

# Integrated Pipeline for Systems Pharmacology in R/Bioconductor

*Nan Xiao @road2stat*  
*7th China R Beijing*



COMPUTATIONAL BIOLOGY &  
DRUG DESIGN GROUP  
CENTRAL SOUTH UNIV., CHINA

2009

Vulnerability & Security

2013

Web Scraping with R

2014

Systems Pharmacology



时间都去哪儿了



**TIME FLIES... 1994-2009**





**TIME FLIES... 2009-2014**





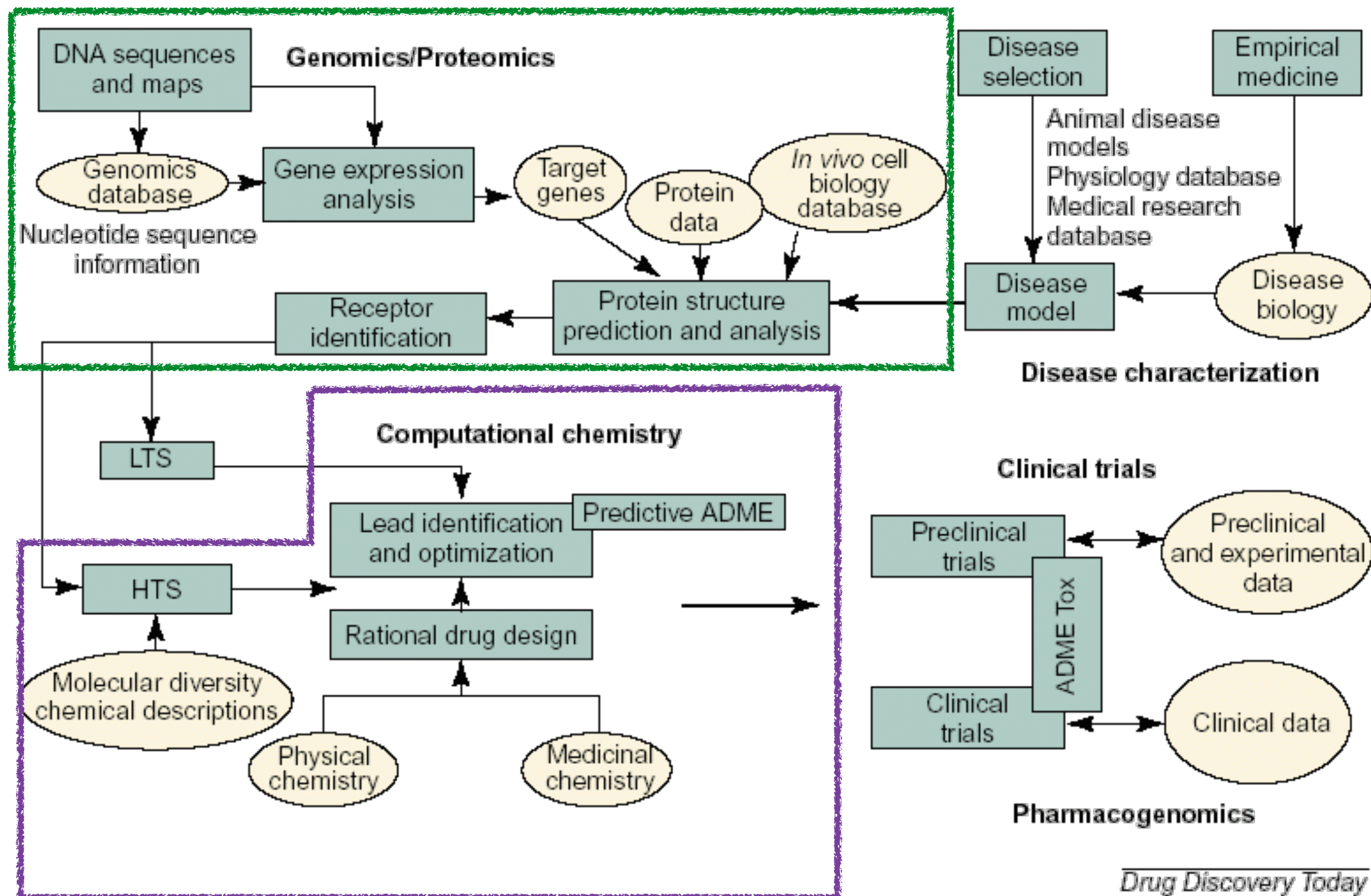
# Outline

- Intro: Systems Pharmacology
- Pipeline: Our solution with R
- Case study: Identify novel drug-ADR associations

