

Systems Biology and Networks

**MMG 835, SPRING 2016
Eukaryotic Molecular Genetics**

George I. Mias
Department of Biochemistry and Molecular Biology
gmias@msu.edu

What is Systems Biology

- Wikipedia: “***Systems biology*** *Systems biology is the computational and mathematical modeling of complex biological systems. An emerging engineering approach applied to biological scientific research, systems biology is a biology-based inter-disciplinary field of study that focuses on complex interactions within biological systems, using a holistic approach (holism instead of the more traditional reductionism) to biological research.*

What is Systems Biology

- nature.com : “*Systems biology is the study of biological systems whose behaviour cannot be reduced to the linear sum of their parts’ functions. Systems biology does not necessarily involve large numbers of components or vast datasets, as in genomics or connectomics, but often requires quantitative modelling methods borrowed from physics.*”

What is Systems Biology

- Encyclopedia of Systems Biology (Springer New York, 2013).
- “*Systems biology refers to the quantitative analysis of the dynamic interactions among several components of a biological system and aims to understand the behavior of the system as a whole. Systems biology involves the development and application of systems theory concepts for the study of complex biological systems through iteration over mathematical modeling, computational simulation and biological experimentation. Systems biology could be viewed as a tool to increase our understanding of biological systems, to develop more directed experiments, and to allow accurate predictions.*”

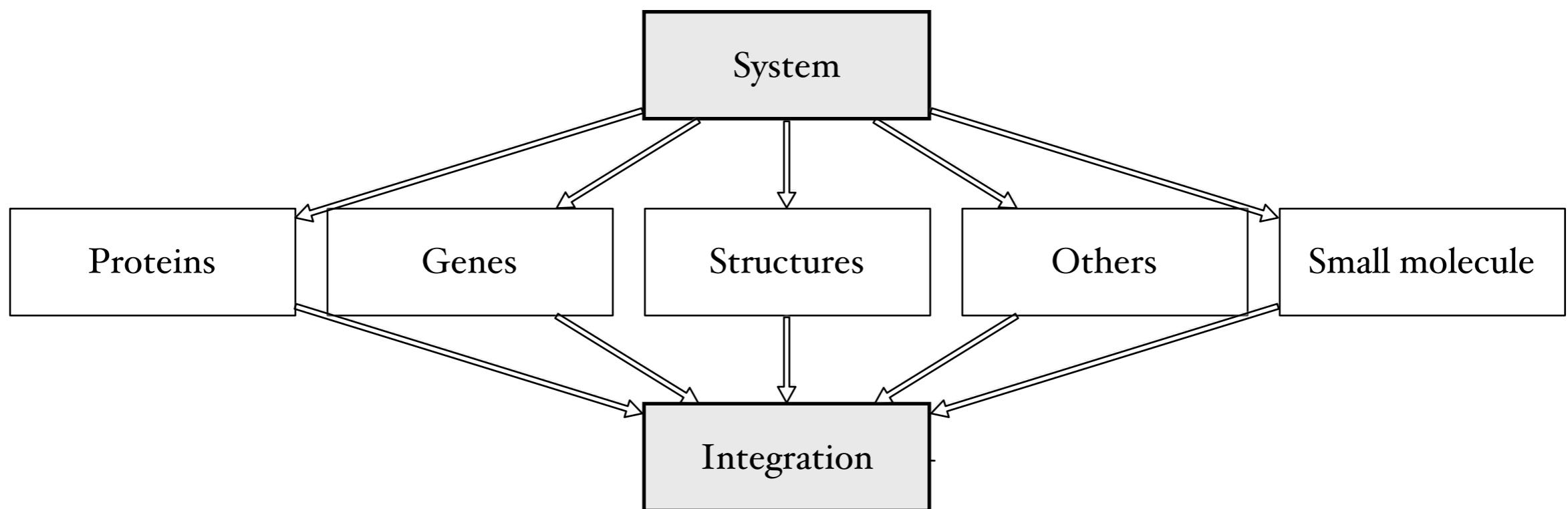
Systems Biology

- Systematic
- Novel approach (in biology at least)
- interdisciplinary
- Non-reductionist
 - Reductionist: Study the subcomponents for a system in detail, each one
 - Whole is greater the sum of its parts

Systems Biology

- Multiple inputs of information in a complex system
- More mathematical than traditional biology

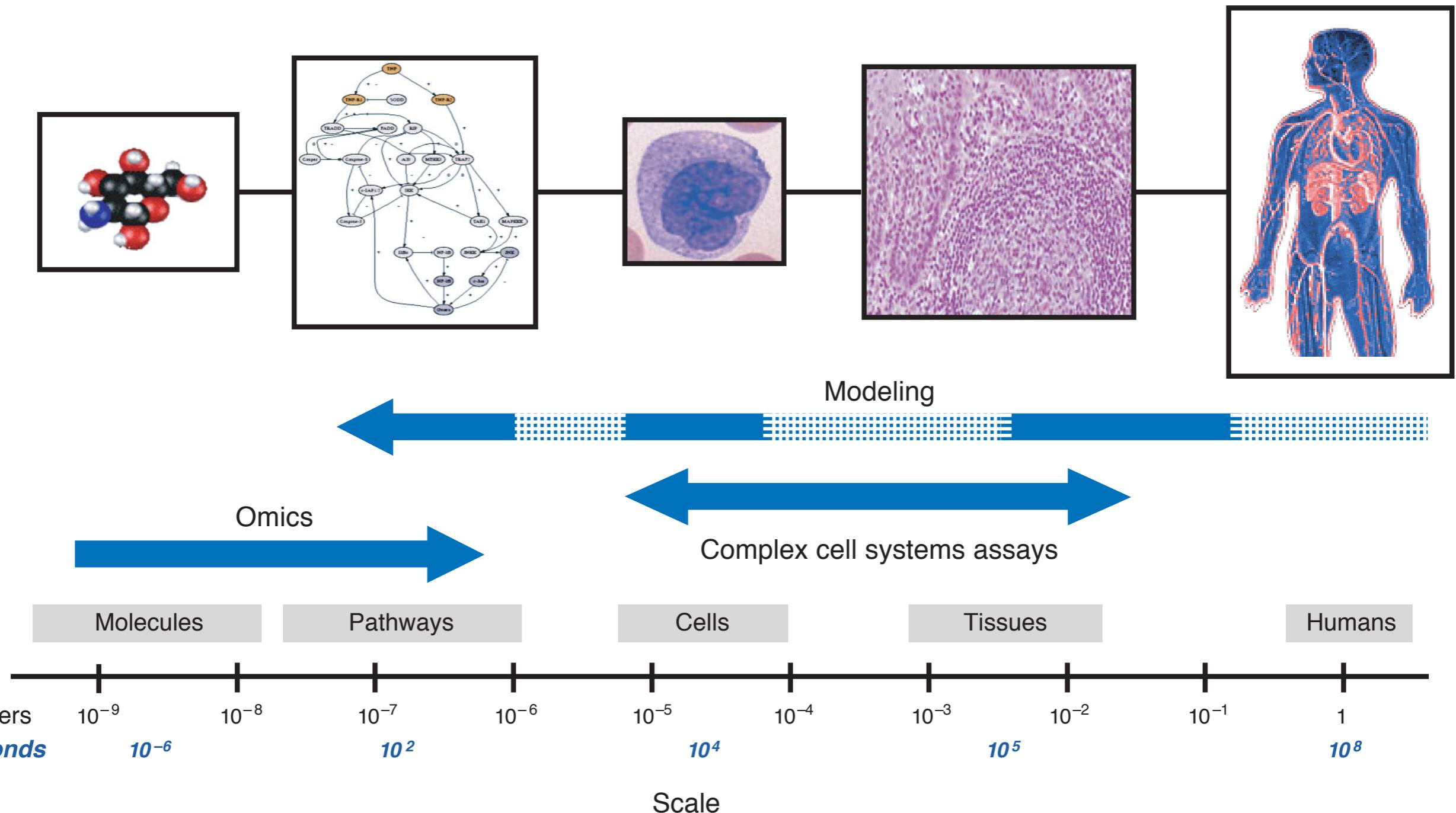
Systems Biology



Systems Biology

- Molecular components
- Cell subsystems
- parts of an organism
- the organism
- the environment
- set of environments
- **Study of**
 - Function
 - Networks
 - Signals
 - Interactions of components

Systems Biology

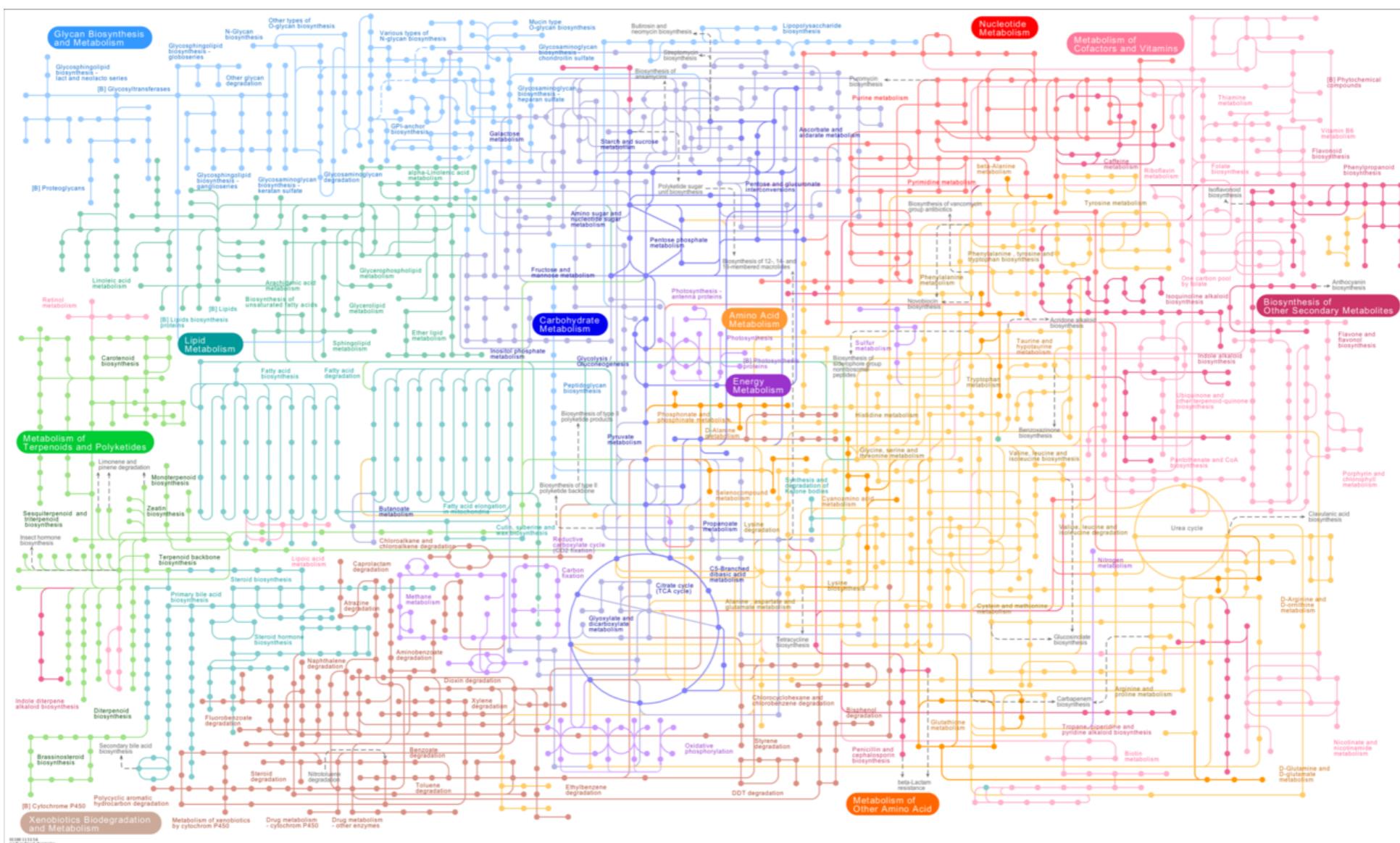


Systems Biology

- Omics approaches
 - Human Genome Project
 - Mass spectrometry
 - Proteomics
 - Metabolomics
 - name-it-omics
- Examples

Systems Biology

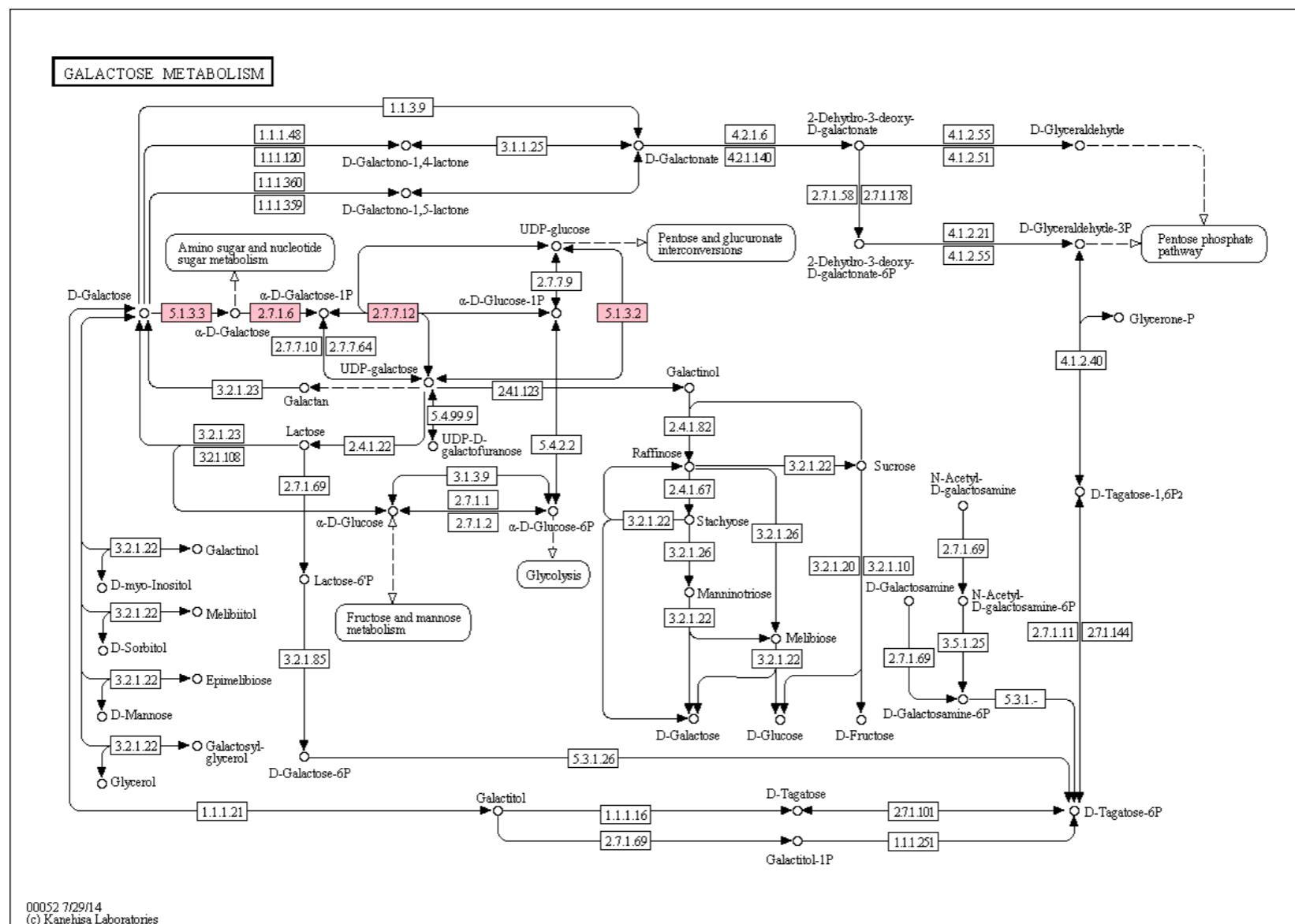
- Example: Metabolism



<http://www.genome.jp/kegg/pathway/map/map01100.html> (11/18/2014)
 KEGG: Kyoto Encyclopedia of Genes and Genomes

Systems Biology

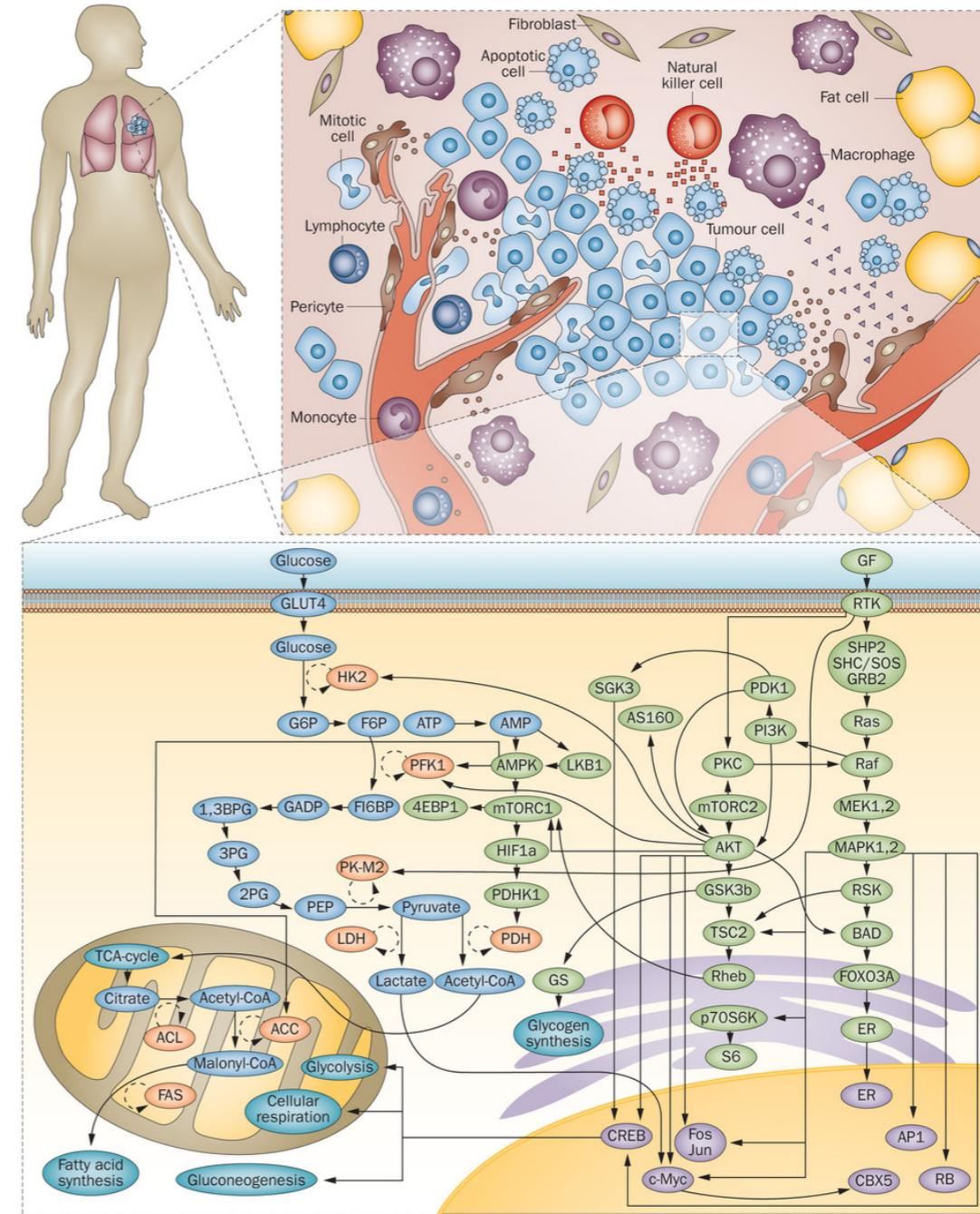
- Example: Metabolism: Galactose Metabolism (degradation)



<http://www.genome.jp/kegg/pathway/map/map01100.html> (11/18/2014)
KEGG: Kyoto Encyclopedia of Genes and Genomes

Systems Biology

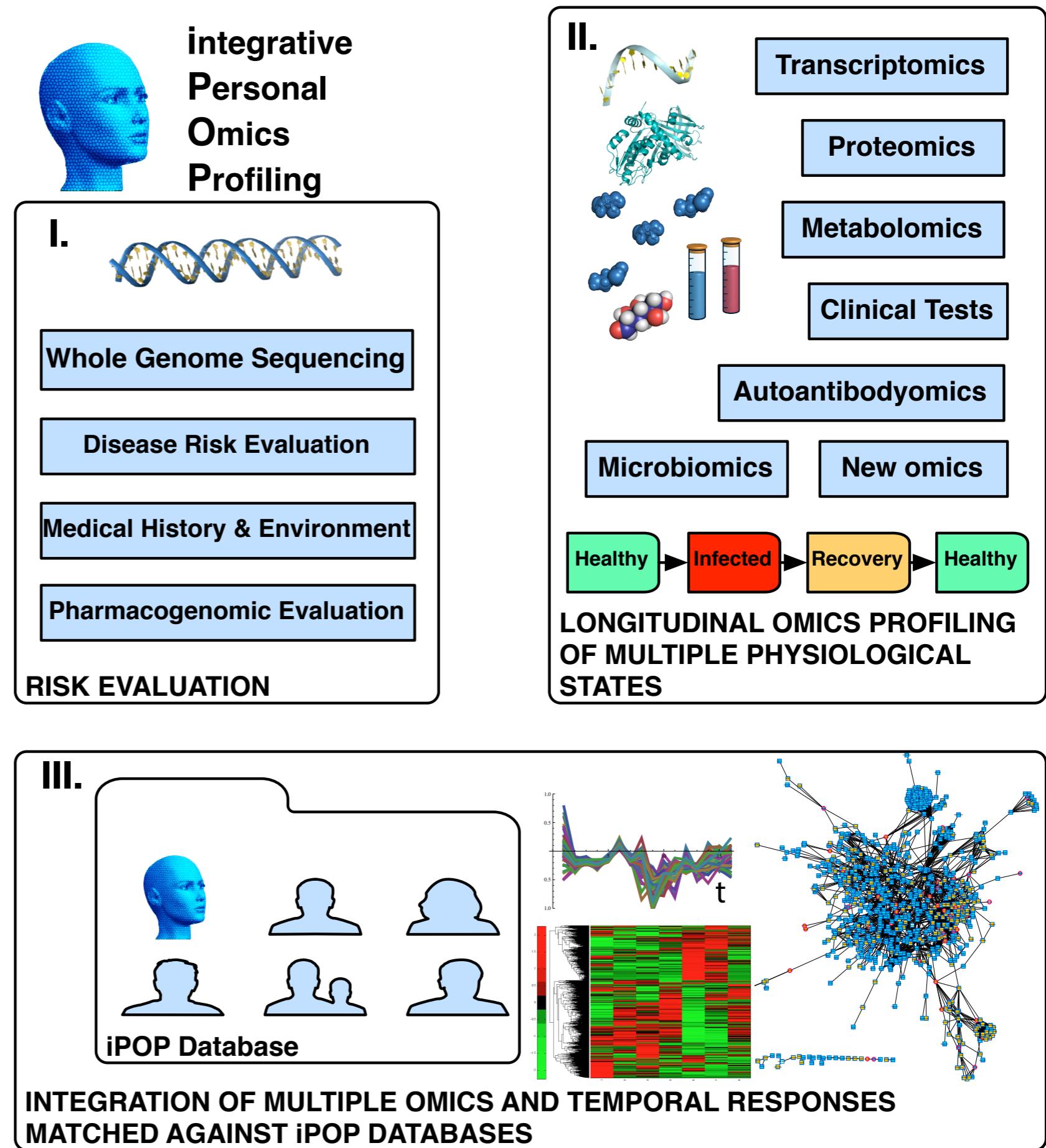
- Example: Cancer
 - Different Networks
 - Homeostasis
 - Molecular components



Werner, H. M. J. et al. (2014) *Nat. Rev. Clin. Oncol.* doi:10.1038/nrclinonc.2014.6

Systems Medicine

- Personalized
 - Determine risks
 - Monitor
 - Integrate



Mias and Snyder, Quantitative Biology 1(1) p. 71 (2013)
 Chen*, Mias*, Li-Pook-Than*, Jiang* et al Cell 148, 1293 (2012)

Systems Biology

- Models
- Experiments
 - data
 - theory
 - computation
- **Robert Wiener (1894-1964)**
- **cybernetics and systems controls**
- **20th century biochemistry**
- **21st century Leroy Hood and others**

Systems Biology

- Experiments
- **BIG DATA!**

Reformulate biological problems in terms of mathematical models.

Models need computational approaches

Big data handling in storing, retrieving useful information and relaying/displaying this information

Data - Omics

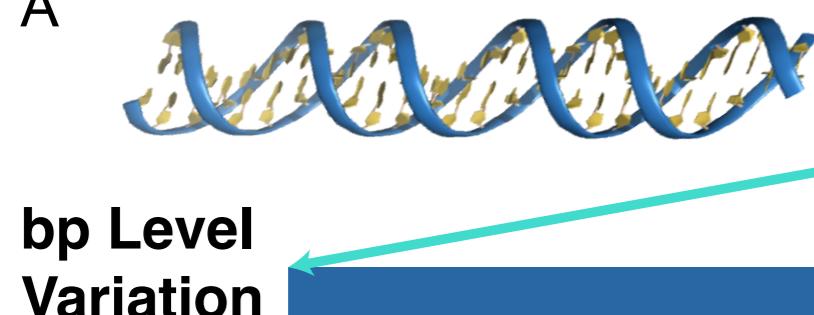
Molecular Components

Nucleic acids
Proteins
Lipids
Carbohydrates

Genomics

DNA VARIANTS

A



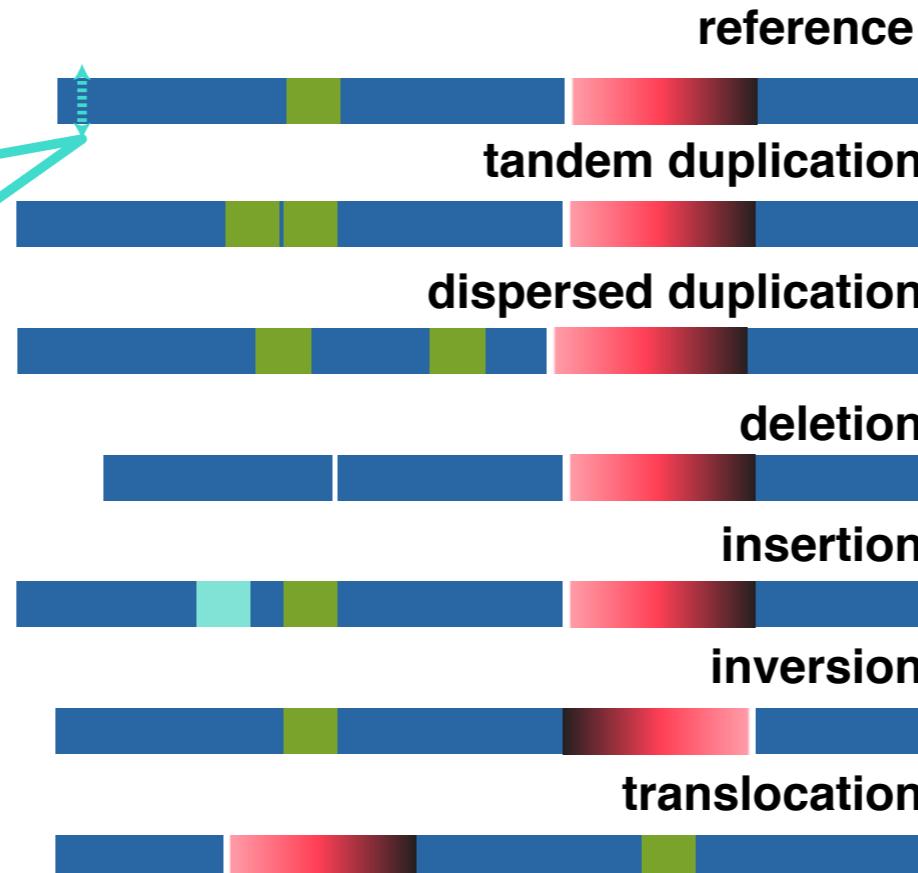
bp Level Variation

reference	<i>ggcttccaggaactc</i>
point mutation	<i>ggcttccaga</i> aactc <i>ggcttccaggaactc</i>
insertion	<i>ggcttccagg</i> gaactc <i>ggcttccaggaactc</i>
deletion	<i>ggcttccaggactc</i> <i>ggcttccagga</i> Xactc



MinION

Structural Variation [>1000 bp]



Illumina



SOLiD



PacBio

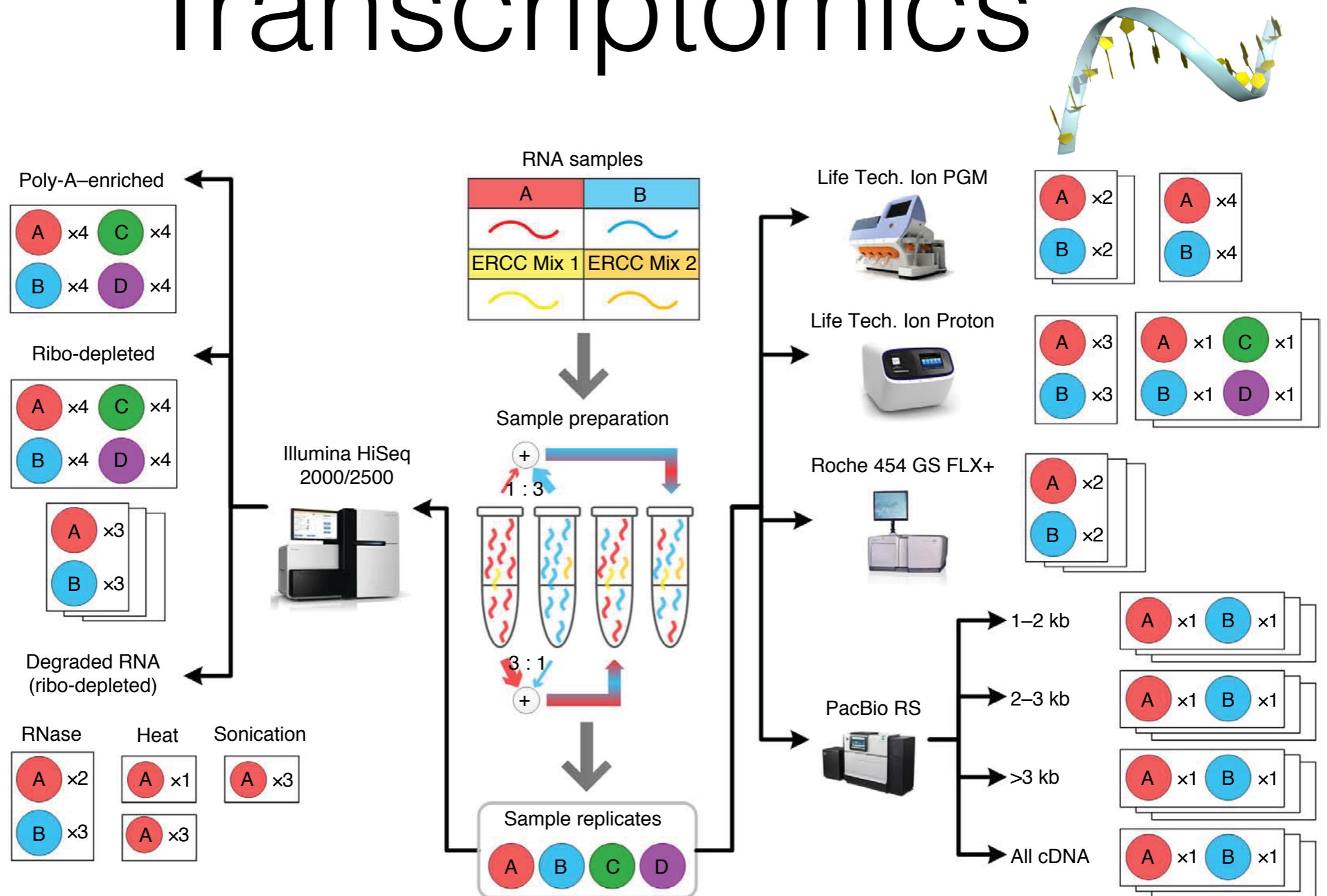


454



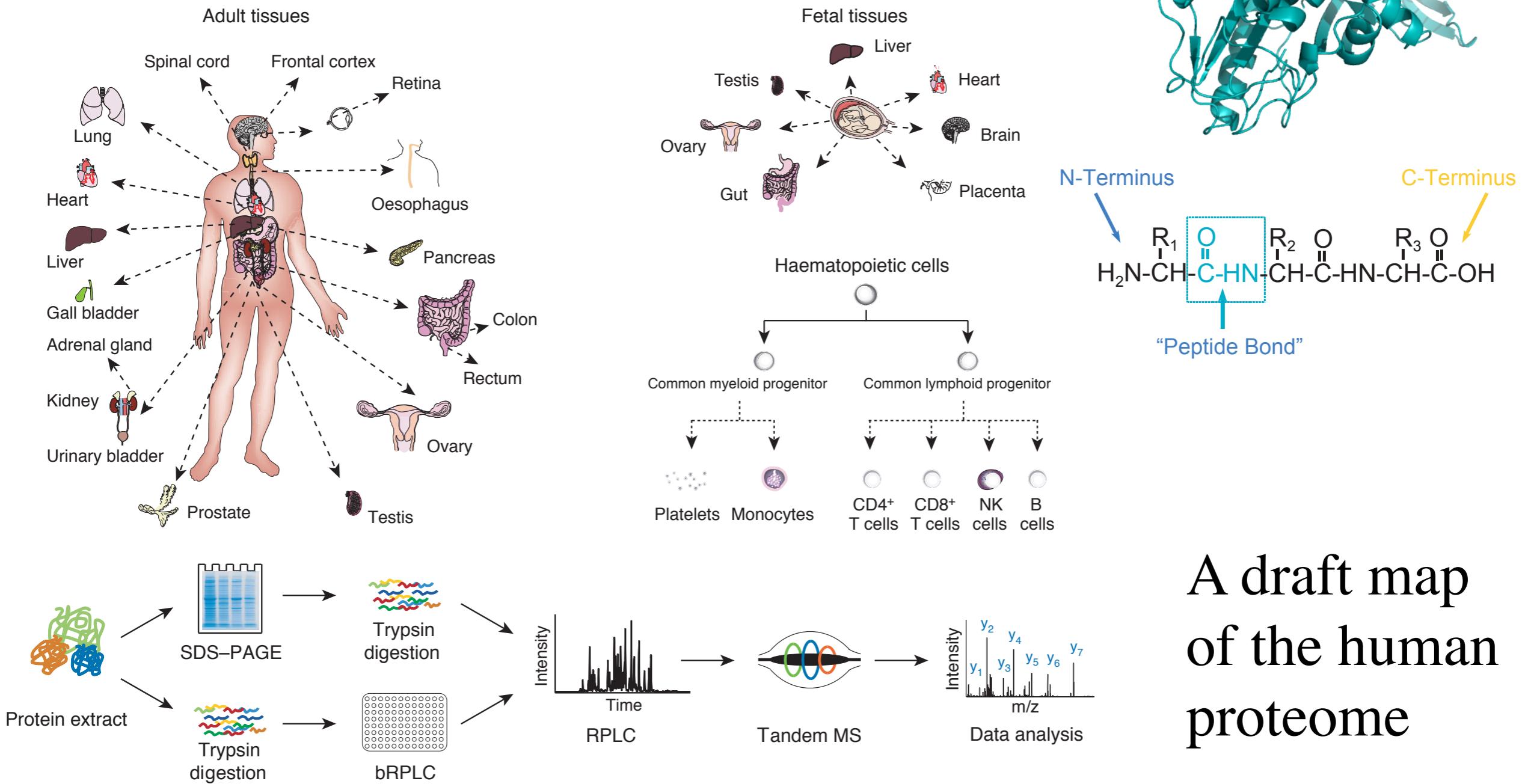
Ion Torrent

Transcriptomics



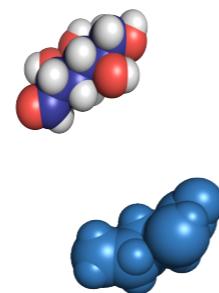
Li, Tighe et al., Nature Biotechnology 12(9), p. 915 (2014)

Proteomics



Kim et al., Nature 509, p. 575 2014

Metabolomics



thermofisher.com

NMR

Other varieties



agilent.com

all metabolites in cells

- small molecules
- lipids
- peptides
- amino acids
- nucleic acids
- organic acids

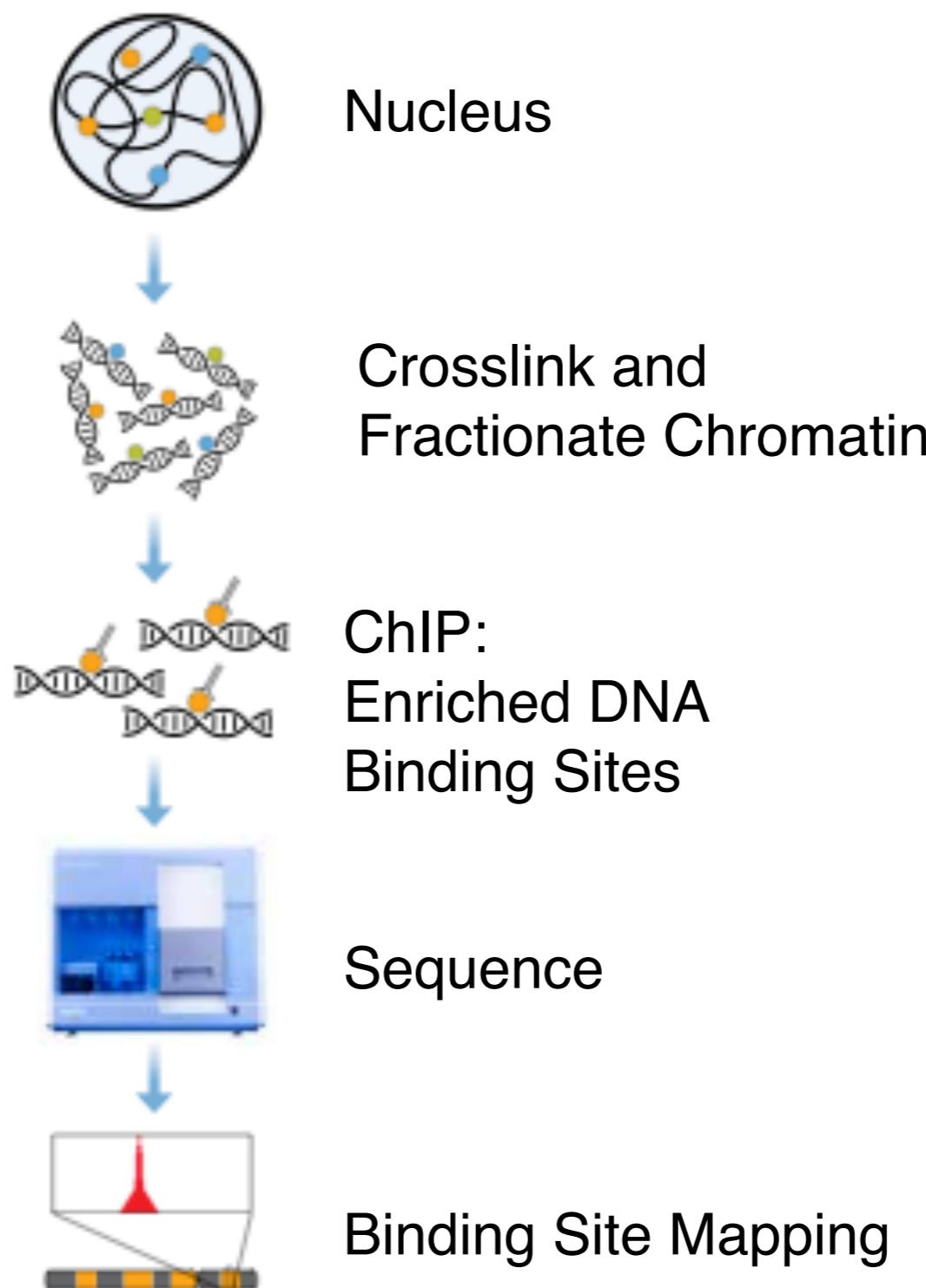
Lipidomics

- lipids
- metabolic signaling
- energy storage
- cell proliferation
- cell migration
- apoptosis
- cellular membrane

Interactions

Whole-Genome Chromatin IP Sequencing (ChIP-Seq)

identify binding sites of DNA-associated proteins



Interactions

Protein Arrays

Peptide Array

Bead Methods

Two Hybrid Methods

Many more!

