MANCHESTER 1824
The University of Manchester

Modelling approaches in biology

Jean-Marc Schwartz

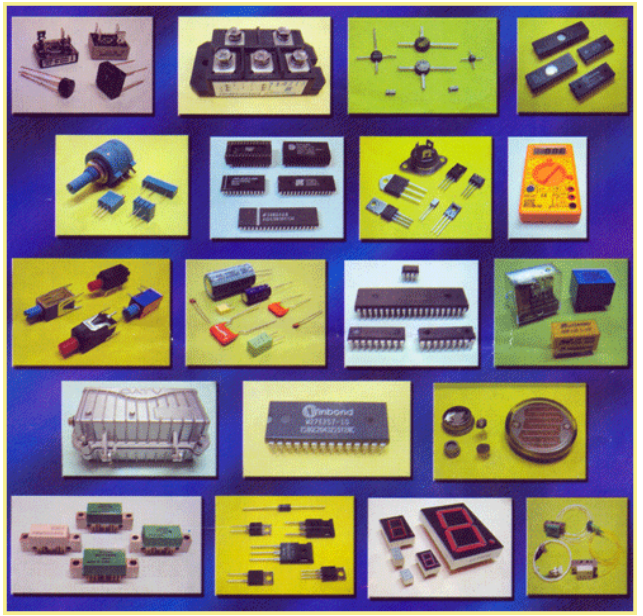
Faculty of Life Sciences
University of Manchester

jean-marc.schwartz@manchester.ac.uk

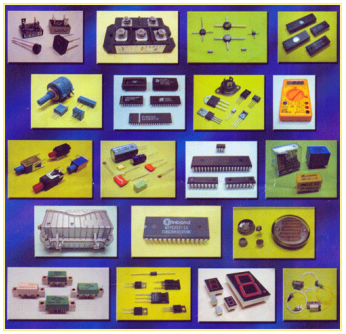
TGAC, Norwich, UK
27 January 2016

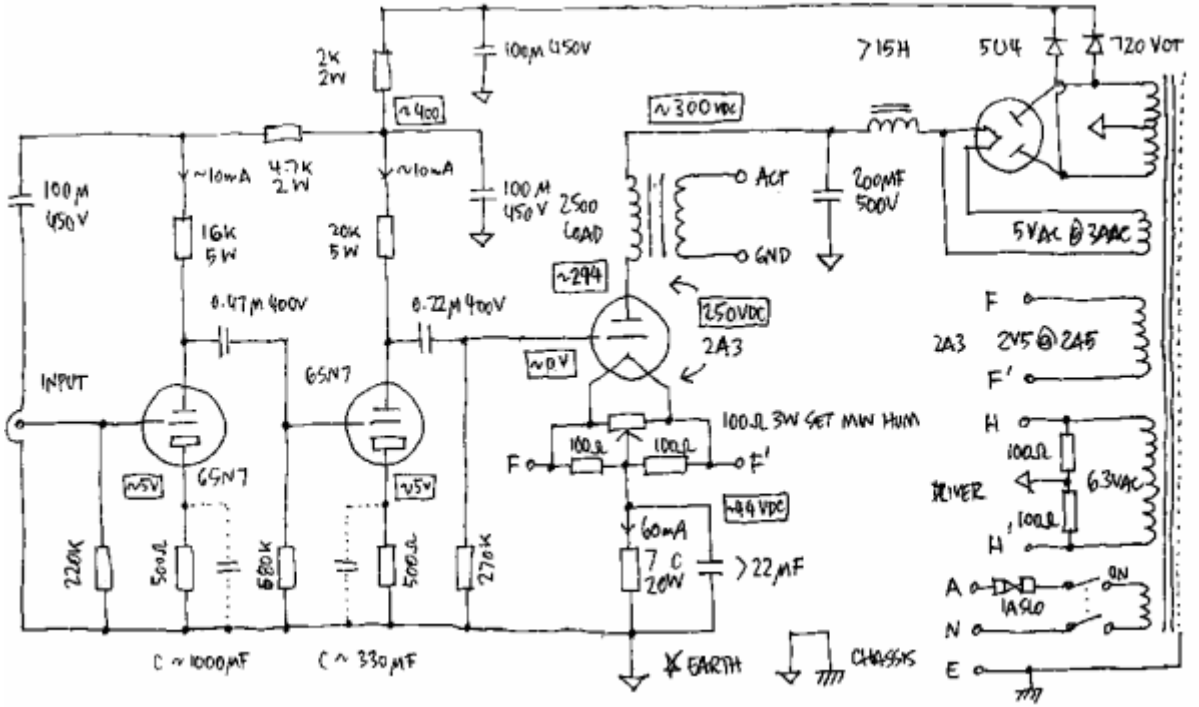
Why modelling?

- We can expect that all molecules will become measurable in the future...
- Will this be sufficient to understand a living system?



Why modelling?





Why modelling?

- A model is a simplified representation of reality, not a duplication of the real system.
  - A biological object or process can be described by different models.
  - The choice of a particular model reflects the purpose of the study and the desired level of precision.