# Part I

# Systems Pharmacology

# Flow of Information in a Drug Discovery Pipeline

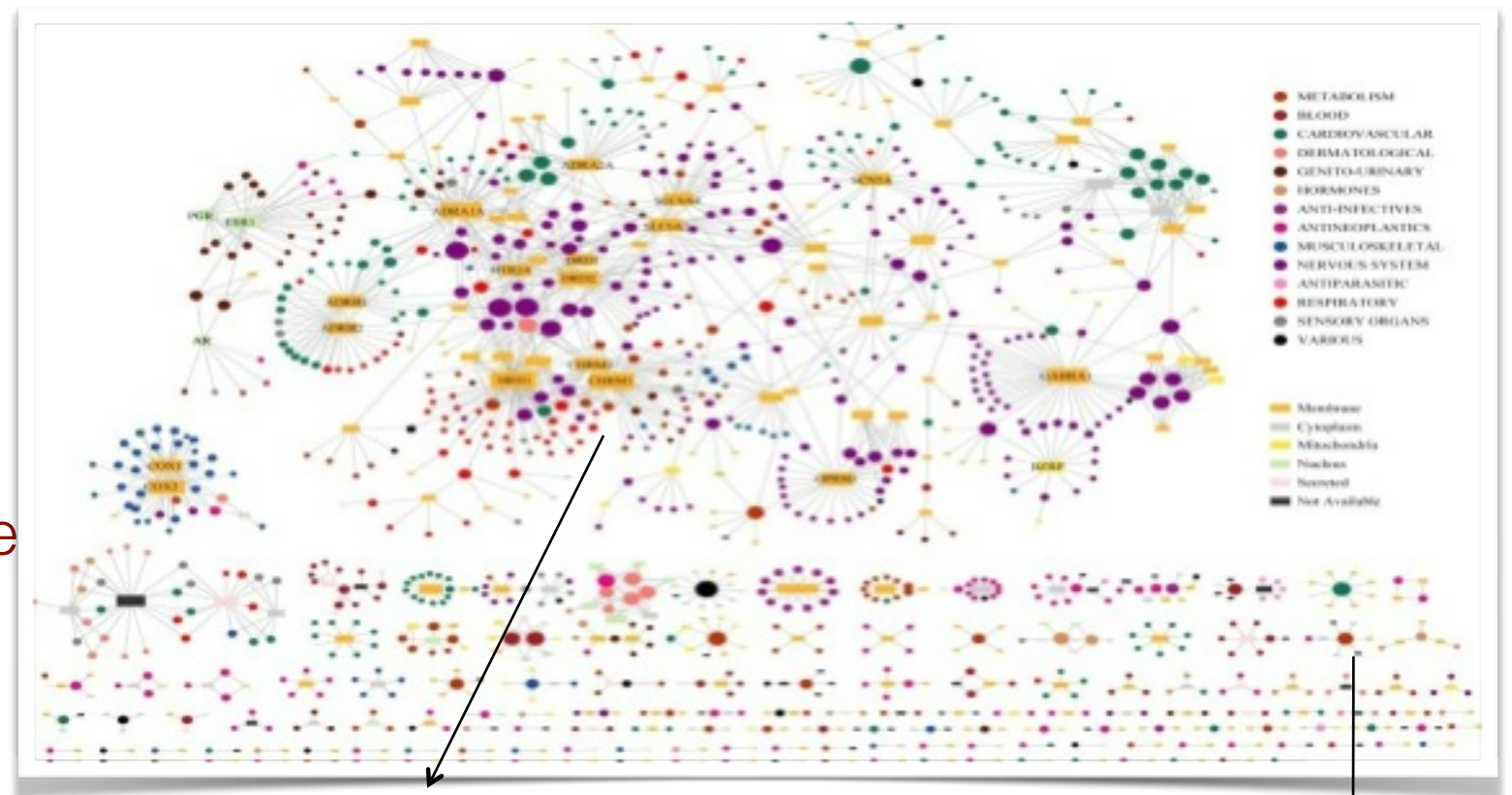


# The Evolution of the Innovation

## The Facts

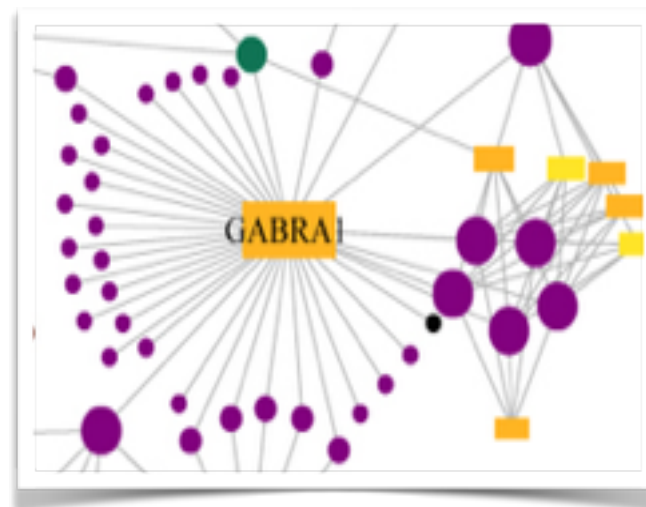
### Reductionism

- Key - Lock Model
- Clean Drug
- One drug, one target, one disease



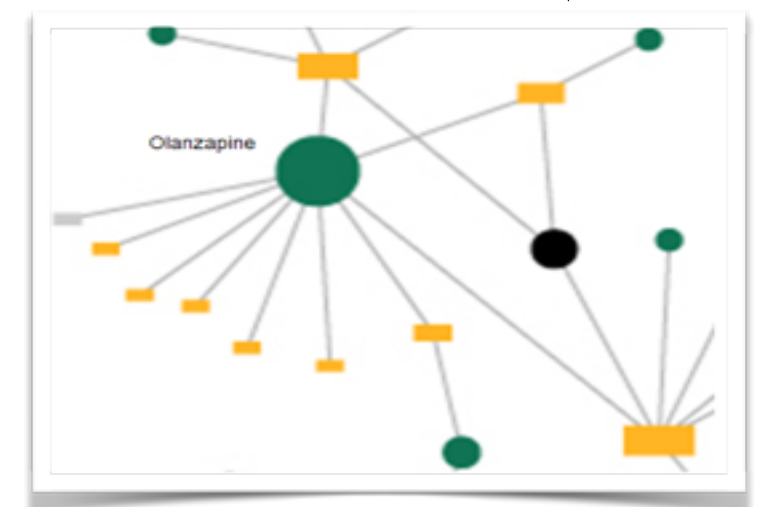
### System Theory

- Systems Pharmacology
- Network Pharmacology
- Systems Biology



Drug *Olanzapine* link to ~11 targets

Target *GABRA1* link to ~40 drugs



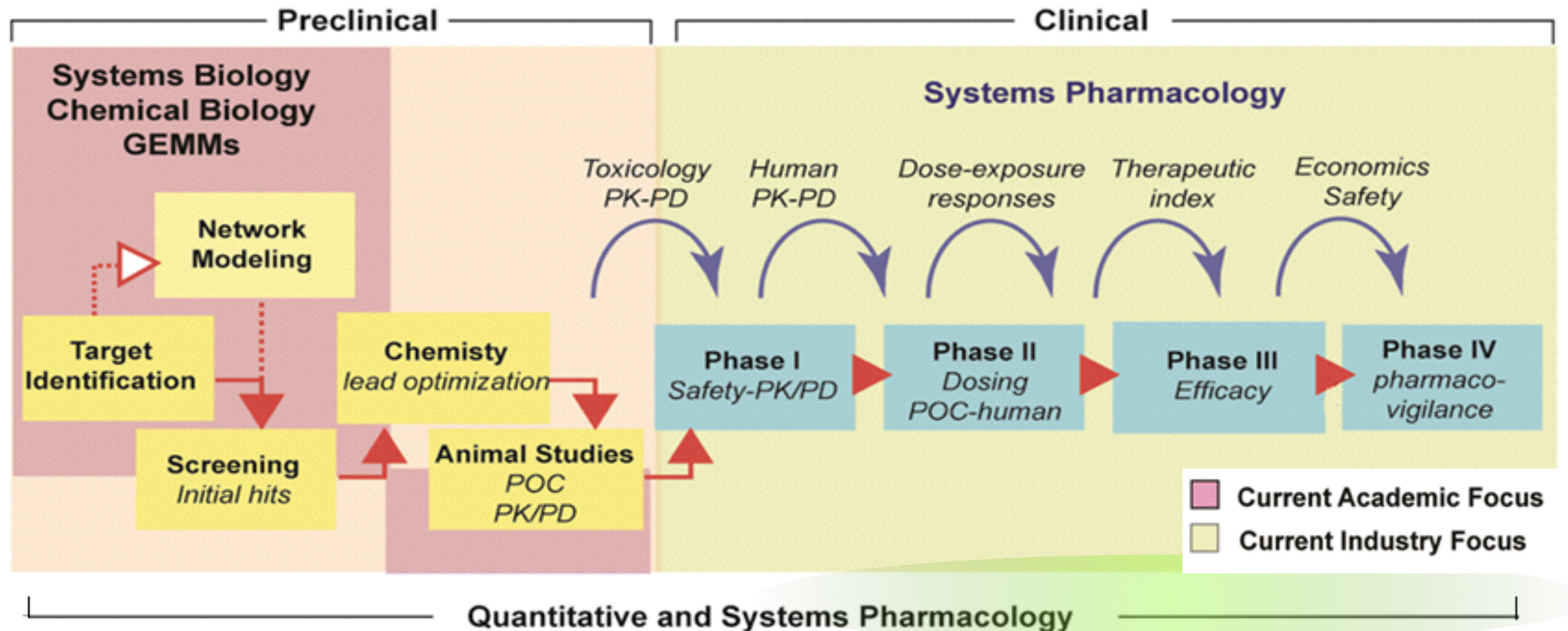
# Big Data, Small Details

M. Sirota et al., Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3 (2011).





# Pipeline of Systems Pharmacology



Academic

Industry

- Application of systems biology approaches
- Combining large-scale experimental studies
- Model-based computational analyses to study drug activities, targets, and effects

- Using pharmacodynamic (PD) and pharmacokinetic (PK) modelling
- Predicting dose-exposure responses and evaluating market potential

NIH White Paper by the QSP Workshop Group (Oct, 2011)

# Biology's Dry Future

The explosion of publicly available databases housing sequences, structures, and images allows life scientists to make fundamental discoveries without ever getting their hands “wet” at the lab bench

*Science* (2013) 342, 186-189.



“I’m like a **kid in a candy store**.  
There is so much we can do.”

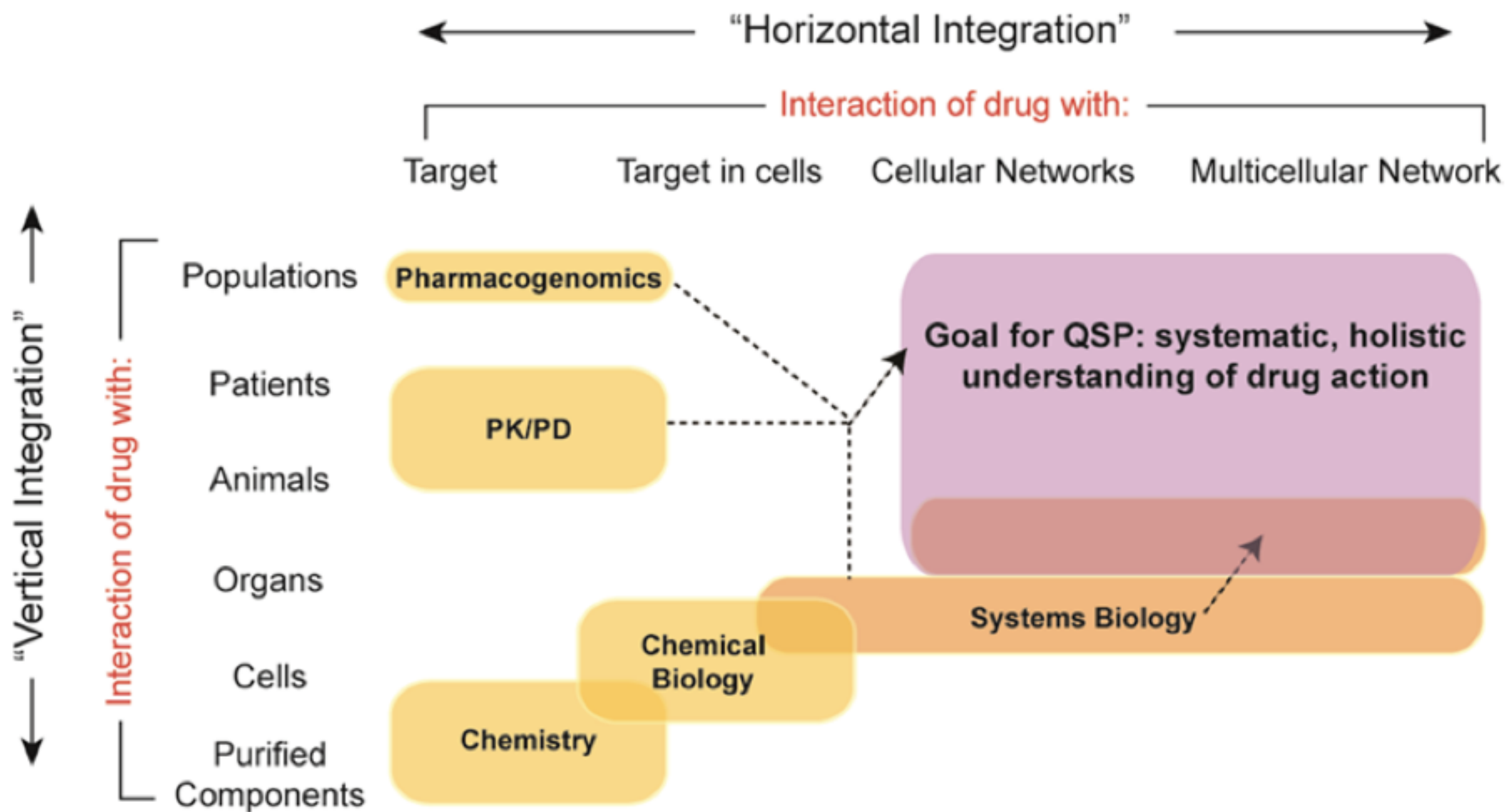
—Atul Butte, Stanford University School of Medicine



“You basically **don’t need a wet lab** to  
explore biology.”

—David Heckerman, Microsoft Research





Integration-based systematic thinking, is the core of QSP.

