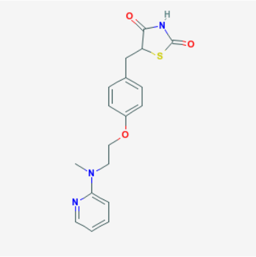


Dijkstra’s algorithm

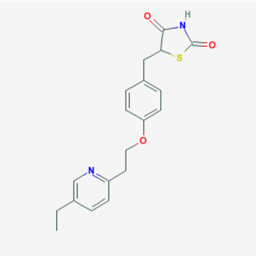
Thiazolinediones (TZDs) – revolutionary treatment for type II Diabetes

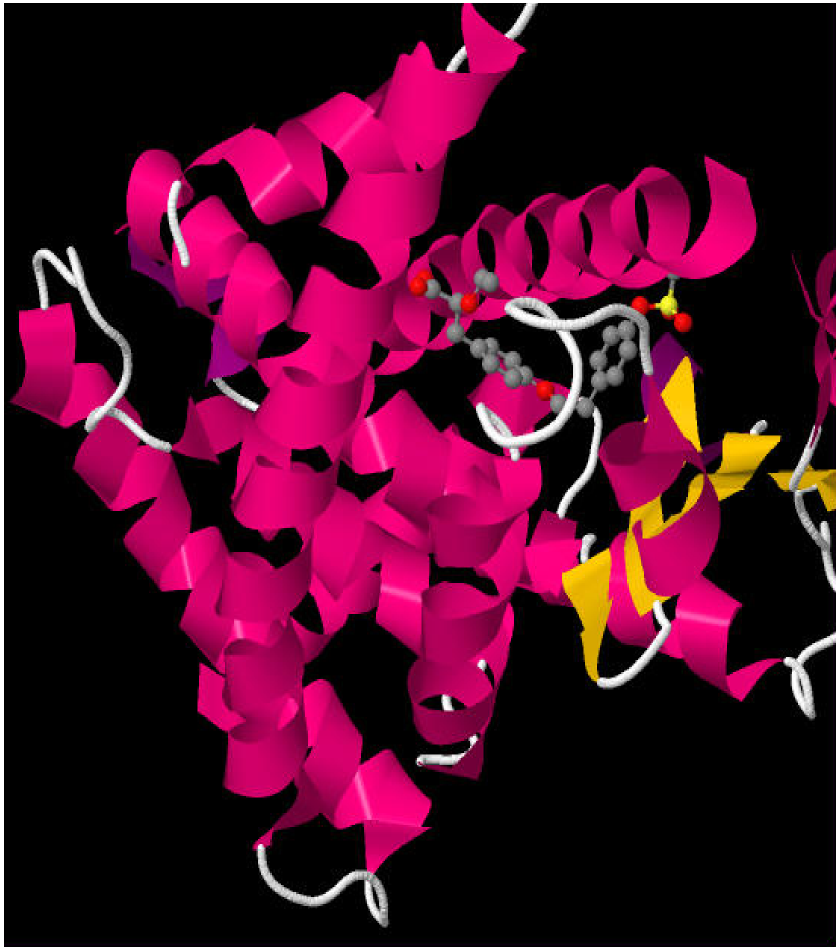


Troglitazone (Rezulin): withdrawn in 2000 (liver disease)



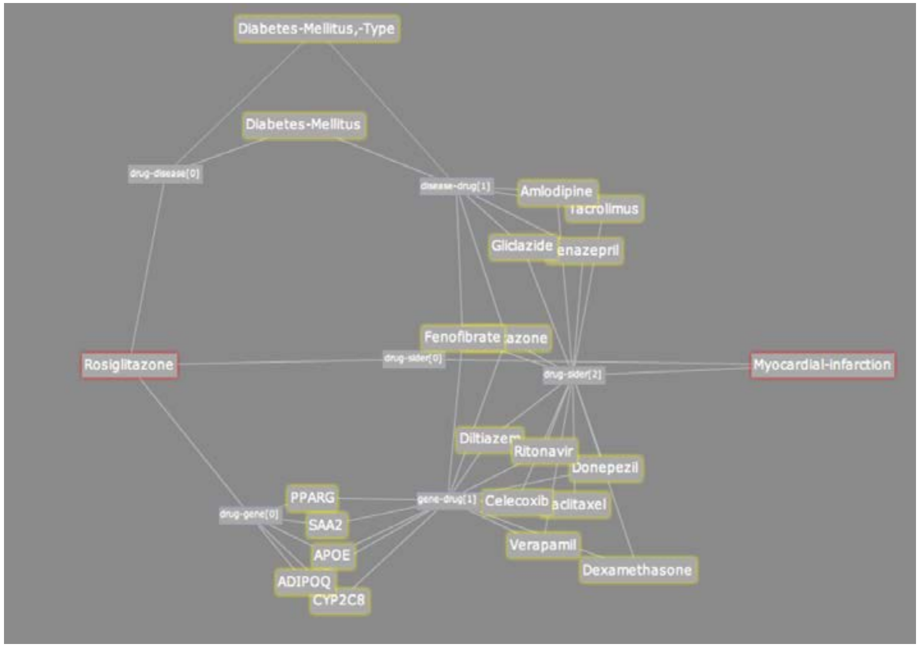
Rosiglitazone (Avandia): restricted in 2010 (cardiac disease)

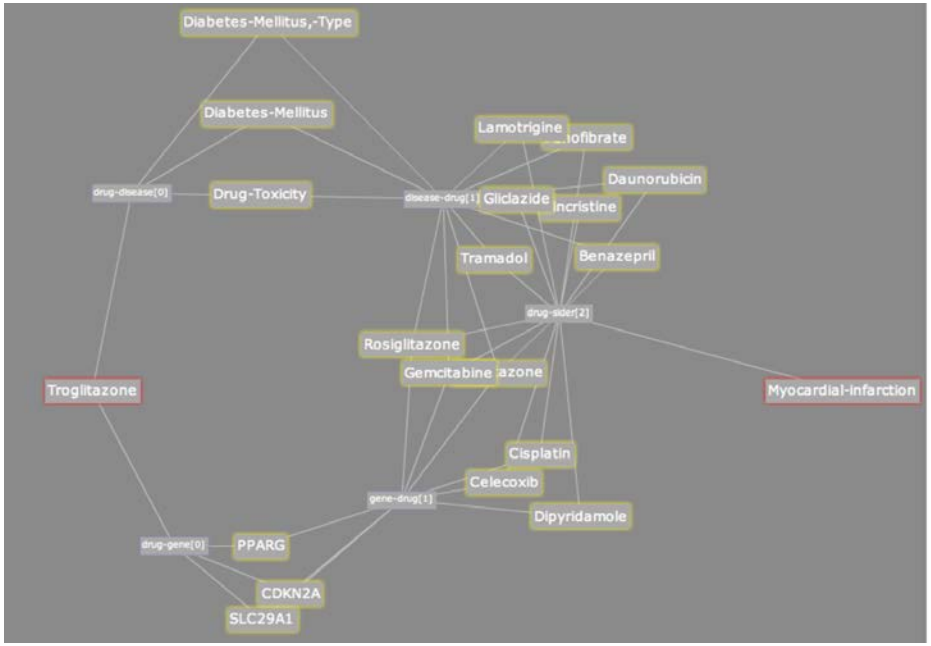




Rosiglitazone bound into PPAR-γ

Pioglitazone: ???? (does decrease blood sugar levels, was associated with bladder tumors and has been withdrawn in some countries.)





PPARG: TZD target

SAA2: Involved in inflammatory response implicated in cardiovascular disease (Current Opinion in Lipidology 15,3,,269-278 2004)

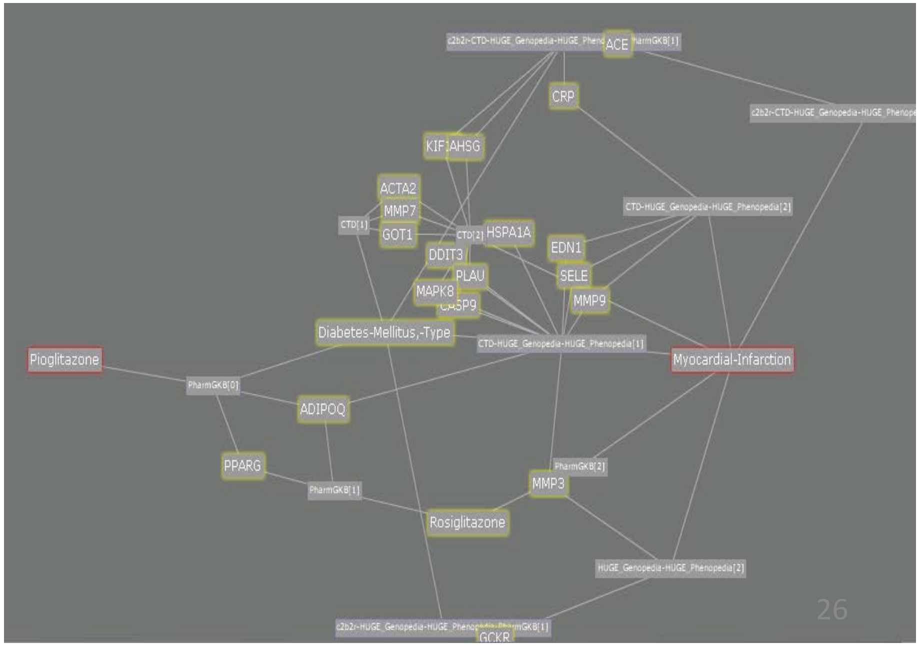
APOE: Apolipoprotein E3 essential for lipoprotein catabolism. Implicated in cardiovascular disease.