- Modelling drives clarification and creates new biological knowledge.
  - Discrepancies between model predictions and experimental observations lead to the formulation of new hypotheses.

Logical modelling

- Precise quantitative data are often difficult to obtain in biology.
- Biological knowledge is often described in qualitative terms.
- Examples:

  "p53 activates the CD95 gene in response to DNA damage by anticancer drugs"

  "E2F1 inhibits MDM2 expression by suppressing its promoter activity"

  "Bcl3 is known to promote cell proliferation and inhibit apoptosis"

Logical modelling

- We can represent this information in binary terms:

  0 ↔ OFF, low amount, inactive, down-regulated...

  1 ↔ ON, high amount, active, up-regulated...

- Interactions can be represented by logical rules:

  Activation ↔ activates, promotes, increases, up-regulates...

  Inhibition ↔ inhibits, represses, decreases, down-regulates...

Model construction

- Interaction databases, e.g.:
  BioGRID, DIP, HINT, HPRD, IntAct, STRING...

- Primary literature papers

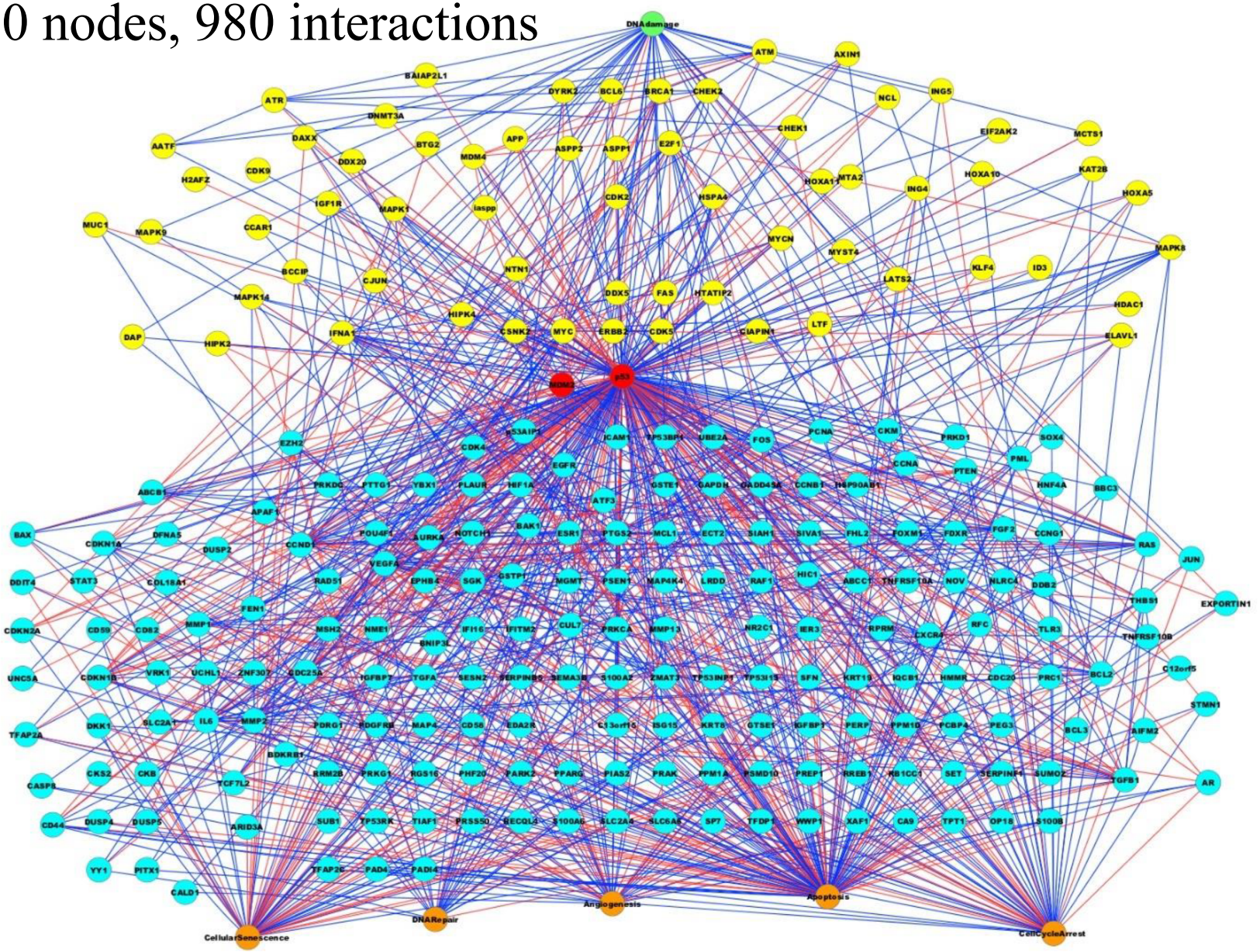
- Filtering and resolution of discrepancies

- Clarification of logical rules, e.g.:
  Database: $A \rightarrow C$, $B \rightarrow C$
  Paper: $A + B \rightarrow C$

- Removal of indirect interactions:
  Direct: $A \rightarrow B$, $B \rightarrow C$
  Indirect: $A \rightarrow C$

Example: Boolean model of DNA damage

- 260 nodes, 980 interactions



Hussain *et al.* (2015) *Genomics Comput Biol* **1**: e20

Model application

- The sequence of states given by the Boolean network model approximates the dynamical behaviour of the system.