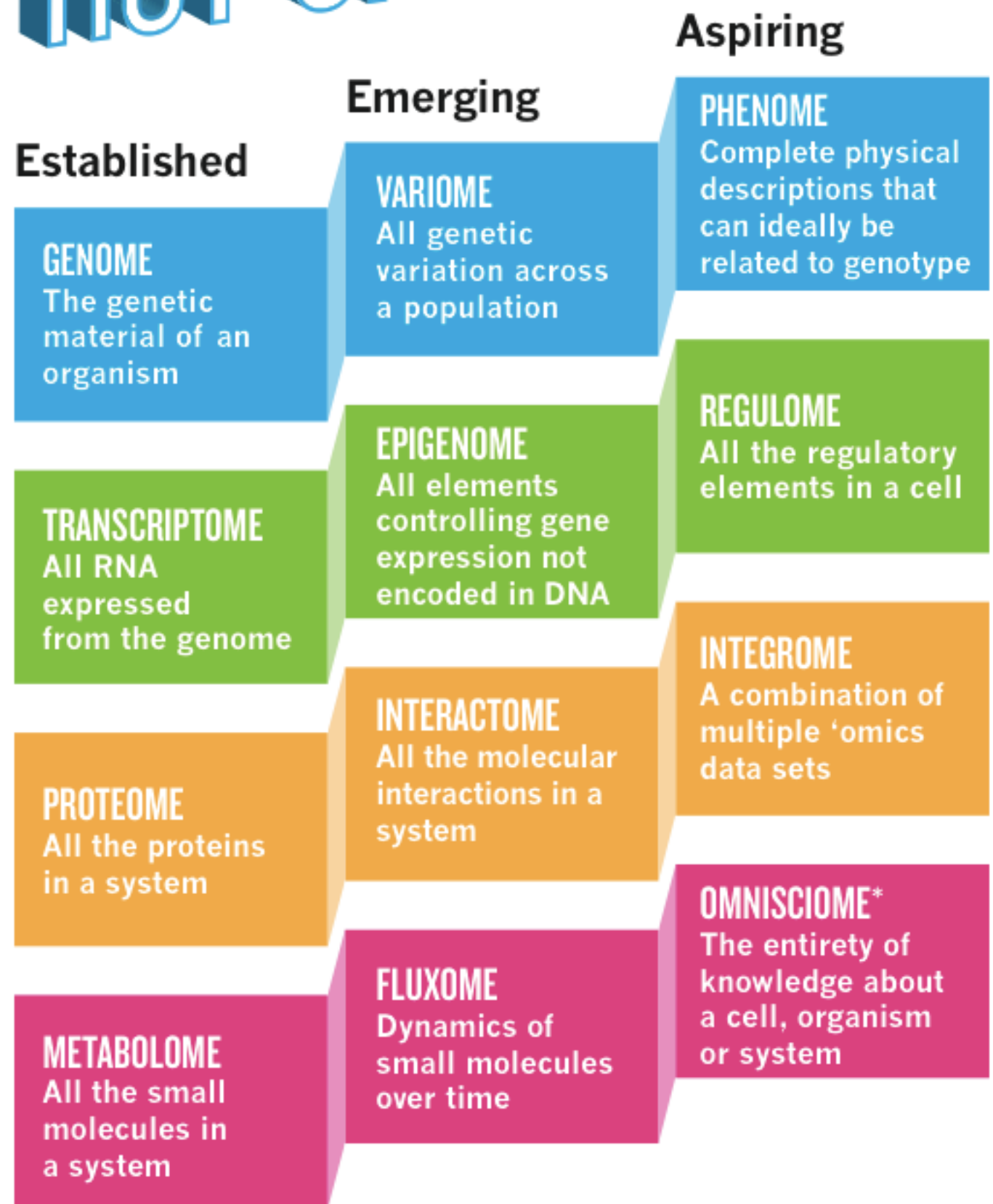
# HOT OR NOT

## THE 'OMES PUZZLE

Where once there was the genome,  
now there are thousands of 'omes.

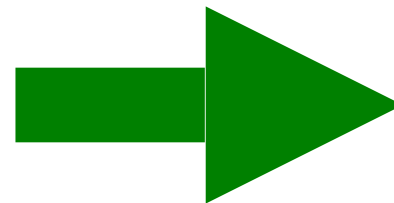
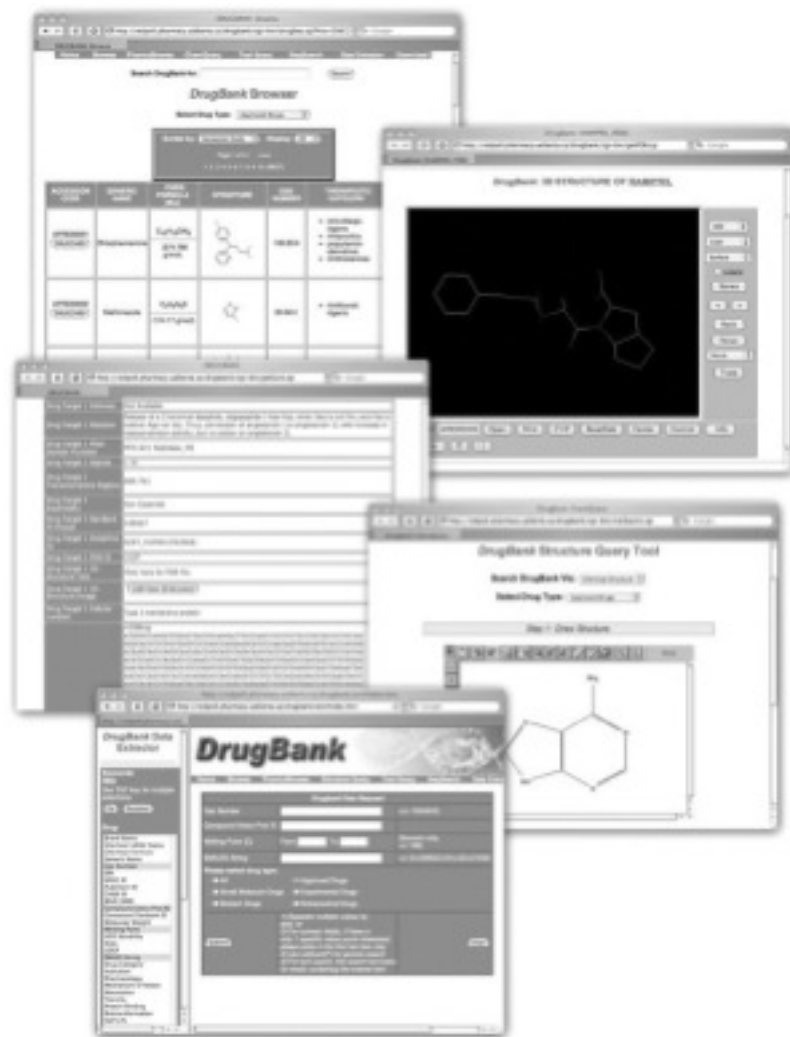
It's a trend to integrate the Omics data, numerical or non-numerical, structural or non-structural, semantic or non-semantic.

*Nature* (2013) 494, 416-419.