ADIPOQ: Adiponectin involved in fatty acid metabolism. Implicated in metabolic syndrome, diabetes and cardiovascular disease

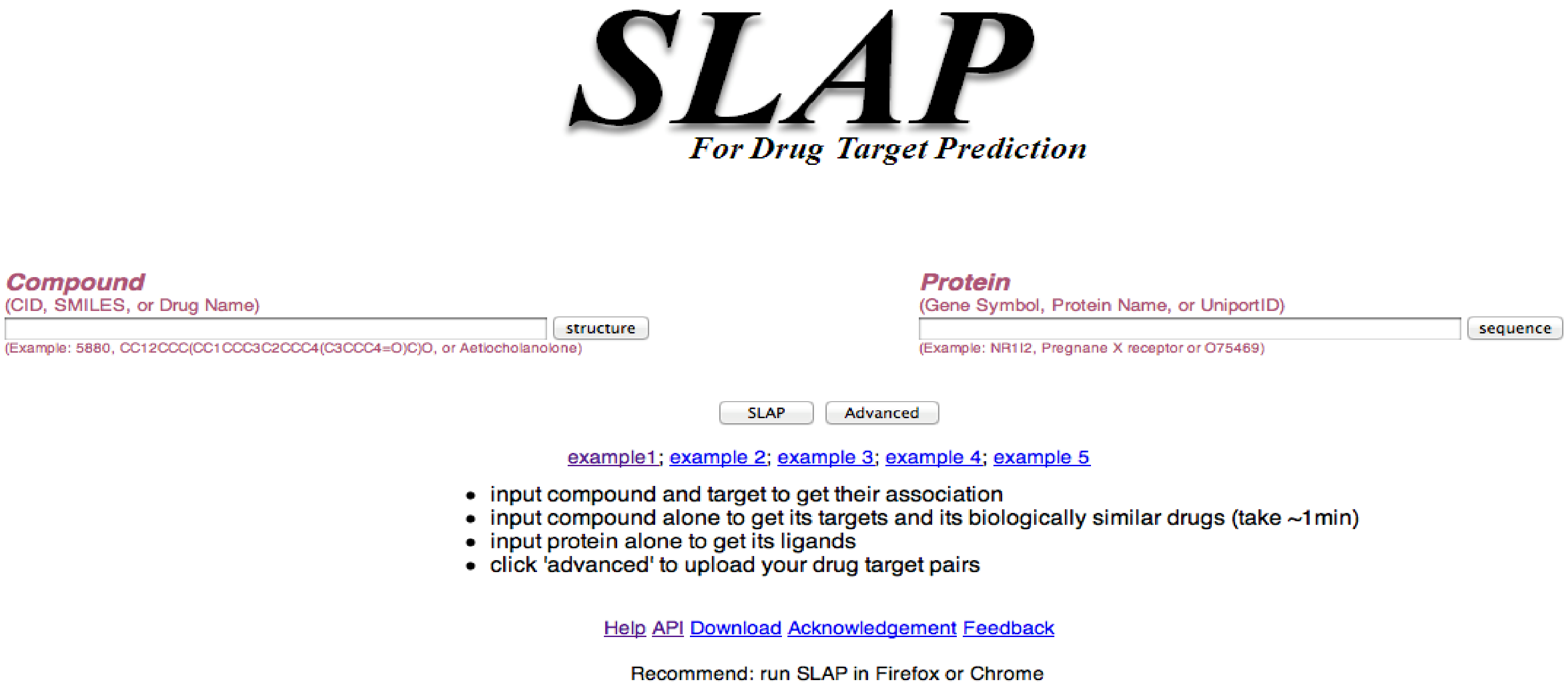
CYP2C8: Cytochrome P450 present in cardiovascular tissue and involved in metabolism of xenobiotics

CDKN2A: Tumor suppression gene

SLC29A1: Membrane transporter



Semantic Prediction
http://chem2bio2rdf.org/slap


SLAP
For Drug Target Prediction
structure
sequence
SLAP
Advanced
example1; example 2; example 3; example 4; example 5
Help API Download Acknowledgement Feedback

Compound
(CID, SMILES, or Drug Name)

(Example: 5880, CC12CCC(CC1CCC3C2CCC4(C3CCC4=O)C)O, or Aetiocholanolone)

Protein
(Gene Symbol, Protein Name, or UniportID)

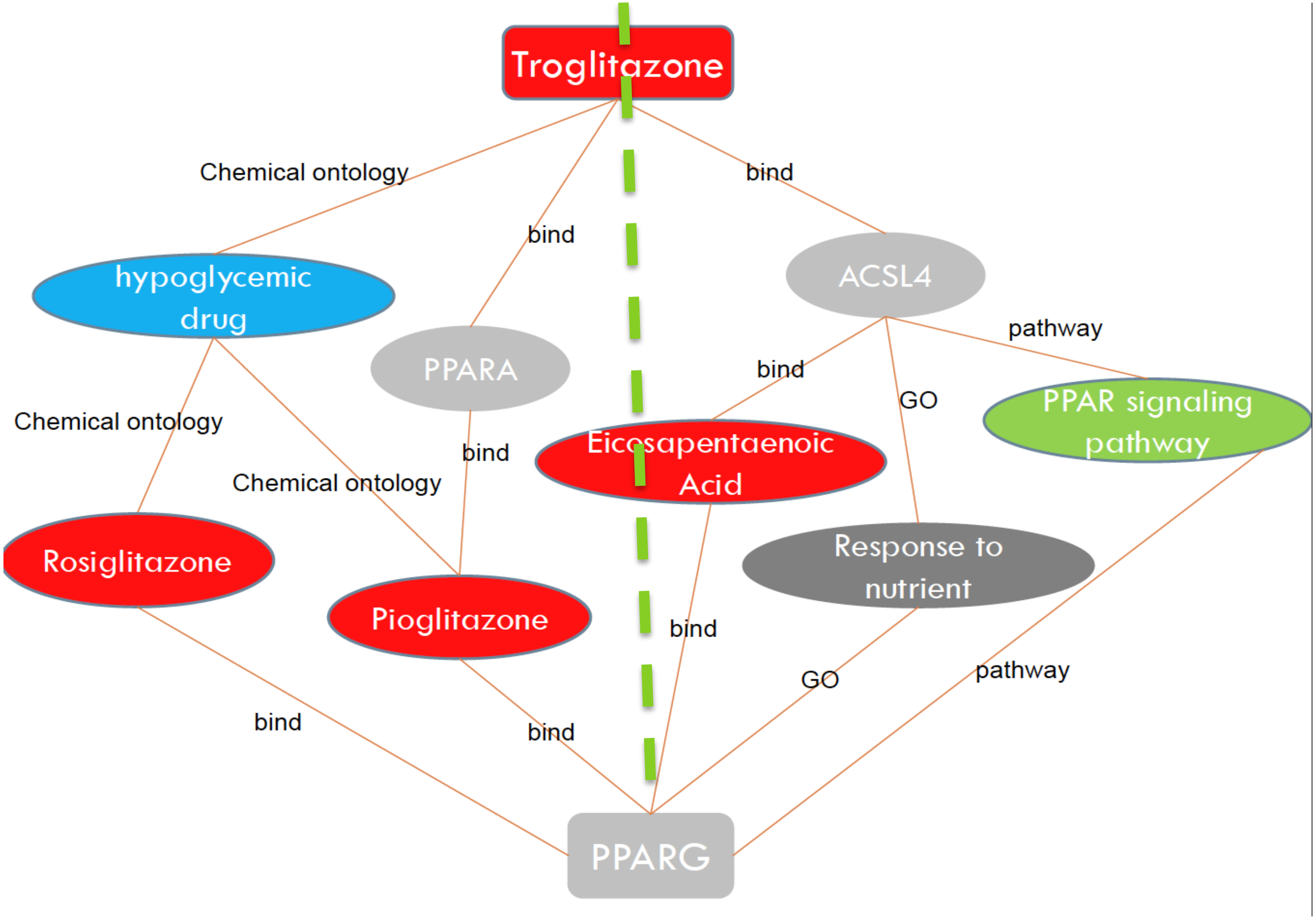
(Example: NR1I2, Pregnane X receptor or O75469)