Practical Issues

Not quantitative enough

Expensive enough

Inconclusive

Not exhaustive

Not dynamic

Systems Biology

- **Applications**
 - Genotype to Phenotype
 - In-silico Cell
 - Physiological Models
 - Personalized Precise and Predictive Medicine

Modeling

What's new?

- Technological Advancements (e.g. mass spectrometry/sequencing/imaging).
- High Performance Computing
- Information Storage abilities
- Information sharing abilities

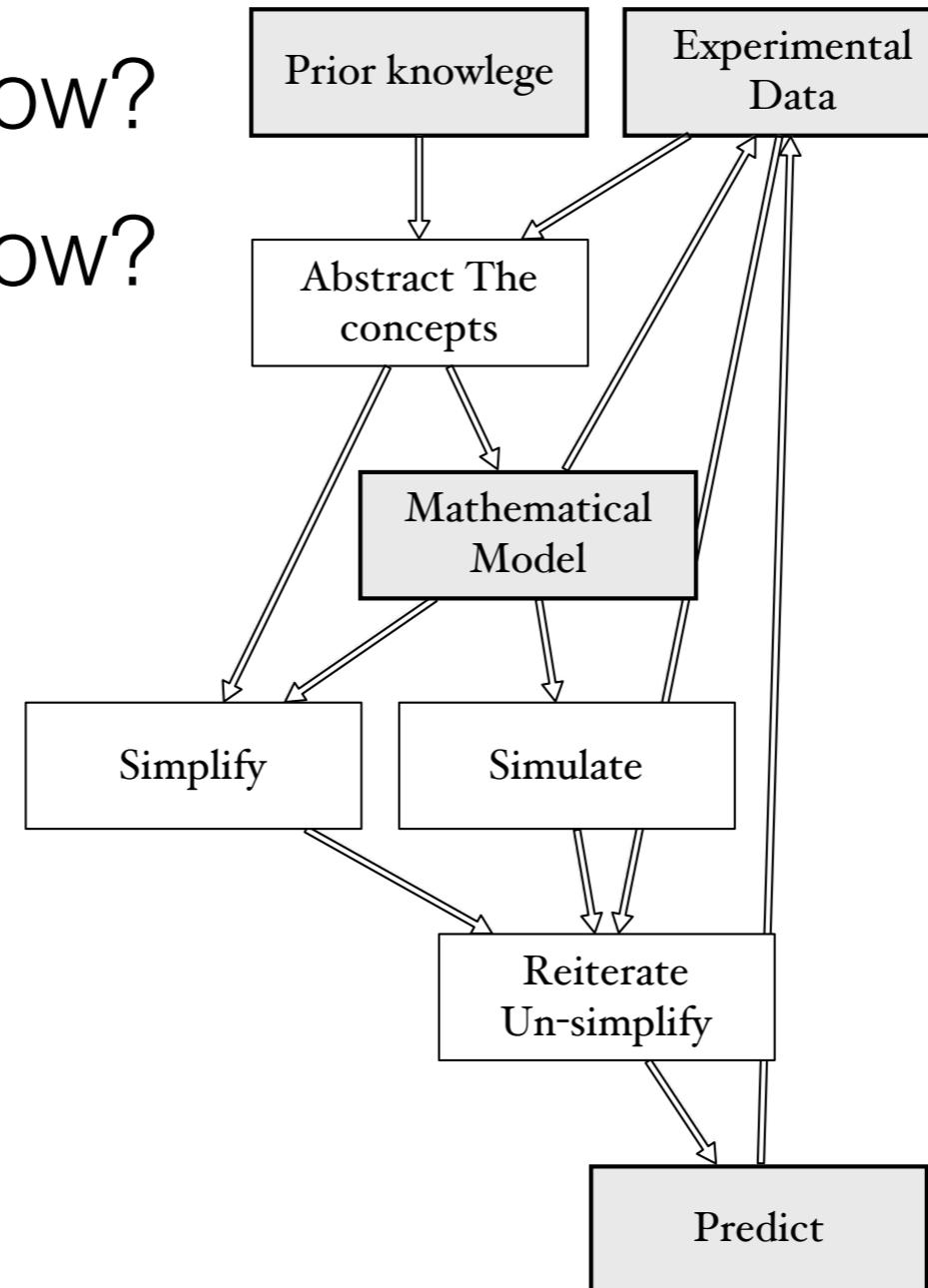
How to approach it

What is the question?

42

What do we know?

What can we know?

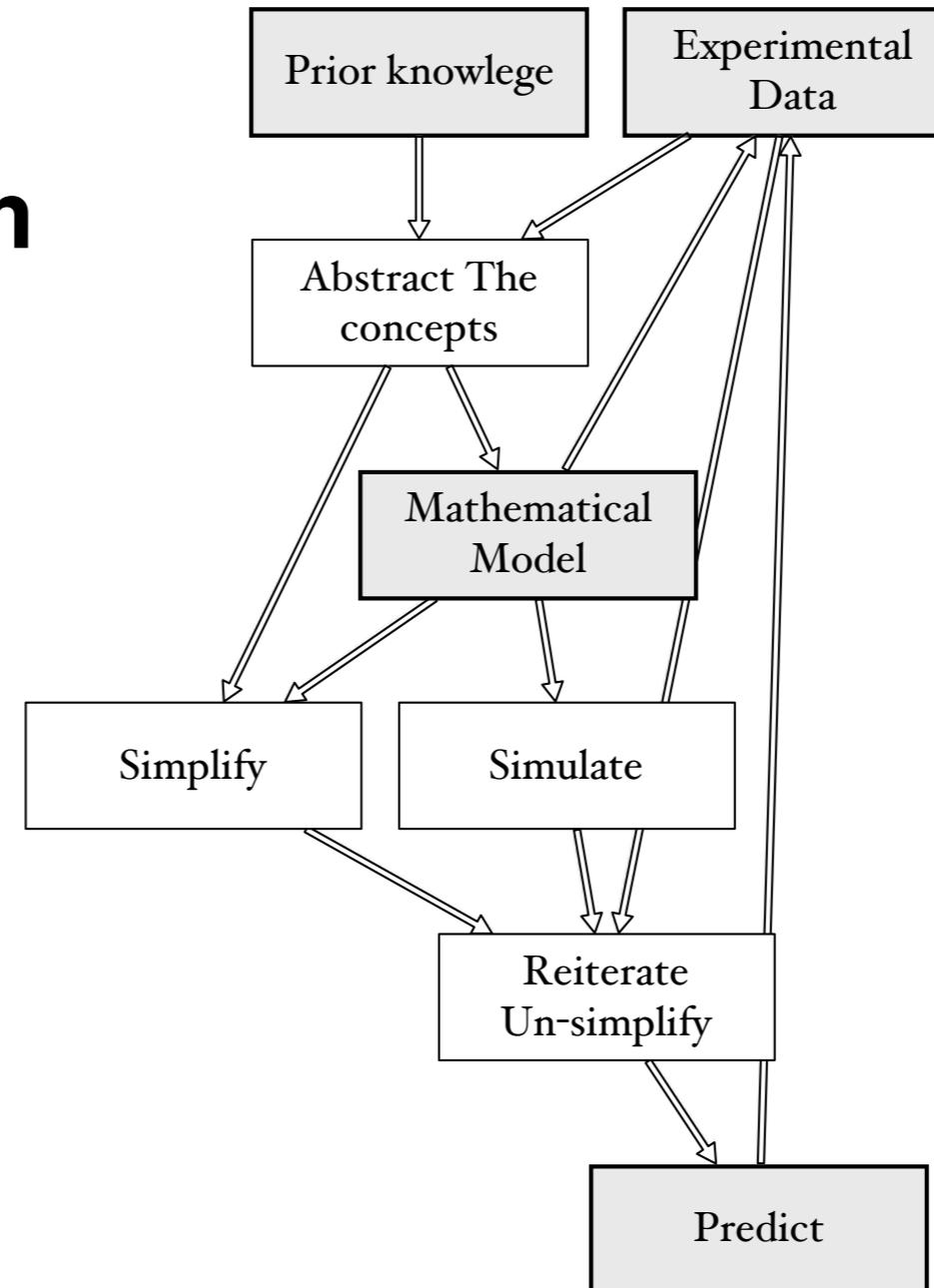


How to approach it

42

Underlying system

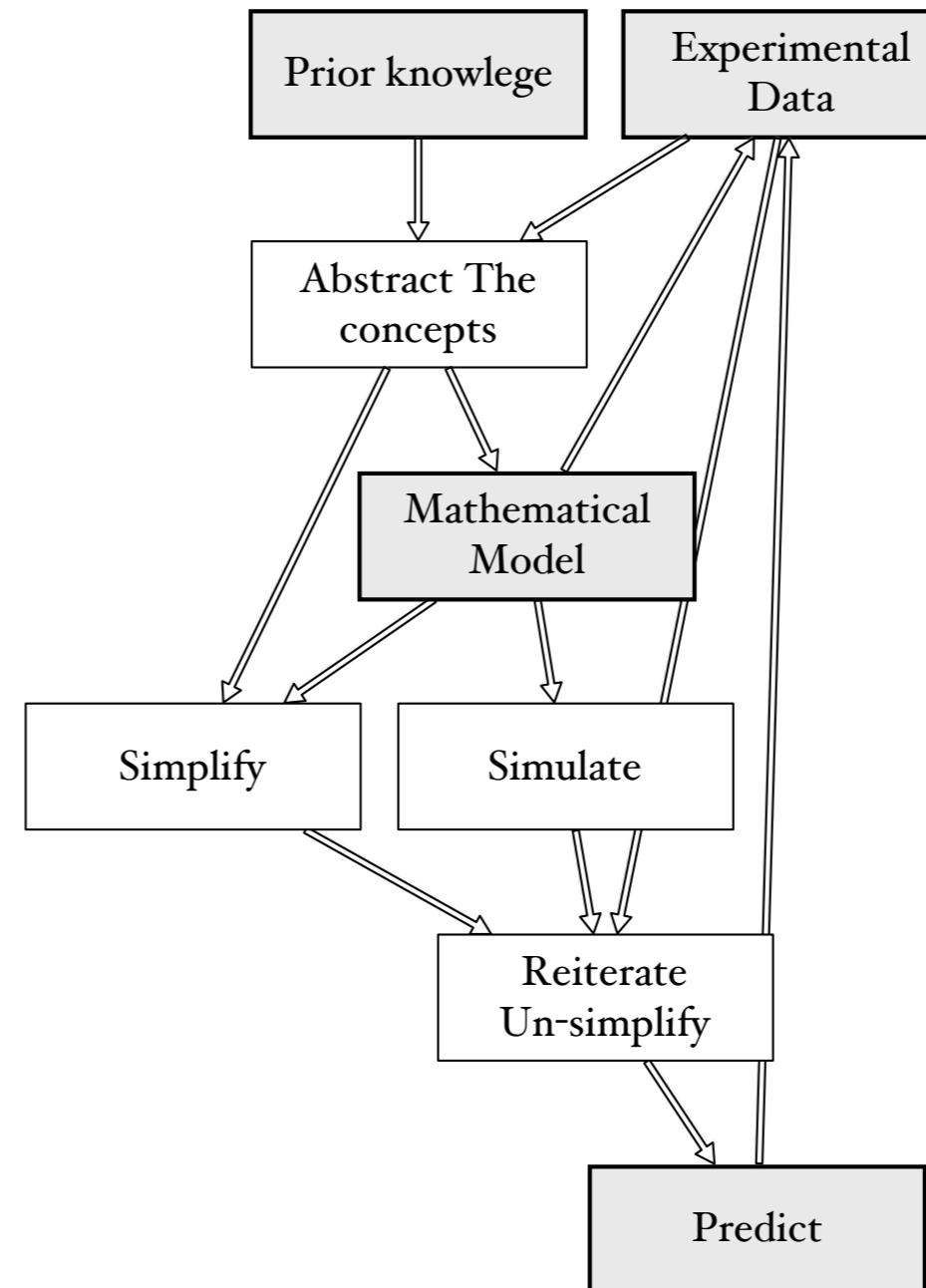
- characteristics
- responses
- interactions



How to approach it

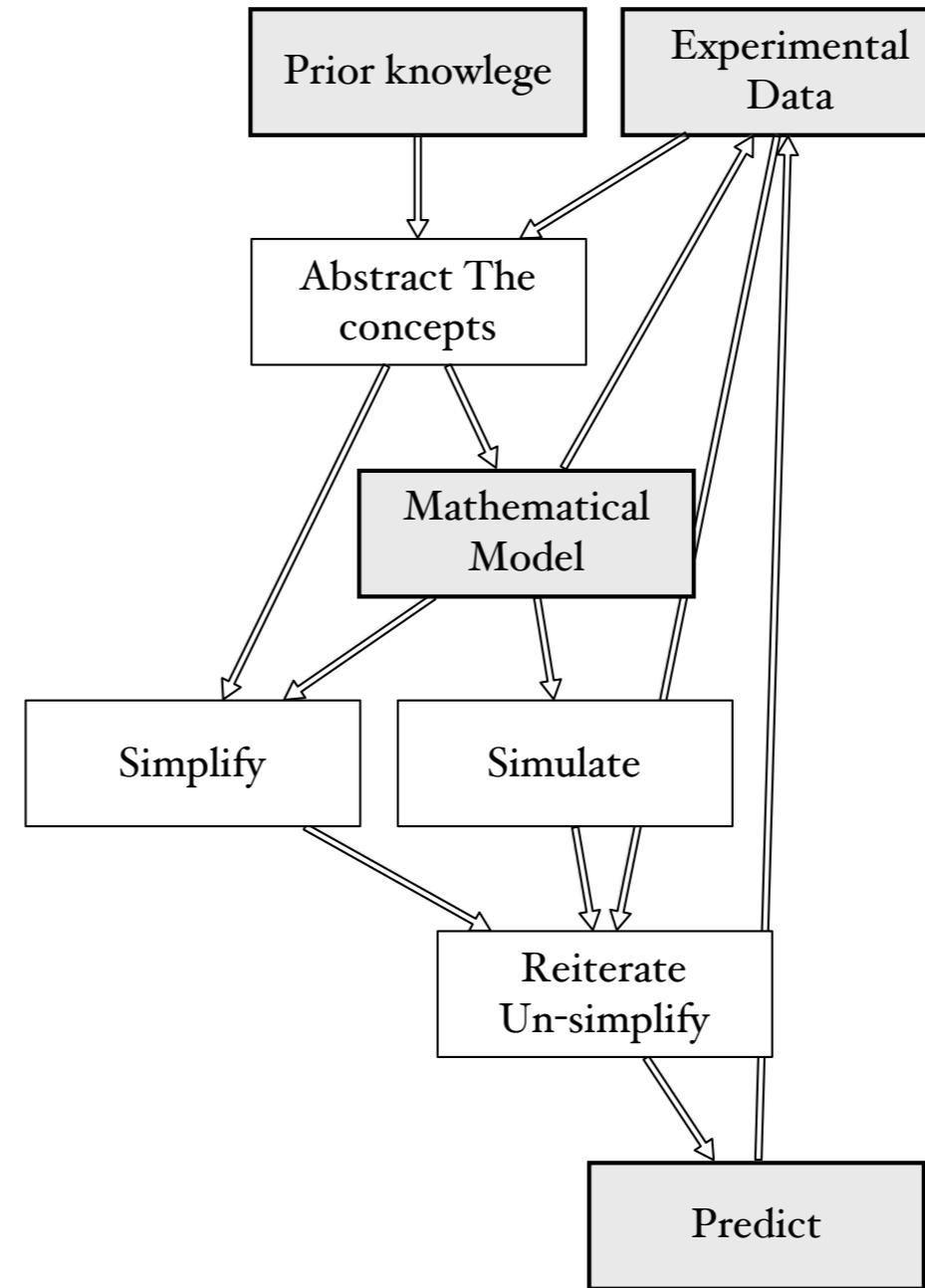
Pick your model (Based on system)

- atomic level
- cells
- molecular components
- dynamic Vs. Static
- How much to coarse grain?
- How many parameters?



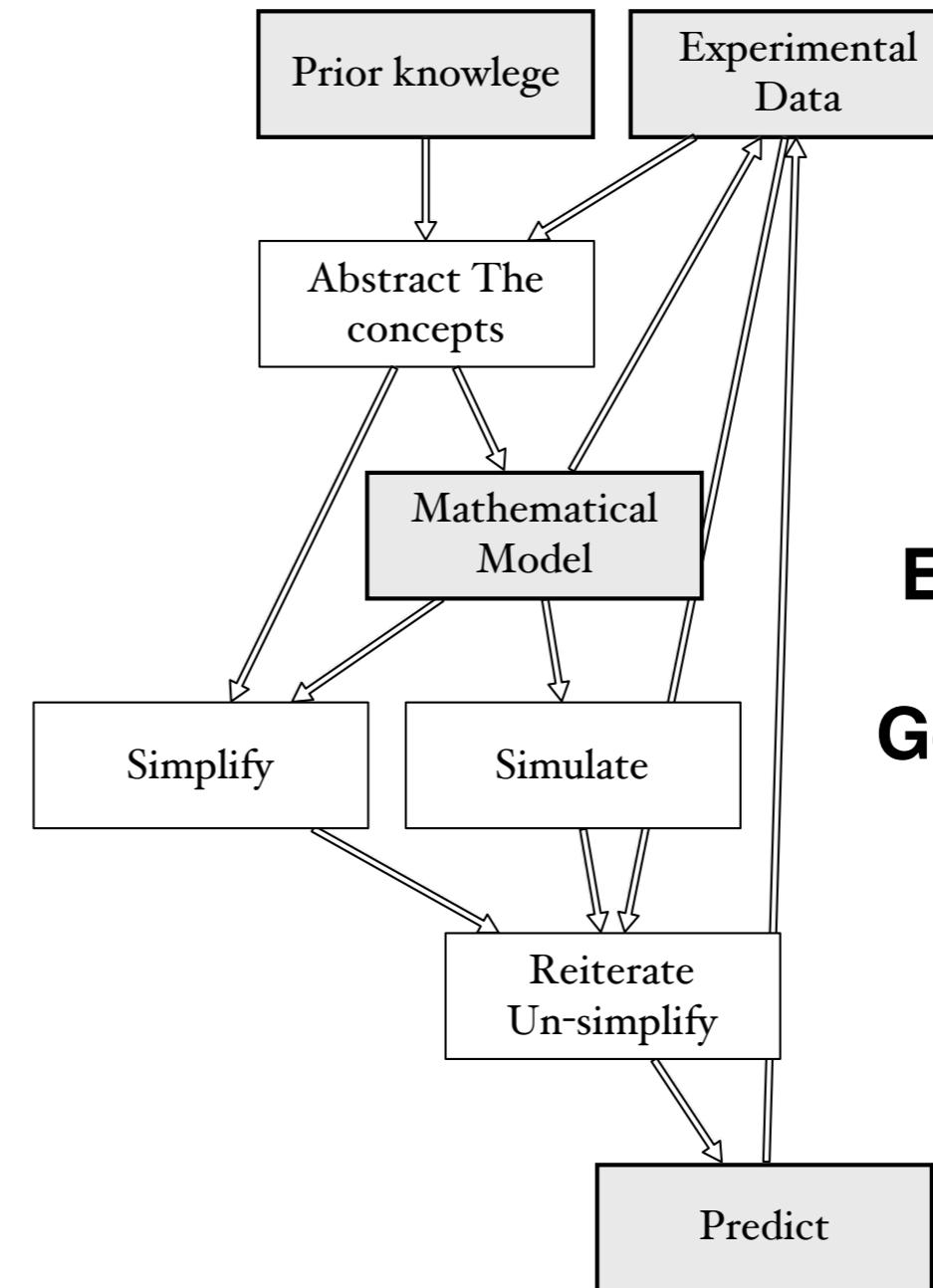
How to approach it

- Interactions and Networks
- Assumptions
- New knowledge required
- Storage of information



How to approach it

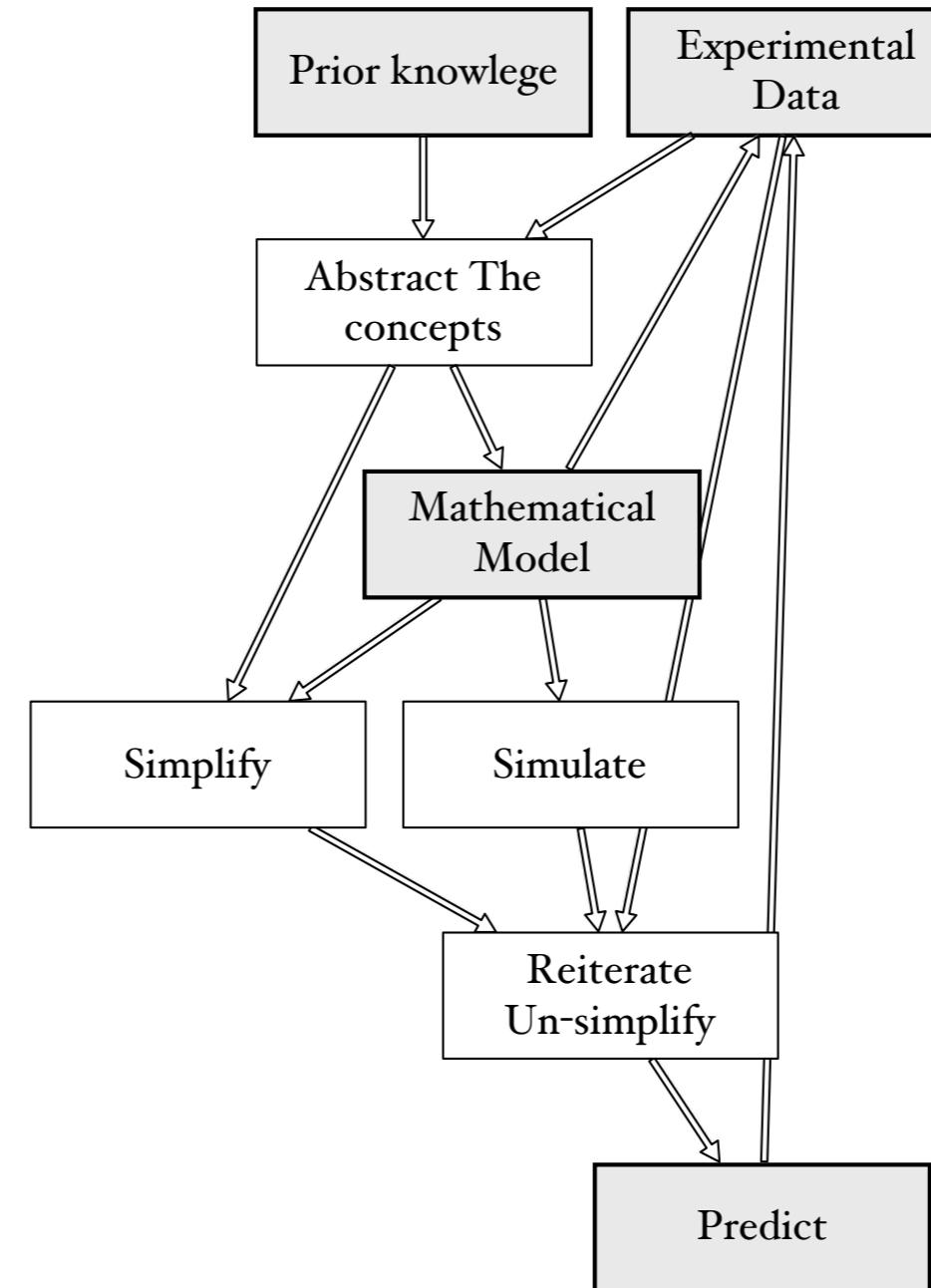
- Interactions and Networks
- Assumptions
- New knowledge required
- Storage of information



Experimental Tests
Generate Hypotheses

How to approach it

- Interactions and Networks
- Assumptions
- New knowledge required
- Storage of information



Validation
may invalidate
cannot confirm

Biochemical Models

Biochemical Models

Chemical Reactions

A transforms to B



- conversion
- modification
- dimerization

A associates with B to form C



- association
- synthesis

C dissociates to A and B



- dissociation
- decomposition

Biochemical Models

Chemical Reactions

null species (e.g. constant abundance)

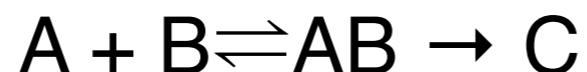
- | | |
|---------------------------|--|
| $A \rightarrow \emptyset$ | <ul style="list-style-type: none">• degradation |
| $\emptyset \rightarrow B$ | <ul style="list-style-type: none">• production• discarded reactants |

Elementary Irreversible Reactions

- Single step
- one direction

Chemical Reactions

Can put it all together: e.g.

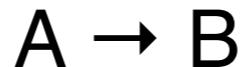


- \rightleftharpoons • 2 elementary reactions
• C covalent modification of AB

Biochemical Models

Chemical Reactions

A transforms to B



Reaction Rate = $k [A]$
mass action

Rate of changes proportional to concentration

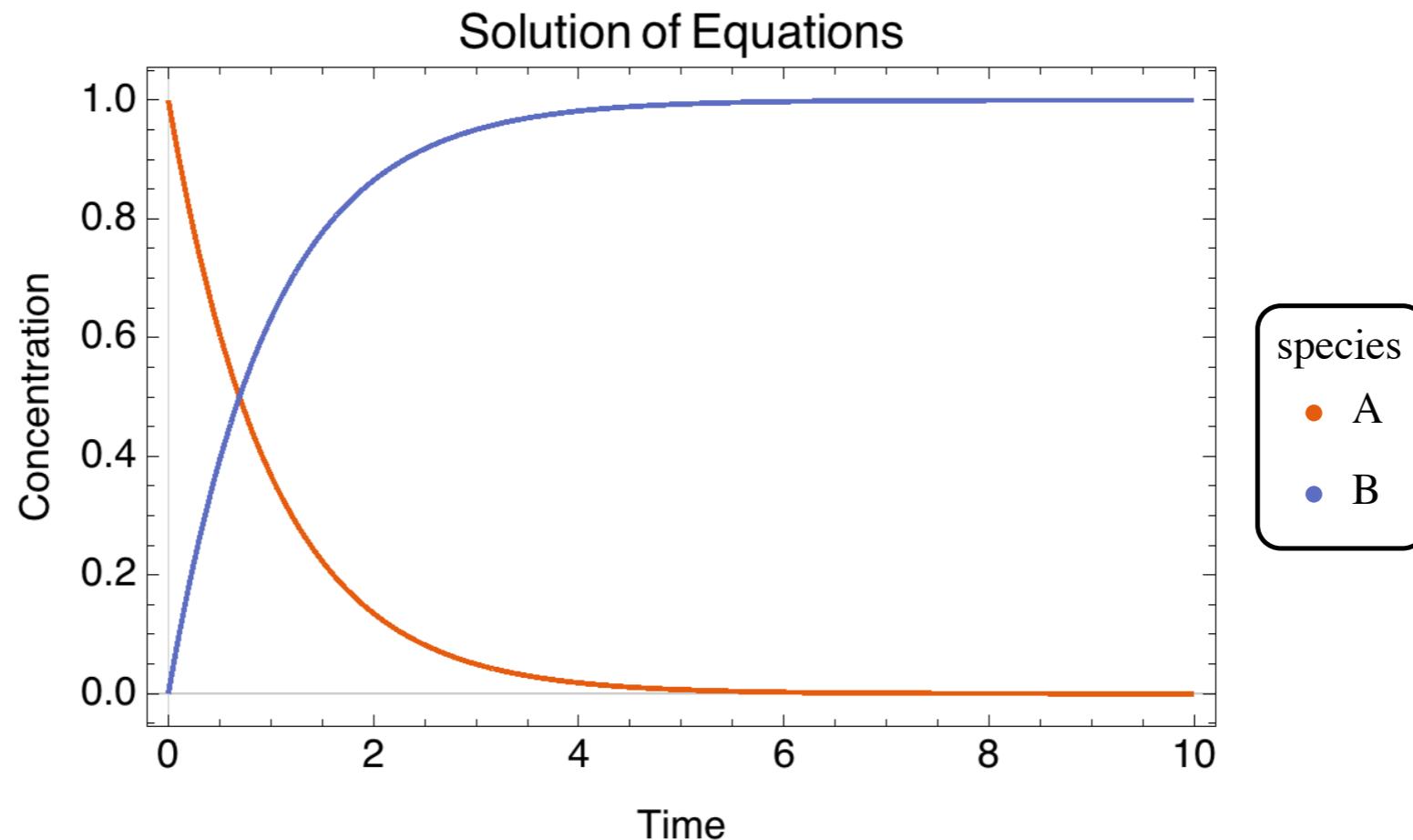
$$\frac{d[A]}{dt} = -k(A); \frac{d[B]}{dt} = k(A)$$