# The Dawn of A New Era: Bio Big Data Blossom

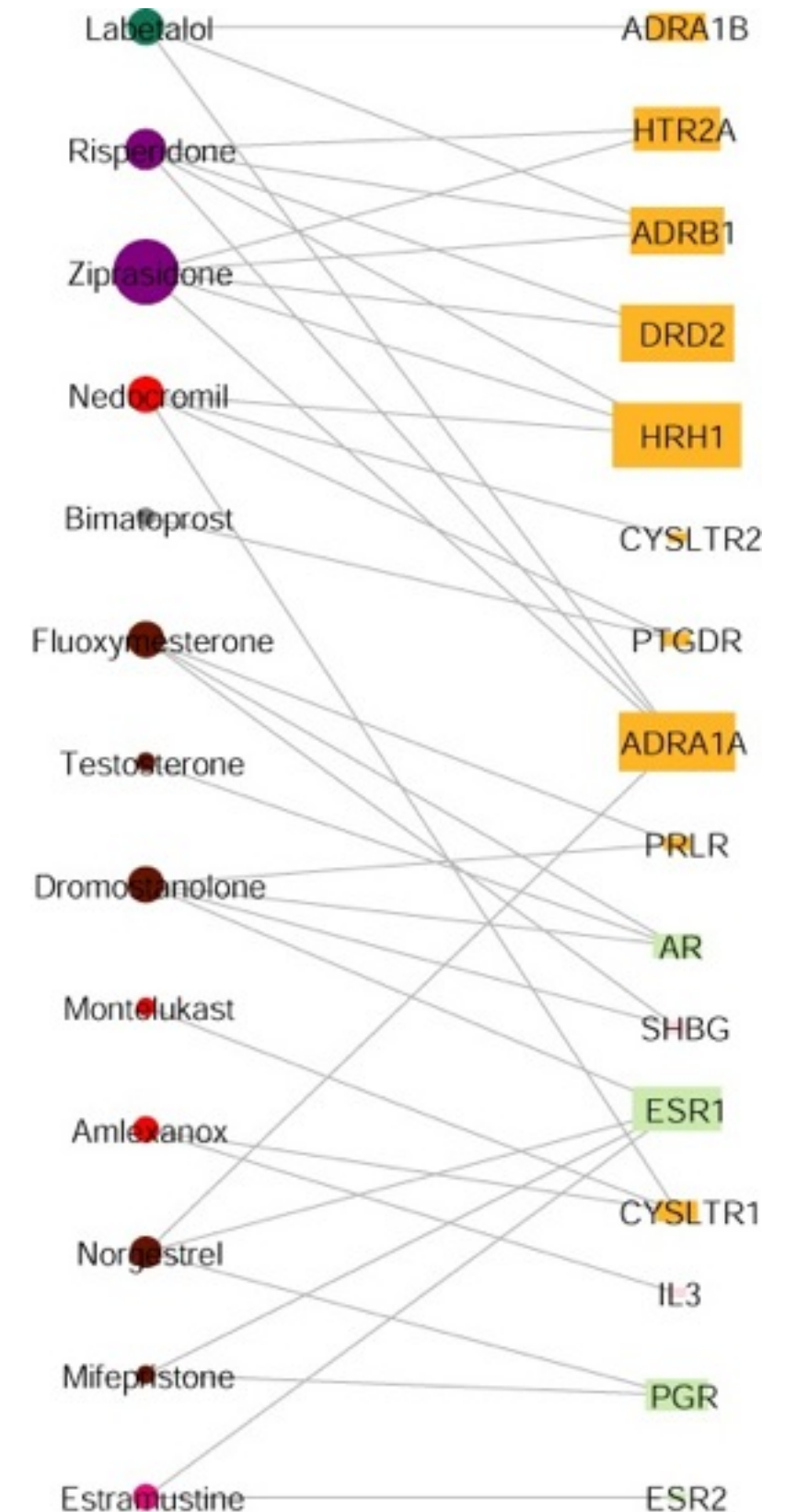
# Drugbank Database



## Drug-Target

### DRUGS

### TARGET PROTEINS



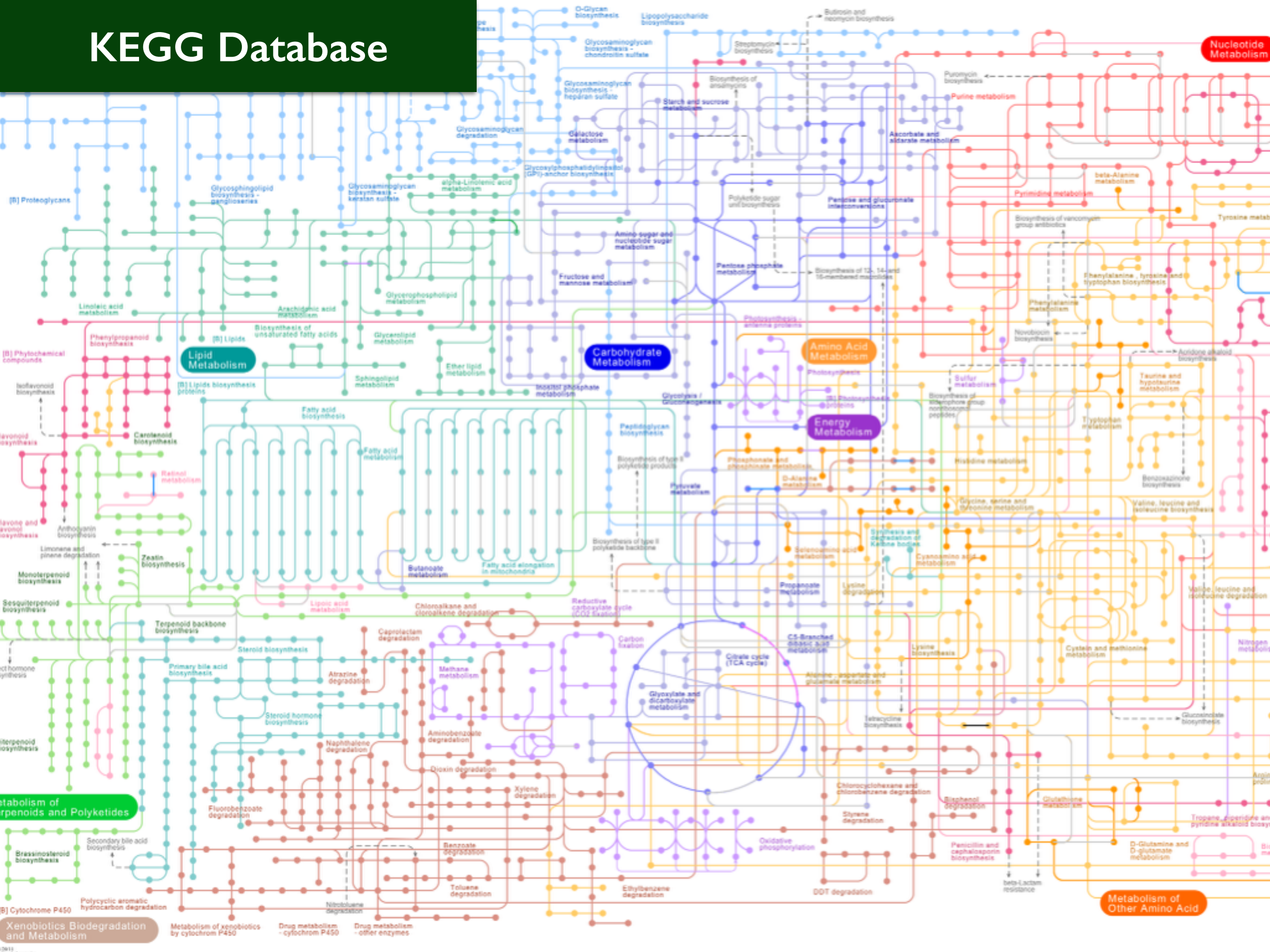
1179 FDA-approved small molecule & biotech drugs (different chemical entities)

890 / 1179 has human protein targets  
390 Human Drug Target Proteins for Approved Drugs.

Wishart DS et al., Nucleic Acids Res. 2006 1;3



# KEGG Database





The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.



# BioGRID Database

**Search the BioGRID**  
Search by identifiers, keywords, and gene names...

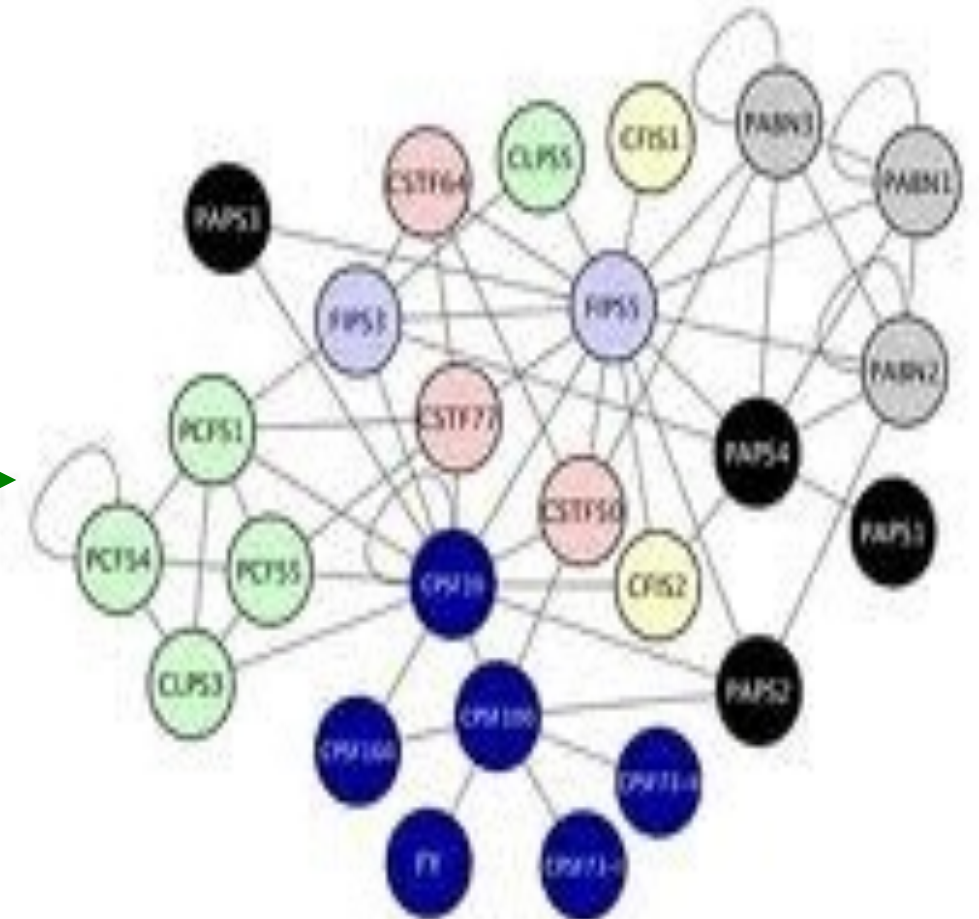
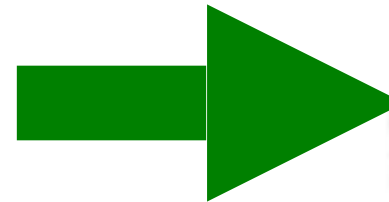
All Organisms

**BIOGRID FUNDING AND PARTNERS**

[more partners](#)

## PPI network



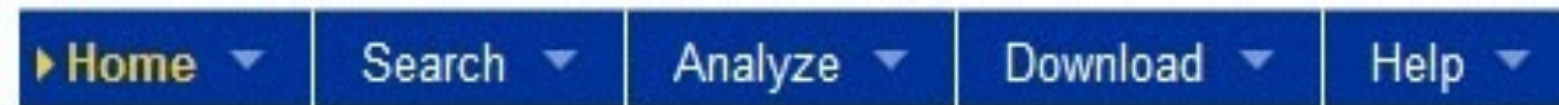
BioGRID is an online interaction repository with data compiled through comprehensive curation efforts. The current index searches 41,785 publications for 722,541 raw protein and genetic interactions from major model organism species.

# CTD Database



*Illuminating how chemicals affect human health.*

Comparative Toxicogenomics Database



**Chemical: Diazinon**

Basic Interactions Genes **GO** Diseases Pathways References Links

The following GO terms are enriched significantly among genes that interact with Diazinon or any of its descendants. The display is limited to GO terms with a corrected p-value less than 0.01, expressed as an "Enrichment Score" (range of 2.00 to >323.31), with higher numbers being more significant than lower.

Showing 1-500 of 546  
First 14 | Prev 4 | Next 4 | Last Page: 1 2

Rank	Ontology	Highest GO Level	GO Term	Enrichment Score	Annotated Genes	Cluster Frequency	Genome Frequency
1.	MF	9	extracellular glutamate-gated ion channel activity	15.43	11	11/215 genes (5.1%)	29/30531 genes (0.1%)
2.	BP						
3.	BP						
4.	MF						
5.	BP						
6.	BP						