- input compound and target to get their association
- input compound alone to get its targets and its biologically similar drugs (take ~1min)
- input protein alone to get its ligands
- click 'advanced' to upload your drug target pairs

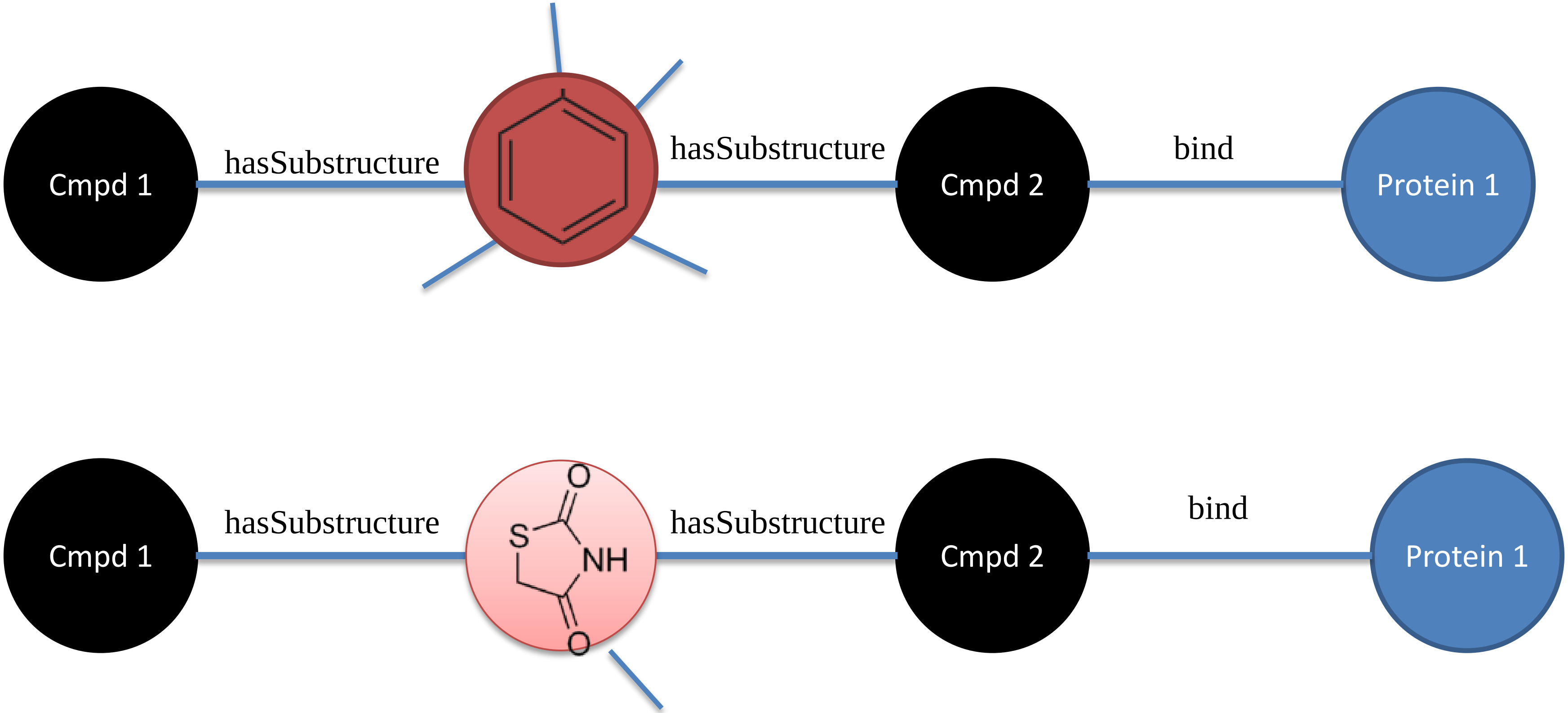
Recommend: run SLAP in Firefox or Chrome

Chen, B., Ding, Y., & Wild, D. (2012). Assessing Drug Target Association using Semantic Linked Data. *PLoS Computational Biology*, 8(7): e1002574. doi:10.1371/journal.pcbi.1002574,

Example: Troglitazone and PPARG



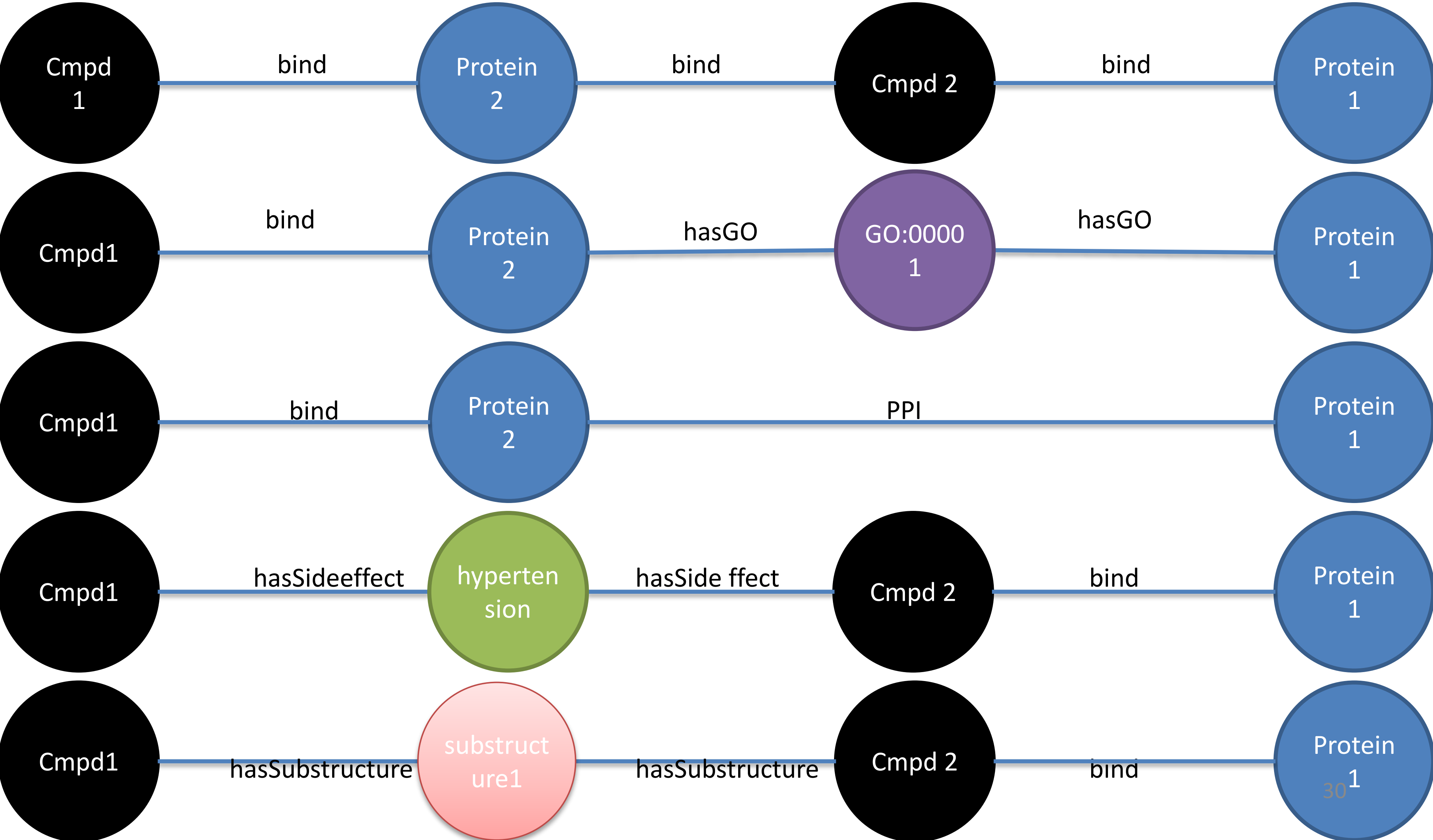
Topology is important for association



Cmpd 1 — hasSubstructure — hasSubstructure — Cmpd 2 — bind — Protein 1

Cmpd 1 — hasSubstructure — hasSubstructure — Cmpd 2 — bind — Protein 1

Semantics is important for association

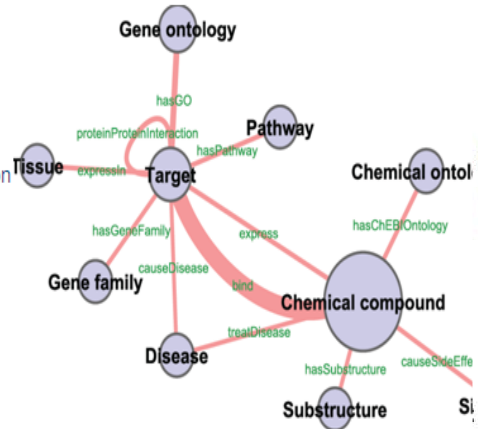


SLAP Pipeline

PubChem	ChEBI	DrugBank
UniProt	UniProtKB-GOA	HGNC
SIDER	OMIM	KEGG
HPRD	ChEMBL	TTD
BindingDB	CTD	PDSP