Biochemical Models

Chemical Reactions

A transforms to B

$$\frac{d[A]}{dt} = -k(A); \frac{d[B]}{dt} = k(A)$$



Rate of changes proportional to concentration

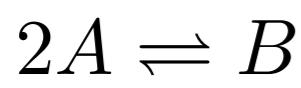
Biochemical Models

Chemical Reactions

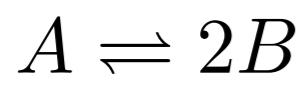
Chemical Reaction Rate Equation for One Species



$$\frac{dc}{dt} = k_f ab = k_f(a_0 - c)(b_0 - c)$$



$$\frac{db}{dt} = k_f(a_0 - 2b)^2 - k_r b$$



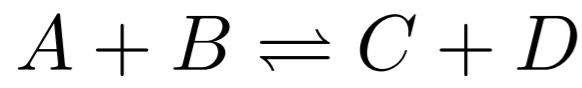
$$\frac{db}{dt} = k_f(a_0 - \frac{b}{2}) - k_r b^2$$



$$\frac{dc}{dt} = k_f(a_0 - c) - k_r(b_0 + c)(c_0 + c)$$



$$\frac{dc}{dt} = k_f(a_0 - c)(b_0 - c) - k_r c$$



$$\frac{db}{dt} = k_f(a_0 - c)(b_0 - c) - k_r(c_0 + c)(d_0 + c)$$

x_0 : initial concentration for x

Biochemical Models

Chemical Reactions

$$X_i(t) = \frac{N_i(t)}{\Omega}$$

X_i : ith species concentration

Ω : system size = $N_A \times$ Volume

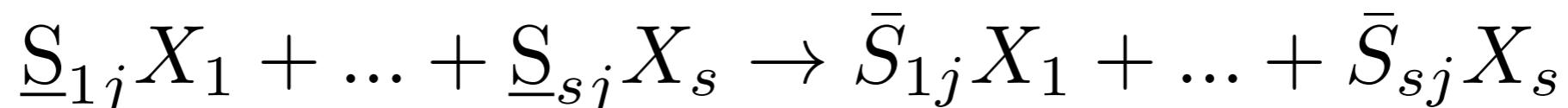
N_i : copy number

N_A : Avogadro's constant

Biochemical Models

Chemical Reactions

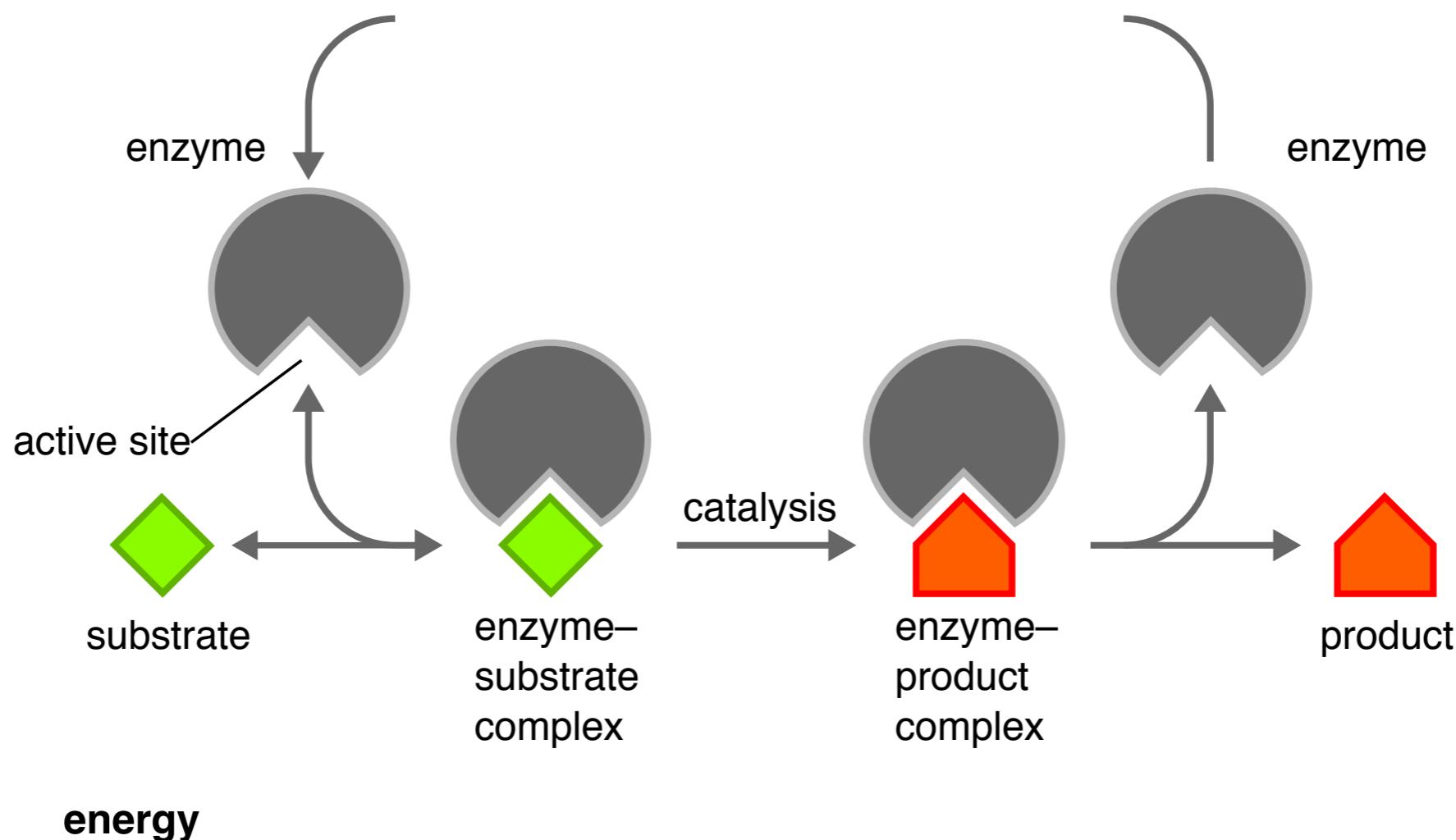
Reaction scheme R_j



S_{ij} stoichiometric coefficient indicates participation of X_i as a reactant

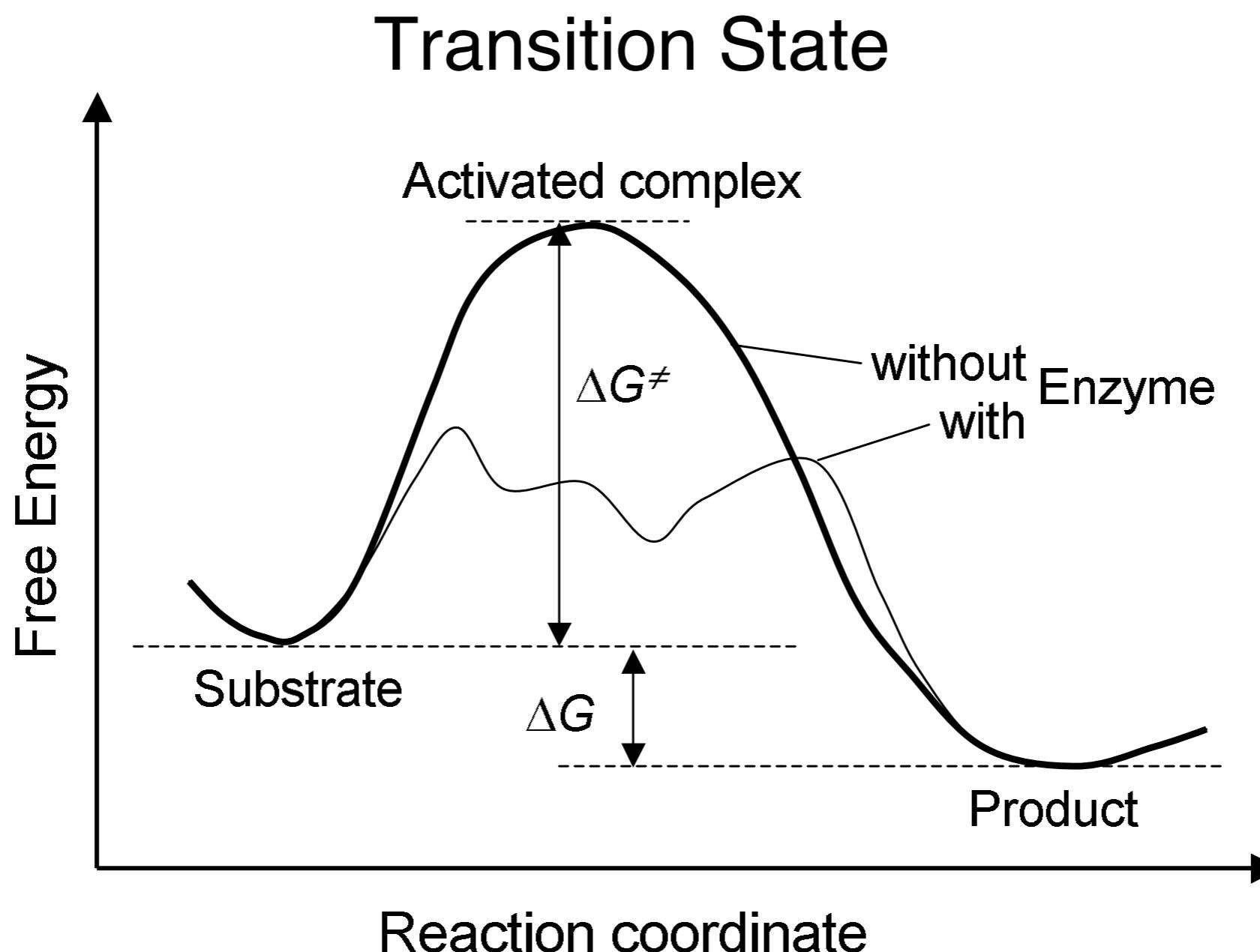
Biochemical Models

Chemical Reactions



Enzyme Catalyzed Conversion of Substrate to Product

Biochemical Models



Biochemical Models

Chemical Reactions

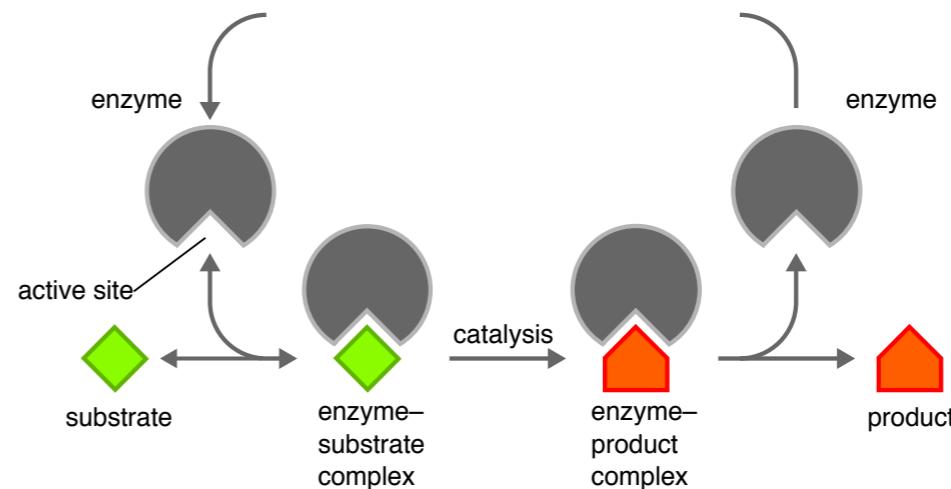


k_1



k_2

- 3 elementary reactions



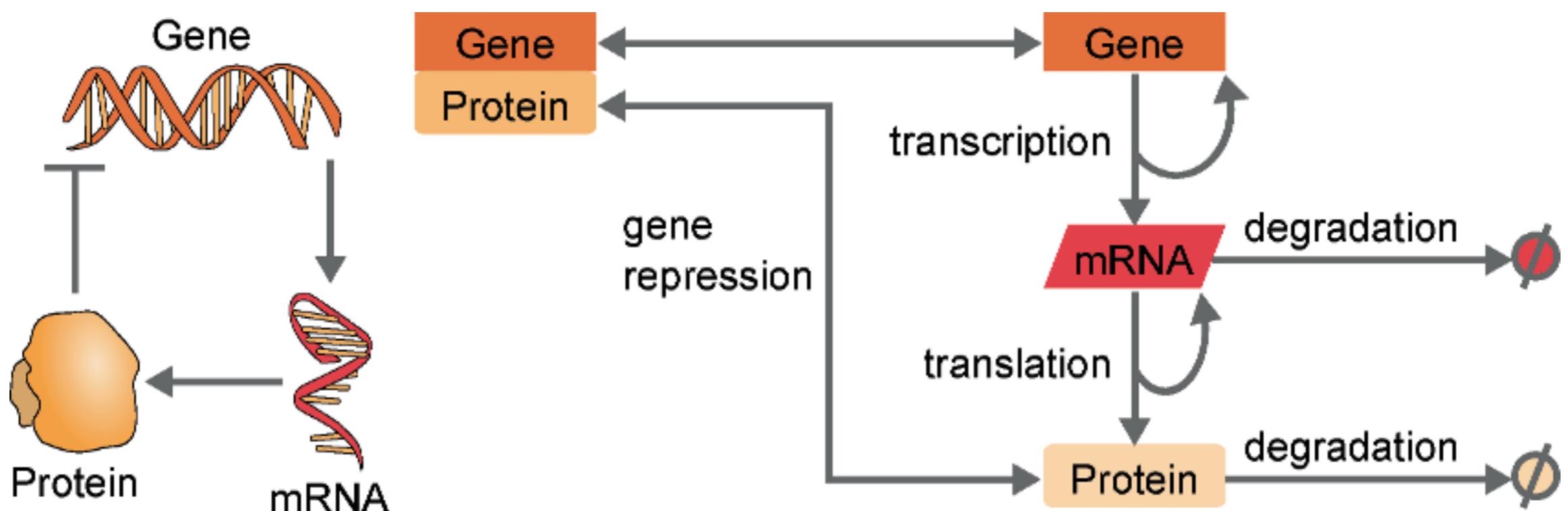
k_{eff}



- Can approximate to single step

Biochemical Models

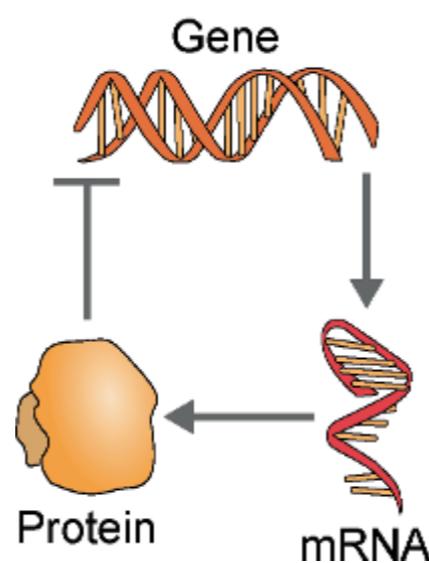
Simplified Gene Regulation



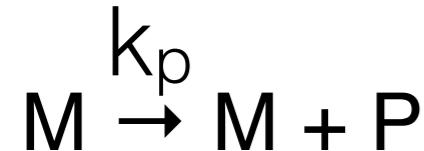
Biochemical Models

Simplified Gene Regulation

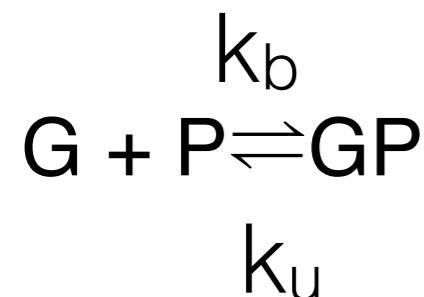
- Transcription



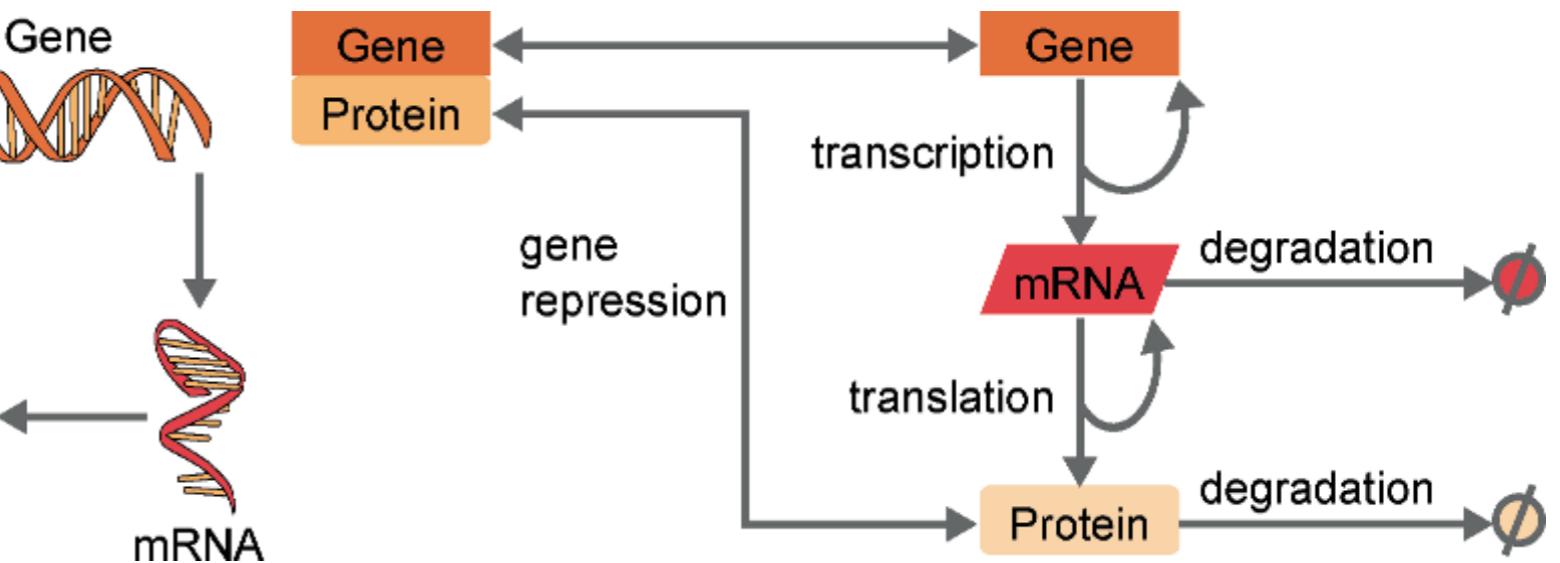
- Translation



- Binding/unbinding



- Degradation

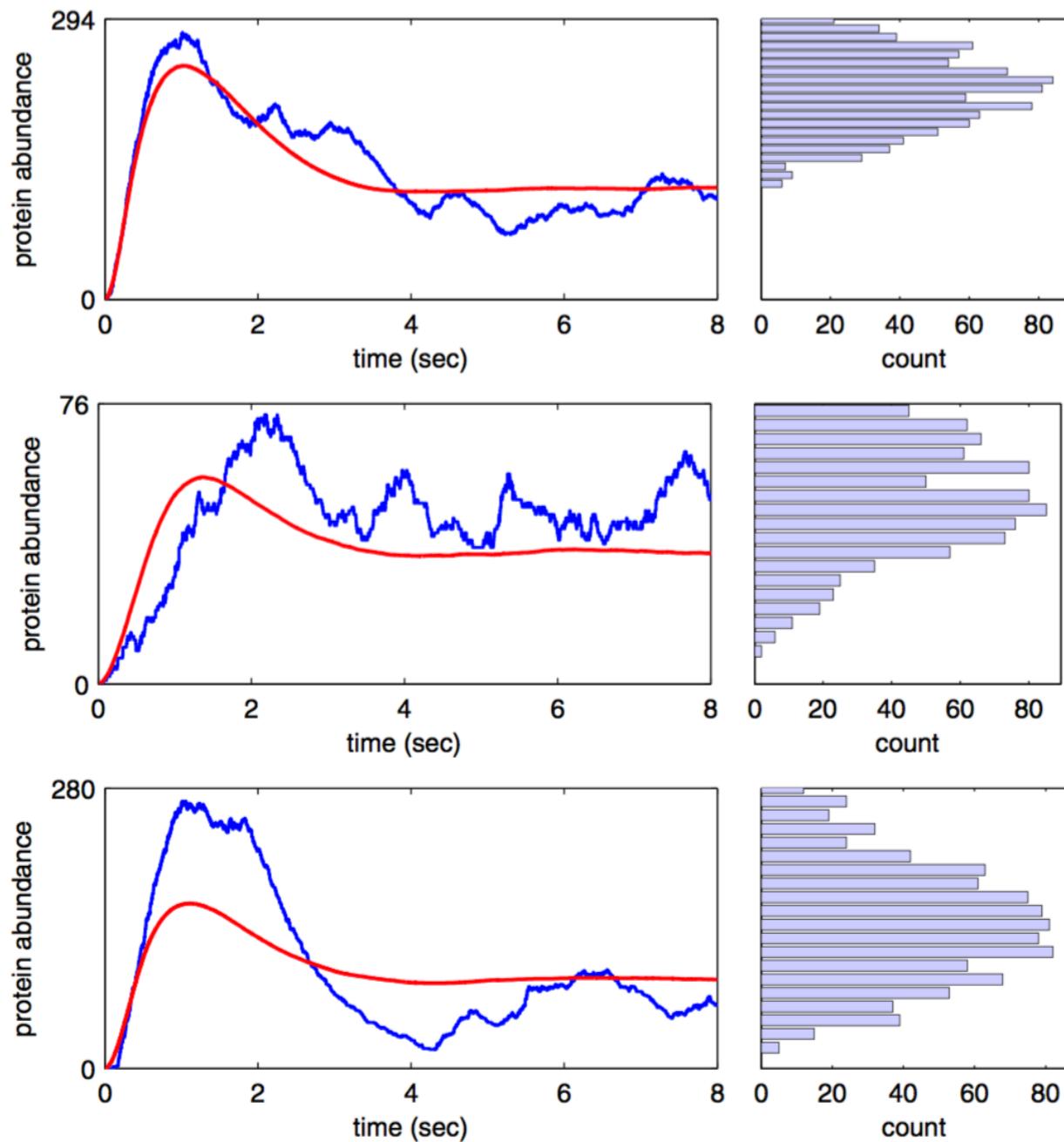


G: gene
M: mRNA
P: protein

Biochemical Models

Simplified Gene Regulation

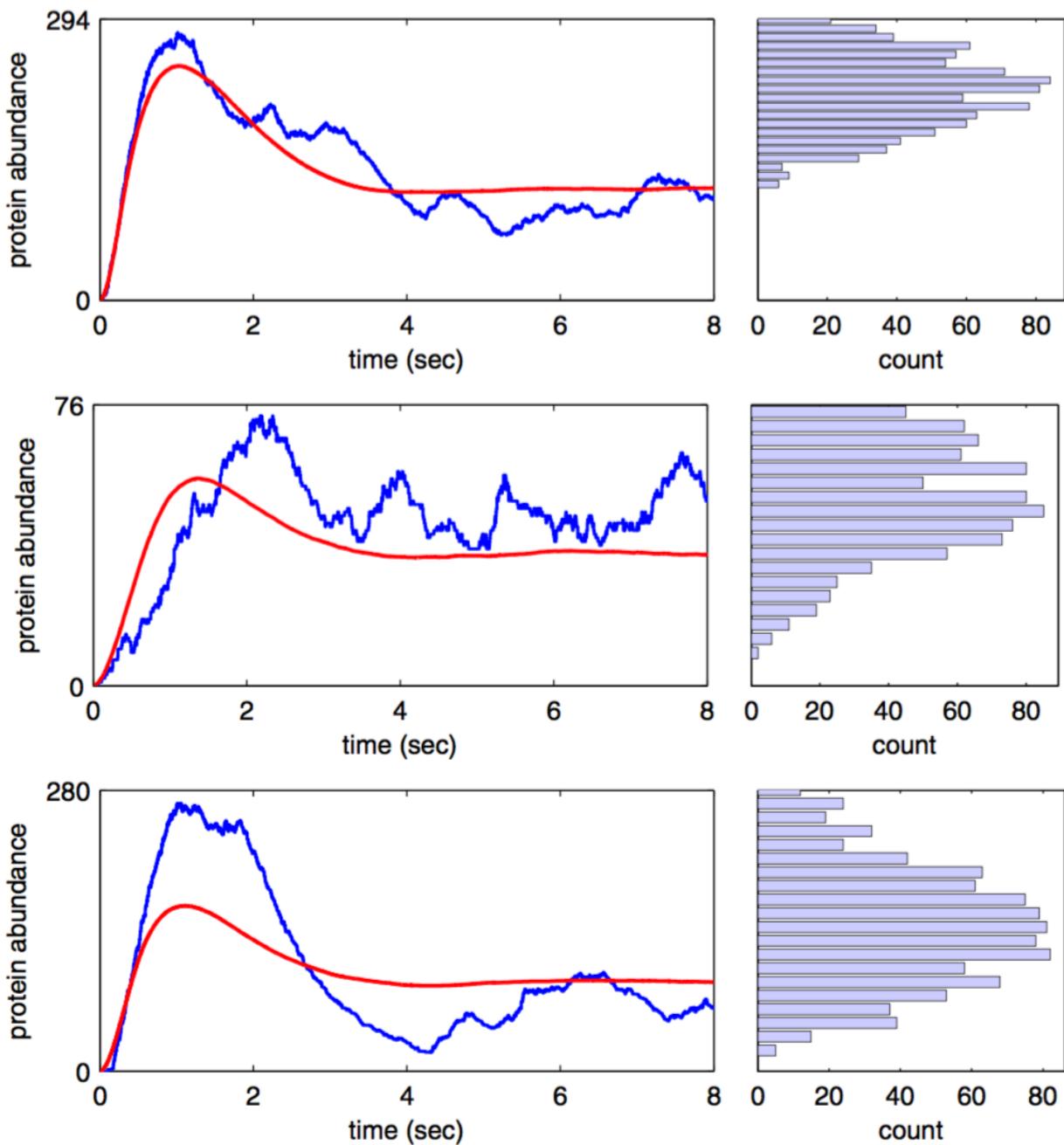
- Add stochasticity
- Time courses of a single stochastic simulation algorithm run (blue)
- Mean over 1000 runs (red curve).
- Initial conditions
 - ▶ 10 copies gene G
 - ▶ 0 copies of other species
- Endpoint histogram shows empirical probability that a cell will have a given protein abundance.



Biochemical Models

Simplified Gene Regulation

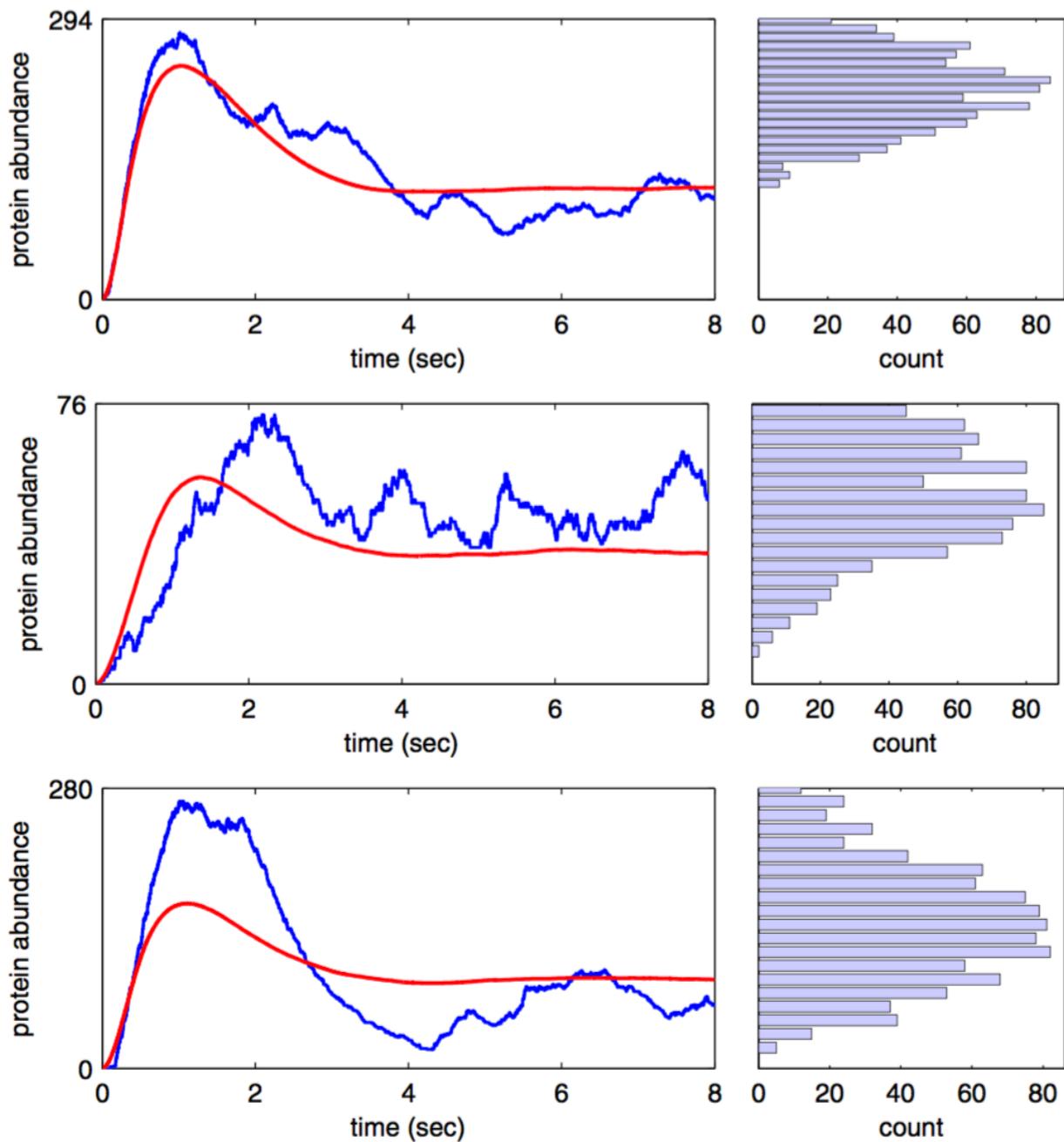
- Top: Small fluctuations with high copy numbers of expressed mRNA and protein.



Biochemical Models

Simplified Gene Regulation

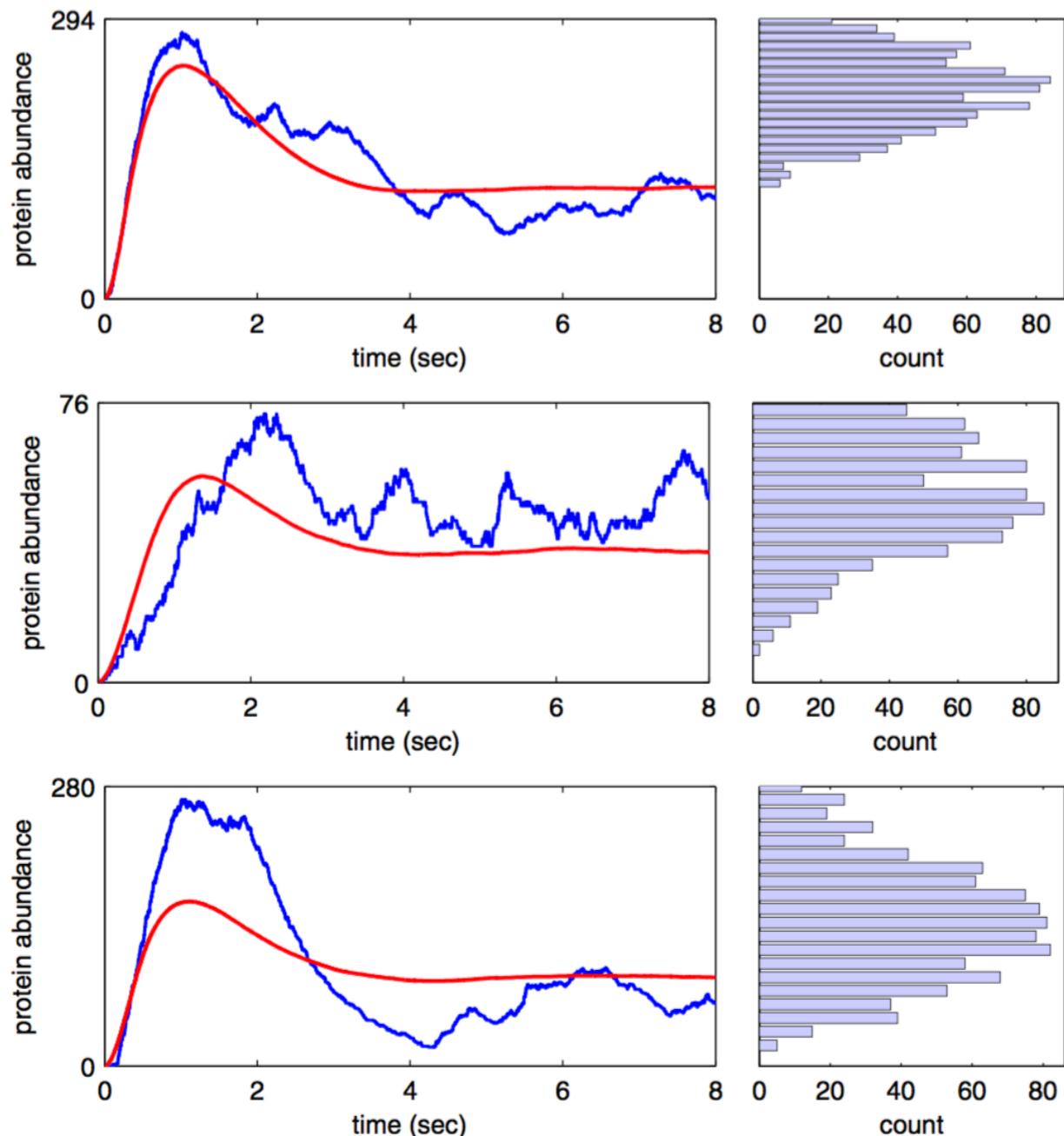
- Middle: A tenfold decrease in the transcription rate k_m leads to
 - ▶ Decrease in the expressed mRNA abundance
 - ▶ An associated decrease in protein abundance
 - ▶ Large fluctuations in the protein abundance.



Biochemical Models

Simplified Gene Regulation

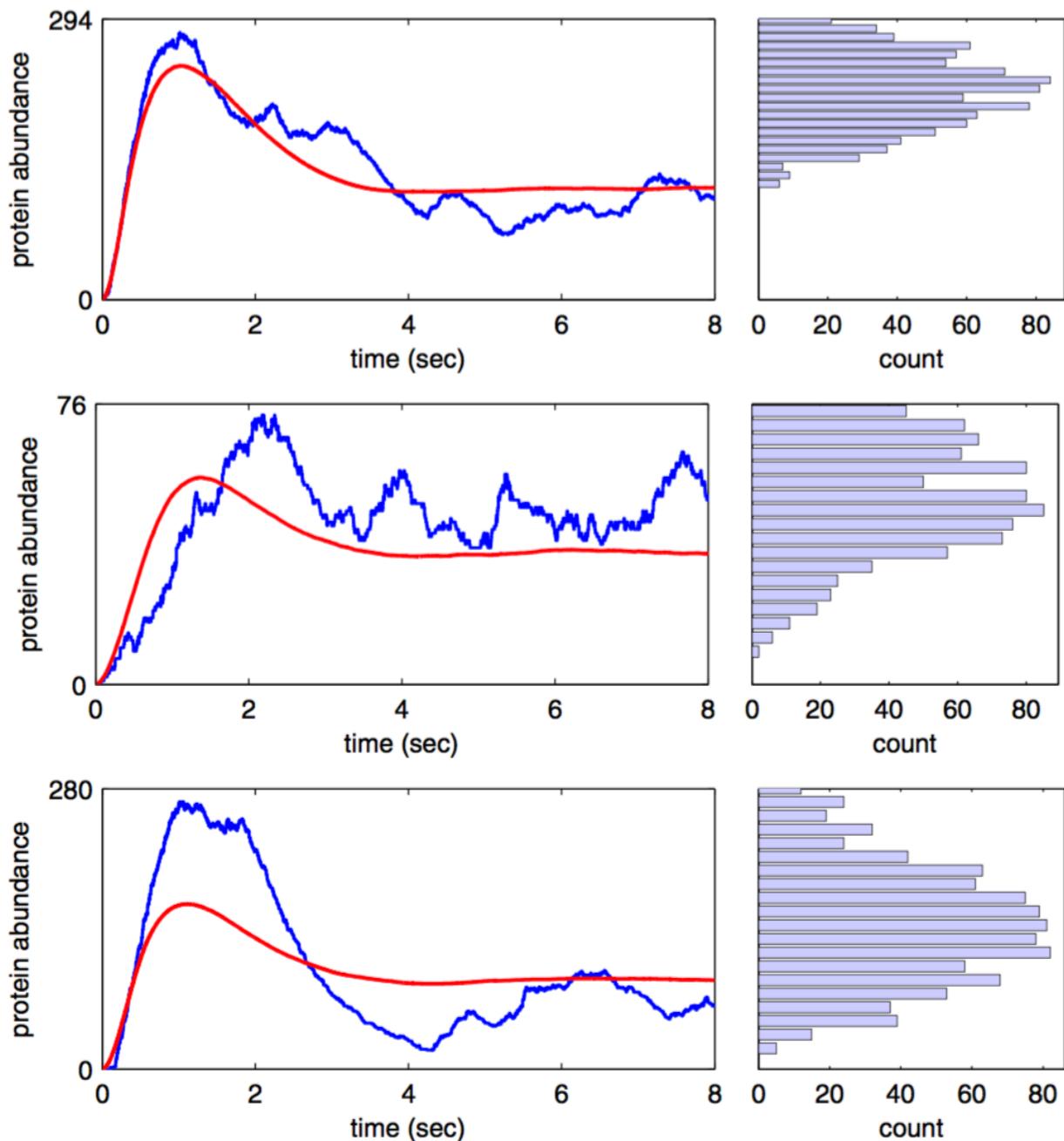
- Bottom: Fluctuations in mRNA abundance at the transcription level are a second important factor contributing to gene-expression noise.
 - ▶ Tenfold decrease rate of transcription
 - ▶ Rate of translation is increased fivefold to keep the protein abundance more or less the same as in the first case



Biochemical Models

Simplified Gene Regulation

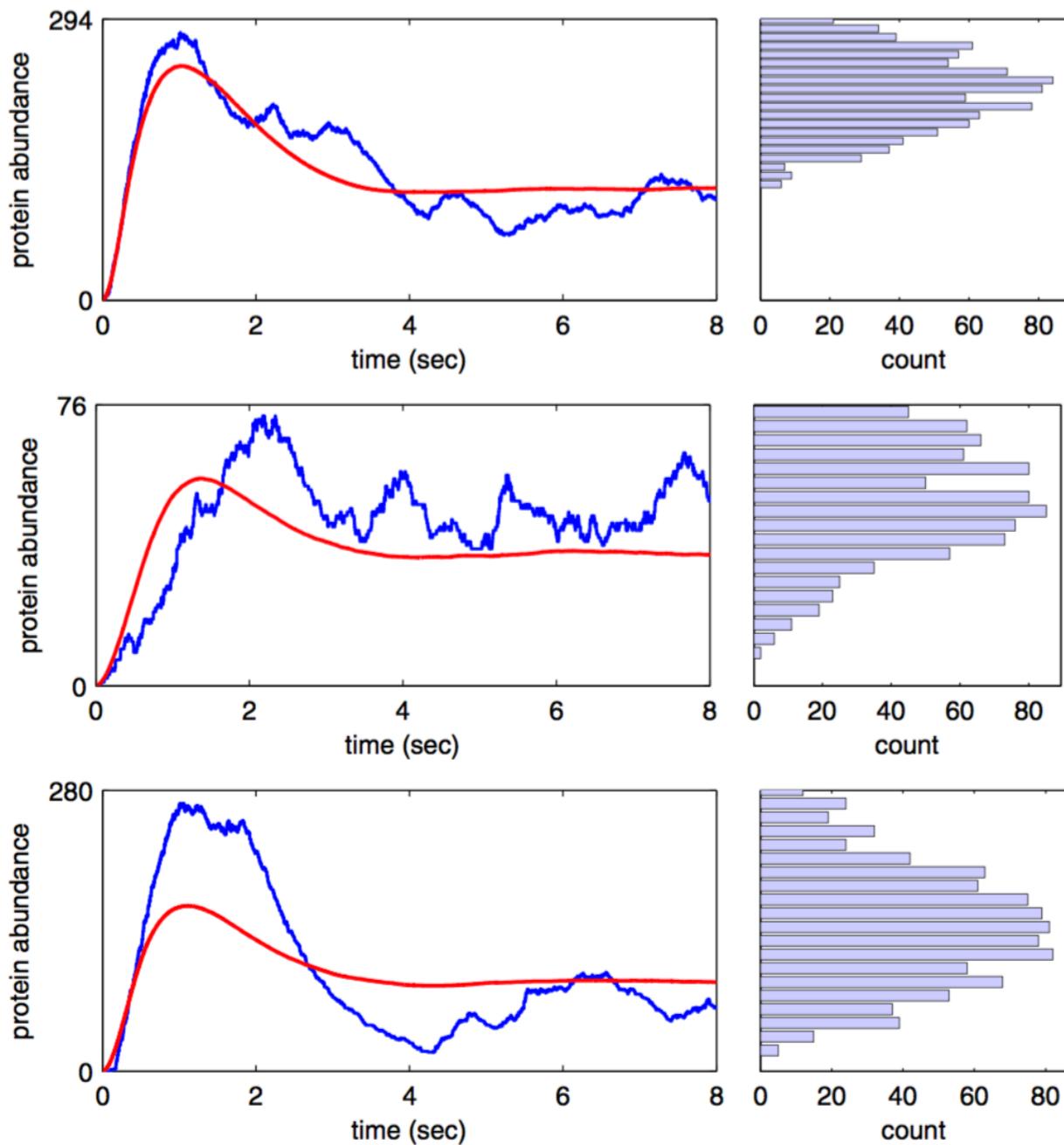
- Bottom:
 - ▶ increased gene-expression noise in spite of large protein abundance
 - ▶ Noise attributable to increased fluctuations in mRNA abundance
 - Causing increased fluctuations in the rate of protein synthesis.