**Chemical: Tetrachlorodibenzo-dioxin**

Basic Interactions Genes **Diseases** Pathways References Links

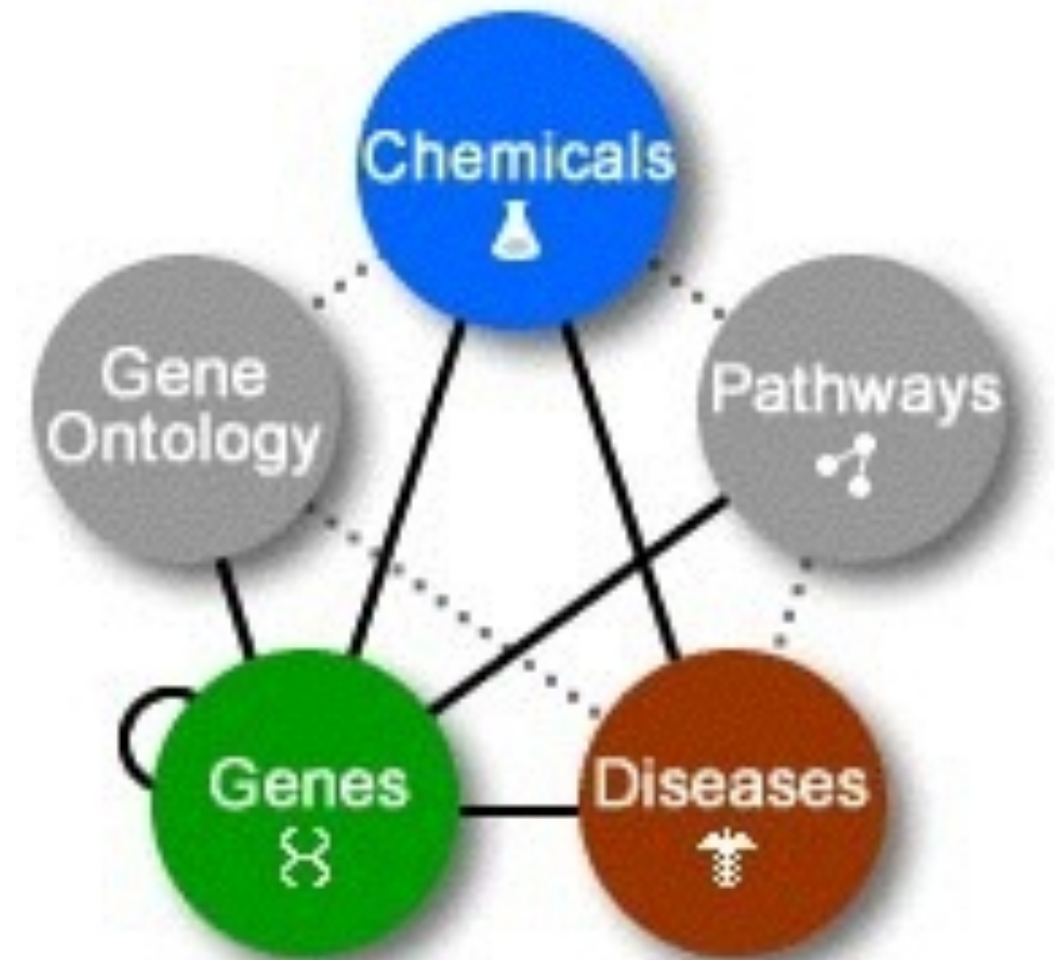
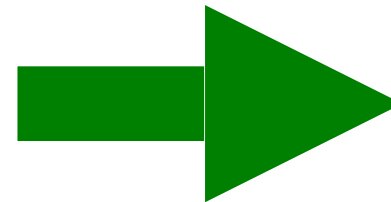
The following diseases are associated with Tetrachlorodibenzo-dioxin or at least one of its descendants. Each association is direct (marking mechanism and/or therapeutic) and/or inferred (via a curated gene interaction).

Showing 1-500 of 595  
First 14 | Prev 4 | Next 4 | Last Page: 1 2

Rank	Chemical	Disease	Direct Evidence	Inference Network	Inference Score	References
1.	Tetrachlorodibenzo-dioxin	Adenoma, Liver Cell	<input checked="" type="checkbox"/>			1
2.	Tetrachlorodibenzo-dioxin	Hydronephrosis	<input checked="" type="checkbox"/>			1
3.	Tetrachlorodibenzo-dioxin	Heart Defects, Congenital	<input checked="" type="checkbox"/>			1
4.	Tetrachlorodibenzo-dioxin	Craniofacial Abnormalities	<input checked="" type="checkbox"/>			1
5.	Tetrachlorodibenzo-dioxin	Diabetes Mellitus, Type 2	<input checked="" type="checkbox"/>	Via 12 genes: ANKRD1, BMP11A, CYP1B1, CYP1A1, CYP1A2, CYP1B1, CYP1B2, CYP1B3, CYP1B4, CYP1B5, CYP1B6, CYP1B7	81.38	14
6.	Tetrachlorodibenzo-dioxin	Diabetes Mellitus, Type 2	<input checked="" type="checkbox"/>	Via 7 genes: GSK3, HNF1B, IL6, LIPC, MBOAT1, SLC6A11, TNSI	42.13	8
7.	Tetrachlorodibenzo-dioxin	Cleft Palate	<input checked="" type="checkbox"/>	Via 4 genes: CYP1B1, CYP1A1, CYP1A2, CYP1B2	38.87	5
8.	Tetrachlorodibenzo-dioxin	Cholangiocarcinoma	<input checked="" type="checkbox"/>	Via 3 genes: CYP1B1, CYP1A1, CYP1A2	30.32	4
9.	Tetrachlorodibenzo-dioxin	Cholangiocarcinoma	<input checked="" type="checkbox"/>	Via 3 genes: CYP1B1, CYP1A1, CYP1A2	30.32	4
10.	Tetrachlorodibenzo-dioxin	Cholangiocarcinoma	<input checked="" type="checkbox"/>	Via 3 genes: CYP1B1, CYP1A1, CYP1A2	30.32	4

**Observed: Comps Network**

Pathway view of top 10 Comps:  
[XGML file](#) | [Visualize \(requires Flash\)](#)



The Comparative Toxicogenomics Database (CTD) provides information about interactions between environmental chemicals and gene products and their relationships to diseases. Chemical-gene, chemical-disease and gene-disease interactions manually curated from the literature are integrated.



# SIDER Database

## Browse the drugs by name:

| aba-ami | aml-bec | ben-cab | caf-cef | cel-clo | coc-den |  
lev-mef | meg-met | mex-nap | nar-olm | olo-per | phe-pra

## Browse the side effects by name:

| 5q-abn | abo-acr | act-acu | add-agi | agn-alo | alt-ana | anc-ano | ant-a  
bin-ble | bli | blo | blu-bra | bre | bro-bul | bun-cap | car | cas-cen | cer-c  
coo-cox | cra-cut | cya-dea | dec-den | dep-det | dev-dia | dif-diu | div-dry  
era-eva | eve-eva | fca-fat | fca-fib | fic-fca | fic-fca | for-gan | for-g  
jun-lab |

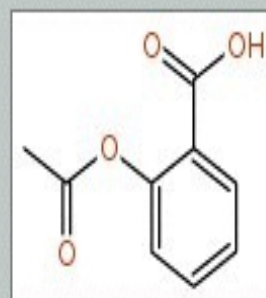
## Downloading data

Here, you can download the current version of the data. Previous versions can be found on the FTP site.

## Mapping of labels

The package inserts contain information about the side effects. The information, labels were mapped to STITCH identifiers. (These compound identifiers might change in the future years.)

## Information



More information: [STITCH](#), [PubChem](#) and possibly [Wikipedia](#) or [Medpedia](#)

ATC Codes: [A01AD05](#), [B01AC06](#), [N02BA01](#)

Side effect	Data for drug	Placebo Labels
		1 2 3 4 5 6
Dyspnoea def	28%	
Headache def	18%	
Dizziness def	12%	
Lightheadedness	12%	
Hypotension def	2%	
Arrhythmia def	1%	
Bronchospasm def	postmarketing	
Grand mal convulsion def	postmarketing	
Cardiac arrest def	postmarketing	
Seizures def	postmarketing	
Torsade de pointes def	postmarketing	
Loss of consciousness def	postmarketing	
Ventricular fibrillation def	postmarketing	
Bradycardia def	postmarketing	
Blood pressure increased	postmarketing	

## Database statistics

Number of drugs and side effects					
# of SE	# of drugs	# of drug-SE pairs	Pairs with frequency information		
4192	996	99423	40.8%		
Number of drug-side effect pairs in different frequency ranges					
	frequent (with exact data)	infrequent (with exact data)	rare (with exact data)	postmarketing	total
drug	11475 (10316)	9471 (3236)	6650 (2068)	21664	40603
placebo	4330 (4330)	2043 (2043)	1425 (1425)	0	6370

## Version Information

The current version has been released on October 17, 2012. This release uses the MedDRA dictionary (version 14.0) and provides access to preferred terms and lower-level terms. The number of drugs has increased from 888 to 996. Compared to the release in March 2012, additional side effects have been retrieved by better processing of the labels. Side effects that are mentioned on the label as either potential or not occurring are removed. SIDER 1 is still available via [FTP](#).

SIDER contains information on marketed medicines and their recorded ADRs. The information is extracted from public documents and package inserts. It contains 99423 drug-ADR pairs associated with 996 drugs and 4192 ADRs.

Michael Kuhn et al., Molecular Systems Biology 2010 (6) 1-6

**CHALLENGE ACCEPTED**



# Part II

## Packages & Web Servers

# What do we need for Systems Pharmacology Modeling?

- Information: Multi-scale Representation
- Methodology: Multi-scale Modeling

# What does R need?

- Good at methodology and modeling, state-of-art statistical machine learning methods, Bioconductor
- Lacks of bio/chem data representation
- A good representation is fundamental and critical

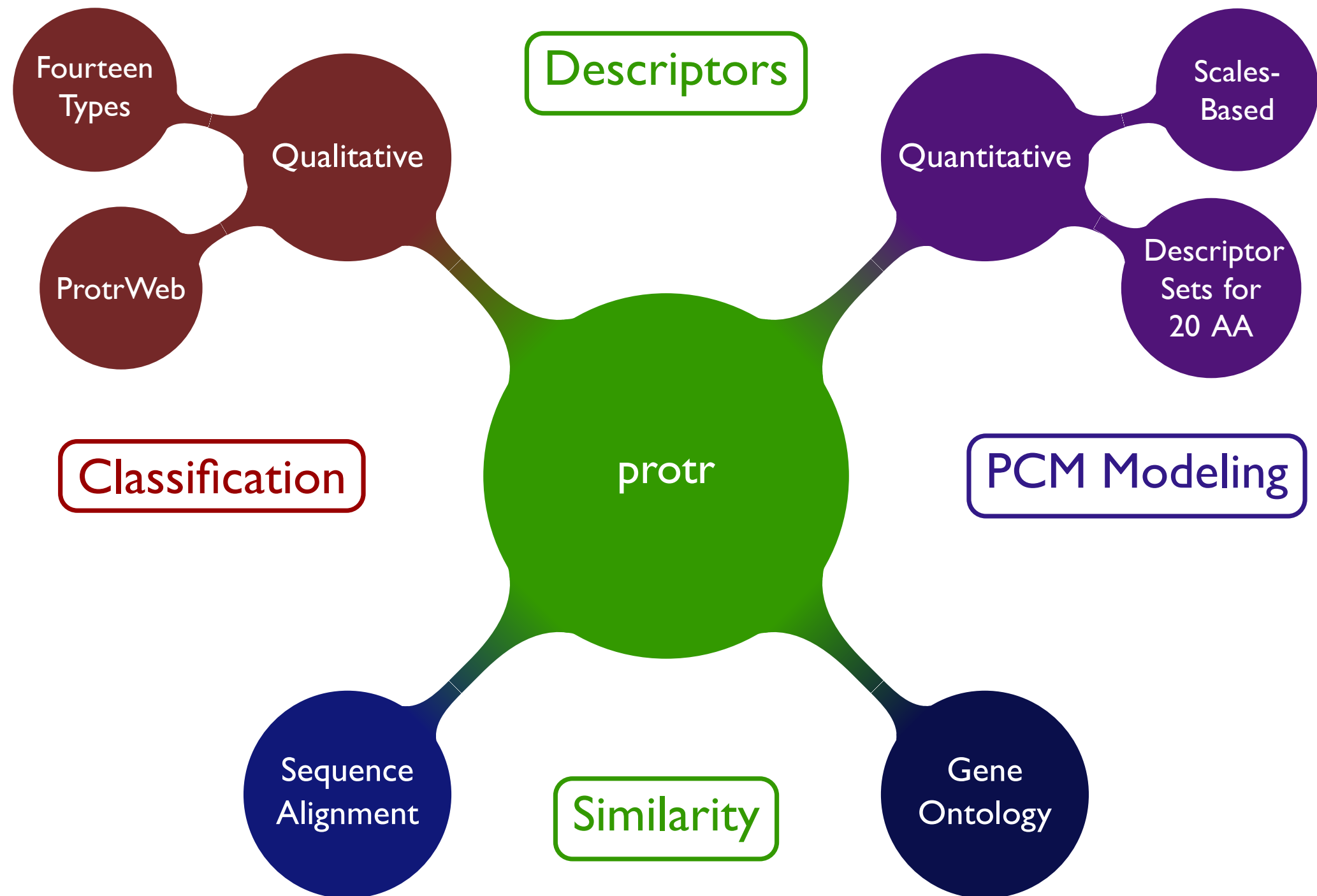
# What we did?