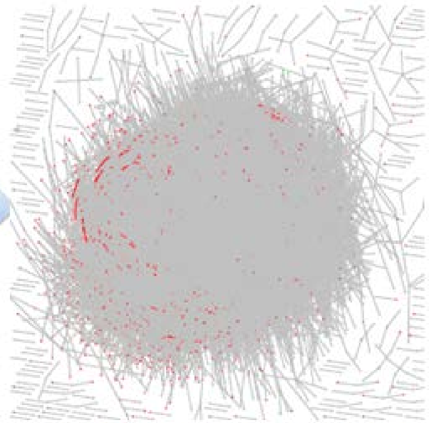
(a) Raw Data Sets

Semantic annotation



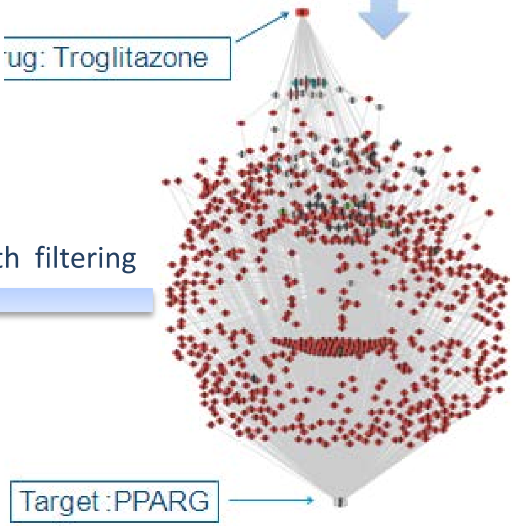
(b) Ontological level schema

Integration



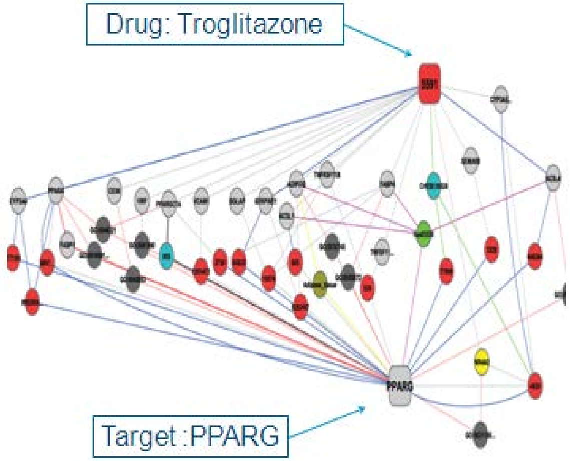
(c) Semantic Linked Data

Path finding



(d) Paths (length <4) between Two Nodes

Path filtering



(e) Significant Paths between Two Nodes

Association Assessing

1. Edge weight:

$$p(e(i \rightarrow j)) = \frac{1}{\sum_{k}^{n=1} R_{i,n} == R_{i,j}}$$

2. Path score:

$$p(P_l(t \rightarrow s)) = p(P_l(e_{m\rightarrow m-1}, ..., e_{3\rightarrow 2}, e_{2\rightarrow 1})) = \prod_{i=1}^{m-1} e_{i+1\rightarrow i}$$

$$log(p(P_l(t \rightarrow s))) = \sum_{i=1}^{m-1} log(e_{i+1\rightarrow i})$$

3. Association score

$$raw\ score(s,t) = \sum_{l}^{n} \frac{log(p(P_l)) - \theta(log(P_l))}{\sigma(log(P_l))}$$

(f) Statistical Models

Cross-check with SEA

- SEA analysis (Nature 462, 175-181, 2009) predicts 184 new compound-target pairs, 30 of which were experimentally tested
- 23 of these pairs were experimentally validated (<15uM) including 15 aminergic GPCR targets and 8 which crossed major receptor classification boundaries
- 9 of the aminergic GPCR target pairings were correctly predicted by SLAP (p<0.05) – for the other 6 compounds were not present in our set
- 1 of the 8 cross-boundary pairs was predicted

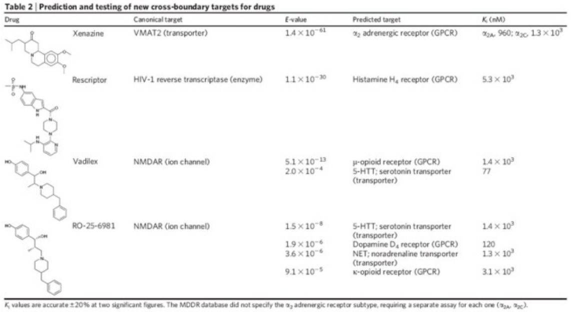


Table 2 | Prediction and testing of new cross-boundary targets for drugs

Drug	Canonical target	*E*-value	Predicted target	K_i (nM)
Xenazine	VMAT2 (transporter)	1.4×10^{-61}	α_2 adrenergic receptor (GPCR)	α_{2A}, 960; α_{2C}, 1.3×10^3
Rescriptor	HIV-1 reverse transcriptase (enzyme)	1.1×10^{-30}	Histamine H_4 receptor (GPCR)	5.3×10^3
Vadilex	NMDAR (ion channel)	5.1×10^{-13}	μ-opioid receptor (GPCR)	1.4×10^3
		2.0×10^{-4}	5-HTT; serotonin transporter (transporter)	77
RO-25-6981	NMDAR (ion channel)	1.5×10^{-8}	5-HTT; serotonin transporter (transporter)	1.4×10^3
		1.9×10^{-6}	Dopamine D_4 receptor (GPCR)	120
		3.6×10^{-6}	NET; noradrenaline transporter (transporter)	1.3×10^3
		9.1×10^{-5}	κ-opioid receptor (GPCR)	3.1×10^3

K_i values are accurate ±20% at two significant figures. The MDDR database did not specify the α_2 adrenergic receptor subtype, requiring a separate assay for each one (α_{2A}, α_{2C}).