Biochemical Models

Simplified Gene Regulation

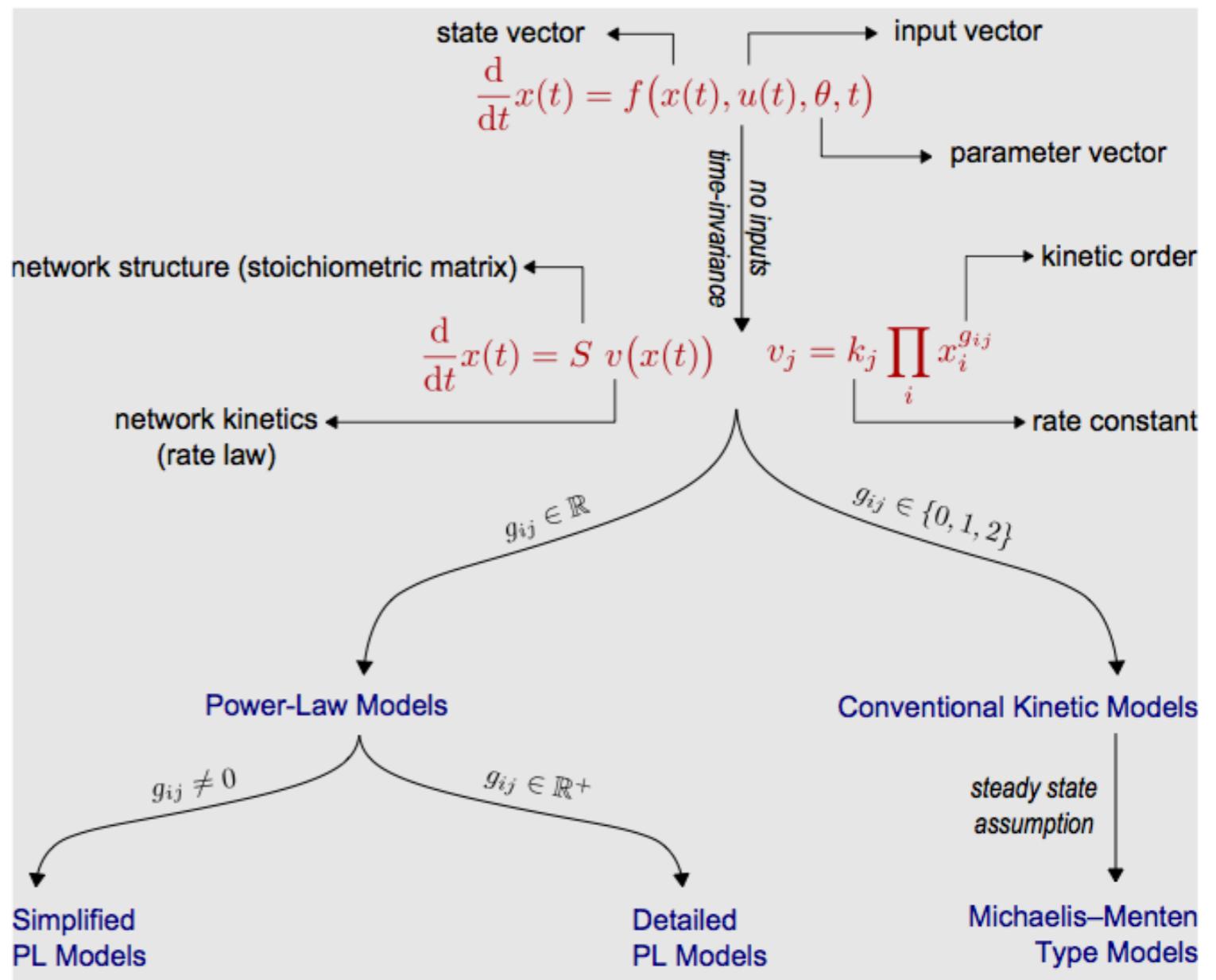
- Noise is propagated from transcription to translation.



Deterministic Models

Chemical Reaction Networks

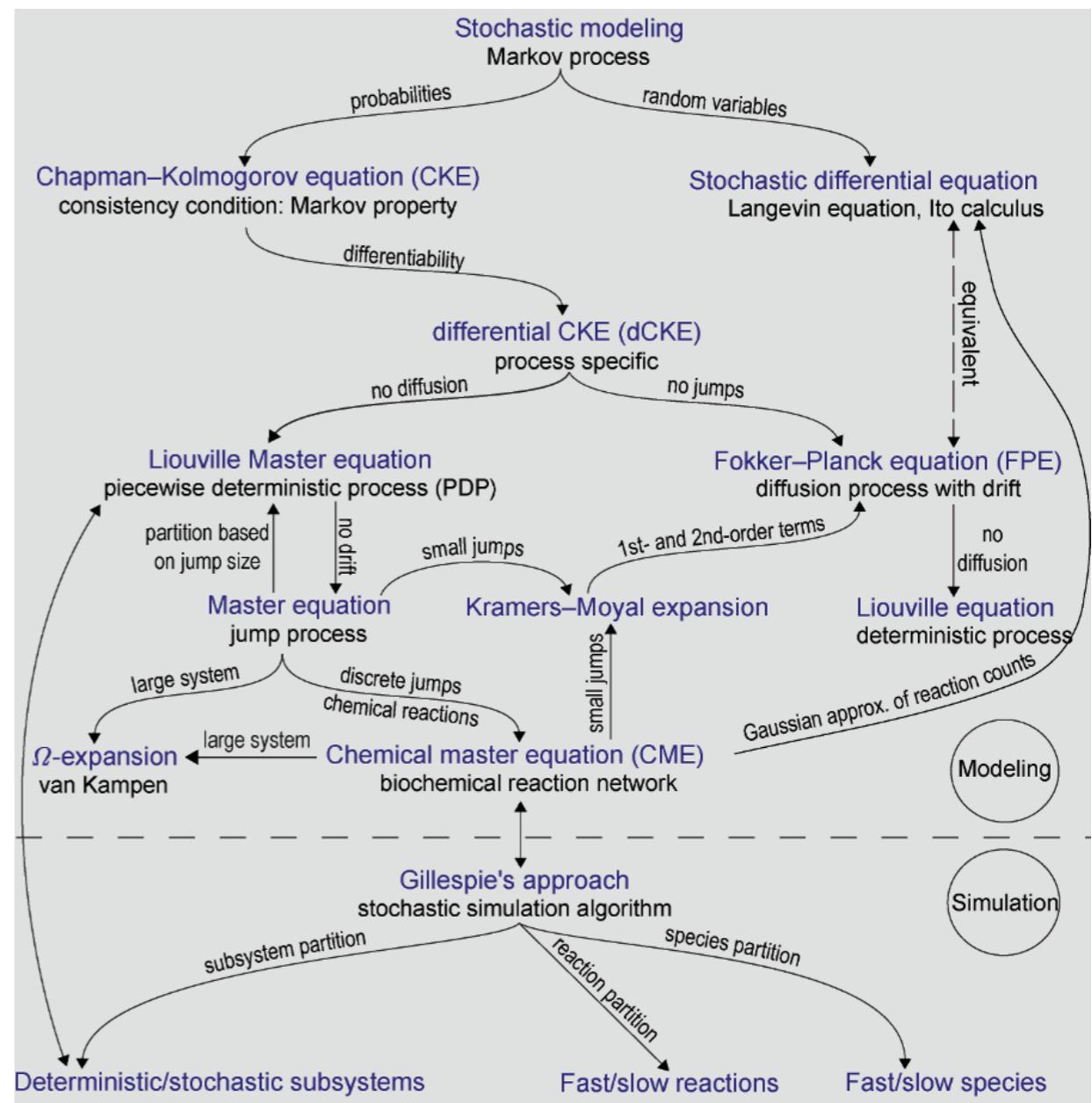
- Differential Equations
 - Stoichiometry matrix
 - (network structure)
 - rate law
 - kinetics
 - collision theory
 - transition state theory



Stochastic Models

Chemical Reaction Networks

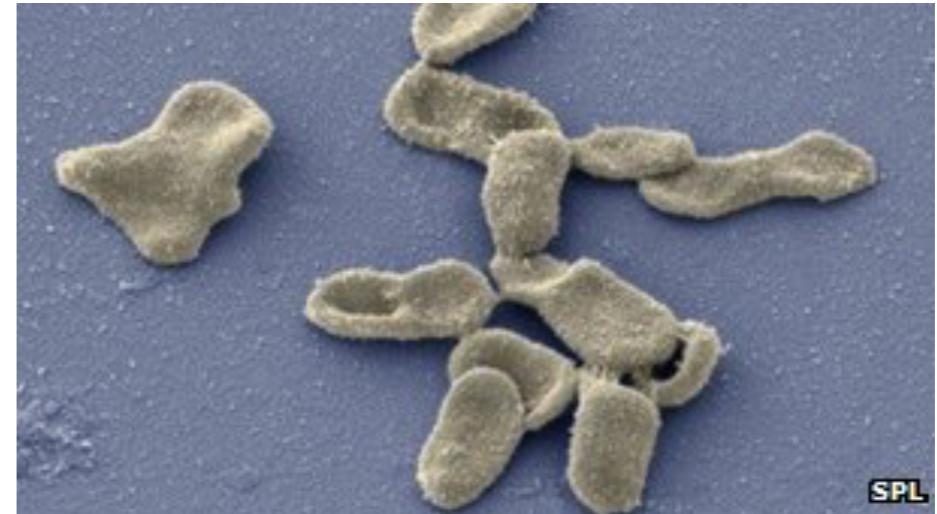
- Stochastic
 - statistical
 - fluctuations
 - noise



Example: Whole Cell Model

- *Mycoplasma genitalium*
parasitic bacterium
synthetic genome 2008 (J. Craig Venter Institute)

- 525 Genes
- 580.07 kb pairs
- 2nd smallest

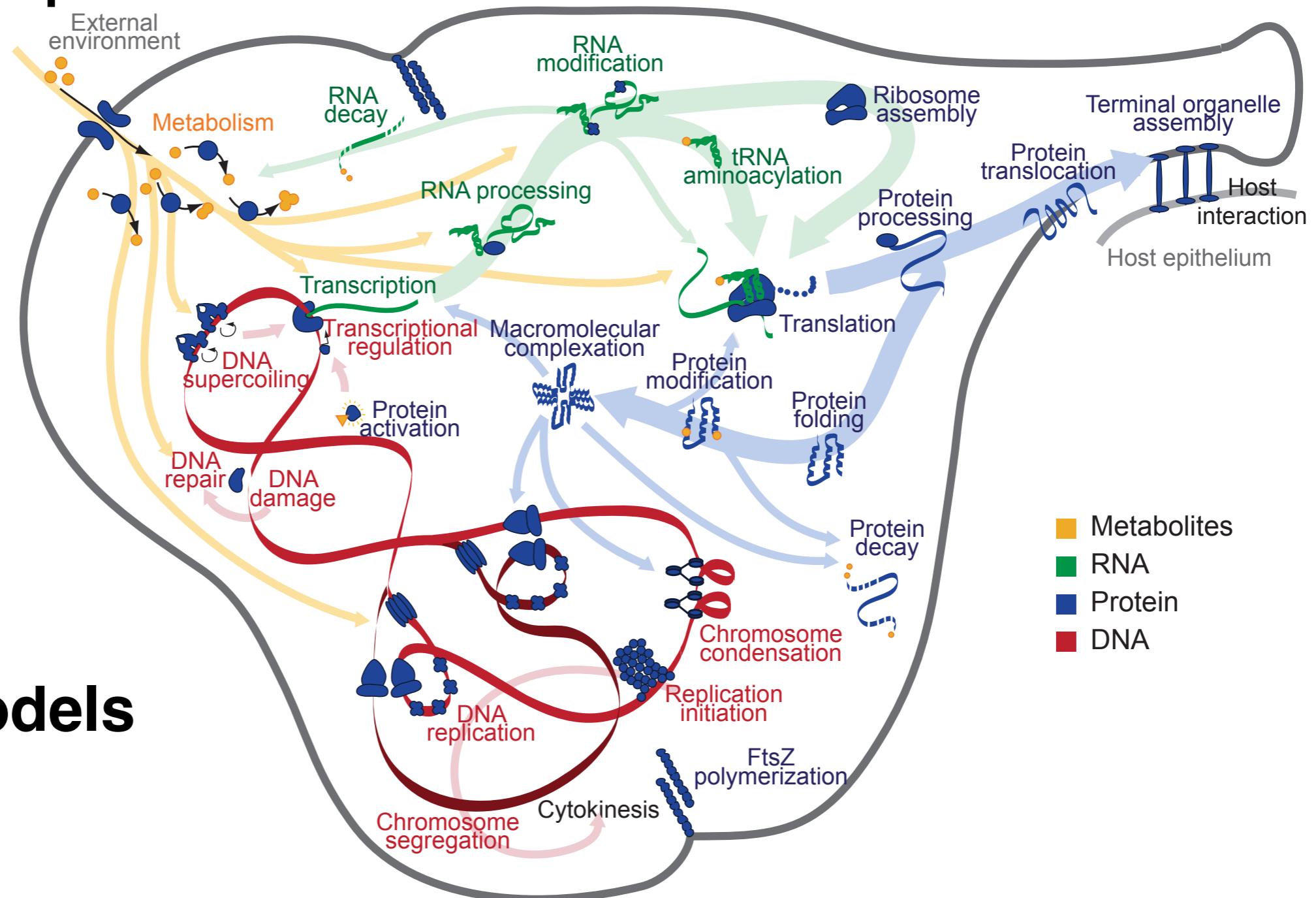


(left) Tully et al, International Journal of Systematic Bacteriology 33 (2): 387 (1983).

(right) <http://www.bbc.com/news/science-environment-19016772>

The virtual cell that simulates life

Example: Whole Cell Model



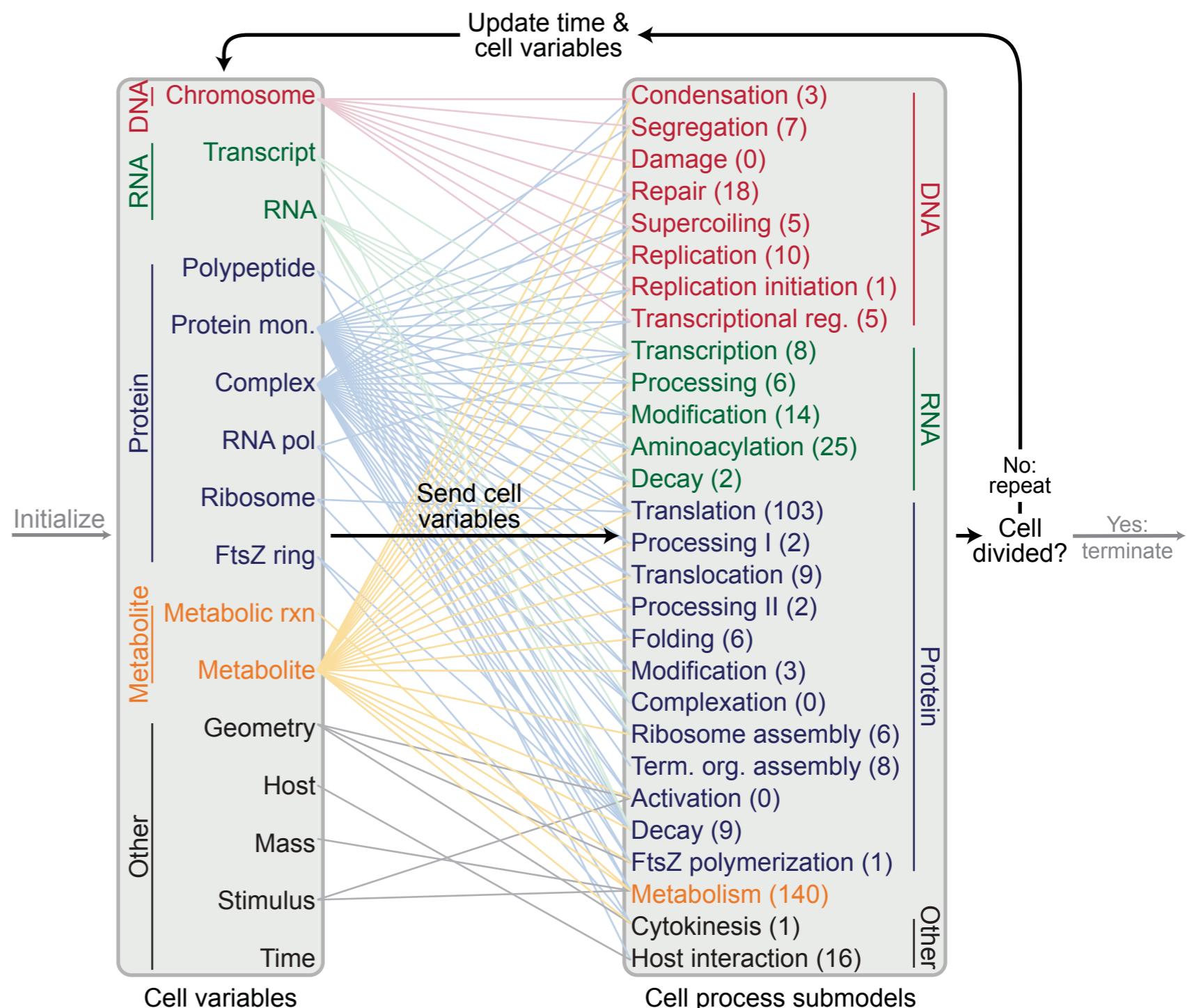
28 Submodels

Example: Whole Cell Model

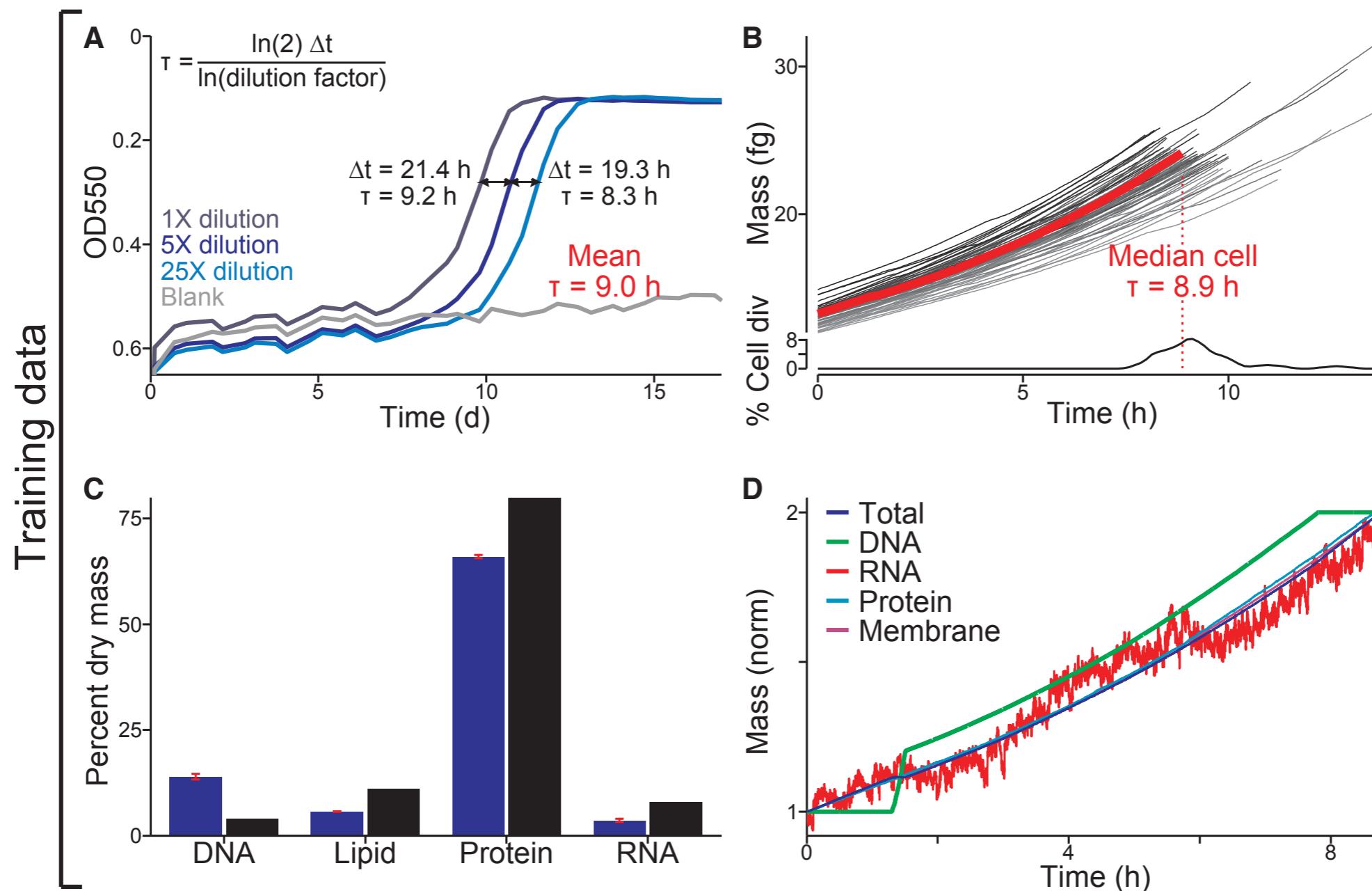
16 Variables

- Random initialize
- 1s time step
- Repeat
- Terminate on cell division

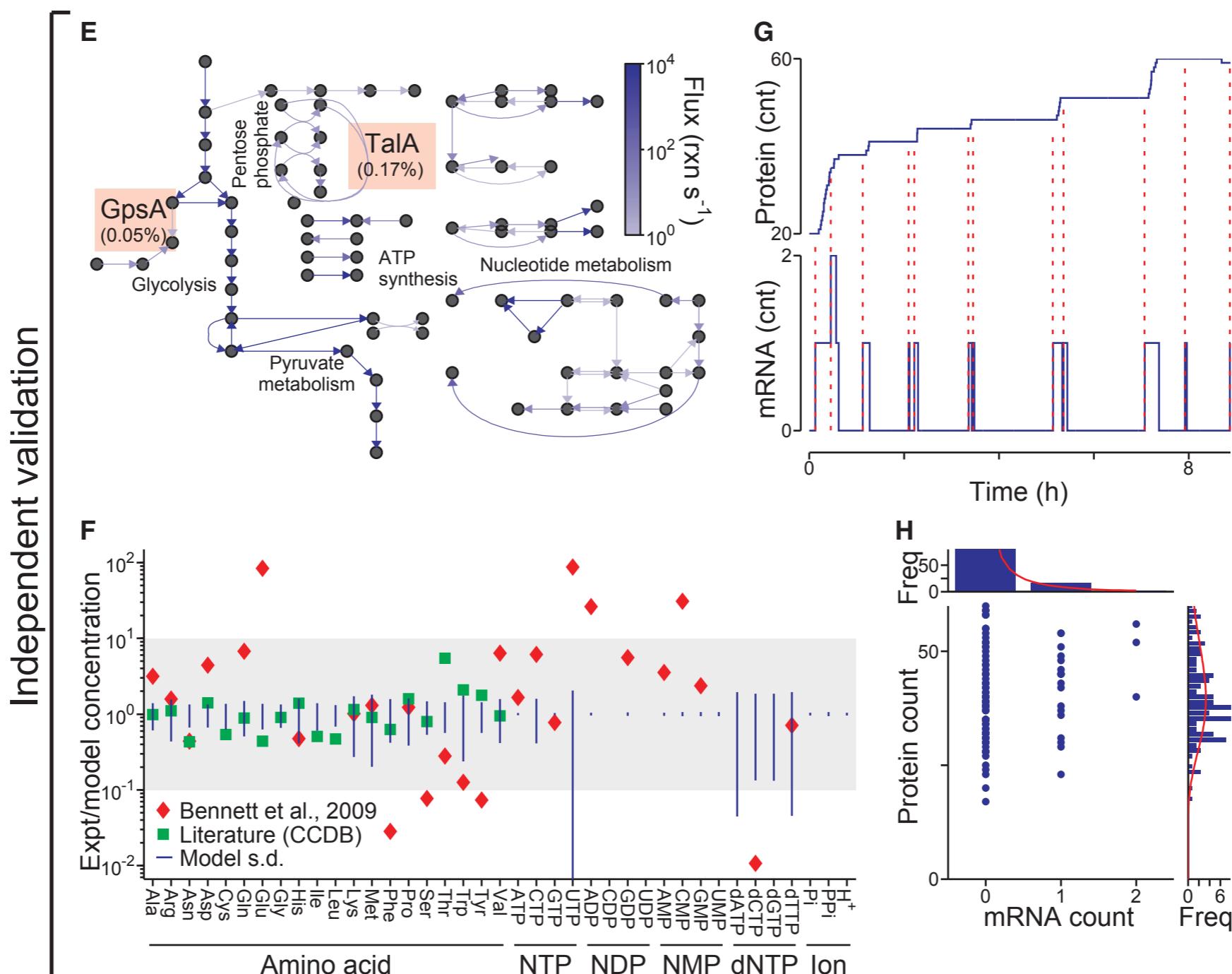
█ Metabolites
█ RNA
█ Protein
█ DNA



Example: Whole Cell Model



Example: Whole Cell Model



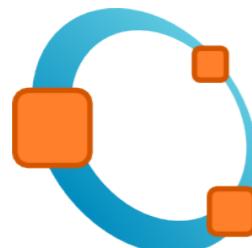
Math tools

- Exploratory data analysis and Data Mining
 - Statistics
 - Distributions
 - Correlations
 - Clustering/Visualization
 - Dimensional Reductions
 - Principal Components Analysis
 - Neural Networks
 - Mathematics
 - Systems of differential Equations
 - different parts require different models

Math tools



<http://www.r-project.org>



<https://www.gnu.org/software/octave/>



<http://www.mathworks.com/products/matlab/>



<http://www.wolfram.com/mathematica/>



<http://www.sagemath.org>

home made varieties

(not an exhaustive list)

Representation

- **Common languages**
 - XML (extensible markup language)
 - <tag attribute="value> content </tag>
 - SBML (Systems Biology Markup Language)
 - CellML
 - SED-ML (Simulation Experiment Description Markup Language)
 - BioPAX (Biological Pathway Exchange - Resource Description Framework [RDF])
- **Databases**
 - KEGG, Gene Ontology (GEO), WikiPathways, IPA, Reactome
- **Networks** (next)
- **Multi-model systems**

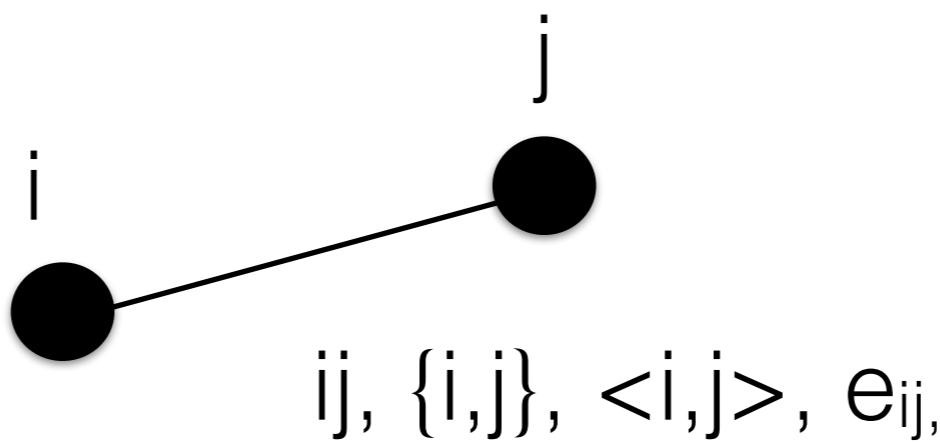
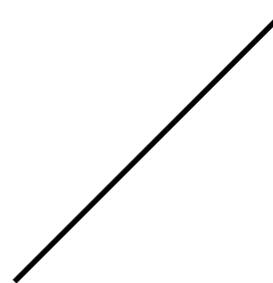
Networks & Pathways

Graphs

Vertices (nodes) {V}



Edges (links) {E}



graph $G = (V, E)$

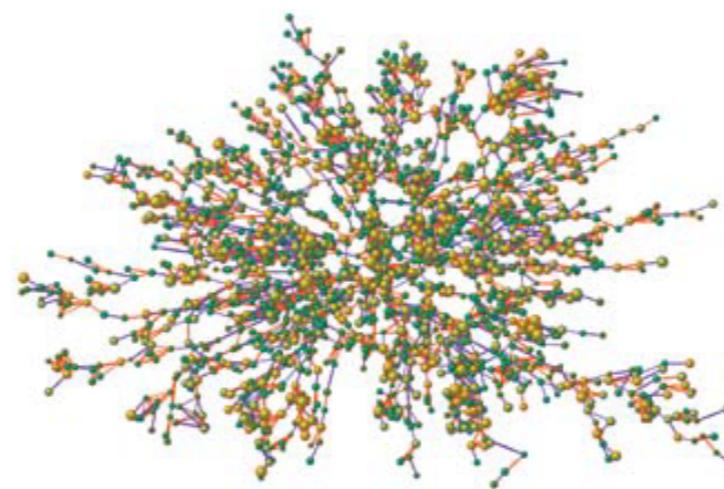
$V(G) = \{\text{set of vertices}\}$

$E(G) = \{\text{set of edges}\}$

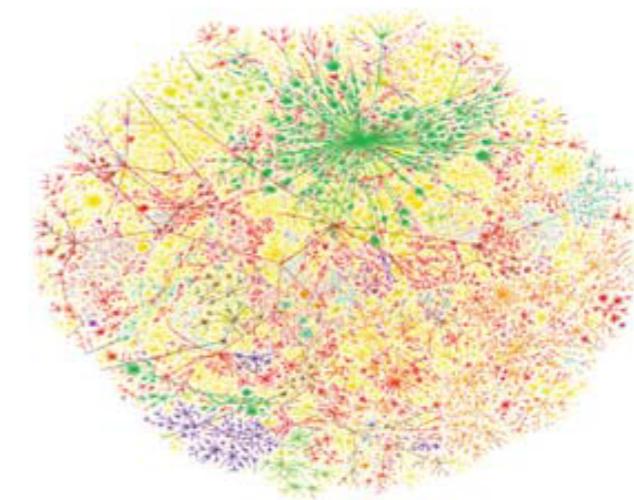
Examples



Adult Human
Brain

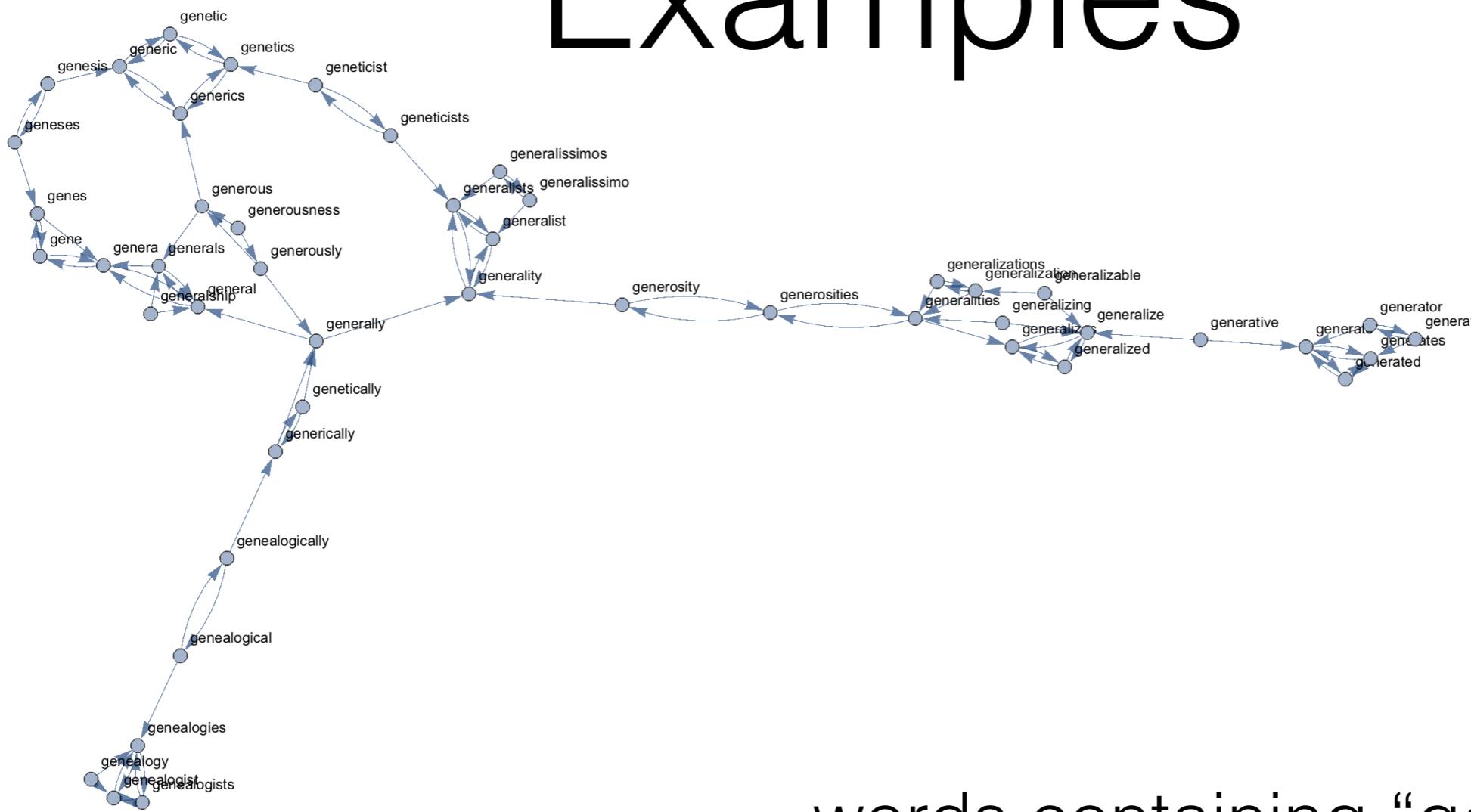


Social Network
Influence of Body
Weight

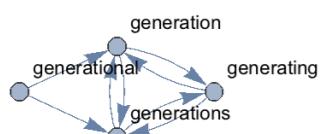


Internet Service
Providers

Examples

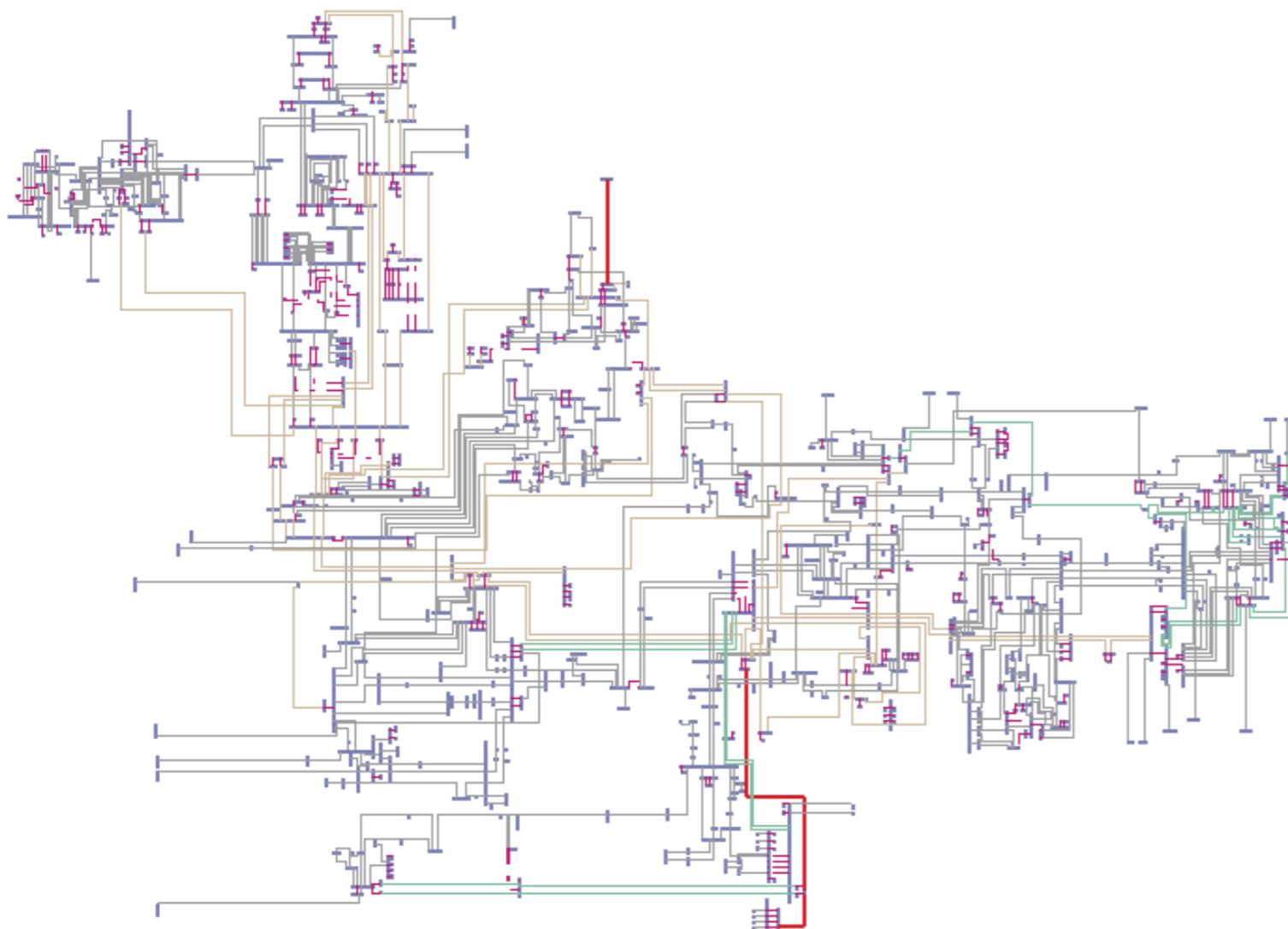


words containing “gene”