R/Bioconductor packages for  
multi-scale molecular representation



protr

Protein Sequence Descriptor Calculation  
and Similarity Computation with R



Schematic diagram of the protr package.  
from Xiao et al., (2014)

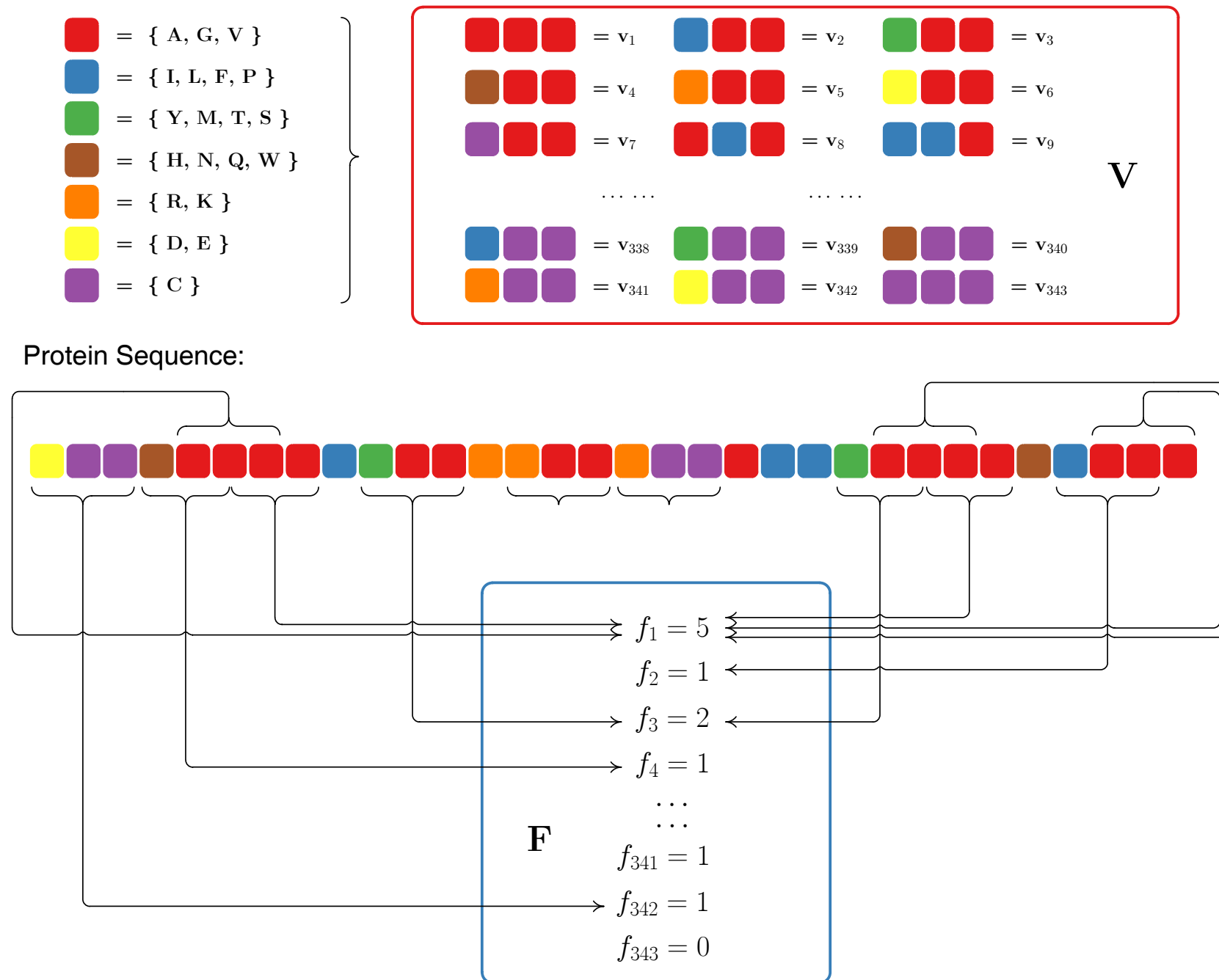
# What could protr do?

- For regular predictive modeling
  - 14 types of commonly used descriptors:
    - Bioinformatics (Classification)
  - 6 types of PCM descriptors:
    - Proteochemometrics (Regression)

# What could protr do?

- For similarity-based modeling methods
  - Similarity derived by sequence alignment & GO:
    - Similarity-based clustering
    - Kernel methods
    - etc.

# Make protein sequence into a numerical vector



- For algorithmic details, see `vignette('protr')`

# ProtrWeb

- Shiny-based
- Fast implementation: 1 Day

ProtrWeb
Home
Get Started
Example Input
Downloads
CBDD Group

# ProtrWeb

### Step 1. Upload Protein Sequence

Upload FASTA File:

Or Upload Raw Sequence File:

### Step 2. Select Descriptor(s)

Descriptor Name (Dim):

- ☐ Amino Acid Composition (20)
- ☐ Dipeptide Composition (400)
- ☐ Tripeptide Composition (8000)
- ☐ Normalized Moreau-Broto Autocorrelation (240)
- ☐ Moran Autocorrelation (240)
- ☐ Geary Autocorrelation (240)
- ☐ C/T/D (21 + 21 + 105)
- ☐ Conjoint Triad (343)

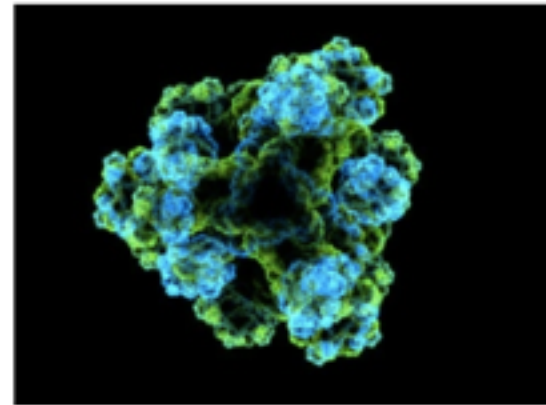
Introduction
Data
Results

## Protein Sequence Feature Extraction with ProtrWeb

Over the past decade, machine learning methods have been successfully employed in the structural, functional and interaction profiles research of proteins and peptides. The structural and physicochemical descriptors have been intensively applied in the research of protein structure and functionalities, including

- Predicting protein structural and functional classes
- Predicting protein-protein interactions
- Predicting protein-ligand interactions
- Predicting protein subcellular locations
- Identifying protein phosphorylation sites
- Predicting protein crystallization propensity and peptides of specific properties

and many more challenging problems. As the fundamental building blocks, sequence-derived structural and physicochemical descriptors extracted from protein and peptide sequences play a highly critical role in the modeling procedure. Here we present *ProtrWeb*, a web server based on our R package *protr*, dedicated to compute such structural and physicochemical descriptors. Currently, *ProtrWeb* offers the functionality for computing 12 different types of qualitative descriptors presented in the *protr* package. The *protr* package offers more quantitative descriptors, miscellaneous tools and datasets, and more customized descriptors could be crafted by accessing the *protr* package directly.