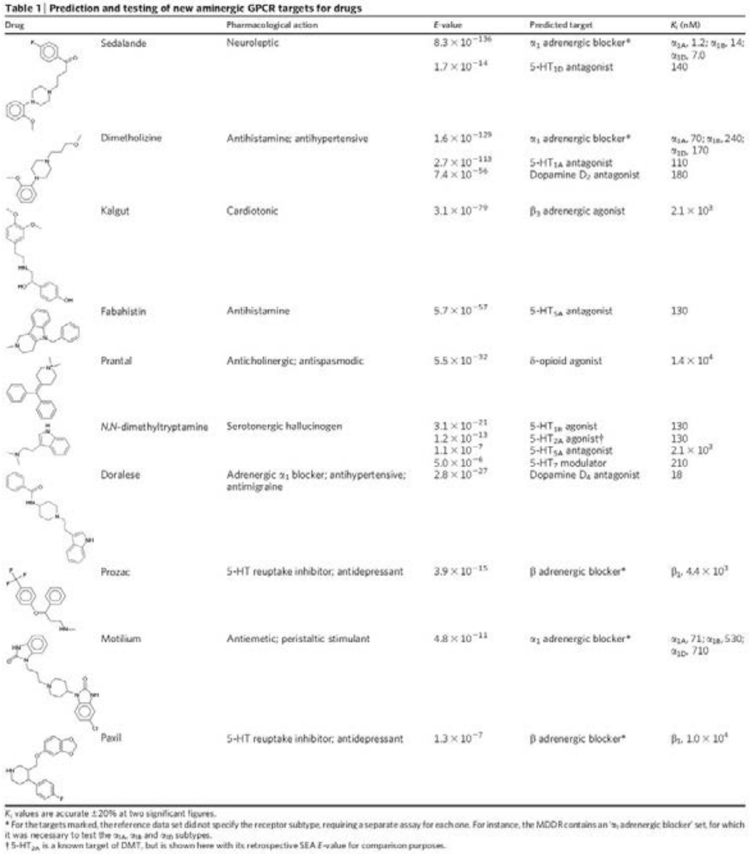


Table 1 | Prediction and testing of new aminergic GPCR targets for drugs

Drug	Pharmacological action	*E*-value	Predicted target	K_i (nM)
Sedalande	Neuroleptic	8.3×10^{-136}	α_1 adrenergic blocker*	α_{1A}, 1.2; α_{1B}, 14; α_{1D}, 7.0
		1.7×10^{-14}	5-HT_{1D} antagonist	140
Dimetholizine	Antihistamine; antihypertensive	1.6×10^{-129}	α_1 adrenergic blocker*	α_{1A}, 70; α_{1B}, 240; α_{1D}, 170
		2.7×10^{-113}	5-HT_{1A} antagonist	110
		7.4×10^{-56}	Dopamine D_2 antagonist	180
Kalgut	Cardiotonic	3.1×10^{-79}	β_3 adrenergic agonist	2.1×10^3
Fabahistin	Antihistamine	5.7×10^{-57}	5-HT_{5A} antagonist	130
Prantal	Anticholinergic; antispasmodic	5.5×10^{-32}	δ-opioid agonist	1.4×10^4
N,N-dimethyltryptamine	Serotonergic hallucinogen	3.1×10^{-21}	5-HT_{1B} agonist	130
		1.2×10^{-13}	5-HT_{2A} agonist†	130
		1.1×10^{-7}	5-HT_{5A} antagonist	2.1×10^3
		5.0×10^{-6}	5-HT_7 modulator	210
Doralese	Adrenergic α_1 blocker; antihypertensive; antimigraine	2.8×10^{-27}	Dopamine D_4 antagonist	18
Prozac	5-HT reuptake inhibitor; antidepressant	3.9×10^{-15}	β adrenergic blocker*	β_1, 4.4×10^3
Motilium	Antiemetic; peristaltic stimulant	4.8×10^{-11}	α_1 adrenergic blocker*	α_{1A}, 71; α_{1B}, 530; α_{1D}, 710
Paxil	5-HT reuptake inhibitor; antidepressant	1.3×10^{-7}	β adrenergic blocker*	β_1, 1.0×10^4

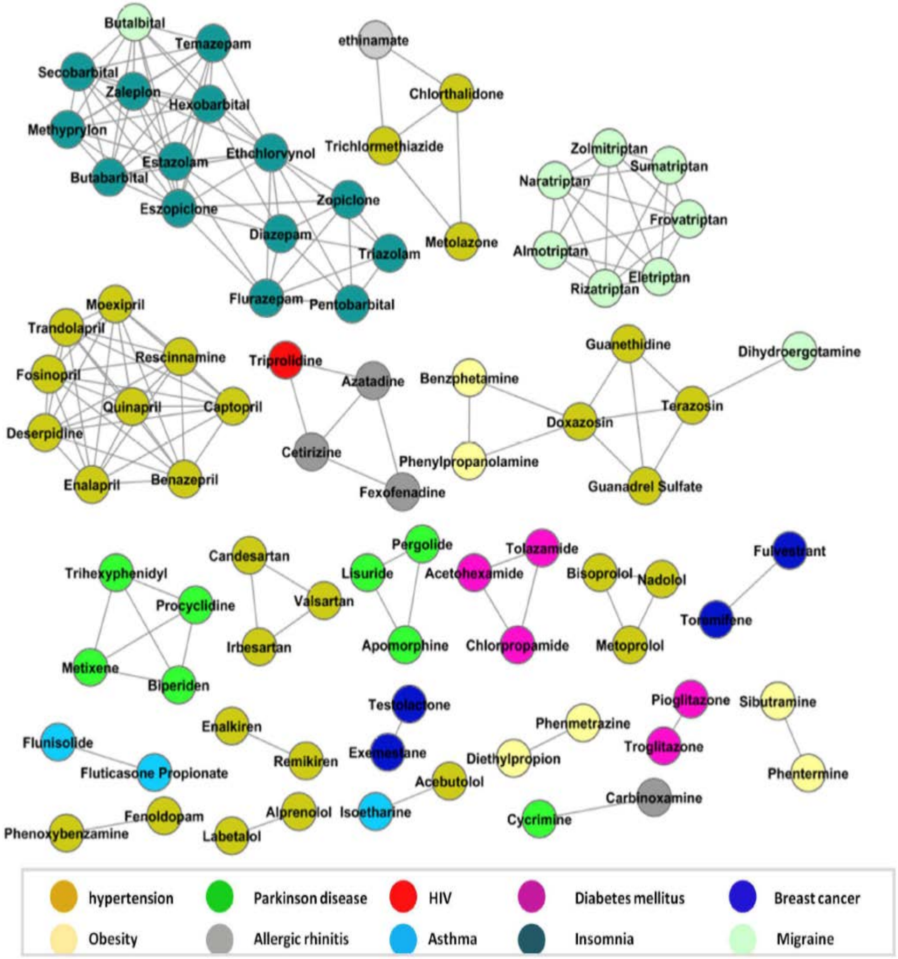
K_i values are accurate ±20% at two significant figures.

* For the targets marked, the reference data set did not specify the receptor subtype, requiring a separate assay for each one. For instance, the MDDR contains an 'α_1 adrenergic blocker' set, for which it was necessary to test the α_{1A}, α_{1B} and α_{1D} subtypes.

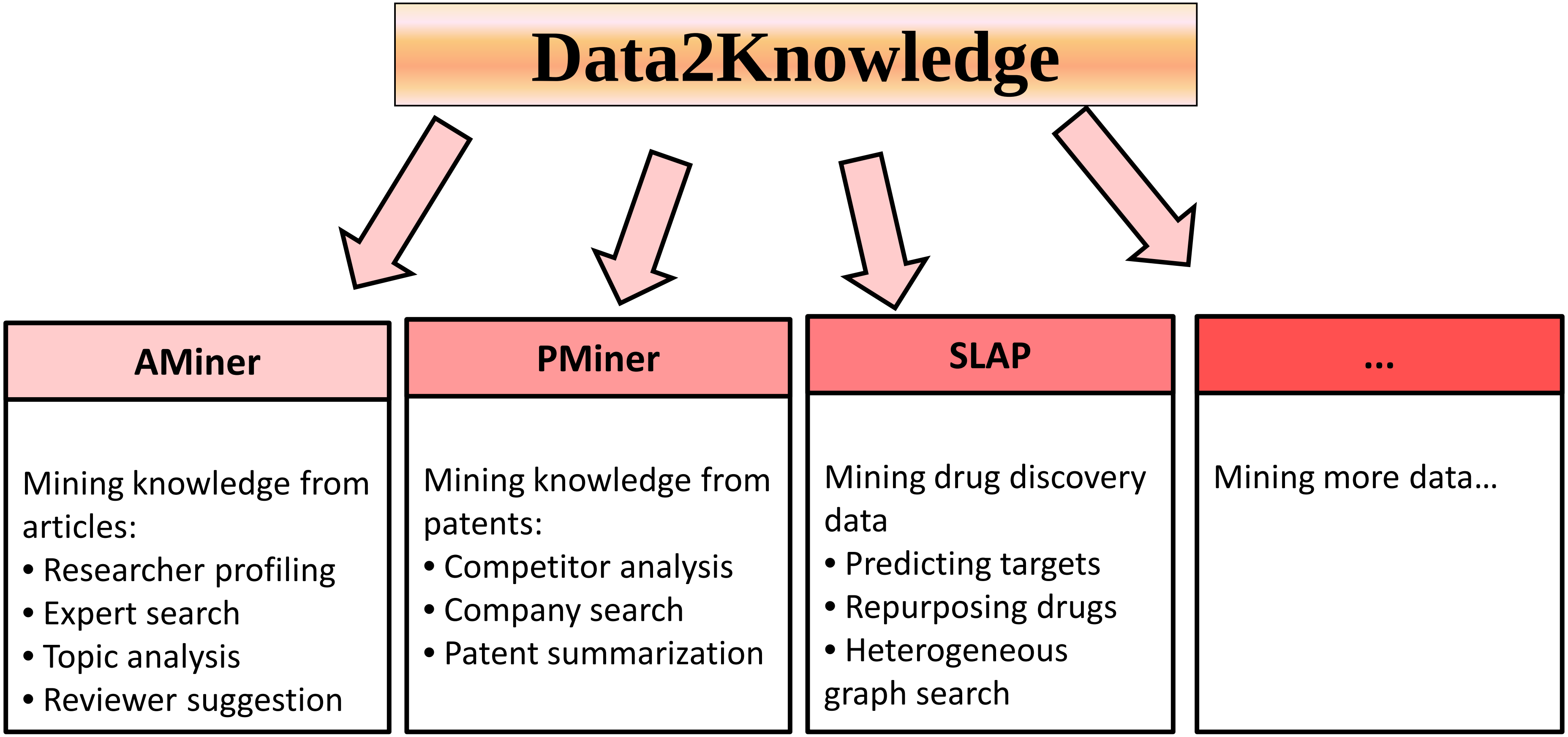
† 5-HT_{2A} is a known target of DMT, but is shown here with its retrospective SEA *E*-value for comparison purposes.