Examples

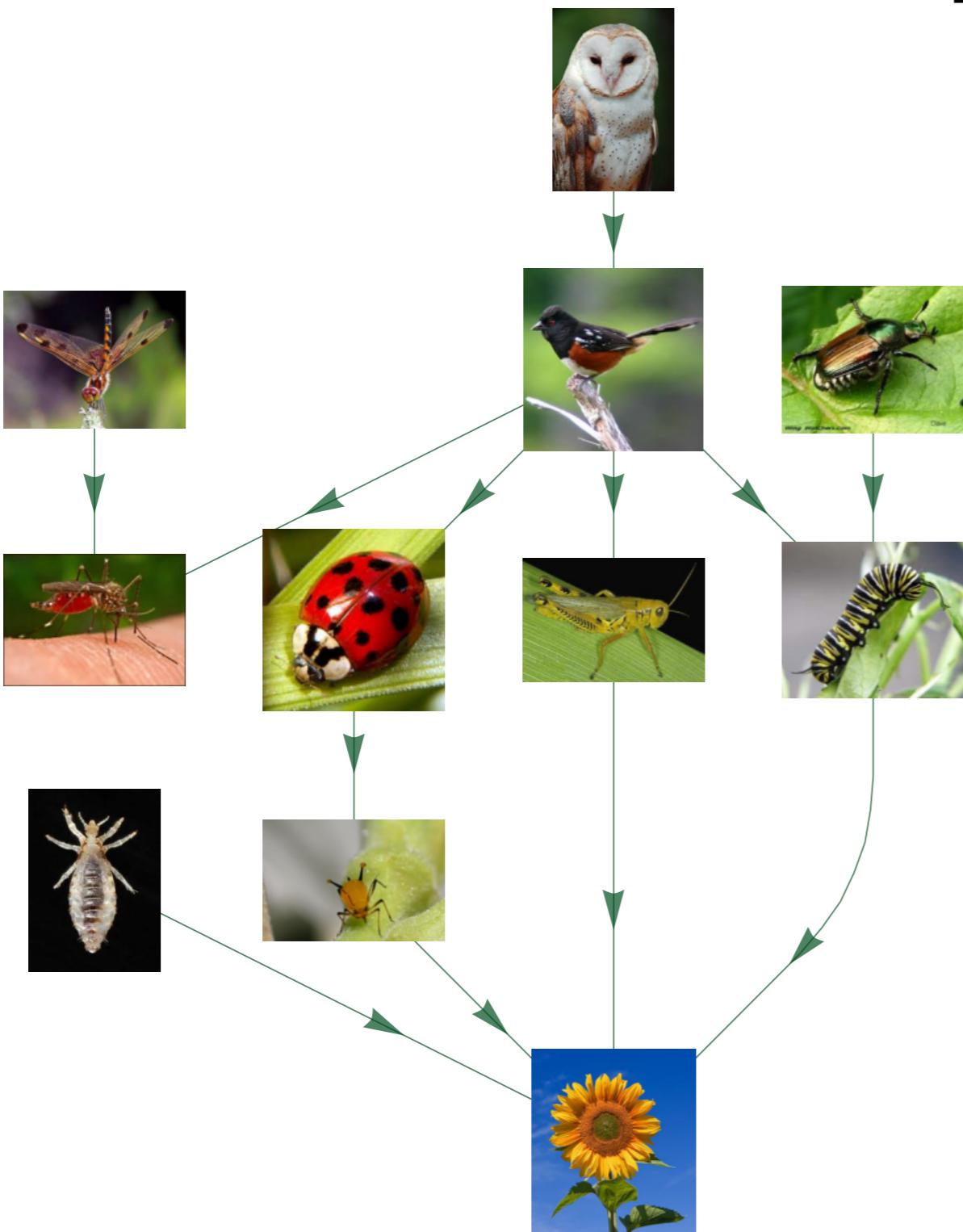
Nodes: Generators and substations



Edges: Transmission lines and transformers. (Line thickness and color indicate the voltage level)

NY Power Grid

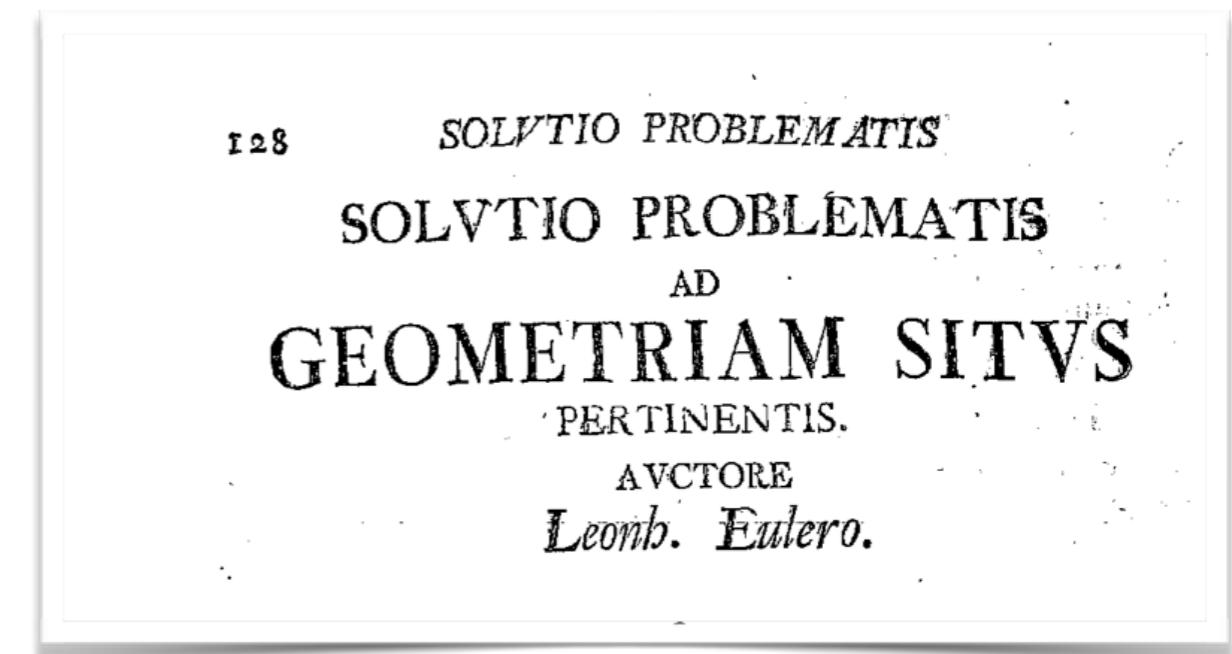
Examples



Food Web Network

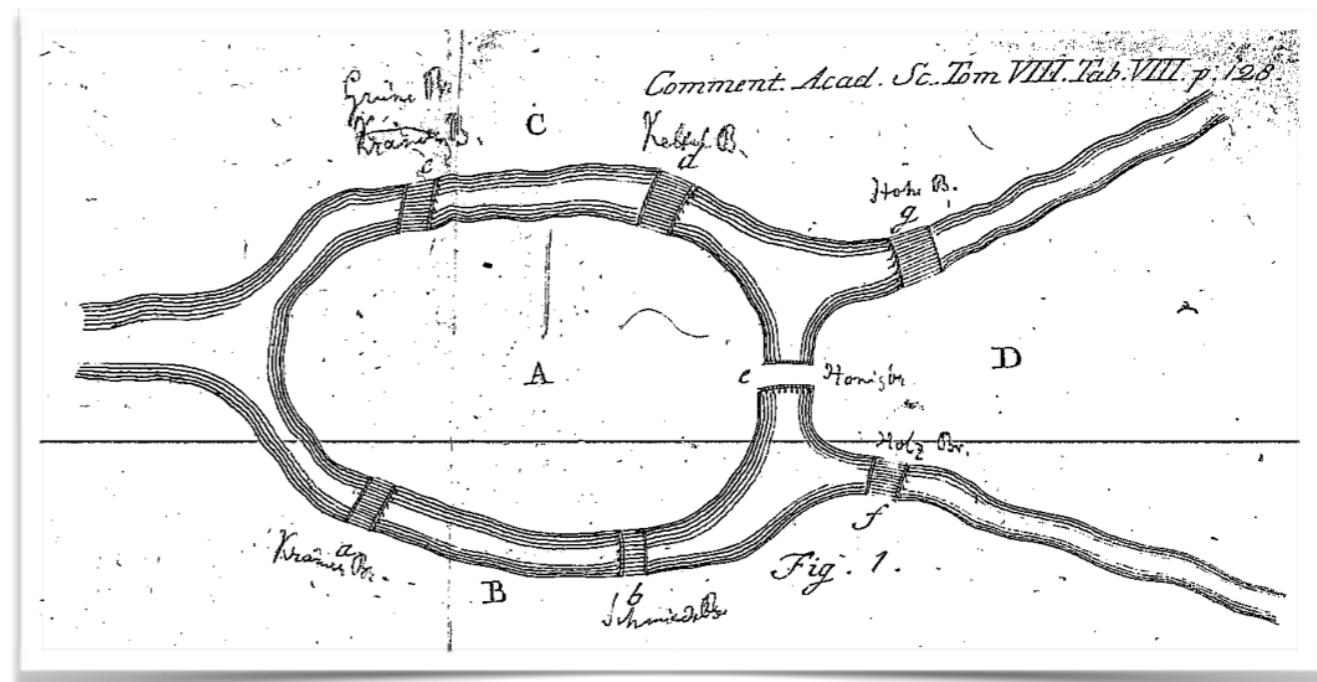
Bridges of Königsberg

- Leonhard Euler
- 1736
- 7 bridges across Pregolya River
- Prussian Königsberg (now Kaliningrad, Russia)
- Is there a route to cross each bridge exactly once?



Bridges of Königsberg

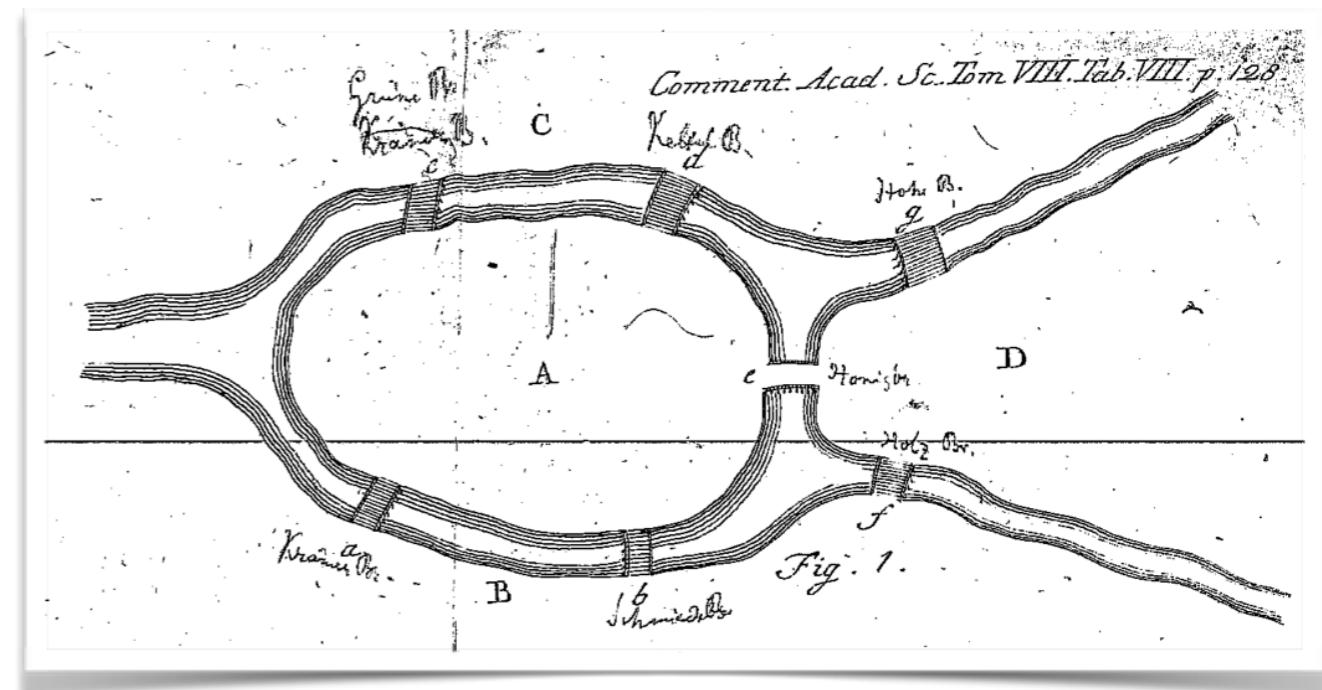
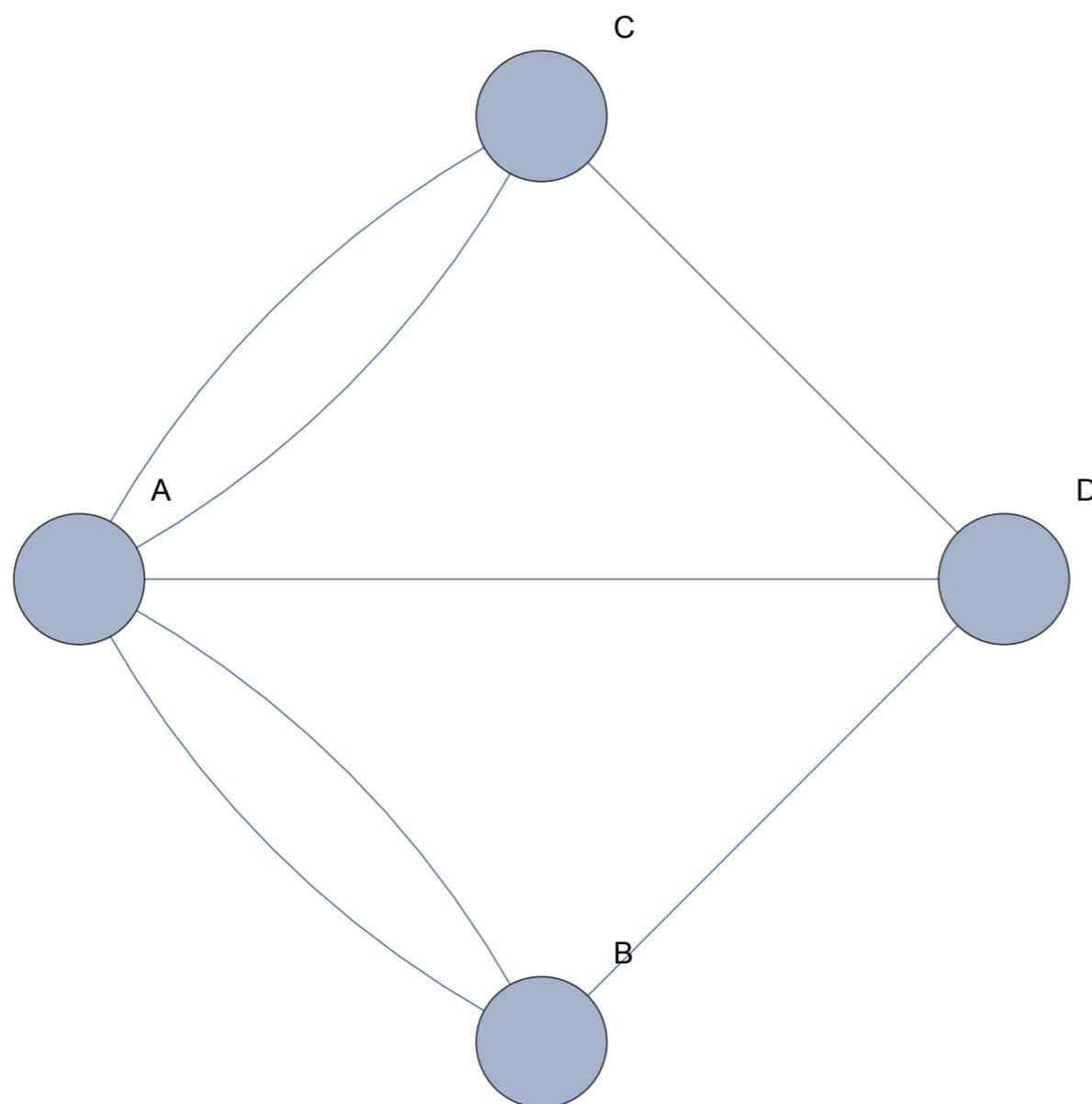
- Leonhard Euler
- 1736
- 7 bridges across Pregolya River
- Prussian Königsberg (now Kaliningrad, Russia)
- Is there a route to cross each bridge exactly once?



e.g. AB, BA, AC, CA, AD, ..., ?

The Euler Archive <http://eulerarchive.maa.org>

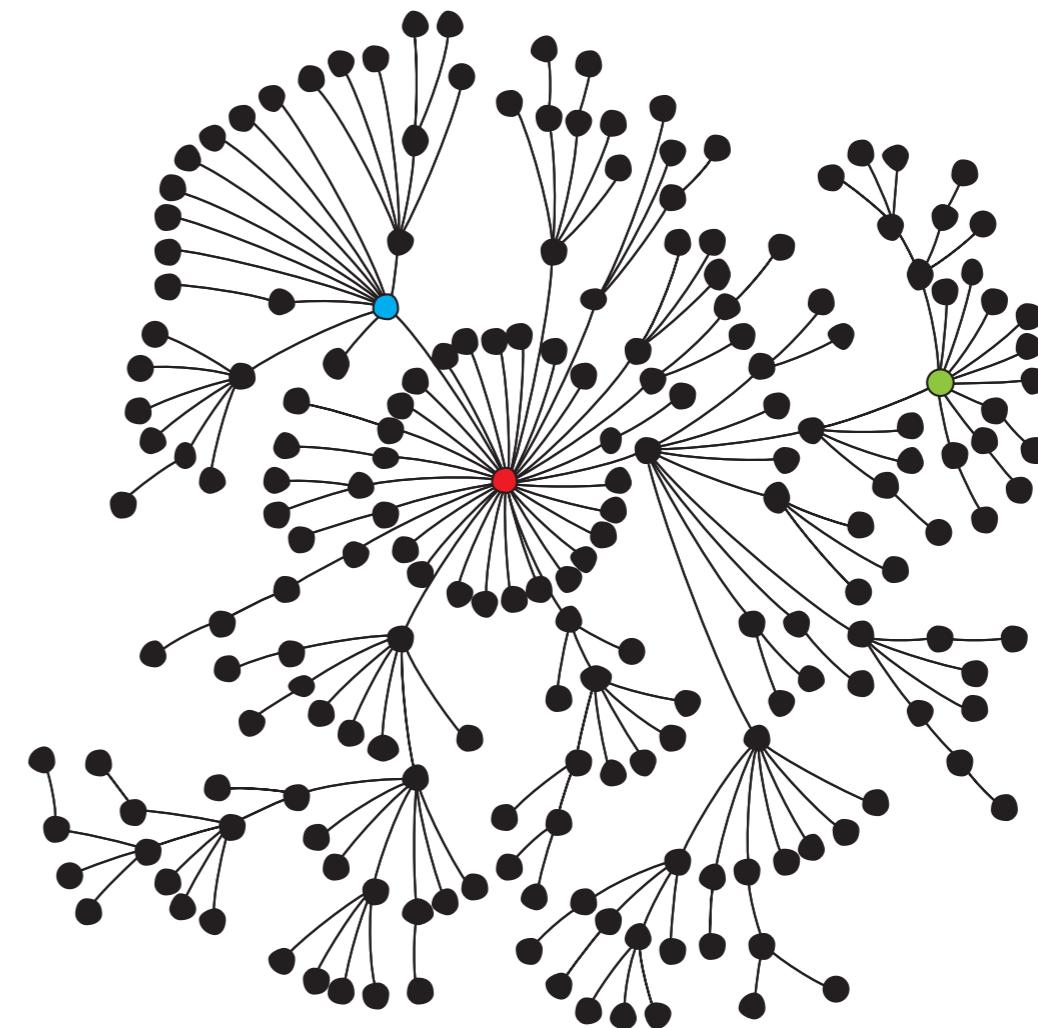
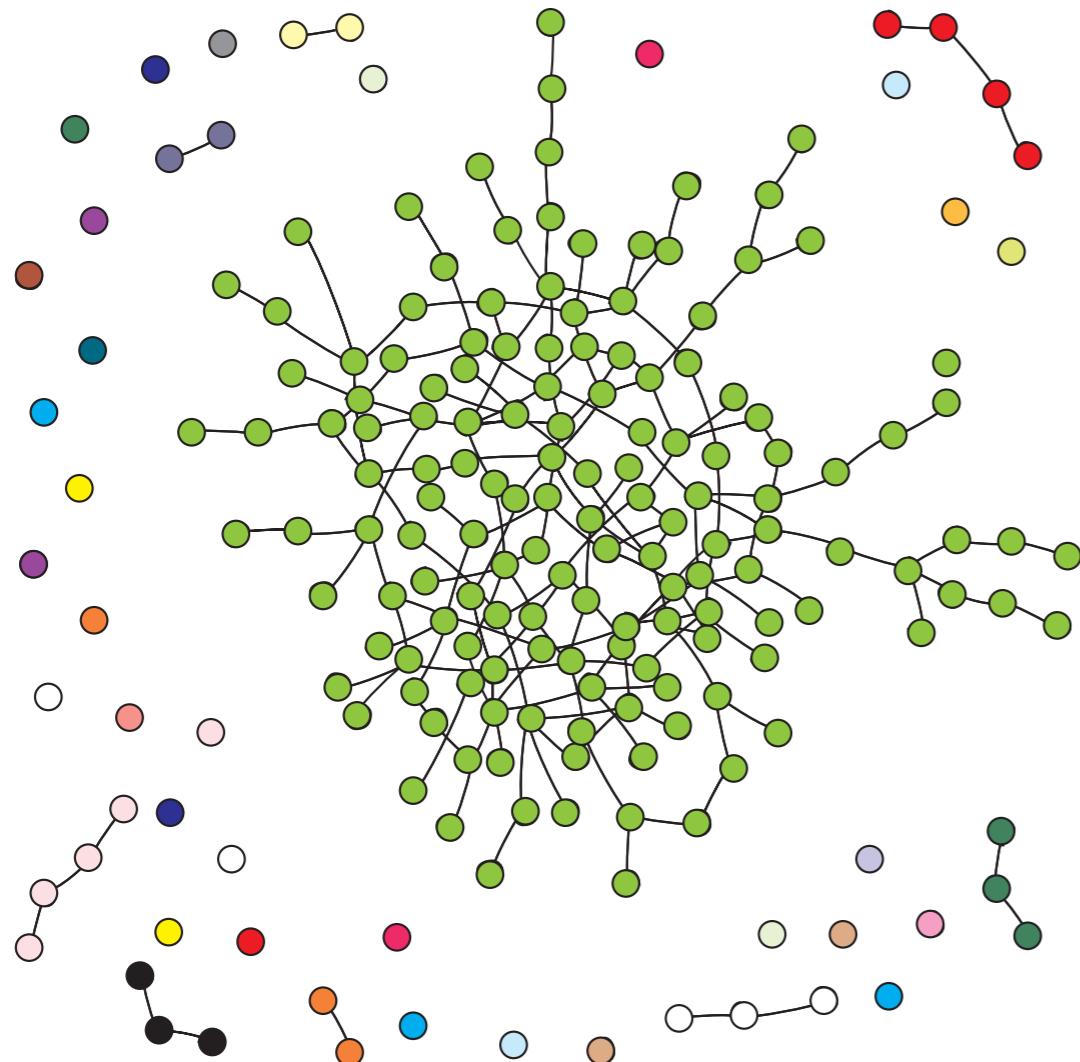
Bridges of Königsberg



e.g. AB, BA, AC, CA, AD, ..., ?

The Euler Archive <http://eulerarchive.maa.org>

Random Networks



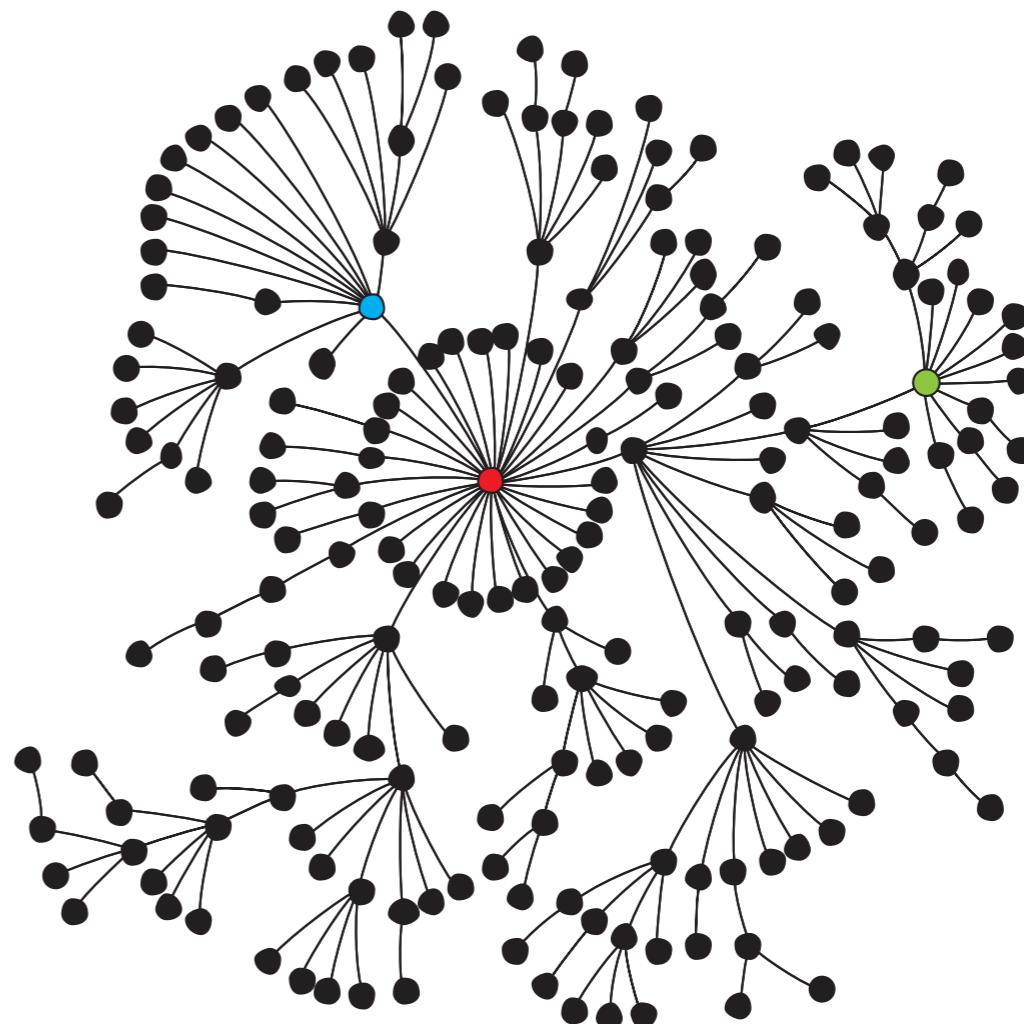
Scale Free Graph

Strogatz, Nature 410, p268 (2001)

Properties

Scale Free Networks

- Robust
 - if vertices removed network still connected
- Vulnerable
 - can target hubs

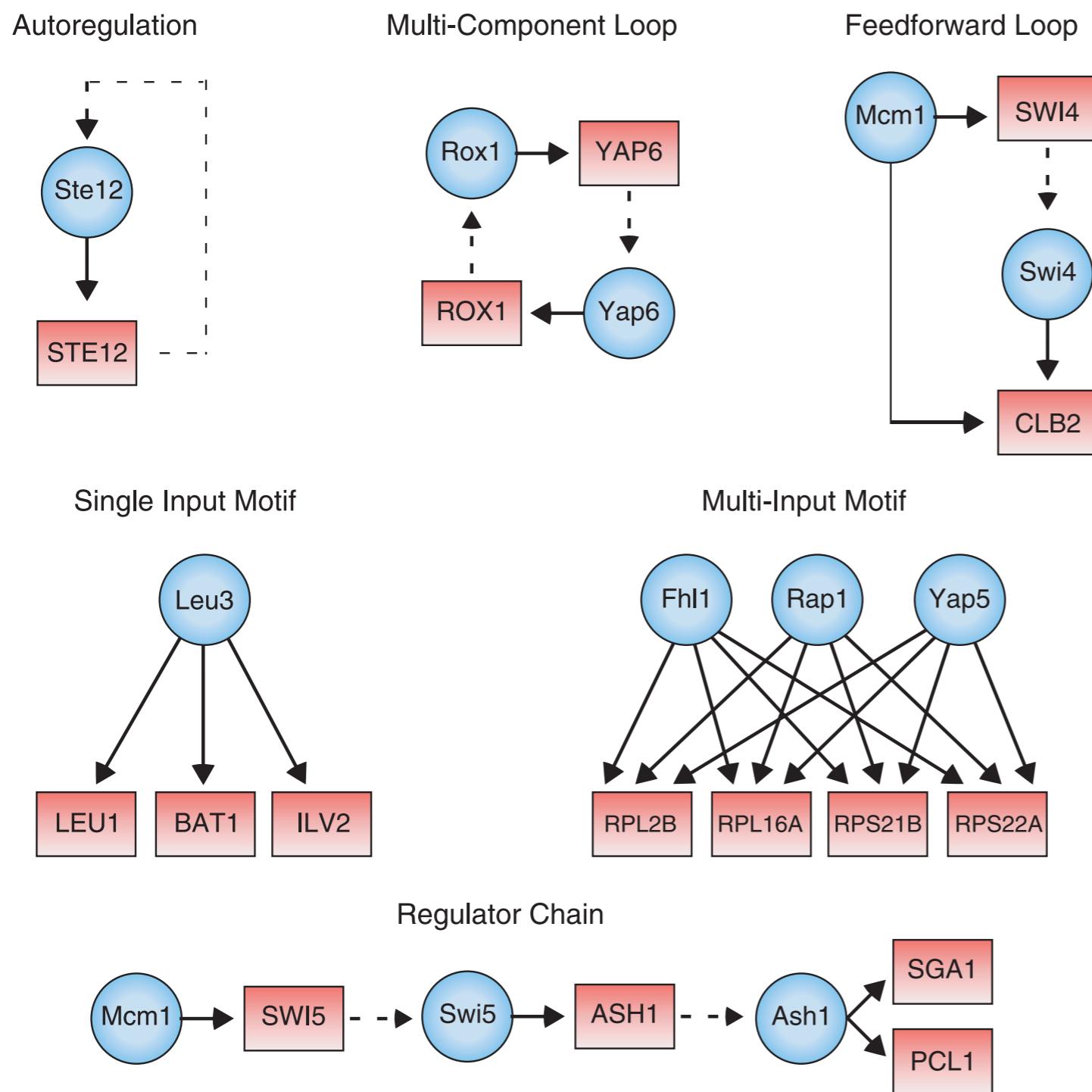


Strogatz, Nature 410, p268 (2001)
Albert et al., Nature 406 p.378 (2000)

Community Structure

- **Social Networks**
 - groups of different people
- **internet pages**
 - various topics
- **Metabolites**
 - various functions
- **genes**
 - various processes

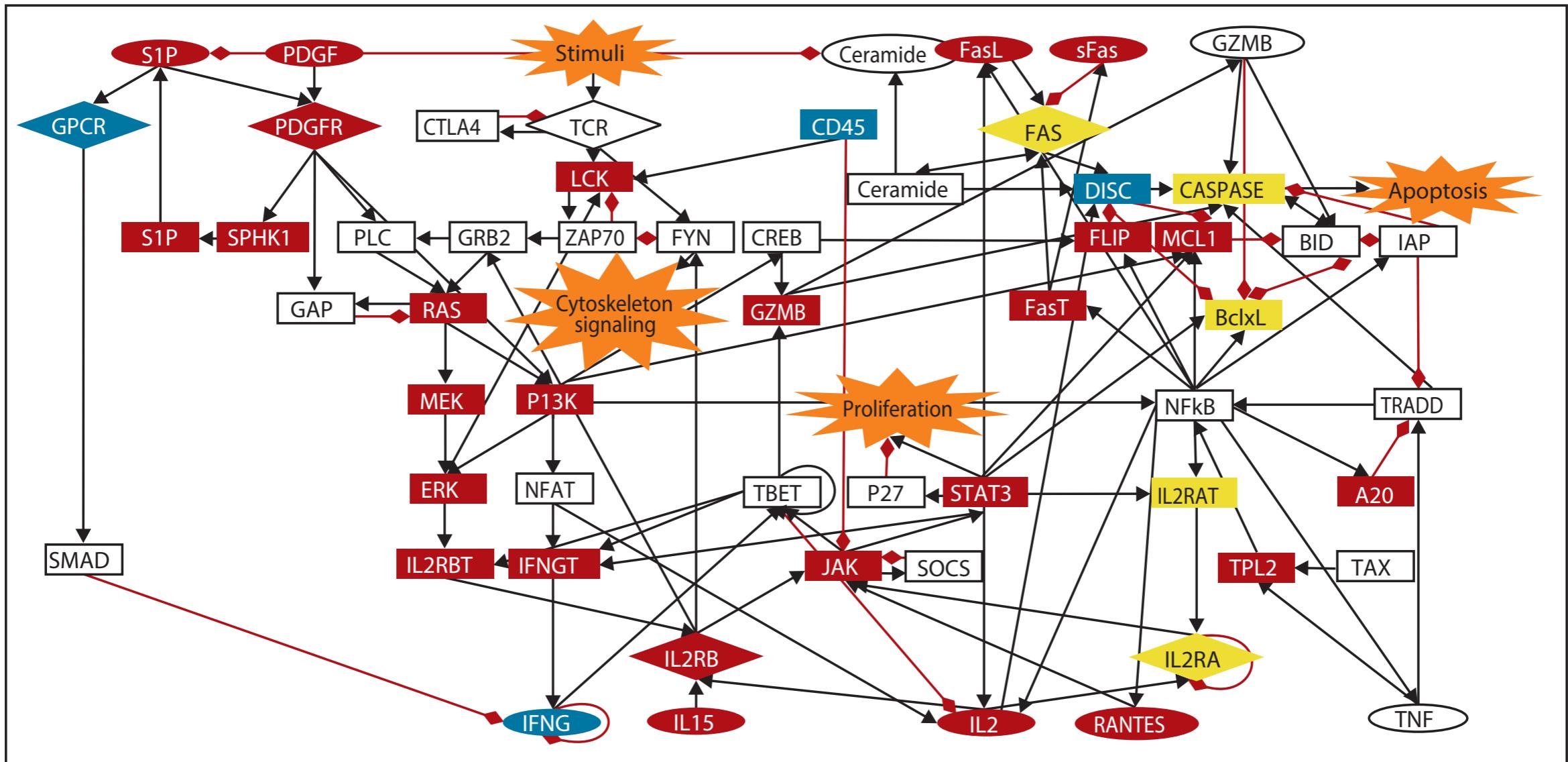
Gene Regulation



- *Saccharomyces cerevisiae*
- Activators (increase)
- Repressor (decrease)
- Feedback loops
- Motifs