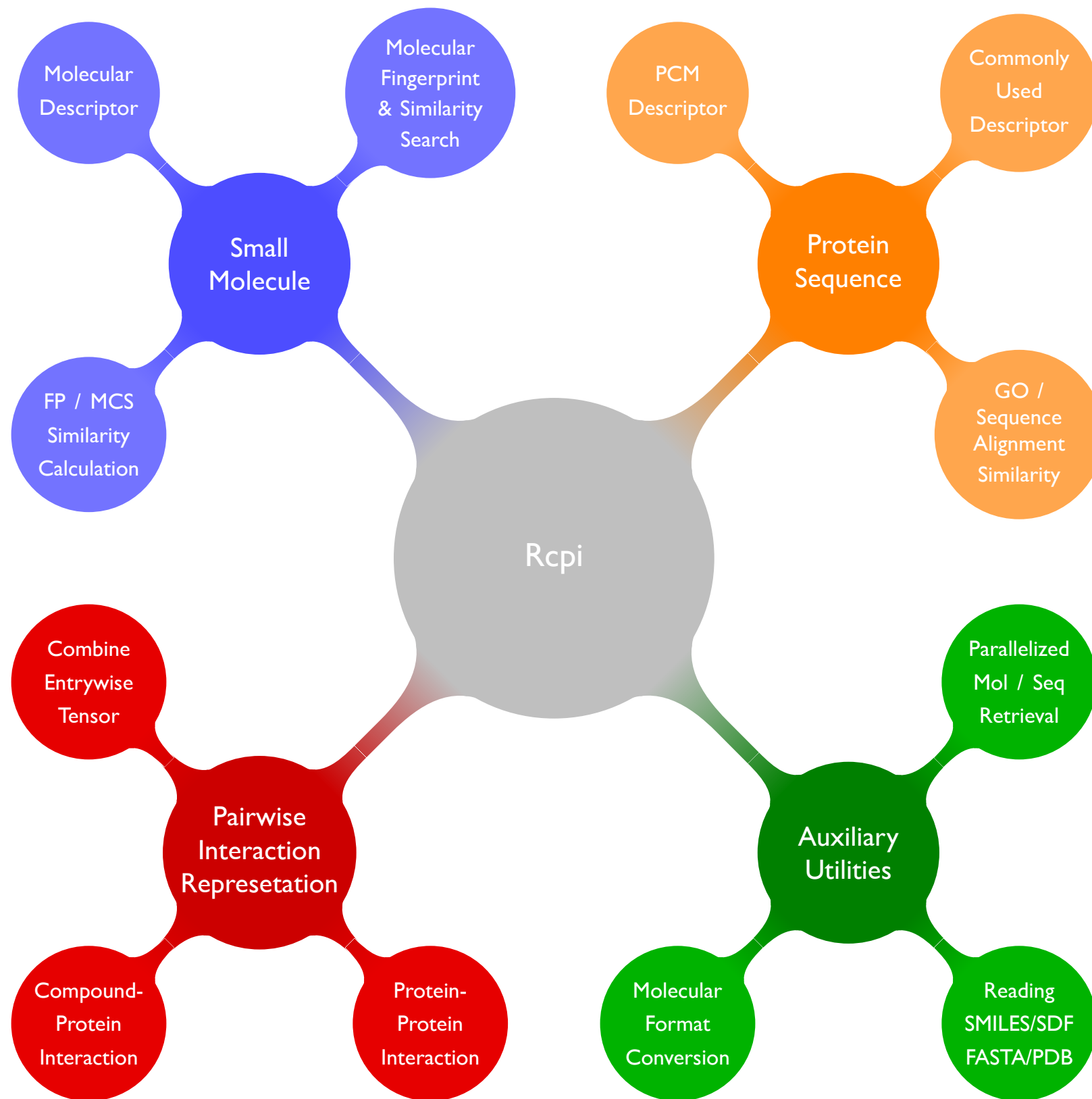


A Screenshot of ProtrWeb

# Rcpi

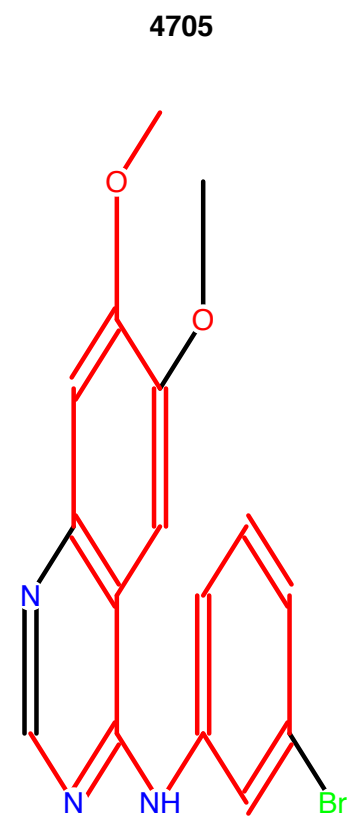
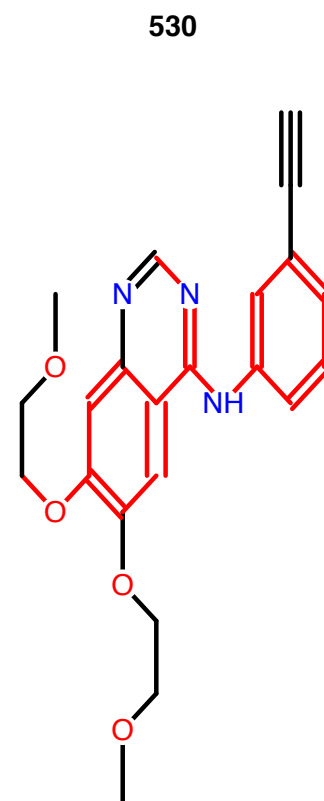
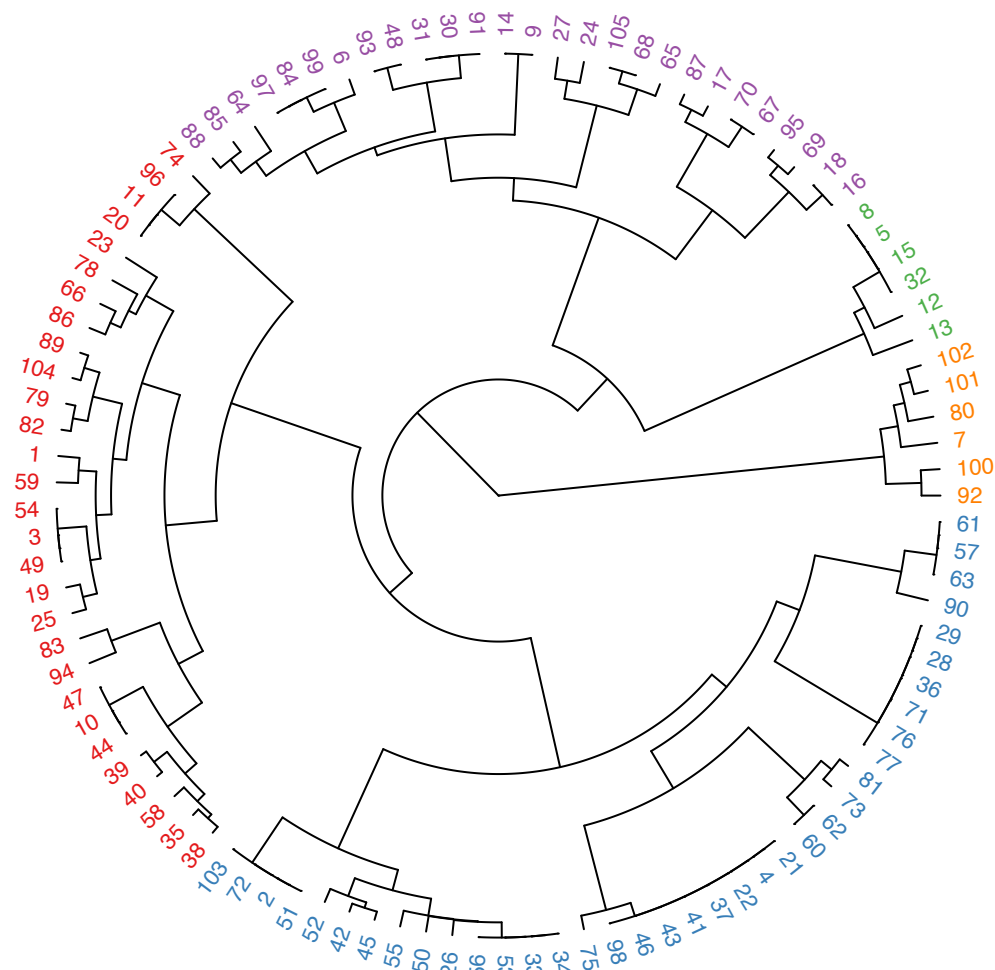
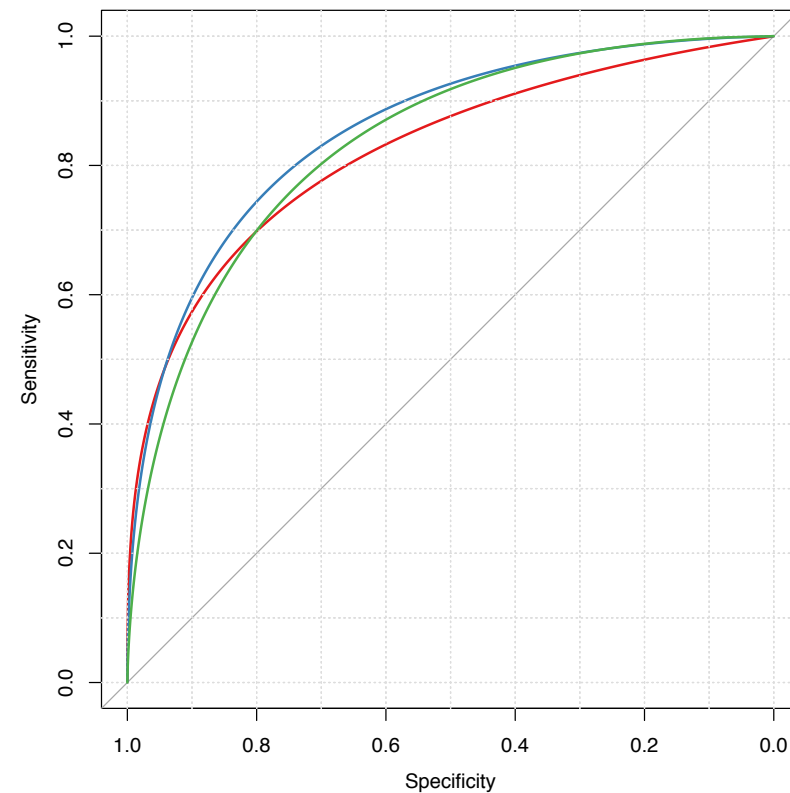
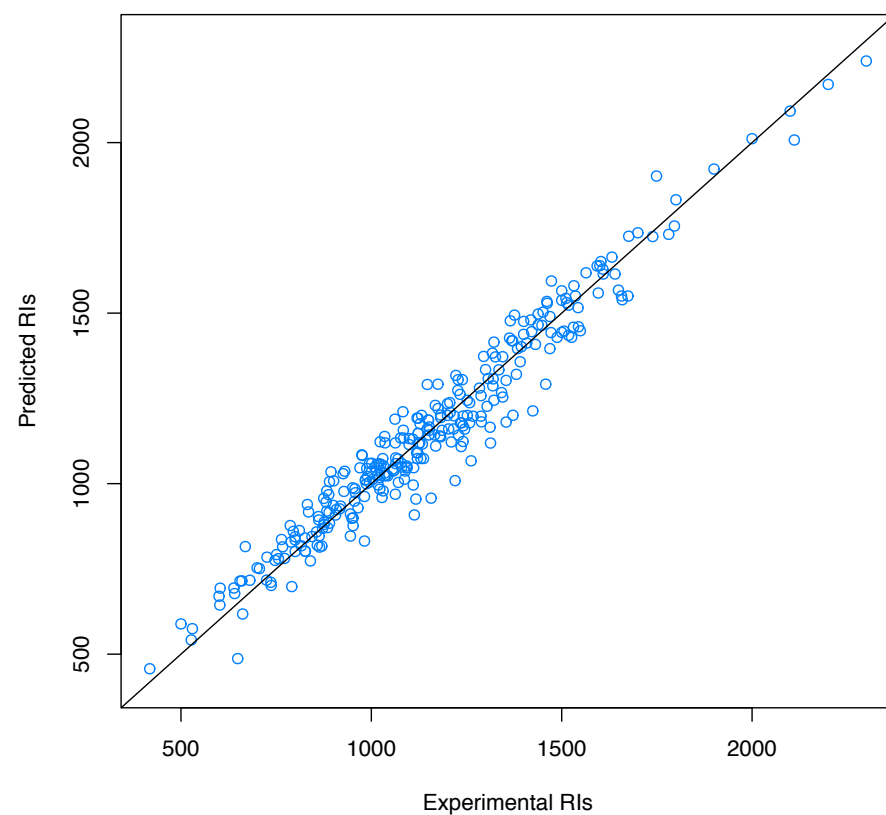
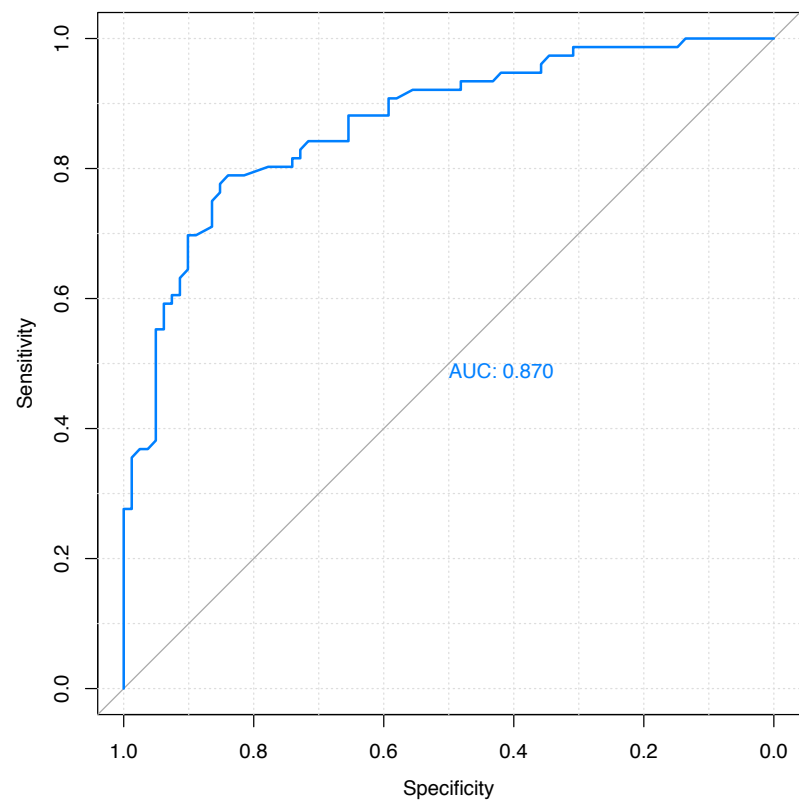
R/Bioconductor Package for Bioinformatics,  
Chemoinformatics & Chemogenomics Research