Assessing drug similarity from biological function

- Took 157 drugs with 10 known therapeutic indications, and created SLAP profiles against 1,683 human targets
- Pearson correlation between profiles > 0.9 from SLAP was used to create associations between drugs
- Drugs with the same therapeutic indication unsurprisingly cluster together
- Some drugs with similar profile have different indications – potential for use in drug repurposing?



Data2Knowledge platform…



Data2Knowledge

AMiner

Mining knowledge from articles:

- Researcher profiling
- Expert search
- Topic analysis
- Reviewer suggestion

PMiner

Mining knowledge from patents:

- Competitor analysis
- Company search
- Patent summarization

SLAP

Mining drug discovery data

- Predicting targets
- Repurposing drugs
- Heterogeneous graph search

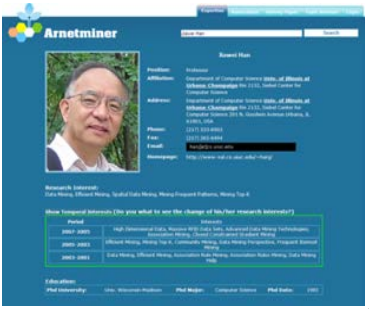
…

Mining more data…

AMiner

Researchers: 31,222,410
Publications: 69,962,333
Conferences/Journals: 330,236
Citations: 133,196,029
Knowledge Concepts: 7,854,301

- Research profiling
- Integration
- Interest analysis



Jiawei Han

Position: Professor
Affiliation: Department of Computer Science Univ. of Illinois at Urbana-Champaign Rm 2132, Siebel Center for Computer Science
Address: Department of Computer Science Univ. of Illinois at Urbana-Champaign Rm 2132, Siebel Center for Computer Science 201 N. Goodwin Avenue Urbana, IL 61801, USA
Phone: (217) 333-6903
Fax: (217) 265-6494
Email: hanj@cs.uiuc.edu
Homepage: http://www-sal.cs.uiuc.edu/~hanj/

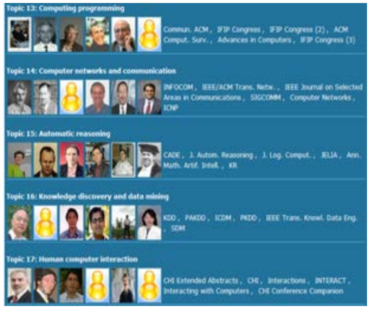
Research Interest:
Data Mining, Efficient Mining, Spatial Data Mining, Mining Frequent Patterns, Mining Top-K

Show Temporal Interests (Do you what to see the change of his/her research interests?)

Period	Interests
2007-2005	High Dimensional Data, Massive RFID Data Sets, Advanced Data Mining Technologies, Association Mining, Closed Constrained Gradient Mining
2005-2003	Efficient Mining, Mining Top-K, Community Mining, Data Mining Perspective, Frequent Itemset Mining
2003-2001	Data Mining, Efficient Mining, Association Rule Mining, Association Rules Mining, Data Mining Help

Education:
PhD University: Univ. Wisconsin-Madison PhD Major: Computer Science PhD Date: 1985

- Topic analysis
- Course search
- Expert search



Topic 13: Computing programming
Commun. ACM , IFIP Congress , IFIP Congress (2) , ACM Comput. Surv. , Advances in Computers , IFIP Congress (3)

Topic 14: Computer networks and communication
INFOCOM , IEEE/ACM Trans. Netw. , IEEE Journal on Selected Areas in Communications , SIGCOMM , Computer Networks , ICNP

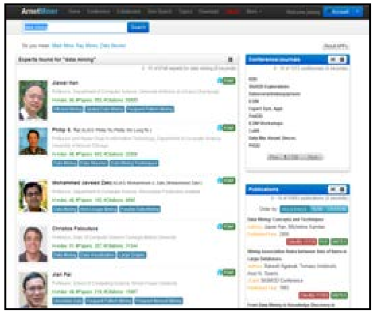
Topic 15: Automatic reasoning
CADE , J. Autom. Reasoning , J. Log. Comput. , JELIA , Ann. Math. Artif. Intell. , KR

Topic 16: Knowledge discovery and data mining
KDD , PAKDD , ICDM , PKDD , IEEE Trans. Knowl. Data Eng. , SDM

Topic 17: Human computer interaction
CHI Extended Abstracts , CHI , Interactions , INTERACT , Interacting with Computers , CHI Conference Companion

- Association
- Disambiguation
- Suggestion

- Geo search
- Collaboration recommendation



What is PMiner?

- Current patent analysis systems focus on search
  - Google Patent, WikiPatent, FreePatentsOnline

- PMiner is designed for an ***in-depth*** analysis of patent activity at the topic-level
  - Topic-driven modeling of patents
  - Heterogeneous network co-ranking
  - Intelligent competitive analysis
  - Patent summarization

* Patent data:
> 3.8M patents
> 2.4M inventors
> 400K companies
> 10M citation relationships

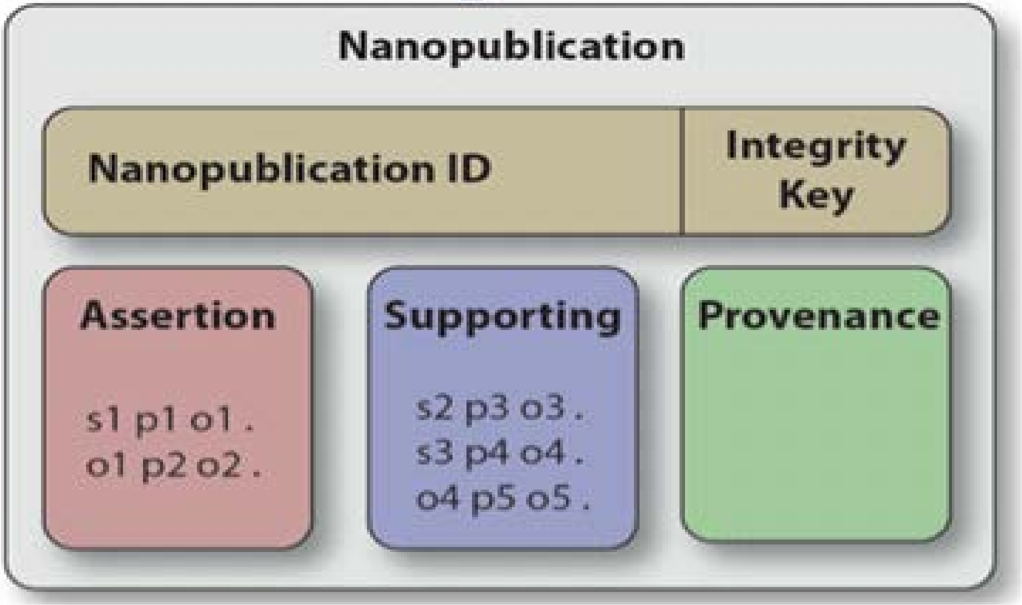
* Journal data:
> 2k journal papers
> 3.7k authors

The crawled data is increasing to >300 Gigabytes.

J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, and A. K. Usadi. PatentMiner: Topic-driven Patent Analysis and Mining. KDD'12, pp. 1366-1374.

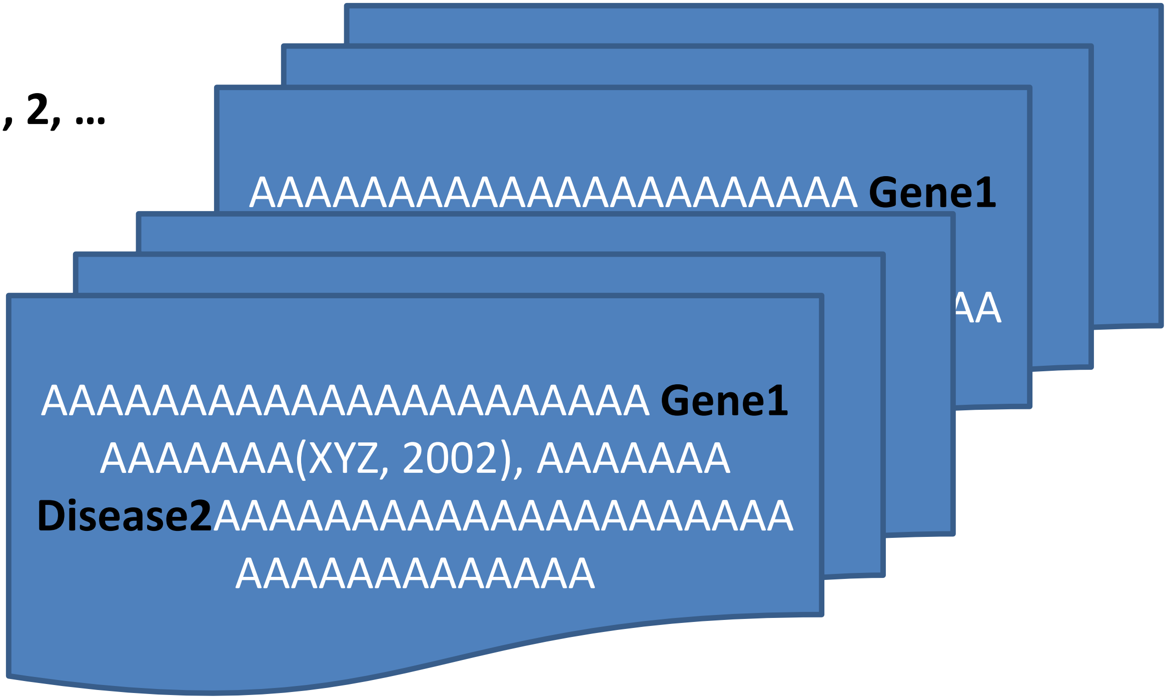
Semantic Publishing

- Turn literature knowledge into actionable data to generate more powerful knowledge