Lee et al., Science 298, p.799 (2002)

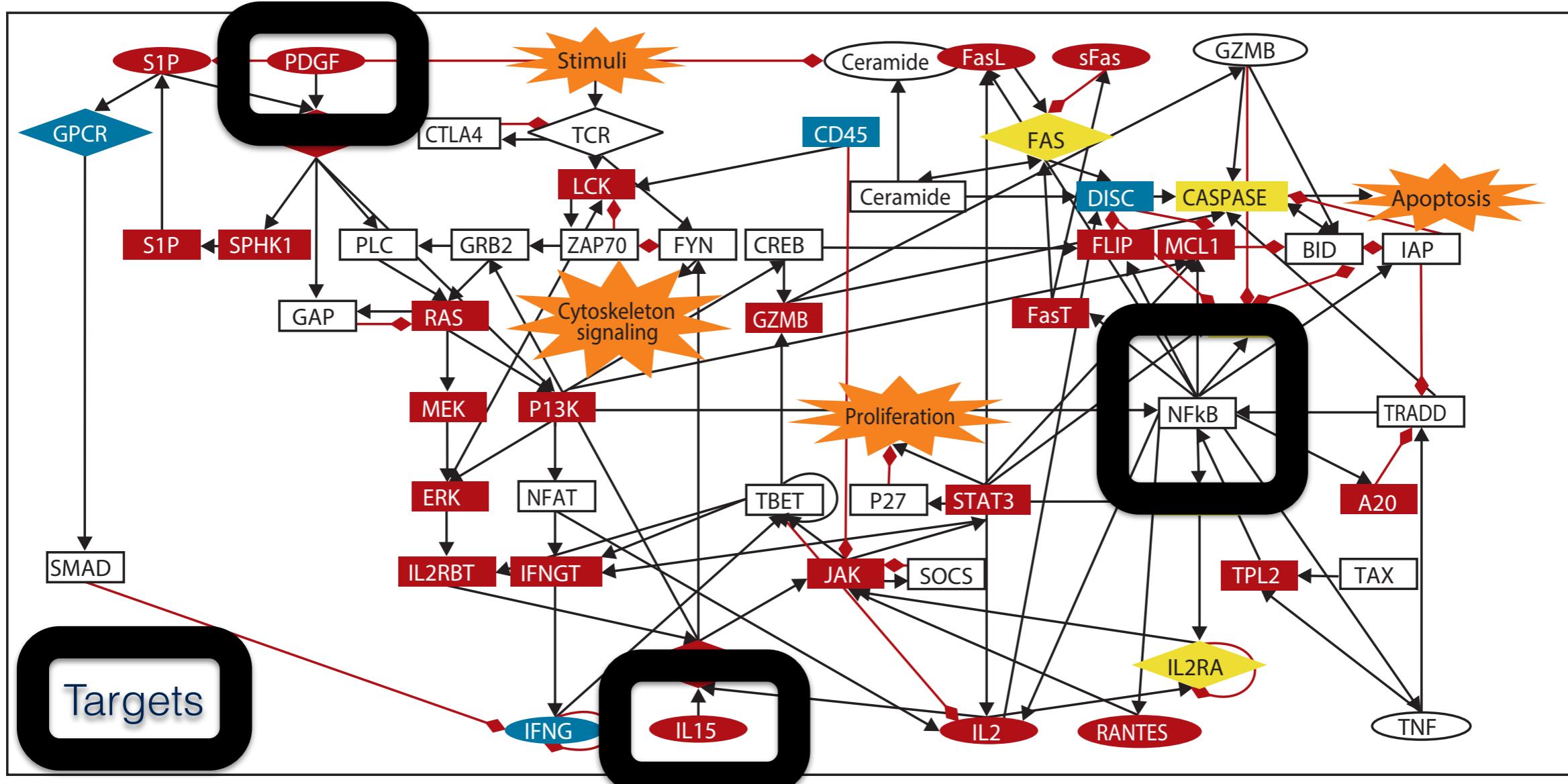
Examples



T cell signaling (Apoptosis) Protein Network

Motter & Albert, Phys. Today 65(4), 43 (2012)

Examples

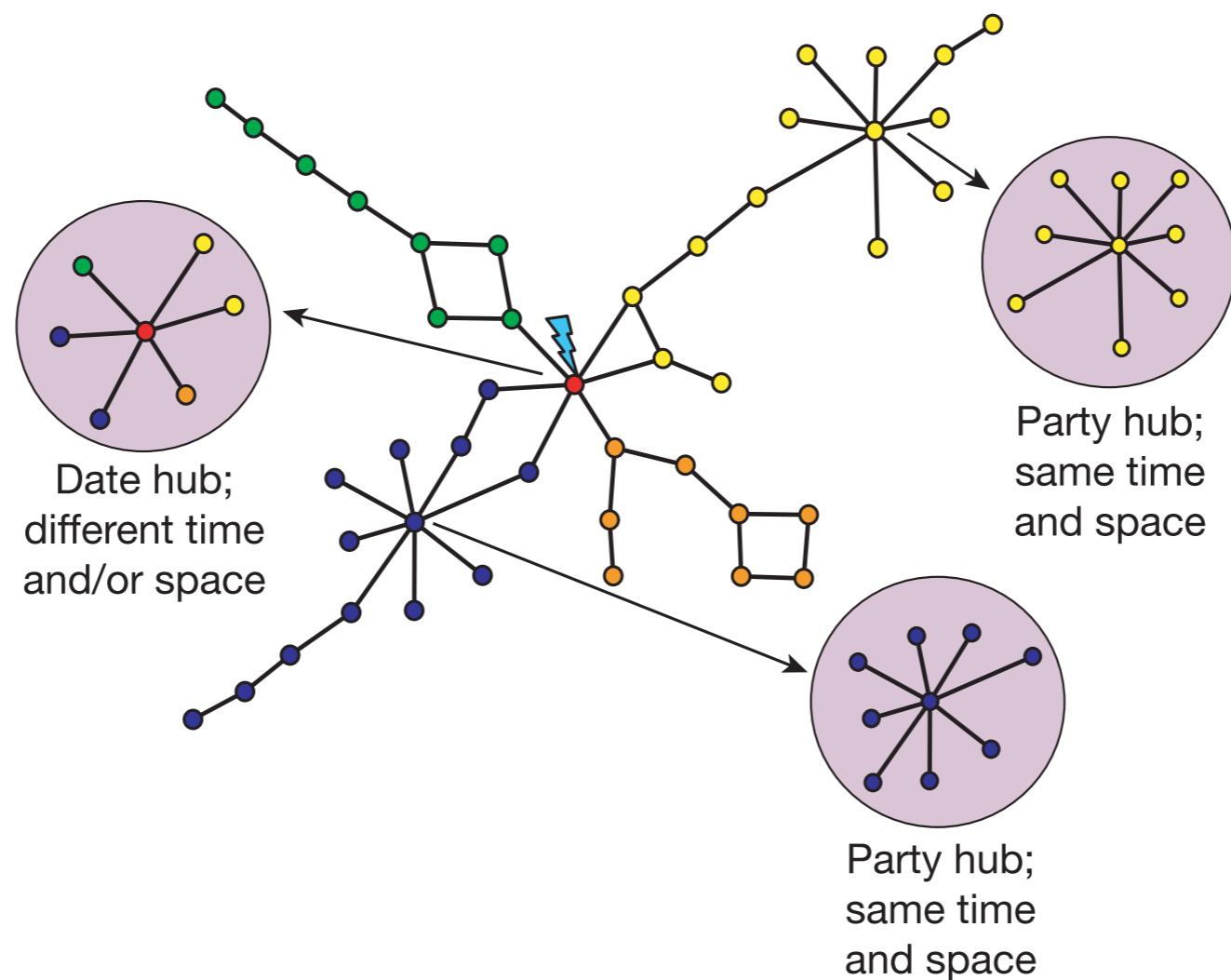


T cell signaling (Apoptosis) Protein Network

Motter & Albert, Phys. Today 65(4), 43 (2012)

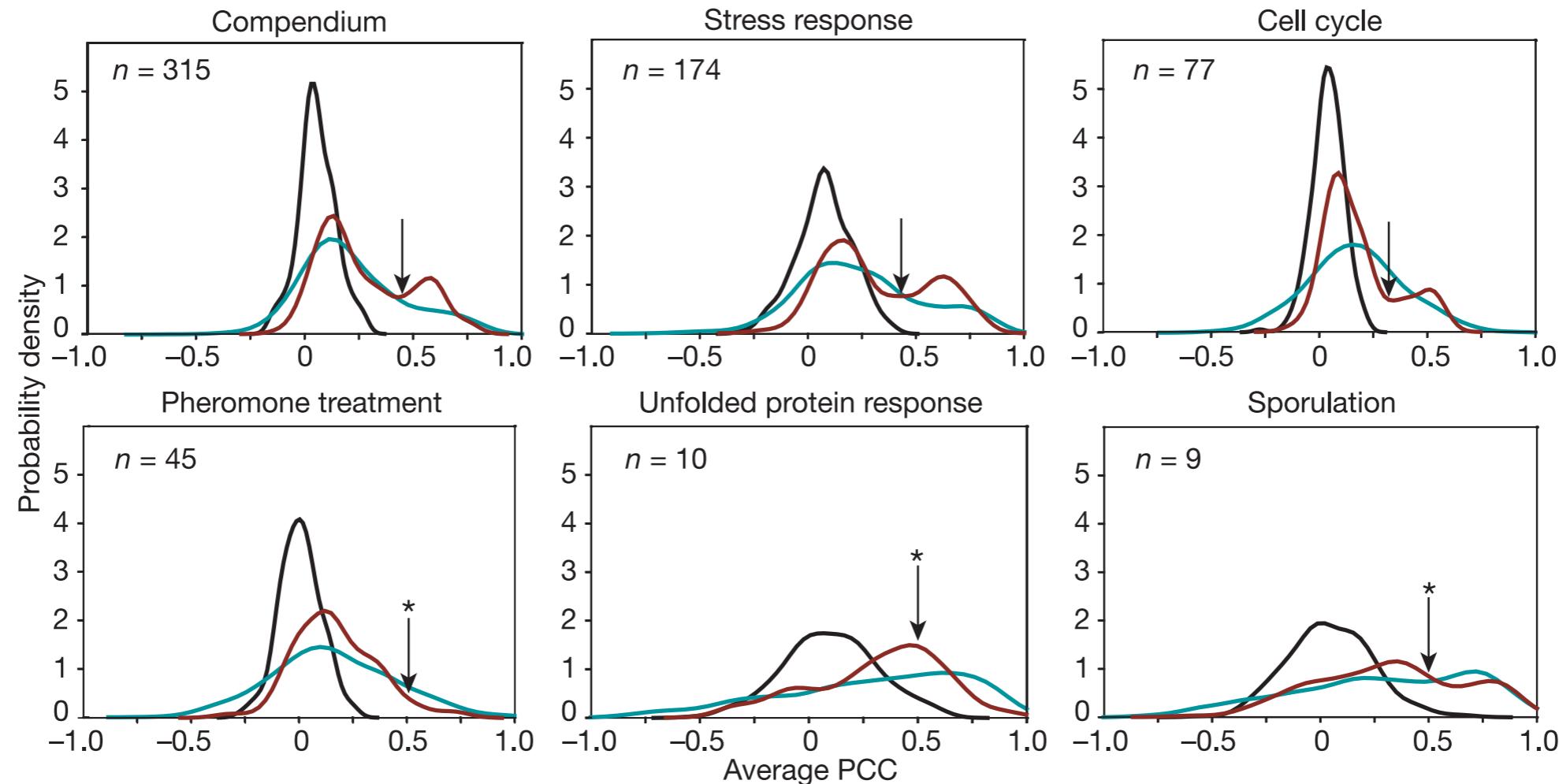
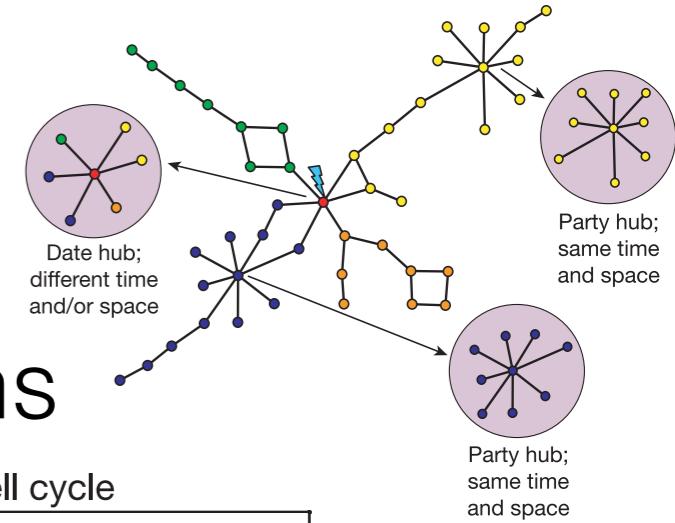
Examples

modularity in Yeast Protein-Protein Interactions



Examples

modularity in Yeast Protein-Protein Interactions

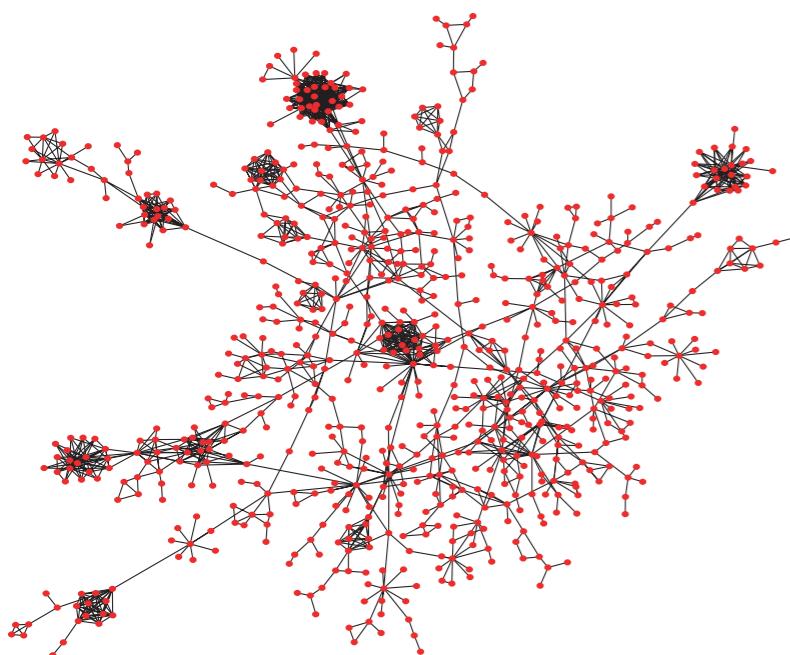


PCC: average Pierson Correlation Coefficient between the hub and each of its respective partners for mRNA expression

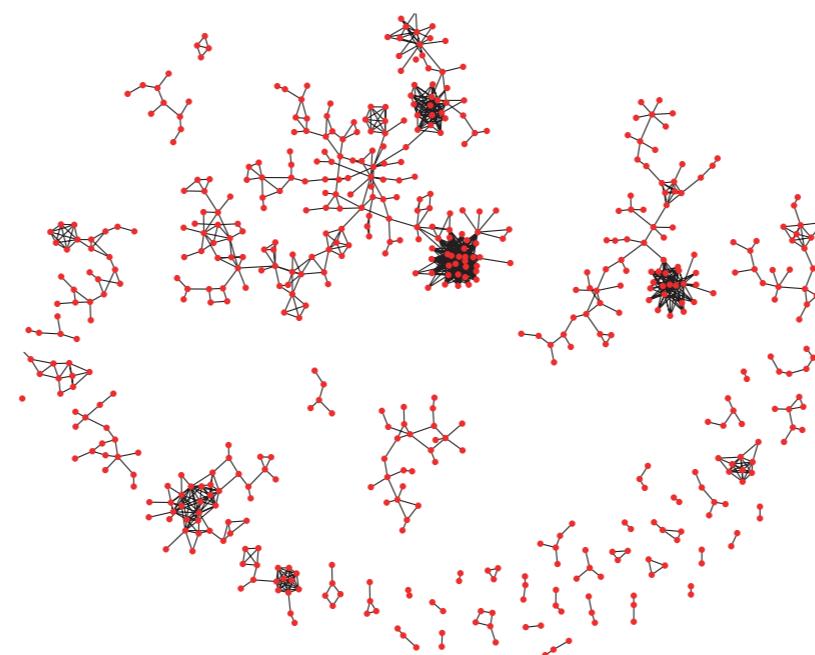
Han et al., Nature 430, p. 88 (2004)

Examples

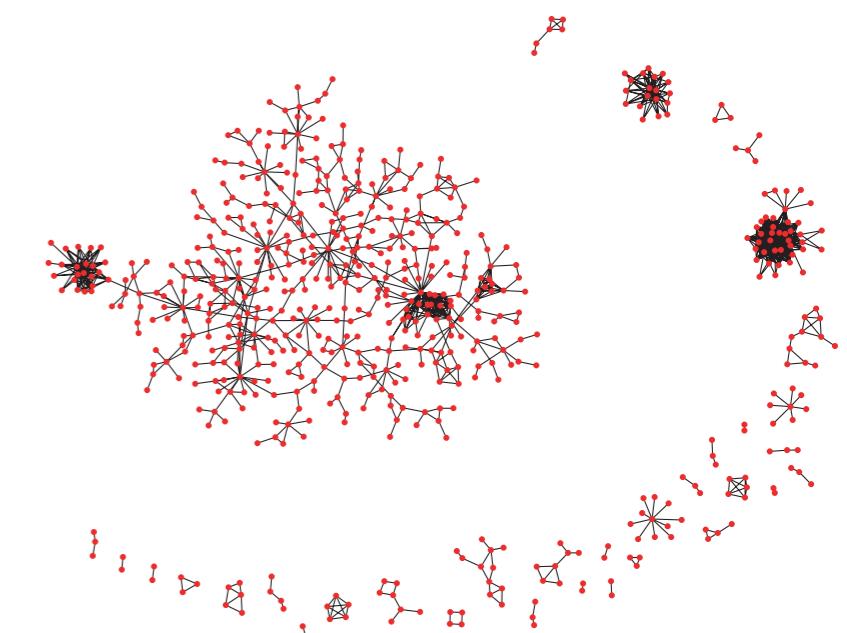
modularity in Yeast Protein-Protein Interactions



main component



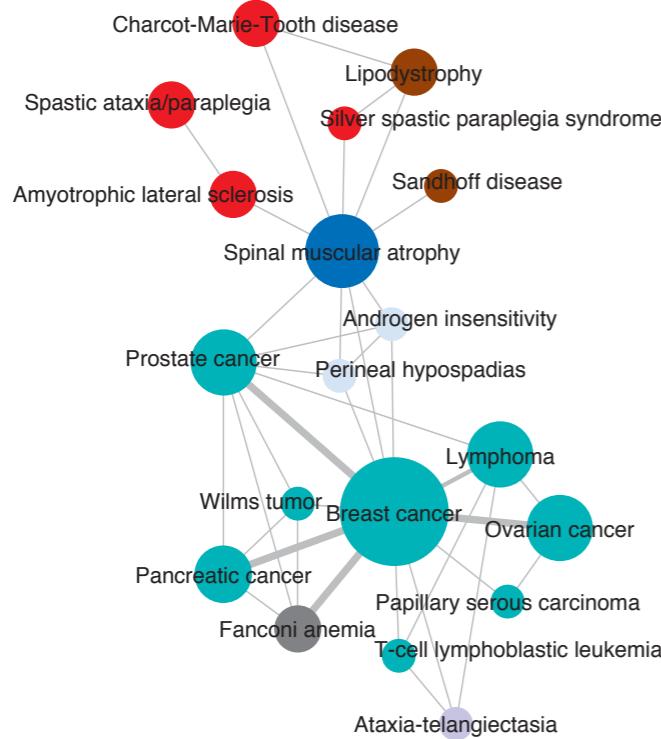
removal of date hubs
(small subnetworks)



removal of party hubs
(intact)

Examples

Human Disease Network
(HDN)



bipartite
OMIM based

DISEASOME

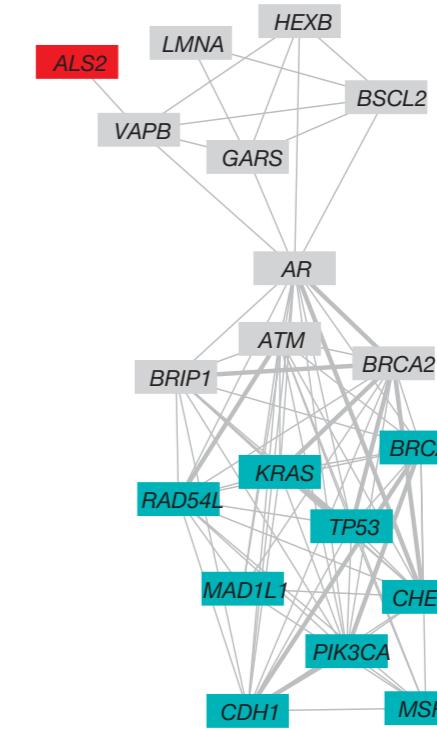
disease phenotype

Ataxia-telangiectasia
Perineal hypospadias
Androgen insensitivity
T-cell lymphoblastic leukemia
Papillary serous carcinoma

disease genome

AR
ATM
BRCA1
BRCA2
CDH1
GARS
HEXB
KRAS
LMNA
MSH2
PIK3CA
TP53
MAD1L1
RAD54L
VAPB
CHEK2
BSCL2
ALS2
BRIP1

Disease Gene Network
(DGN)



Goh et al., PNAS 104(21) p.8685 (2007)

The human disease network

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) Proc Natl Acad Sci USA 104:8685-8690



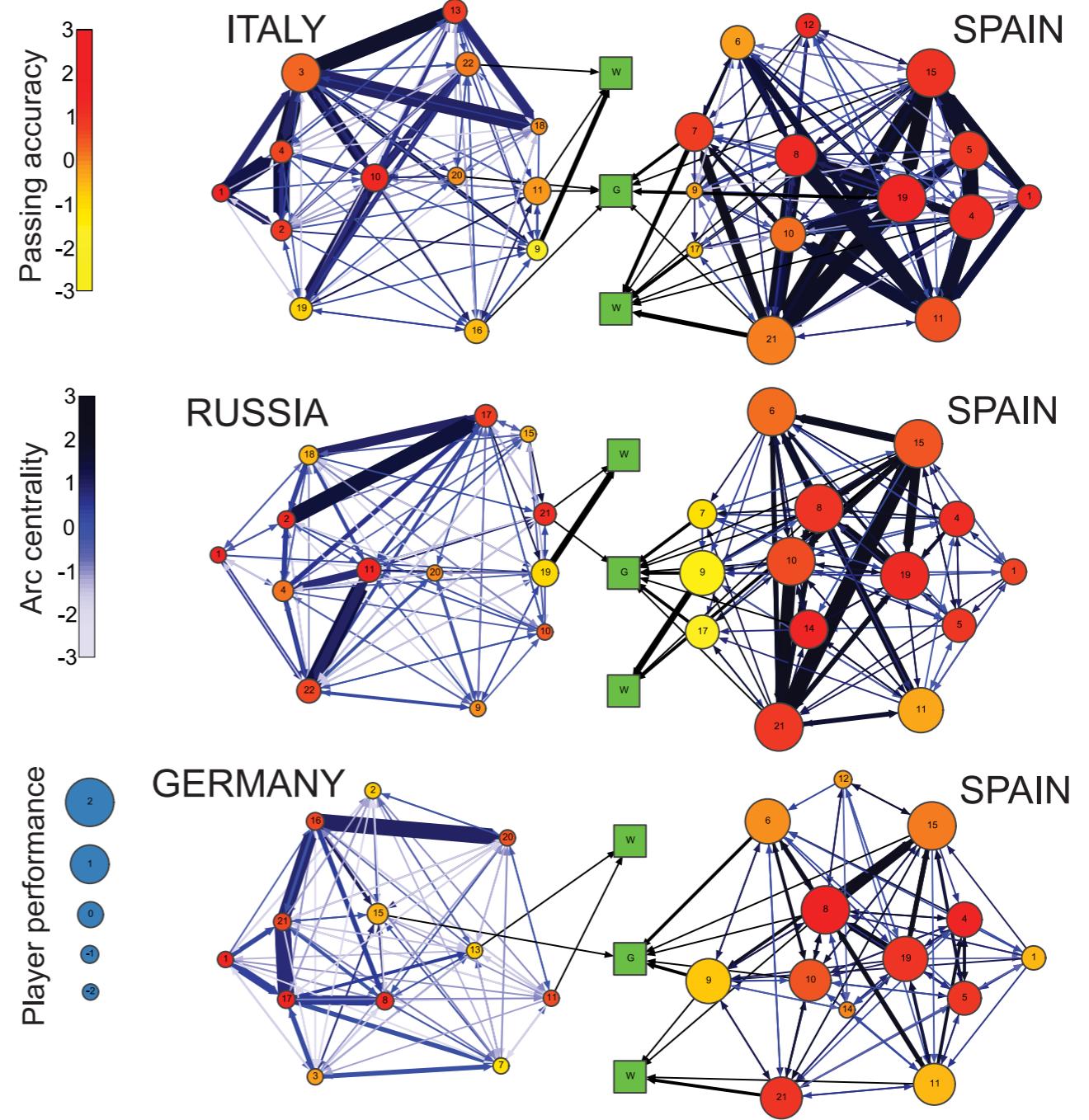
Disorder Name

18	Acromicricardia
25	Acrocephalopolysyndactyly, type II
53	Adrenal hyperplasia, congenital
77	Aldosterone to renin ratio raised
97	Alpha-thalassemia/mental retardation syndrome, hemoglobinopathy
98	Alpha-thalassemia/mental retardation syndrome, nonhemoglobinopathy
107	Anesthesia from kappa-opioid receptor agonist, female-specific
117	Angiotensin I-converting enzyme deficiency
129	Anterior segment cataract
137	Arrhythmogenic right ventricular dysplasia
144	Arrhythmogenic right ventricular dysplasia
151	Athetosis-brachismus dyskinesia syndrome
182	Banayan-Riley-Ruvalcaba syndrome
192	Basal cell carcinoma, sporadic
198	Beta-2-adrenergic receptor antagonist, reduced response to
210	Bone marrow aplasia, epinephrinus, and piosis
217	Carpal tunnel syndrome, familial
237	Central hypothyroidism and retinopathy
287	Central hypothyroidism syndrome
292	Cerebrovascular disease, occlusive
294	Cerlobiotin excretion storage disease
313	Cholesterin ester storage disease
320	Chondrocalcinosis, idiopathic, and respiratory distress
329	Chymotrypsinogen deficiency, familial
344	Colon aganglionosis, total, with small bowel involvement
357	Conotruncal anomaly face syndrome
377	Cranial deafness, progressive, progressive
379	Cystic fibrosis, with epinephrinuria and piosis
396	Cyclic ichthyosis with epidermolytic hyperkeratosis
416	Dentinogenesis imperfecta, Sheldén type
422	Dilated cardiomyopathy with woolly hair and keratoderma
434	Dilated cardiomyopathy with woolly hair and keratoderma
439	Dissection of cervical arteries
451	Dopamine beta-hydroxylase deficiency
452	Dysautonomia, with orthostatic hypotension
453	Dysalbuminemic hyperthyroxinemia
463	Dystyrosinase-related periodic paroxysms
471	Elite sprint athlete performance
474	Endothelial progenitor cell dysfunction
527	Fatty liver, acute of pregnancy
533	Fibular hypoplasia and complex brachydactyly
539	Fibular hypoplasia and complex brachydactyly
544	Focal cerebral dysplasia
545	Focal cortical dysplasia, Taylor balloon cell type
553	Foveal dysplasia, adult-onset, with choroidal neovascularization
584	Giant platelet disorder, isolated thrombocytopathy
594	Glutathione synthase deficiency, hepatic
604	Glycogen storage disease, type II
626	Gregg cephalopelvisynostosis syndrome
646	Hemidysgenesis, systemic, due to acetoxyplasmannia
679	Hemochromatosis, systemic, due to HFE gene mutation
699	Homozygous 2p16 deletion syndrome
701	Homozygous 2p16 deletion syndrome
727	Hypertrophic periodic paralysis
733	Hypokalemic periodic paralysis
780	Hypoparathyroidism-retardation-dysmorphism syndrome
785	Hypoparathyroidism, localized pitressin, localized
792	Hyperparathyroidism, primary, with variable inheritance
803	Immunodeficiency, polyendocrinopathy, and enteropathy, X-linked
809	Inflammatory bowel disease, with associated immunologic disorder
830	Jervell and Lange-Nielsen syndrome
843	Juvenile polyposis/hereditary hemangioma telangiectasia syndrome
845	Keratoderma, palmoplantar, with deafness
847	Laryngopharyngeal dysphonia
868	Laryngopharyngeal dysphonia
891	Laryngopharyngeal dysphonia, with white matter
913	Lower motor neuron disease, progressive, without sensory symptoms
942	Lynch cancer family syndrome II
945	Mandibuloacral dysplasia with type B hypothyroidism
952	Mandibuloacral dysplasia with associated neurologic disorder
969	Methionine synthase deficiency, chb1 type
1010	Methionine synthase deficiency, chb1 type
1050	Methionyl-tRNA methyltransferase deficiency
1056	Miyoshijuvenile/hereditary disease due to PFK deficiency
1057	Miyoshijuvenile/hereditary disease due to PFK deficiency
1080	Nephropathy of inappropriate antidiuresis
1096	Neurofibromatosis-Nosan syndrome
1104	Neurofibromatosis-epidemiologic hyperkeratotic type
1105	Noncompaction of left ventricular myocardium
1113	Noncompaction of left ventricular myocardium
1120	Oculofacciodental syndrome
1140	Oligodonta-colorectal cancer syndrome
1153	Osteopetrosis-pseudogliomas syndrome
1164	Osteopetrosis-pseudogliomas syndrome
1171	Papillary serous carcinoma of the peritoneum
1183	Papillary serous carcinoma of the peritoneum
1227	Pigmentation of hair, skin, and eyes, variation in
1232	Pituitary ACTH-secreting adenoma
1238	Pituitary tumor, with diabetes insipidus
1239	Pneumothorax, primary spontaneous
1263	Prion disease with protracted course
1265	Proteasome, hyperparathyroidism, carcinoid syndrome
1267	Rhinomelic chondrodyplasia punctata
1325	Rhinomelic chondrodyplasia punctata
1343	Robins syndrome, autosomal recessive
1347	Rothmund-Thomson syndrome, childhood and adult forms
1361	Schwart-Zampel syndrome, type I
1378	Schwart-Zampel syndrome, type II
1383	Severe combined immunodeficiency
1443	Skin fragility-woolly hair syndrome
1446	Staphylococcal mastitis central vasculitis
1454	Staphylococcal mastitis central vasculitis
1456	Subcortical laminar heterotopia
1466	Sweat chloride elevation without CF
1474	Tauopathy and respiratory failure
1499	Transient bulbus of the newborn, I and II
1518	Transient bulbus of the newborn, I and II
1519	Transposition of great arteries, debré-lezenfeld
1528	Trisomy-pseudoischadotyly syndrome
1542	Unna-Thost disease, nonepidermolytic
1543	Unna-Thost disease, nonepidermolytic
1563	WAGR association with hydronephrosis
1569	Warthin's resistance/sensitivity
1611	XLA and isolated growth hormone deficiency
1611	Yemenite deaf-blind hypogonadism syndrome
2327	Genetic defect in the brain
2354	Genital atrophy absence of vas deferens
2785	Hypoplastic left heart syndrome, cyanomatia
3032	Hypoplastic left heart syndrome, cyanomatia
3144	Optic nerve colicoma with renal disease
3212	Persistent hyperinsulinemic hypoglycemia of infancy
3260	Premature chromosome condensation w/ microcephaly, mental retardation
3271	Ventricular fibrillation, idiopathic
3558	Ventricular fibrillation, idiopathic
4291	Cerebral cavernous malformations
5170	Placental steroid sulfatase deficiency
5233	Placental steroid sulfatase deficiency

Supporting Information Figure 13 | Bipartite-graph representation of the disome. A disorder (circle) and a gene (rectangle) are connected if the gene is implicated in the disorder. The size of the circle represents the number of distinct genes associated with the disorder. Isolated disorders (disorders having no links to other disorders) are not shown. Also, only genes connecting disorders are shown.