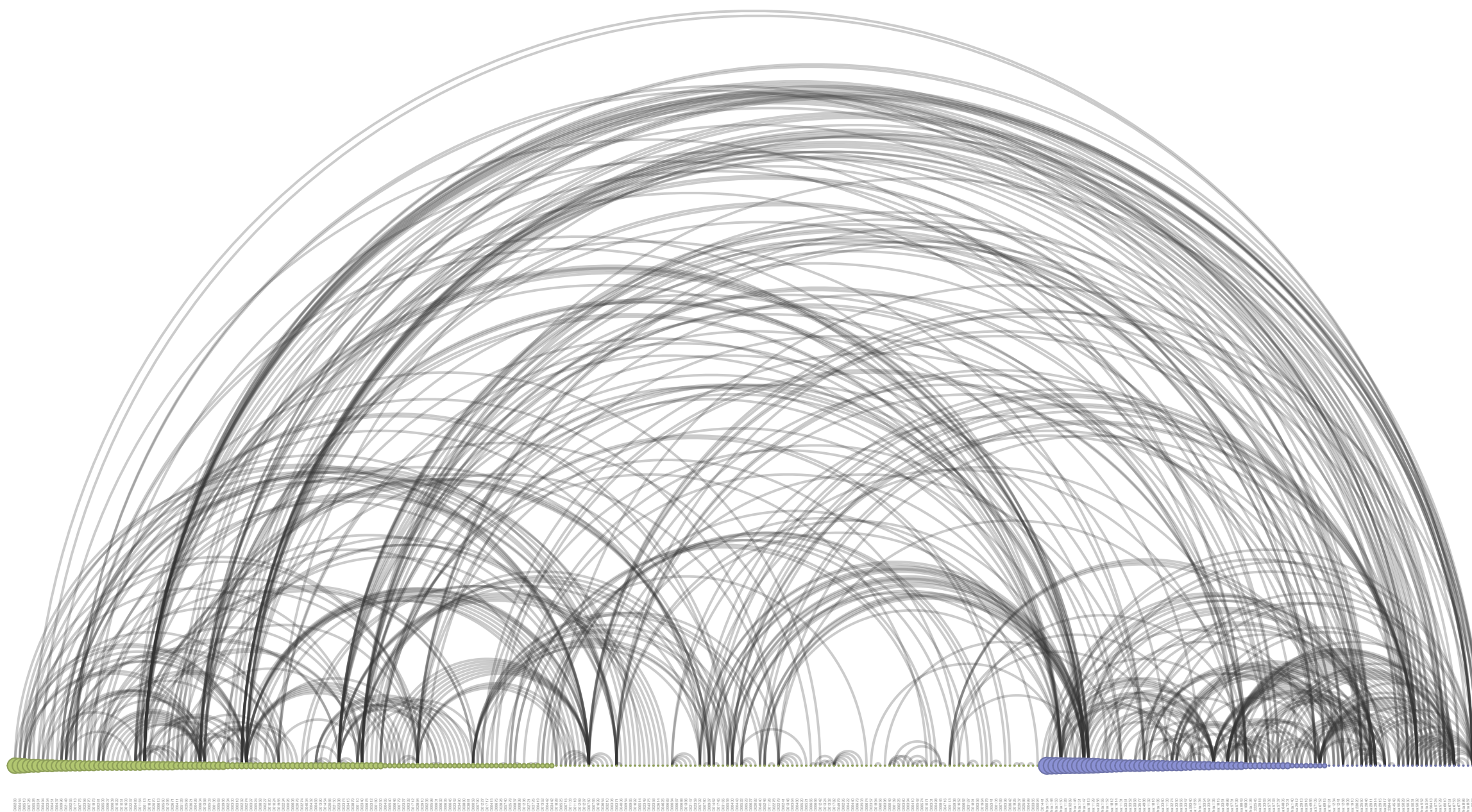




Schematic diagram of the Rcpi package.  
from Xiao et al., (2014)

What could Rcpi do?





Arc diagram of the GPCR drug-target interaction network  
from Xiao et, al. (2014)

# Experience & Pitfalls



# Dependency Hell

- foreach / doParallel / doMC
- Biostrings
- GOSemSim
- ChemmineR
- ChemmineOB
- fmcsR
- rcdk
- RCurl



# Checking Hell

- R CMD check
- BiocCheck





# Experiences

- Use Roxygen2 to generate docs and NAMESPACE
- Cross-platform availability: doParallel / doMC
- Unit Tests

# Part III

## Drug-ADR Prediction

# Identify Novel Drug-ADR Associations

- Integrated multiple evidence from multiple levels
- Collaborative filtering and link prediction
- Mainly done by R, some done by Python

# Summary

# Summary

- Integrating only in the molecular structure level for now
- With R's modelling capability, applications promised.



# Future Works

- protr: Incorporate protein 3D information
- Rcpri: Integration of RDKit, ChemoPy
- Omics Information (Genome / Proteome / Phenome)
- **Network-based** representations

# Our Vision

- Systematic integration
- Comprehensive pipeline

# Resources

- protr

<http://cran.r-project.org/web/packages/protr/>

- Rcp

<http://bioconductor.org/packages/release/bioc/html/Rcpi.html>

- ProtrWeb

<http://cbdd.csu.edu.cn:8080/protrweb/>

有时，整个地球结盟促进某些幸运的学科发展，而那些学科也随之绽放出新思想的花蕾、取得惊人的进展。而关键在于，哪里有大量累积起来的关于这个领域的有意义的问题，并且总有新技术应用于该领域，使得更加贴近的观察那些过程成为可能。

*Efron, B. (2005). Bayesians, frequentists, and scientists. JASA, 100(469).*



现在这个星球也许正在联合起来促进统计学的发展。新技术——电子计算技术，打破了曾限制了传统统计理论发展的计算瓶颈。同时，一类重要问题的洪流正奔向我们，其表现形式为大型数据集以及大规模推断问题。我相信，这一代统计学家将投身于一个新的统计创新年代，一个可与 Fisher、Neyman、Hotelling 以及 Wald 的黄金时代相媲美的时代。

*Efron, B. (2005). Bayesians, frequentists, and scientists. JASA, 100(469).*

Sometimes, not very often, the planets align for some lucky discipline, which then blossoms with new ideas and breath-taking progress. Microbiology is a perfect current example. The key there was a buildup of interesting questions concerning cellular processes, followed by new technology that enabled a much closer look at those processes in action.

*Efron, B. (2005). Bayesians, frequentists, and scientists. JASA, 100(469).*

Now the planets may be aligning for statistics. New technology, electronic computation, has broken the bottleneck of calculation that limited classical statistical theory. At the same time an onrush of important new questions has come upon us, in the form of huge data sets and large-scale inference problems. I believe that the statisticians of this generation will participate in a new age of statistical innovation that might rival the golden age of Fisher, Neyman, Hotelling, and Wald.

*Efron, B. (2005). Bayesians, frequentists, and scientists. JASA, 100(469).*

Q & A