http://nanopub.org/

Paper 1, 2, …



AAAAAAAAAAAAAAAAAAAAAA **Gene1**

AA

AAAAAAAAAAAAAAAAAAAAAA **Gene1**
AAAAAAA(XYZ, 2002), AAAAAAA
Disease2AAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAA

Paper n: Gene 1, Disease 2, (XYZ, 2002)

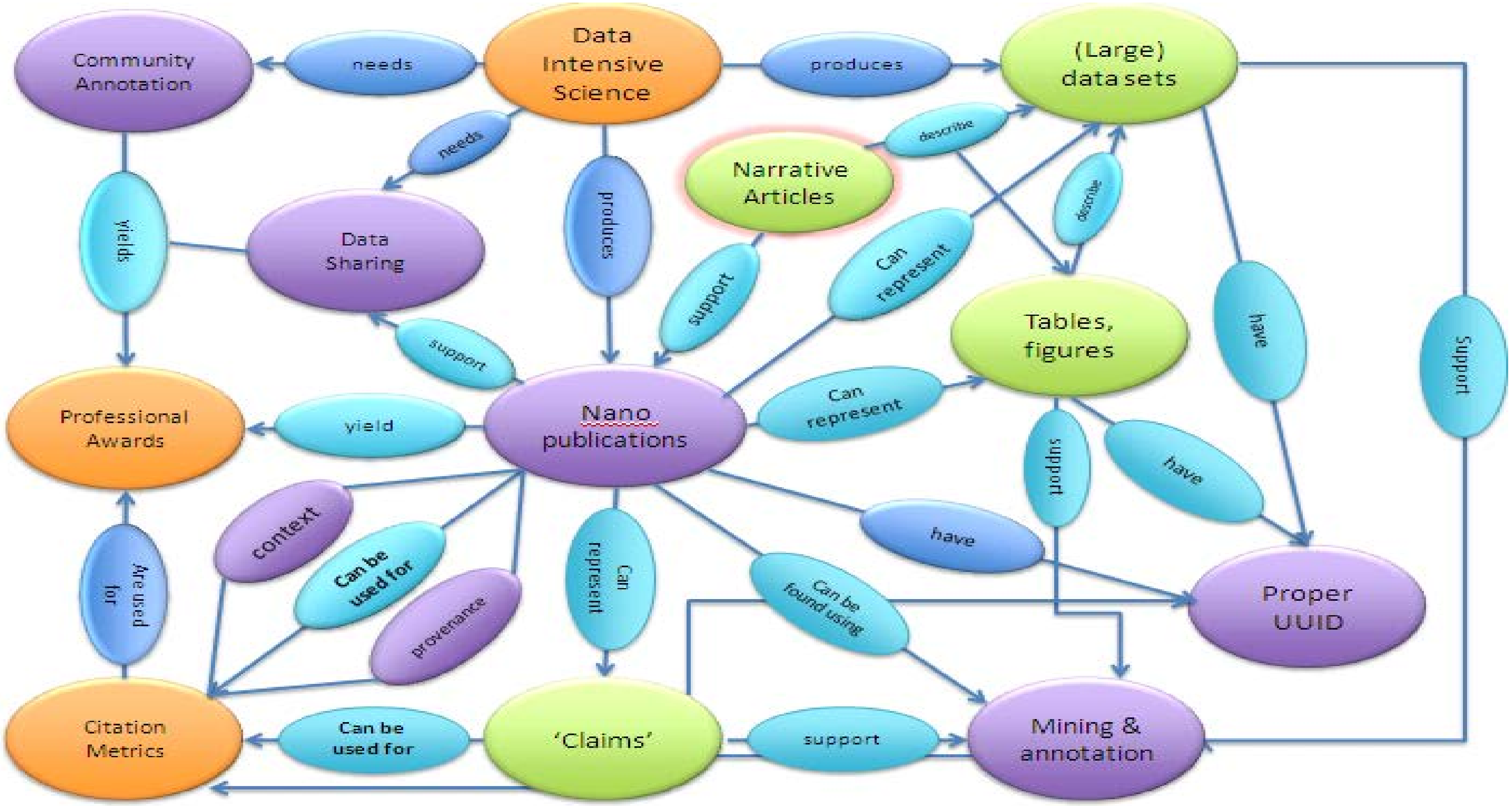
……

Paper 2: Gene 1, Disease 2, (XYZ, 2002)

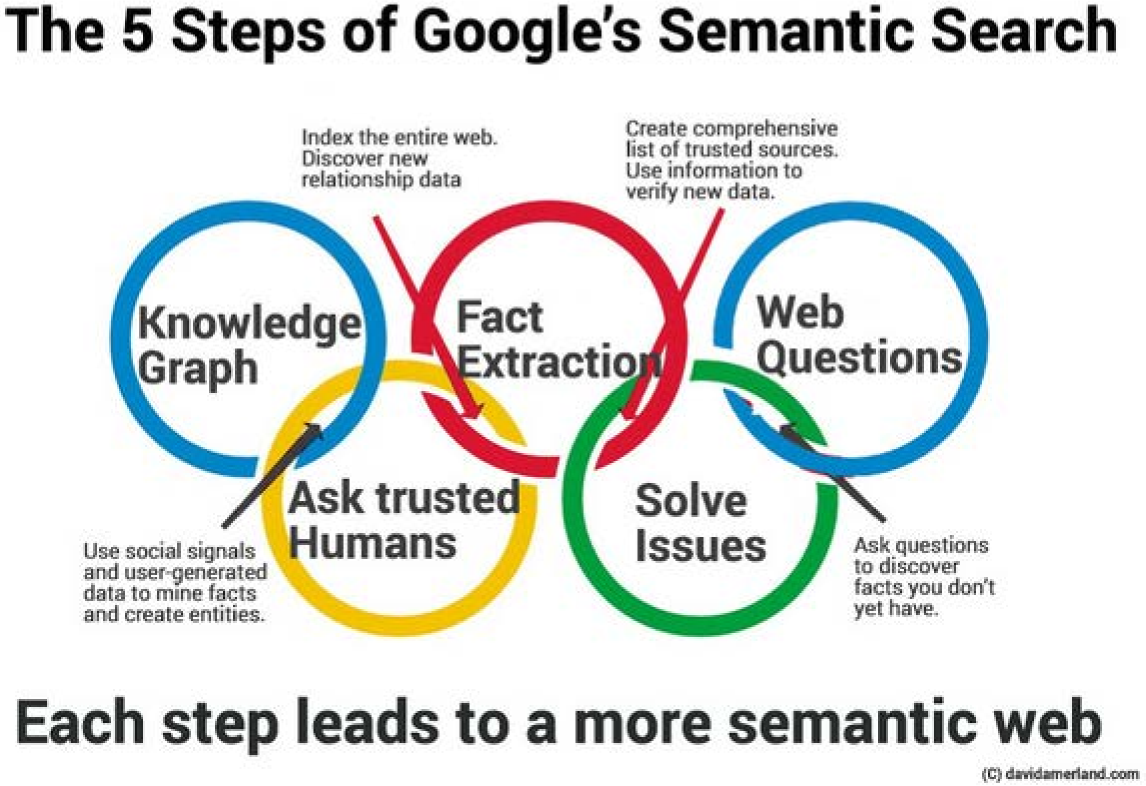
Paper 1: Gene 1, Disease 2, (XYZ, 2002)

Demonstrating strong evidence of the connection/relationship between concept Gene 1 and Disease 2, or concept GENE and DISEASE

Knowledge Graph



Semantic Search



The 5 Steps of Google's Semantic Search

Index the entire web. Discover new relationship data

Create comprehensive list of trusted sources. Use information to verify new data.