Other Applications

Quantifying the Performance of Individual Players in a Team Activity (Euro 2008)



Gene Ontology

Ontology

Definition

- “A set of concepts and categories in a subject area or domain that shows their properties and the relations between them.” *New Oxford American Dictionary, Oxford University Press 2013*

Gene Ontology

Gene Ontology (GO)

“a computational representation of our evolving knowledge of how genes encode biological functions at the molecular, cellular and tissue system levels”

- [Controlled vocabulary of terms](#)
 - ▶ Describe gene product characteristics
 - ▶ [Gene product annotation data](#)
- [Tools for GO](#)

Gene Ontology

Controlled vocabularies

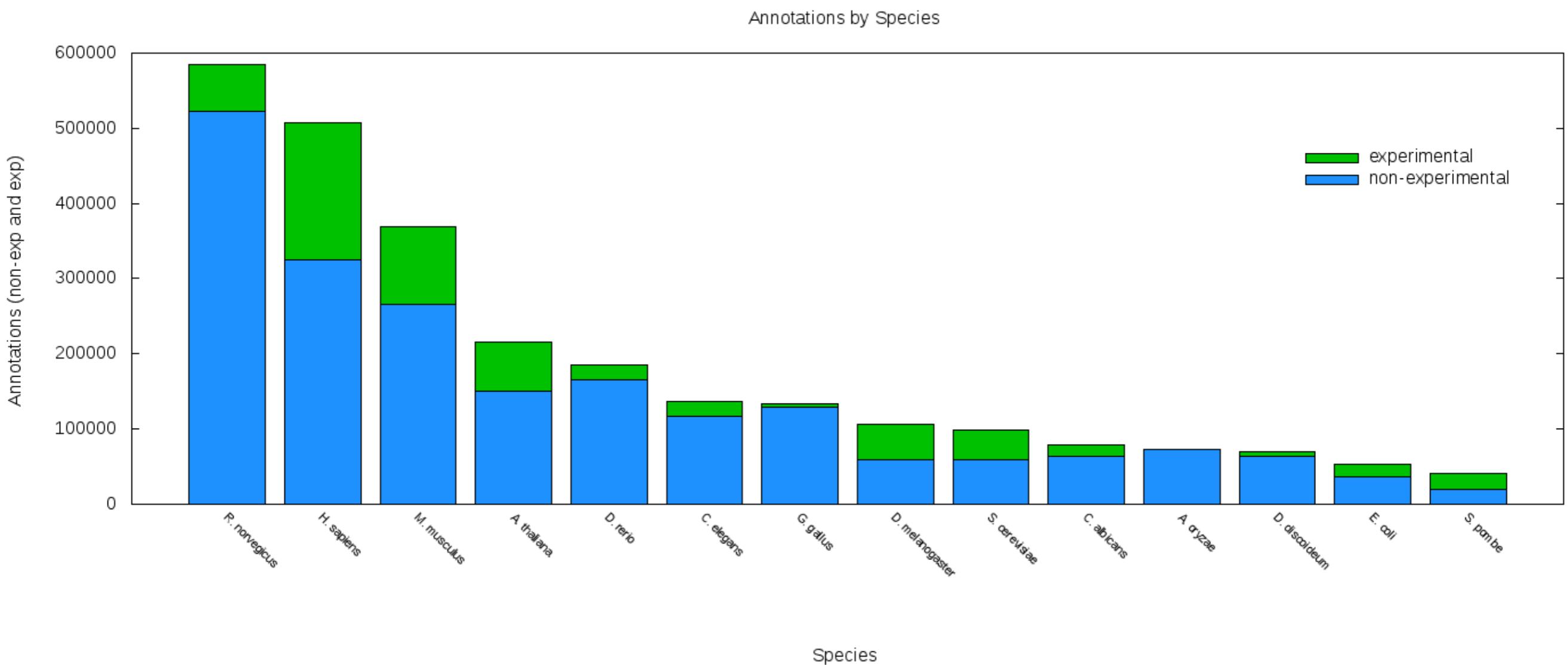
- **Cellular Component (CC)**, the parts of a cell or its extracellular environment
- **Molecular Function (MF)**, the elemental activities of a gene product at the molecular level, such as binding or catalysis
- **Biological Process** operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

Gene Ontology

Example

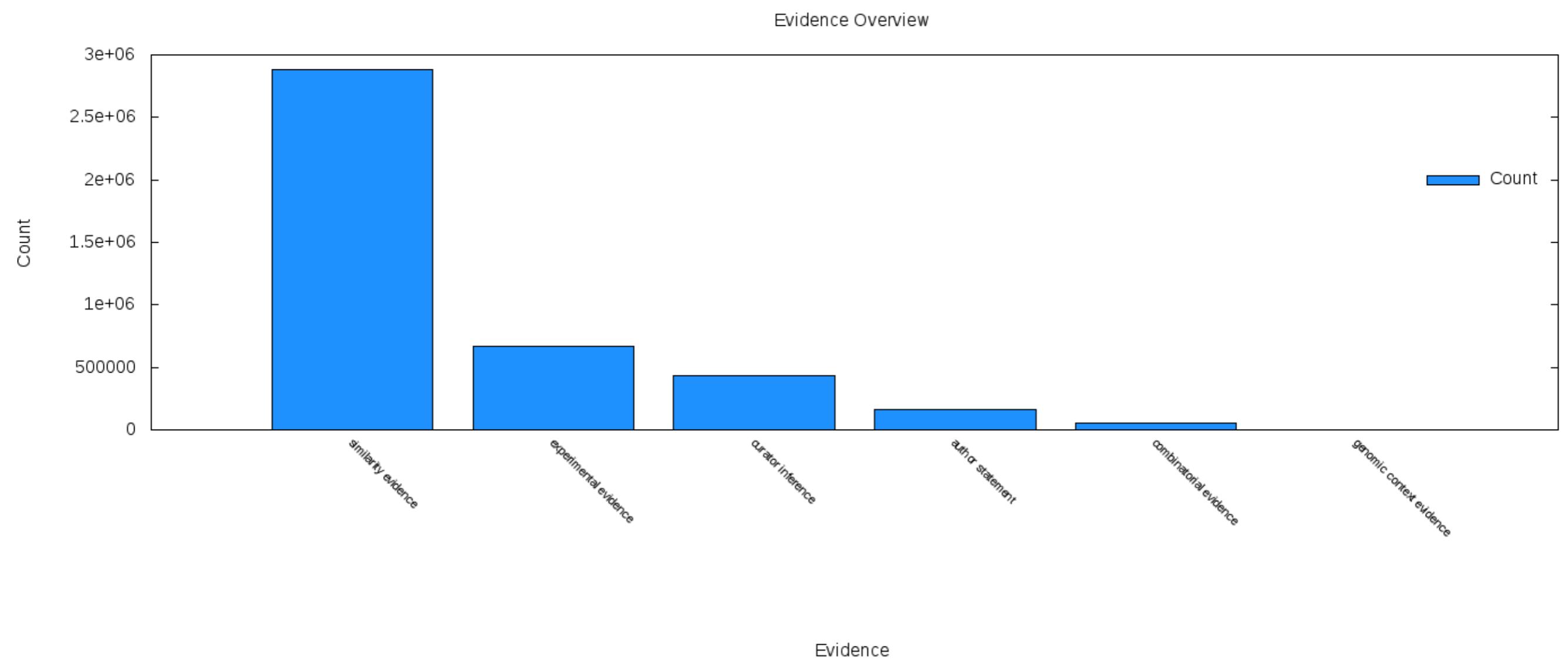
- gene product "cytochrome c"
 - ▶ **Molecular Function**
 - "oxidoreductase activity"
 - ▶ **Biological Process**
 - "oxidative phosphorylation"
 - "induction of cell death"
 - ▶ **Cellular Component**
 - "mitochondrial matrix"
 - "mitochondrial inner membrane".

Gene Ontology



Note: GO vocabulary is designed to be species-agnostic, and includes terms applicable to prokaryotes and eukaryotes, and single and multicellular organisms.

Gene Ontology

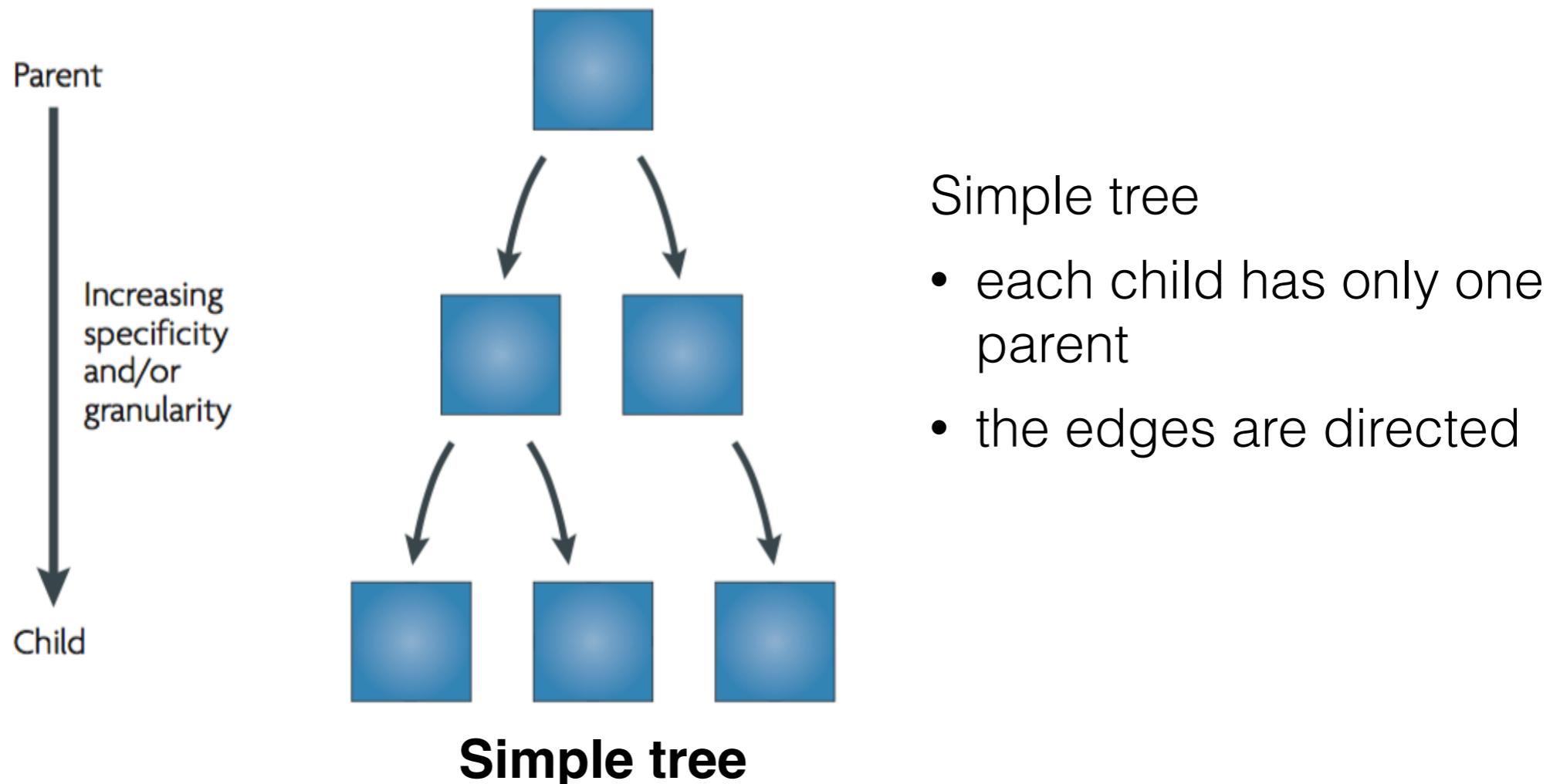


Gene Ontology

evidence code	evidence code description	source of evidence	Manually checked
IDA	Inferred from direct assay	Experimental	Yes
IEP	Inferred from expression pattern	Experimental	Yes
IGI	Inferred from genetic interaction	Experimental	Yes
IMP	Inferred from mutant phenotype	Experimental	Yes
IPI	Inferred from physical interaction	Experimental	Yes
ISS	Inferred from sequence or structural similarity	Computational	Yes
RCA	Inferred from reviewed computational analysis	Computational	Yes
IGC	Inferred from genomic context	Computational	Yes
IEA	Inferred from electronic annotation	Computational	No
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes
ND	No biological data available	No information available	Yes
NR	Not recorded	Unknown	Yes

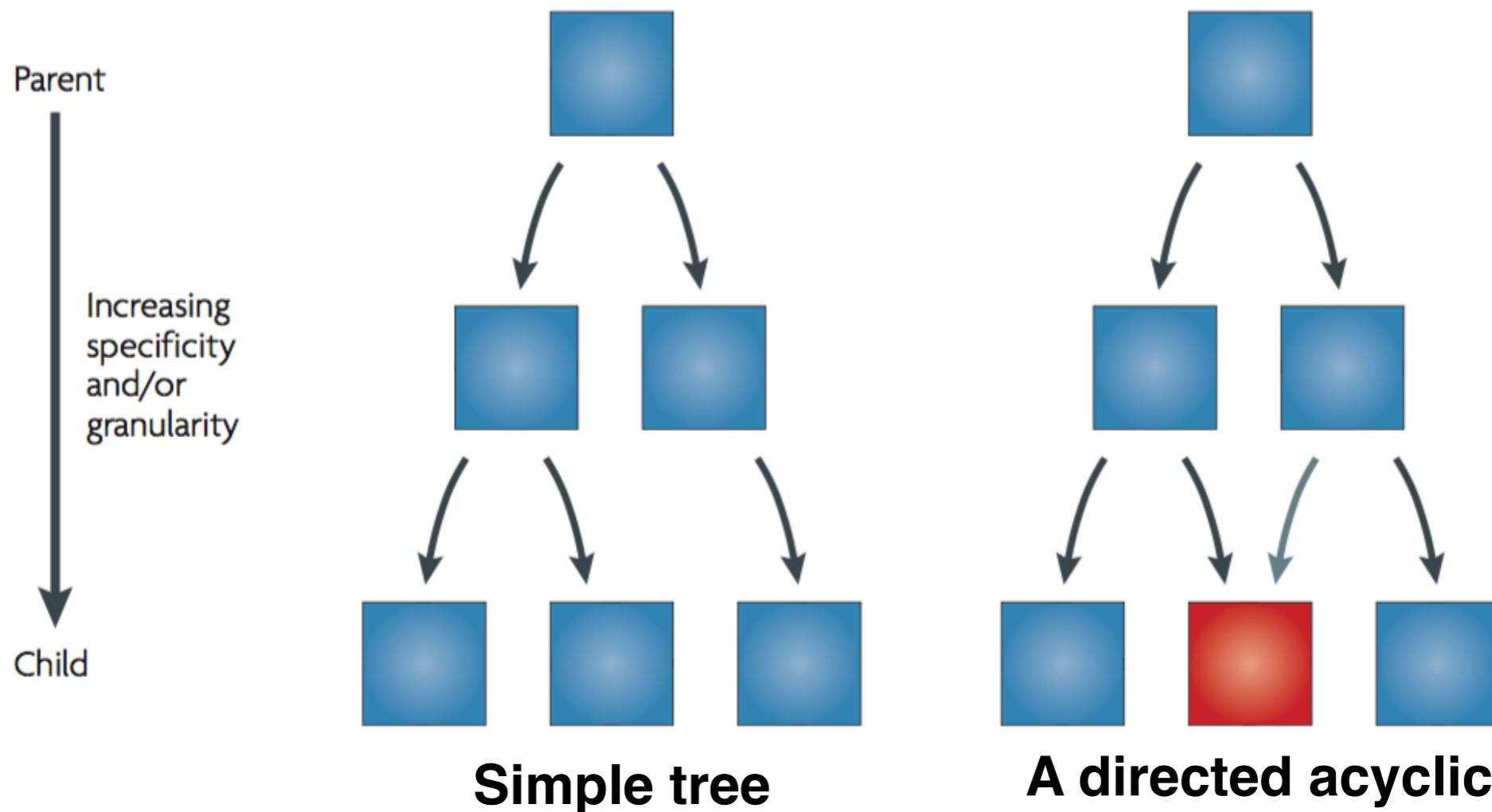
Gene Ontology

GO ontology structured as directed acyclic graph

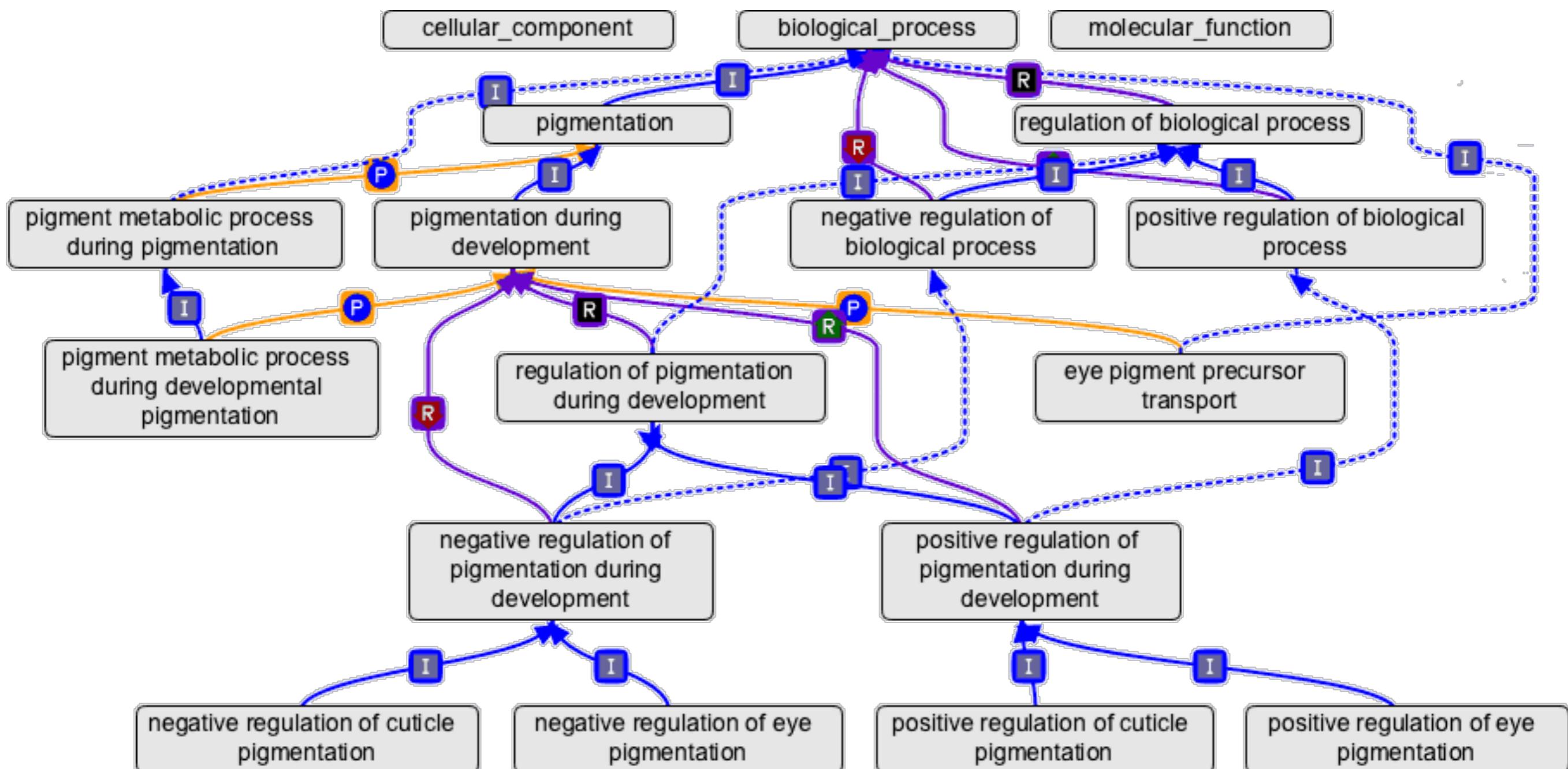


Gene Ontology

GO ontology structured as **directed acyclic graph**.

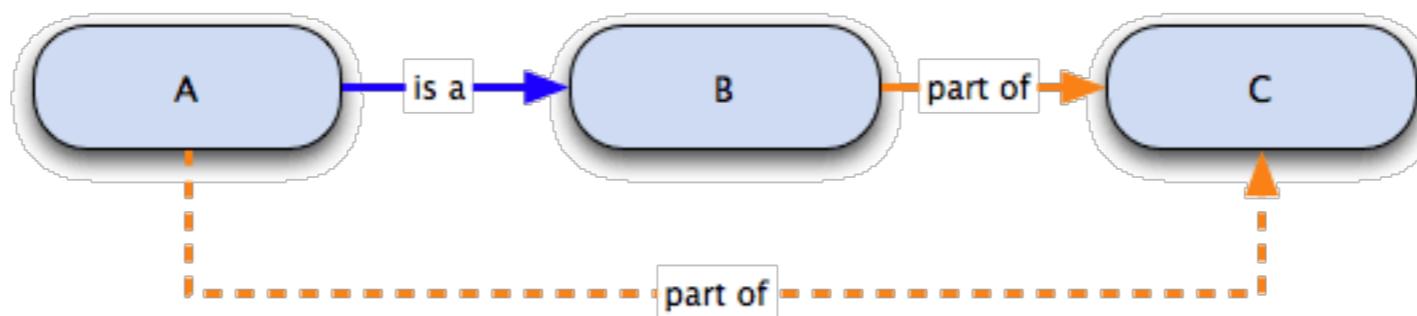


Gene Ontology



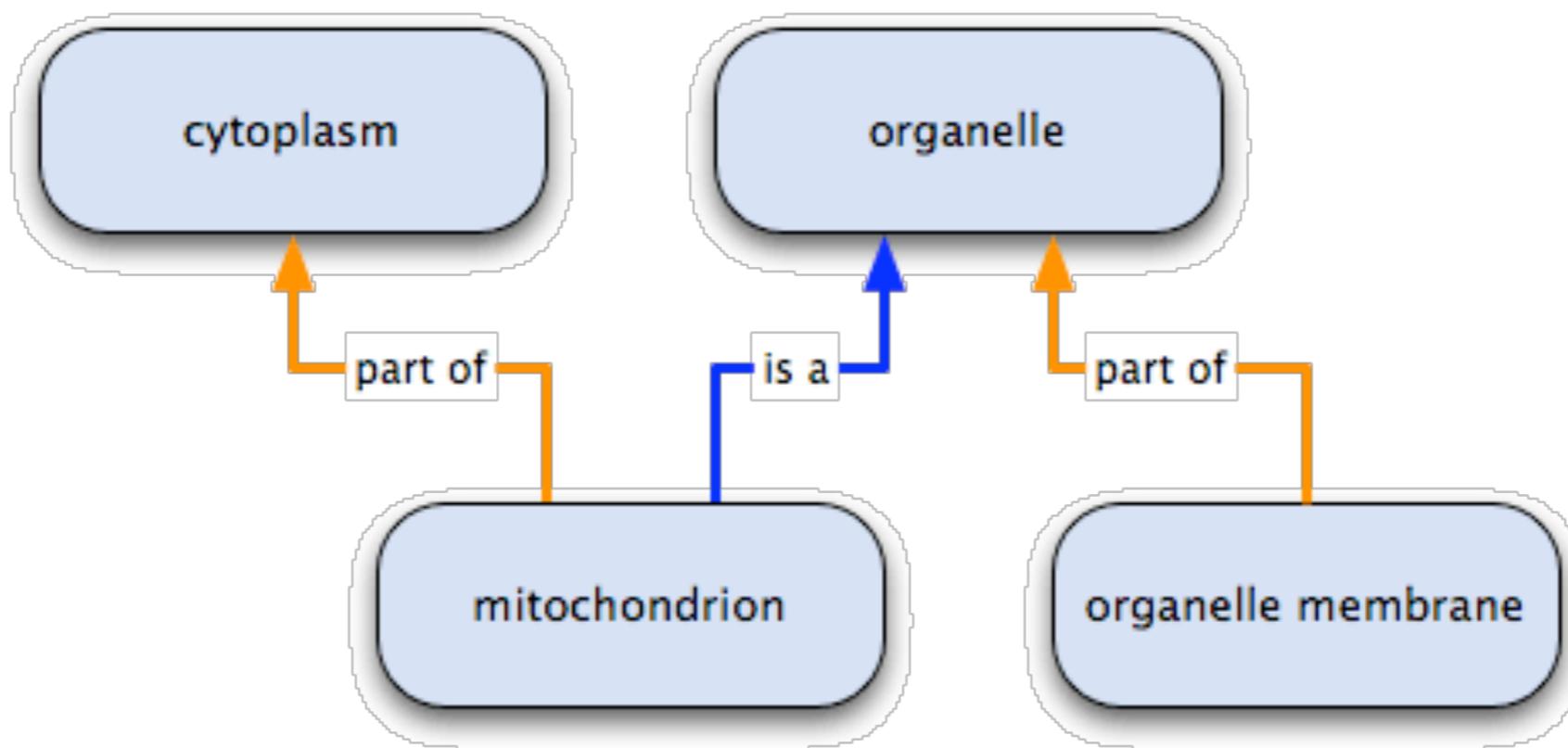
Each term can have relationships to one or more other terms in the same or other domains

Gene Ontology

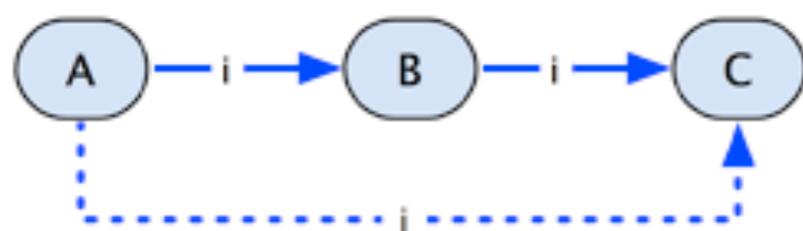


- *A is a B*
- *B is part of C*
- we can infer that *A is part of C*

Gene Ontology



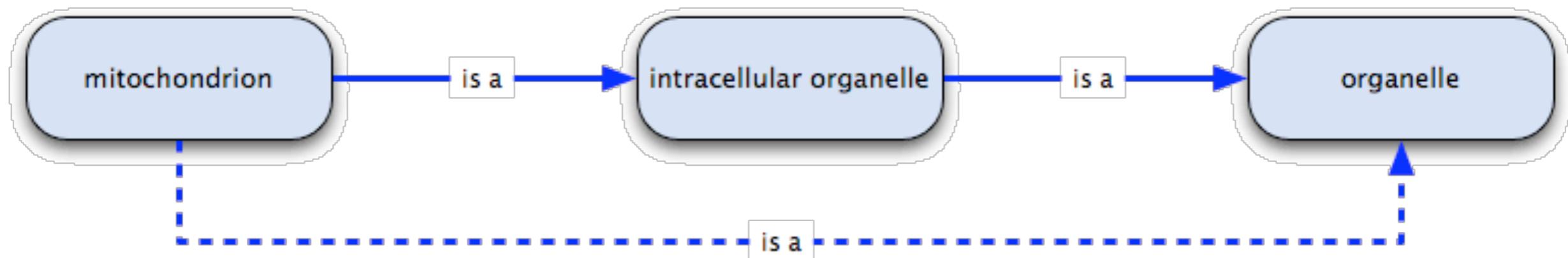
Gene Ontology



The *is a* relation is **transitive**:

If A *is a* B & B *is a* C

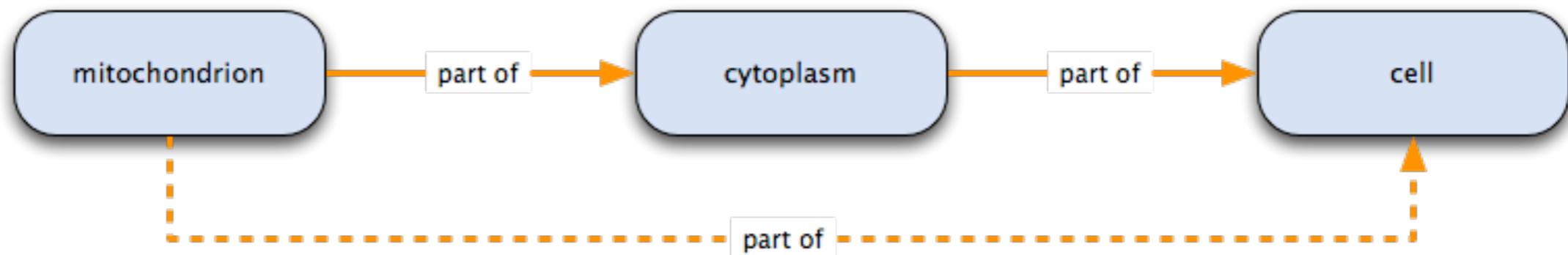
We can infer that A *is a* C



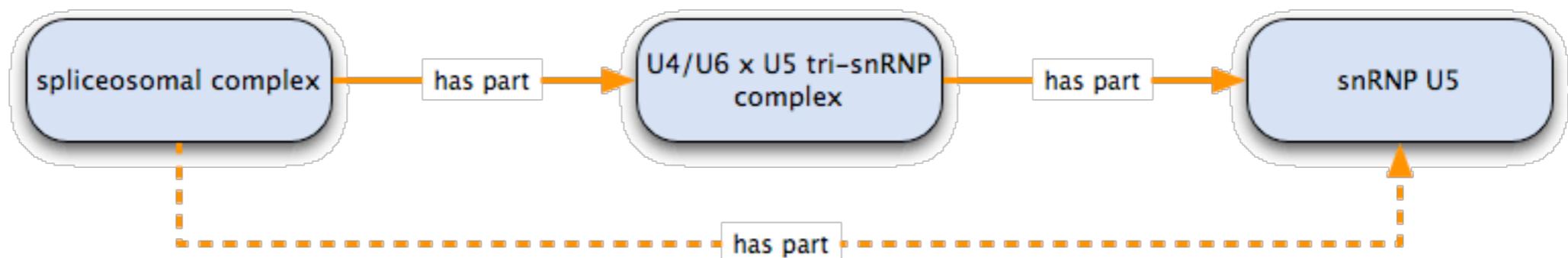
mitochondrion *is an* intracellular organelle and intracellular organelle *is an* organelle
therefore mitochondrion *is an* organelle.

Gene Ontology

part of



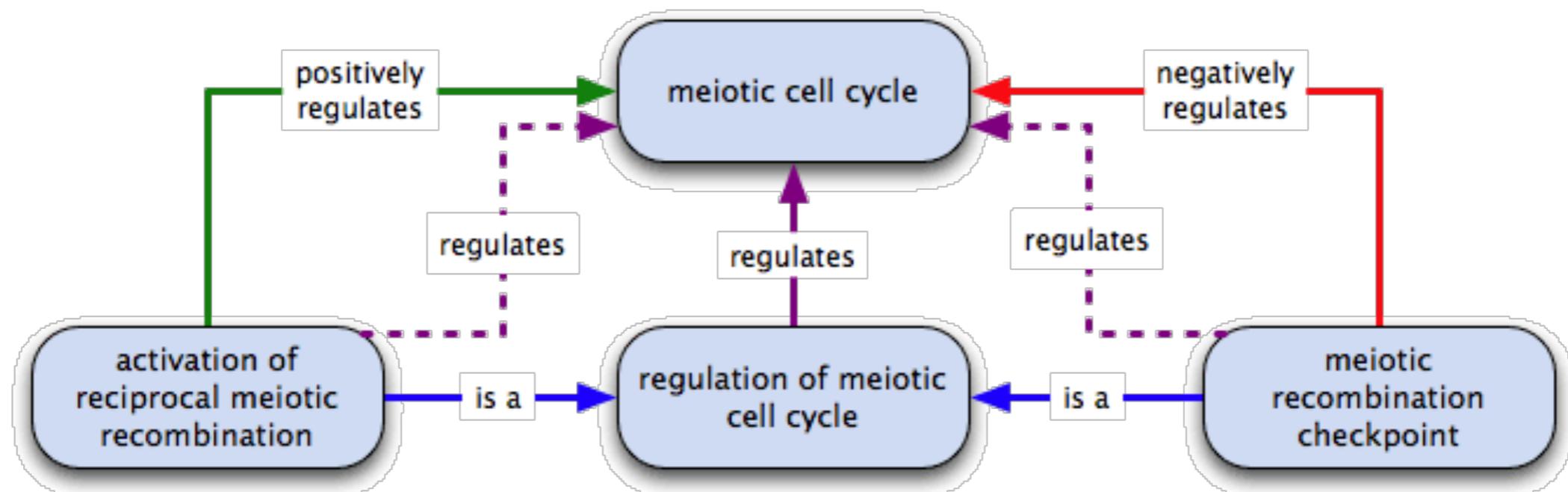
has part



- spliceosomal complex has part U4/U6 x U5 tri-snRNP complex
- U4/U6 x U5 tri-snRNP complex has part snRNP U5
- therefore spliceosomal complex has part snRNP U5

Gene Ontology

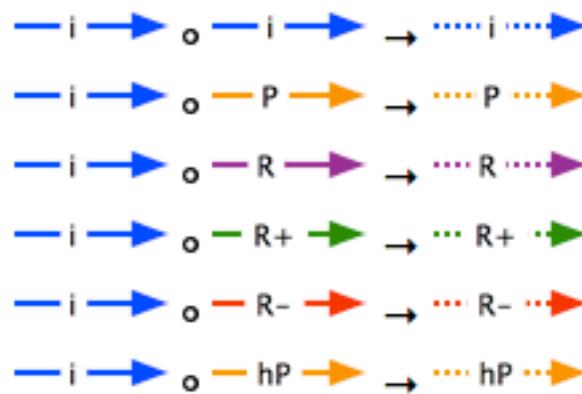
regulates



Gene Ontology

Inferences

is a



part of



regulates



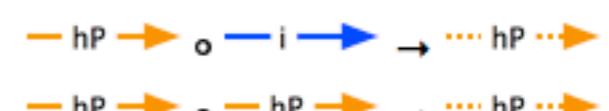
positively regulates



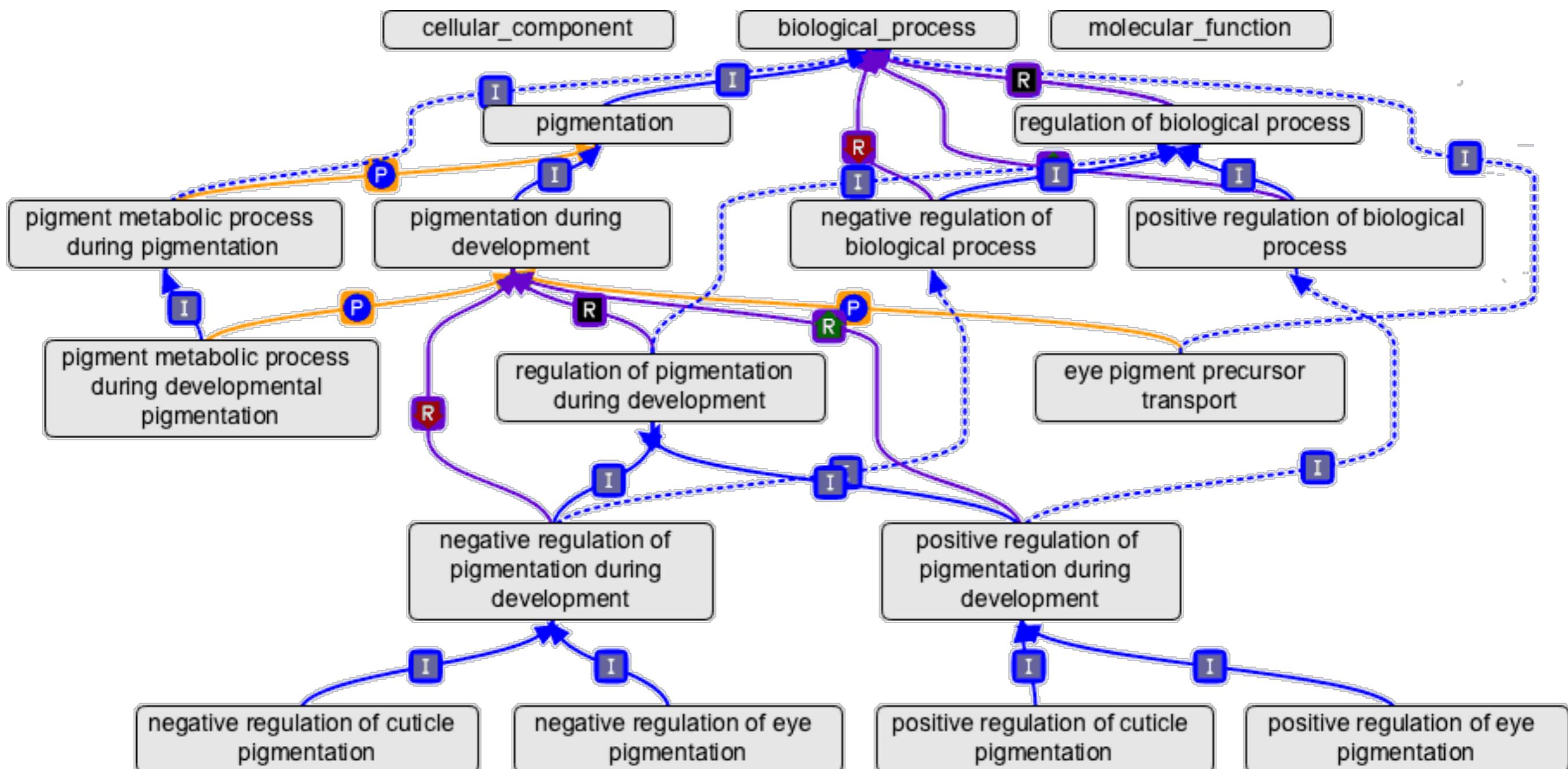
negatively regulates



has part



Gene Ontology



Each term can have relationships to one or more other terms in the same or other domains

Gene Ontology

Ten Quick Tips for Using the Gene Ontology

1. Know the Source of the GO Annotations You Use
2. Understand the Scope of GO Annotations
3. Consider Differences in Evidence Codes
4. Probe Completeness of GO Annotations
5. Understand the Complexity of the GO Structure
6. Choose Analysis Tools Carefully
7. Provide the Version of the Data/Tools Used
8. Seek Input from the GOC Community and Make Use of GOC Resources
9. Contribute to the GO
10. Acknowledge the Work of the GO Consortium