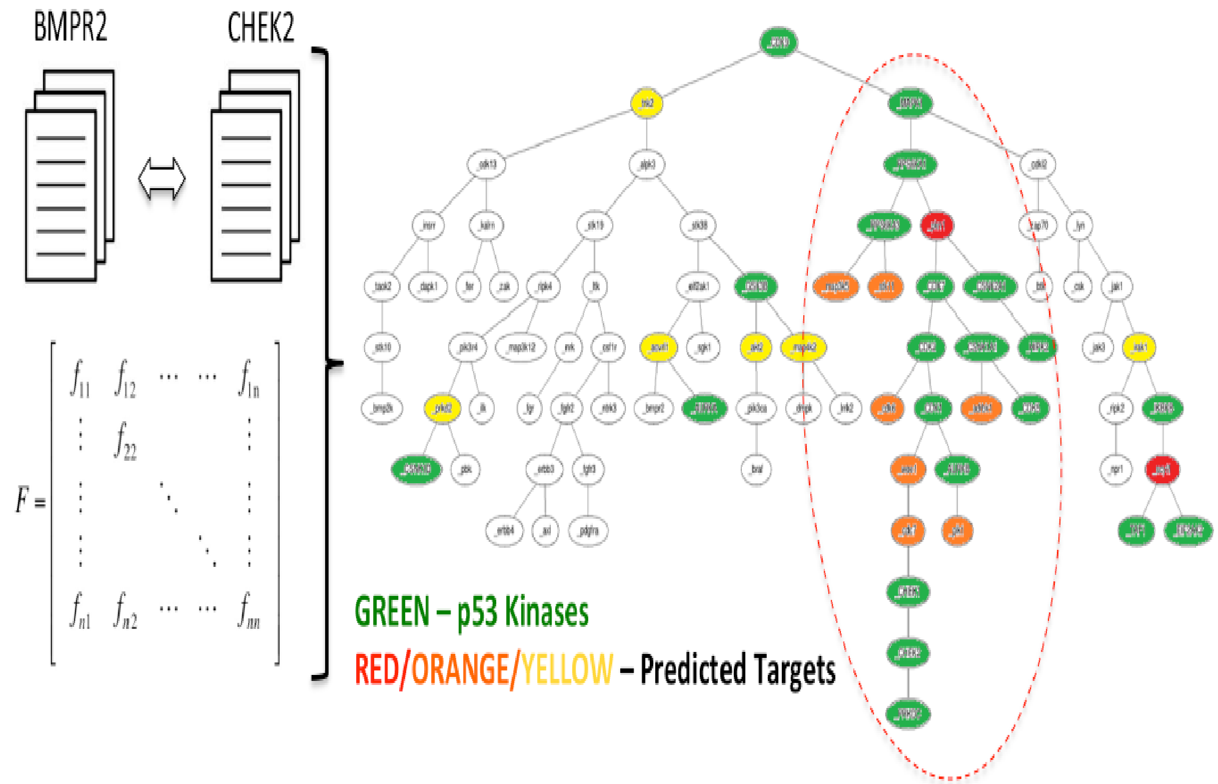
Use social signals and user-generated data to mine facts and create entities.

Ask questions to discover facts you don't yet have.

Each step leads to a more semantic web

(C) davidamerland.com

Digital Discovery


BMPR2
CHEK2
$F =$
f_{11} f_{12} … … f_{1n}
f_{22}
f_{n1} f_{n2} … … f_{nn}

GREEN – p53 Kinases

RED/ORANGE/YELLOW – Predicted Targets

Figure 1 Kinases are clustered based on their literature distance. The clustered p53 kinases (green) suggest new kinases that may also phosphorylate p53.

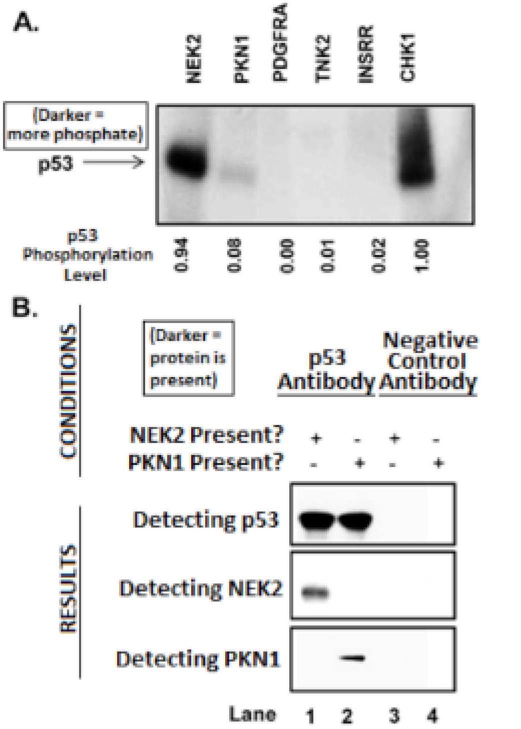


Figure 6 Experimental validation of candidate p53 kinases as *bona fide* p53 kinases. (A) *In vitro* kinase assay demonstrates phosphorylation of p53 by top ranked candidate kinases PKN1 and NEK2. Relative levels of p53 phosphorylation are indicated for each kinase normalized to positive control CHK1. Though the signal is weak for PNK1, subsequent experiments lend further support. (B) PKN1 and NEK2 shown to interact with p53 *in vivo*. A p53 antibody isolates p53 and any proteins bound to it. Antibodies detect the presence of candidate kinases in this isolate.

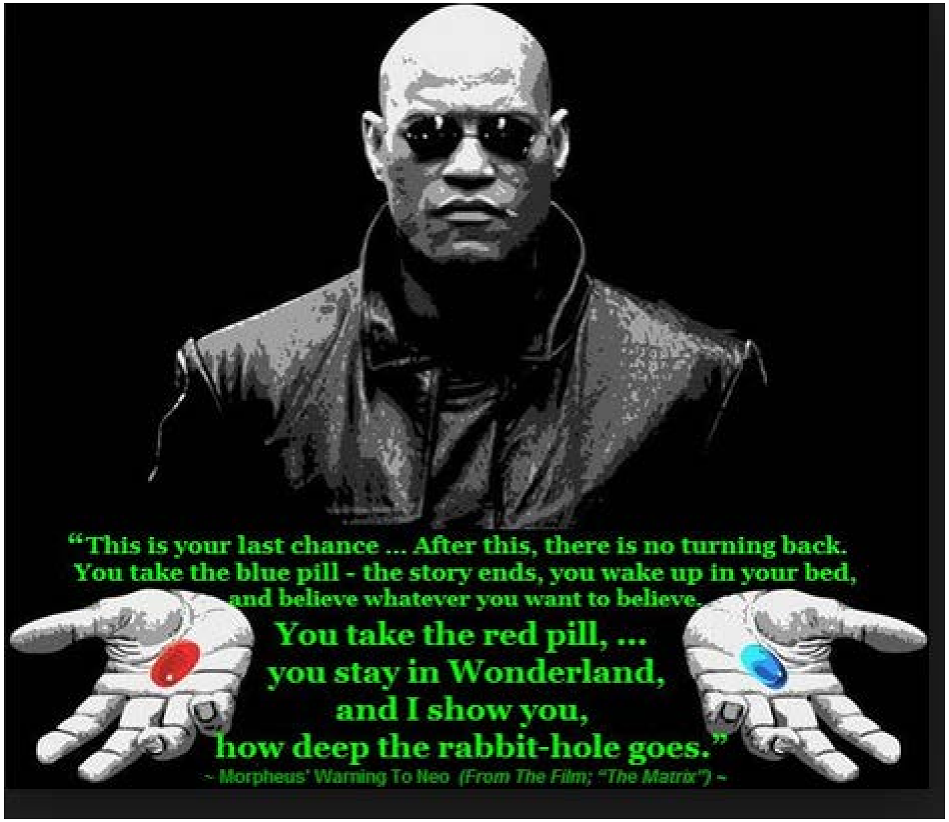
Spangler, S., Wilkins, A.D., Bachman, B. J., et al.(2014). Automated hypothesis generation based on mining scientific literature. Proceedings of the 20th *ACM SIGKDD*, August 24-27, 2014, New York, USA.

More

- Data-Driven Discovery
  - Medicine
  - Neuroscience
  - Math and Material science
  - Social science (education, business, poverty reduction)
  - Digital humanities and Arts (distant reading, digital painting, digital recipes, computational creativity (story, joke and poetry generation)
- Digital Creativity

Future of Knowledge

Questions?


"This is your last chance ... After this, there is no turning back.
You take the blue pill - the story ends, you wake up in your bed,
and believe whatever you want to believe.
You take the red pill, ...
you stay in Wonderland,
and I show you,
how deep the rabbit-hole goes."
~ Morpheus' Warning To Neo (From The Film; "The Matrix") ~

The Matrix