Gene Ontology

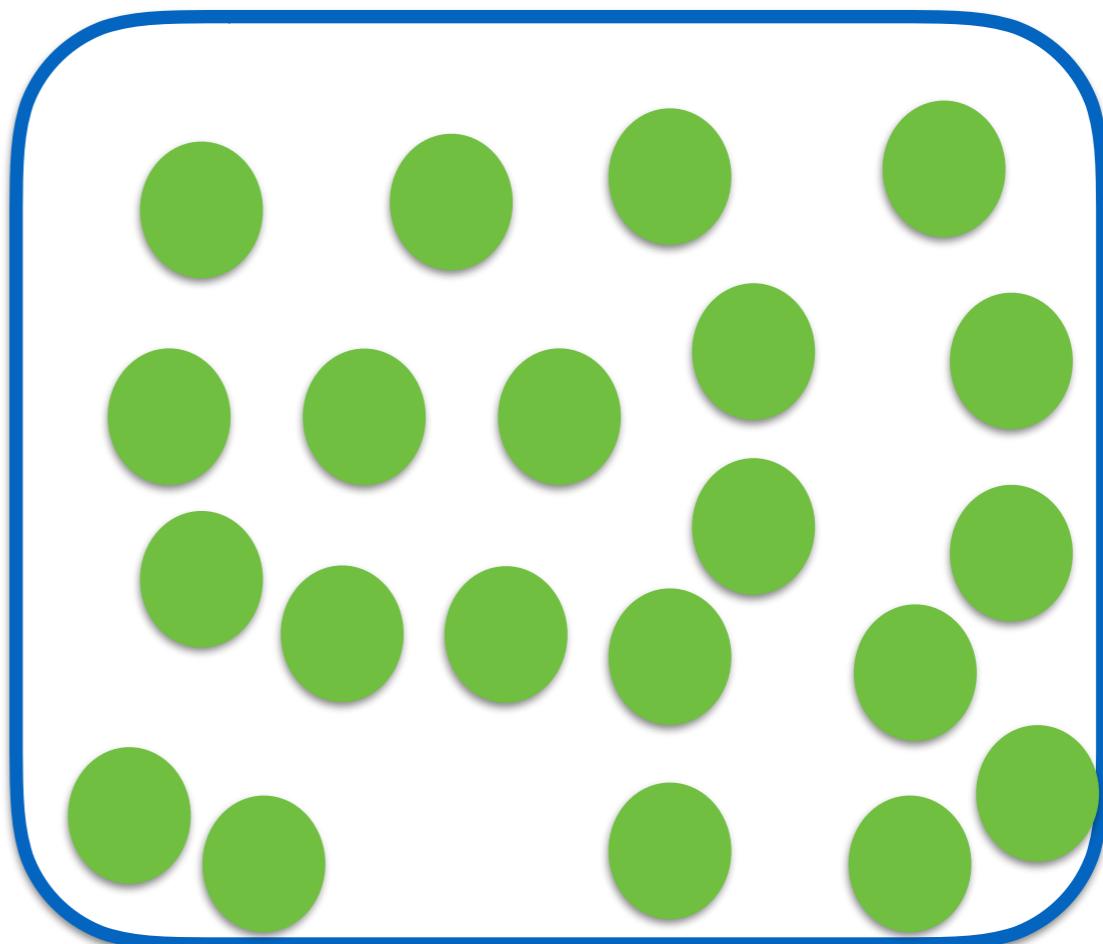
GO Enrichment Analysis

For a set of genes up/down-regulated find which GO terms are over-represented (or under-represented) using annotations for that gene set

Gene Ontology

Basic Enrichment Analysis

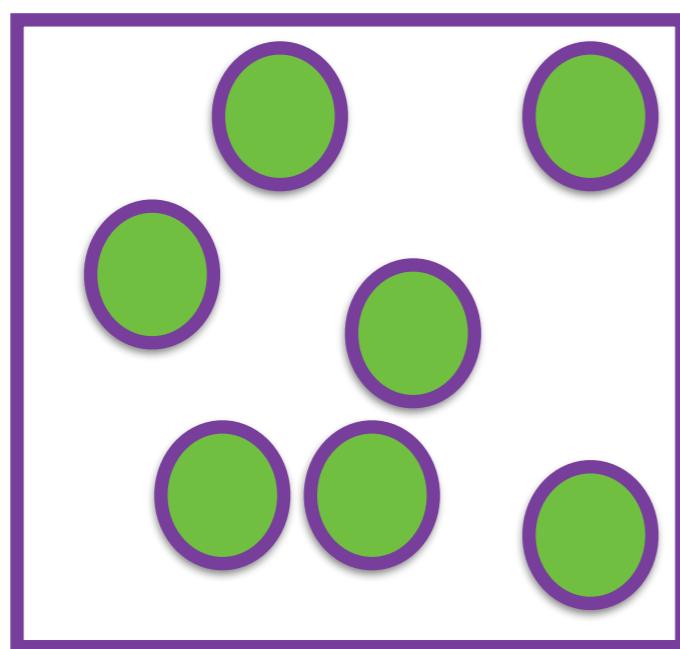
Global Gene Set



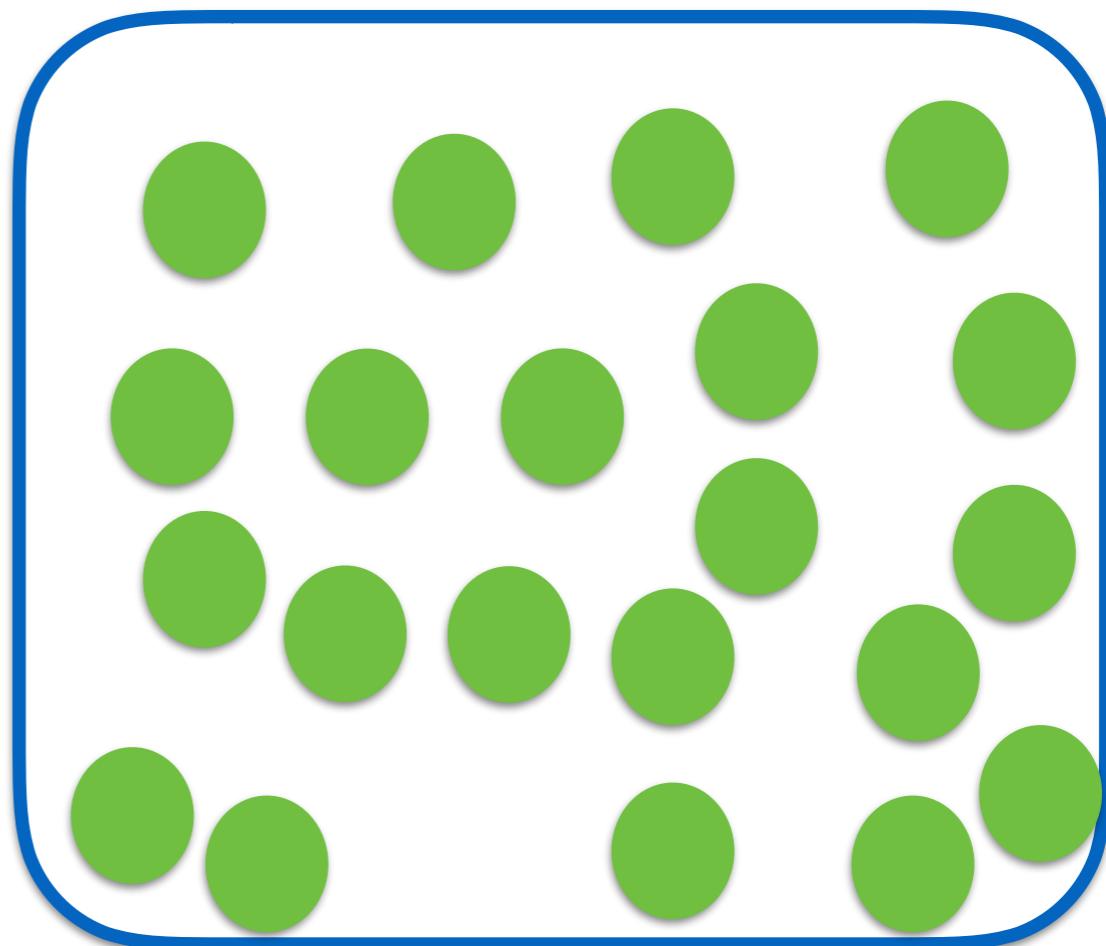
Gene Ontology

Basic Enrichment Analysis

GO Category or Pathway of interest



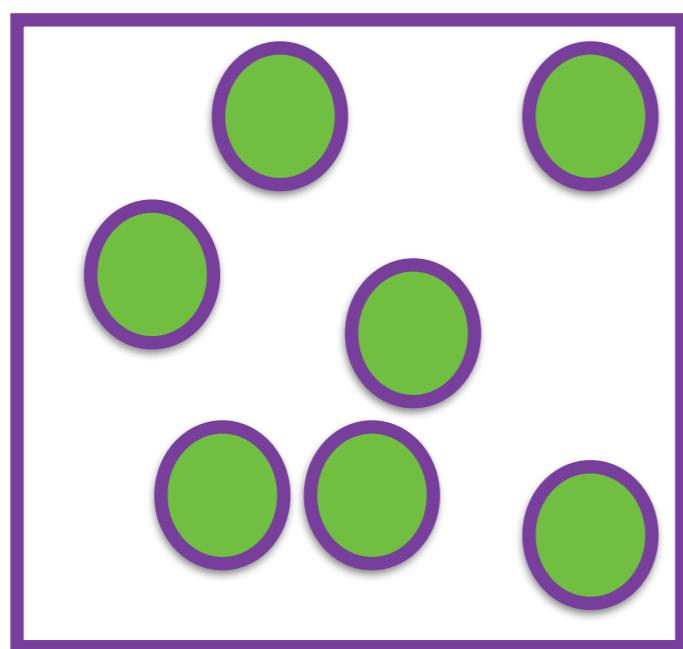
Global Gene Set



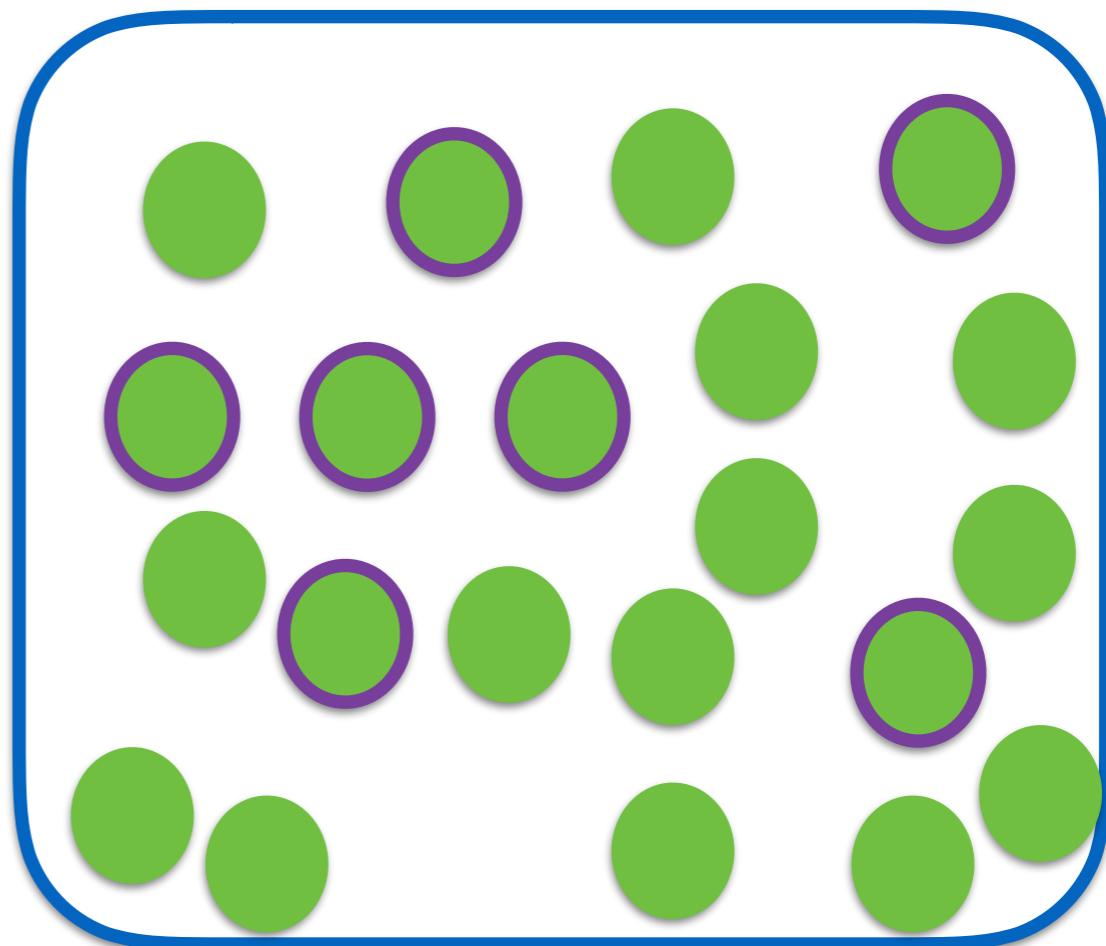
Gene Ontology

Basic Enrichment Analysis

GO Category or Pathway of interest



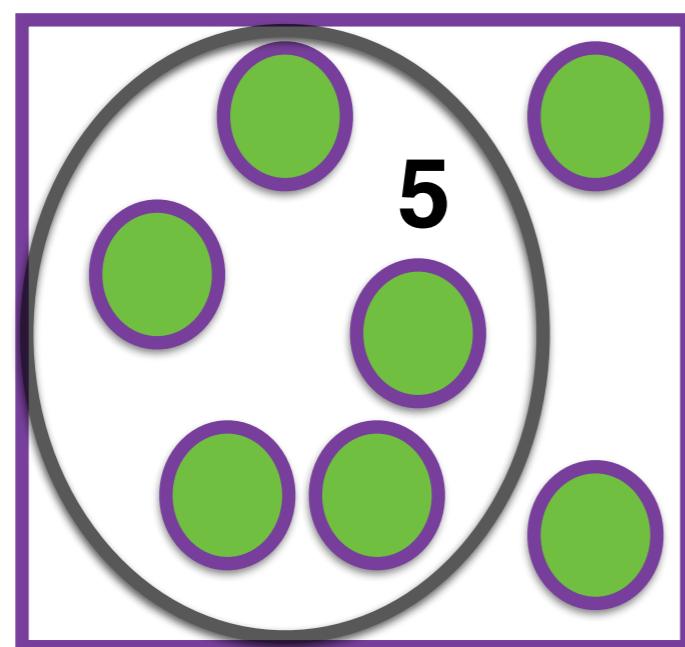
Global Gene Set



Gene Ontology

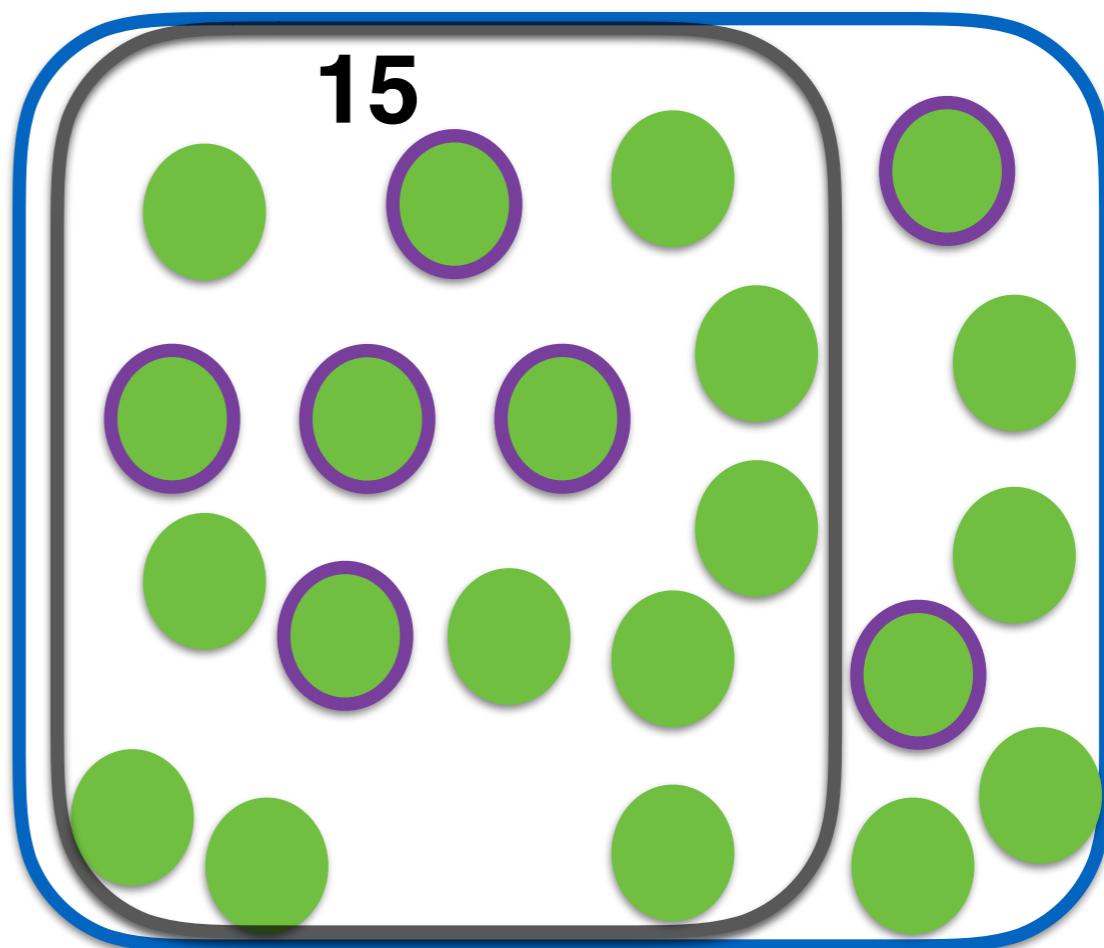
Basic Enrichment Analysis

GO Category or Pathway of interest



query/tested

Global Gene Set

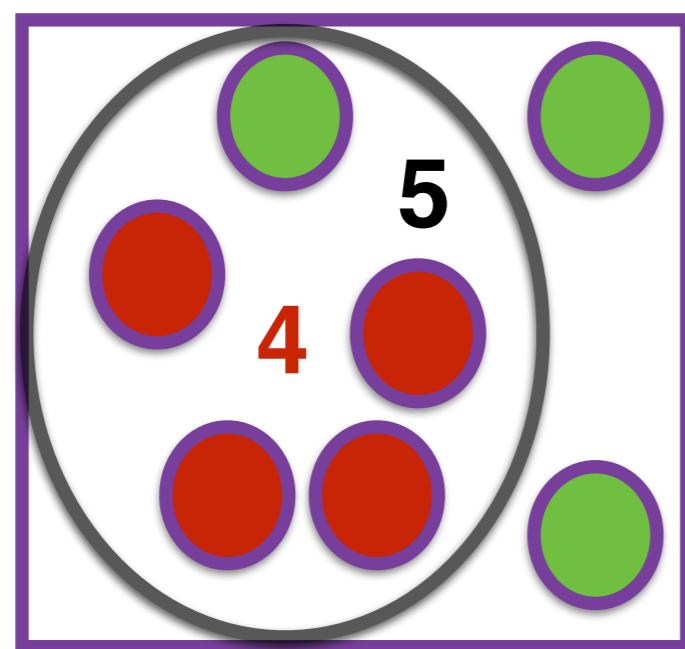


**select the right background set
e.g. right organism**

Gene Ontology

Basic Enrichment Analysis

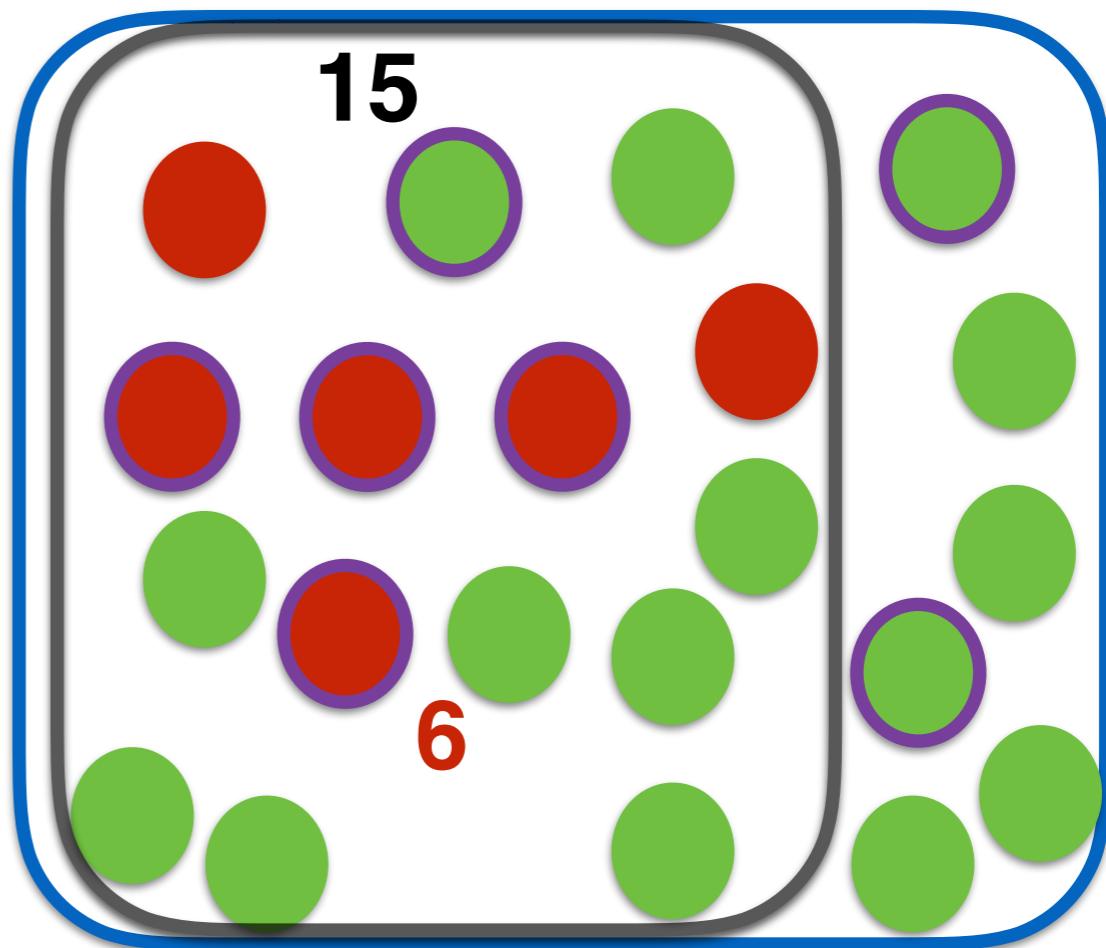
GO Category or Pathway of interest



Experimentally tested

Interesting behavior
E.g. Unregulated

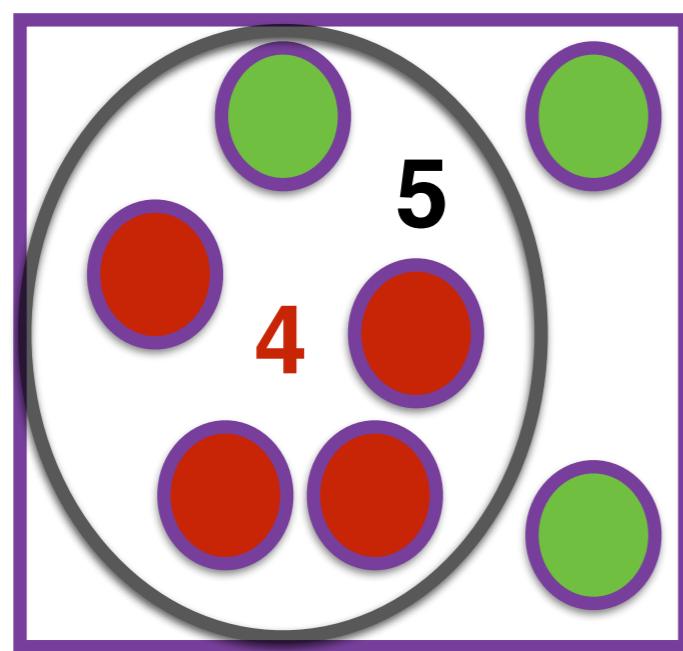
Global Gene Set



Gene Ontology

Basic Enrichment Analysis

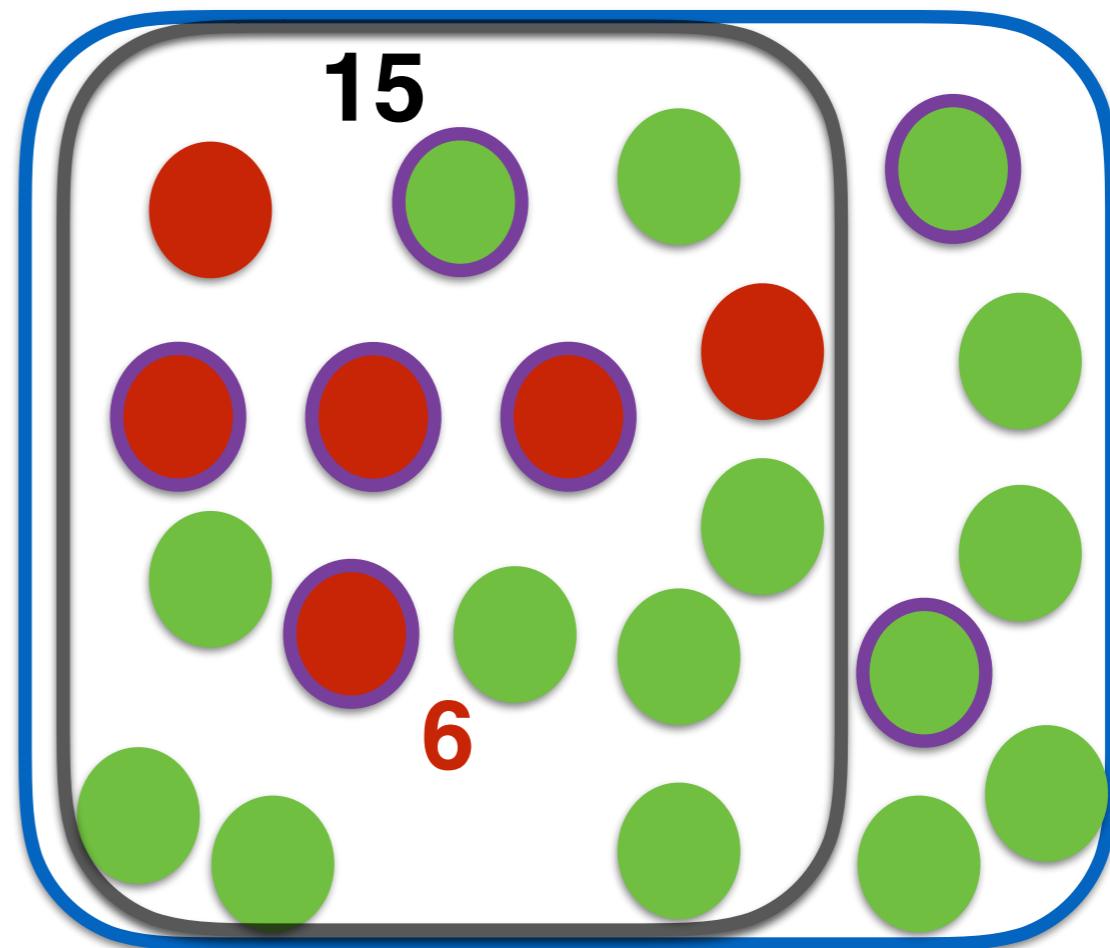
GO Category or Pathway of interest



Experimentally tested

Interesting behavior
E.g. Unregulated

Global Gene Set

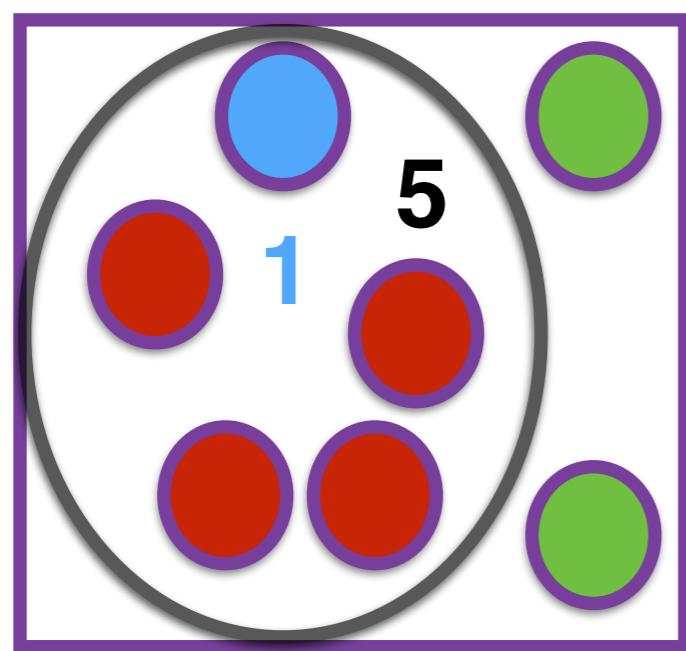


4 out of 6 Vs 5 out of 15

Gene Ontology

Basic Enrichment Analysis

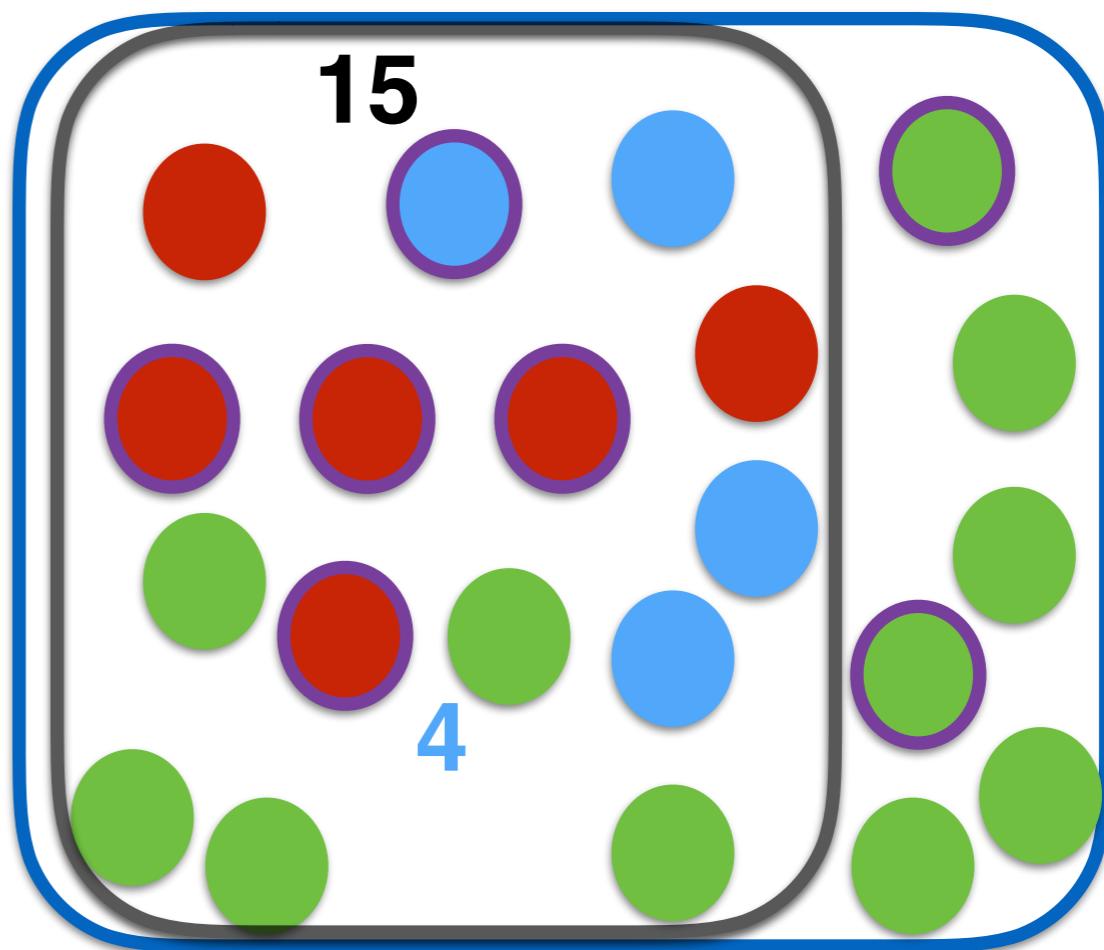
GO Category or Pathway of interest



Experimentally tested

Interesting behavior
E.g. Downregulated

Global Gene Set

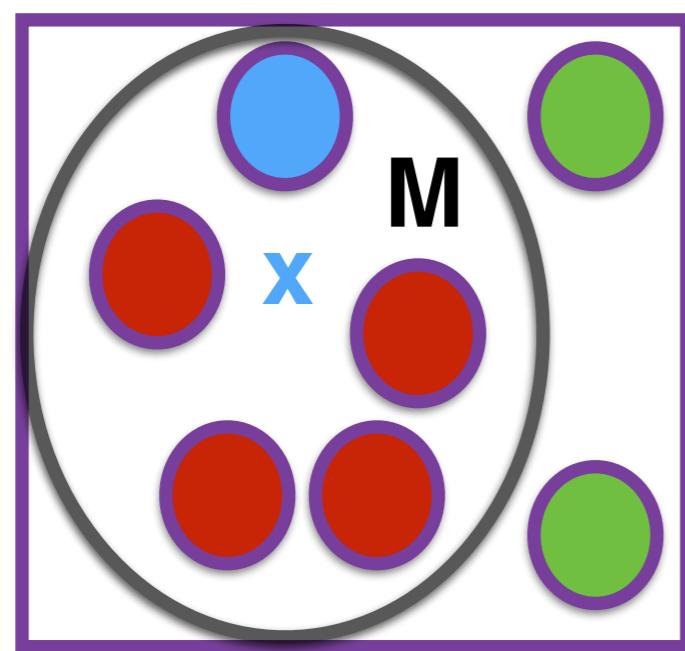


1 out of 4 Vs 5 out of 15

Gene Ontology

Basic Enrichment Analysis

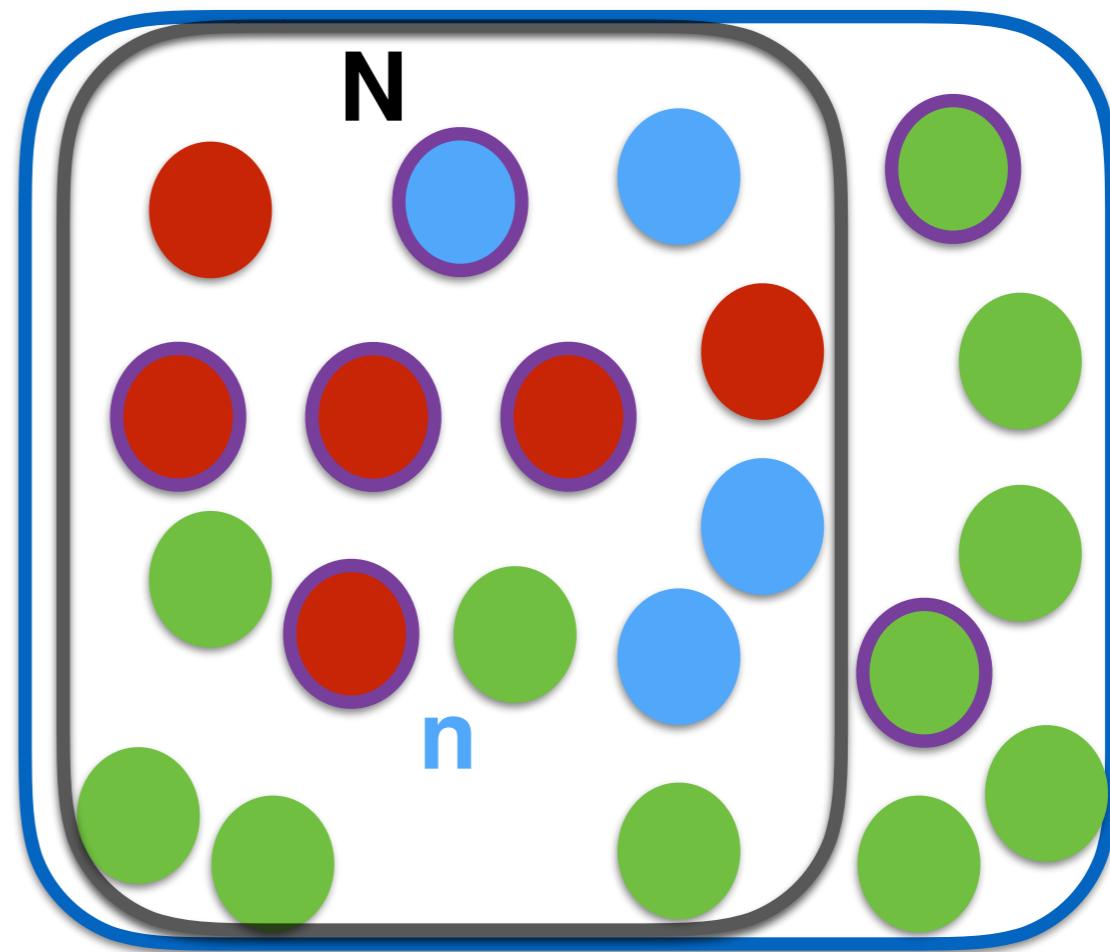
GO Category or Pathway of interest



Experimentally tested

Interesting behavior

Global Gene Set



x out of n Vs M out of N

Gene Ontology

Basic Enrichment Analysis

probability of getting at least x successes, i.e. x or more genes in the GO category (hypergeometric function sum)

$$p = \sum_{i=x}^n \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}$$

where

$$\binom{k}{l} = \frac{k!}{l!(k-l)!}$$

- N genes in big set
- M genes in category in big set
- n genes of interest
- looking of x or more hits

Remember to Correct For Multiple Hypothesis Tests

x out of n Vs M out of N

Gene Ontology

- Panther (pantherdb.org)
- <http://geneontology.org>
- DAVID (<https://david.ncifcrf.gov>)
- GSEA (<http://software.broadinstitute.org/gsea/index.jsp>)
- many others...