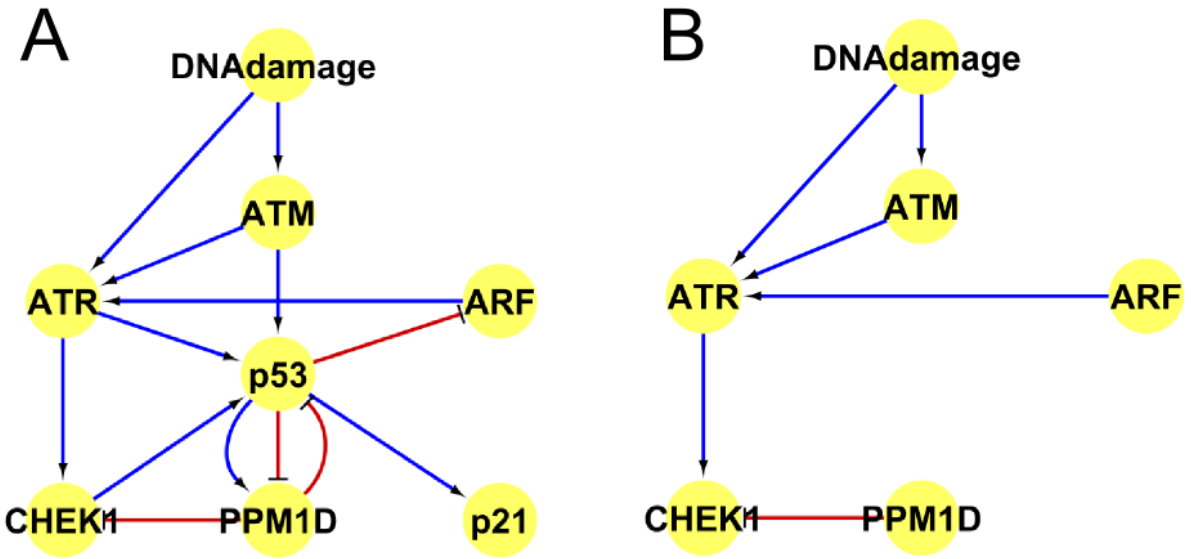
- Iterations are repeated until the systems reaches an end state:
  - When the system reaches a steady-state, the values of nodes no longer change with time.
  - When the system reaches a cycle, nodes recurrently pass through the same values.

- Effects of mutations or treatments can be predicted.

Model application

- Predicted an up-regulation of CHEK1 in the absence of functional p53, which was confirmed by experiments.

- 61 other predictions, with implications for treatment of p53 negative tumours.



Tian *et al.* (2013) *PLoS One* **8**: e72303

Genome-wide experimental validation

- Predictions by logical steady state analysis of p53 wild type model and p53 mutant model were compared with gene expression profiles.
- Two parameters were defined to evaluate the strength of our model prediction: $\mathbf{E_{mod}}$ and $\mathbf{E_{exp}}$
- $\mathbf{E_{mod}}$ represents the node state change in the model prediction, with $\mathbf{E_{mod} \in \{-1,0,1\}}$
- $\mathbf{E_{exp}}$ represents the node state change in microarray expression data, with $\mathbf{E_{exp} \in \{-1,0,1\}}$

Genome-wide experimental validation

- We defined the fold change ***FC*** by:

  $$FC = M1 / M2$$

  where ***M1*** is the median of expression values in sample 1

  where ***M2*** is the median of expression values in sample 2

- A threshold θ was defined using the mean value of $\log_{10}(FC(i))$, μ, and the standard deviation, σ:

  $\theta_{max} = \mu + \sigma$

  $\theta_{min} = \mu - \sigma$

- If $\log(FC(i)) > \theta_{max}$ gene i was considered as up-regulated, $E_{exp} = 1$
- If $\log(FC(i)) < \theta_{min}$ gene i was considered as down-regulated, $E_{exp} = -1$
- If $\theta_{min} < \log(FC(i)) < \theta_{max}$ gene i was considered as unchanged, $E_{exp} = 0$

Genome-wide experimental validation

- The difference between model predictions and microarray expression data is defined by $d = |E_{mod} - E_{exp}|$

- If $d = 0$, the prediction is correct.

- If $d = 1$, there is a small error between model prediction and microarray expression change (e.g. up-regulated vs. unchanged).

- If $d = 2$, there is a large error between model prediction and microarray expression change (e.g. up-regulated vs. down-regulated).

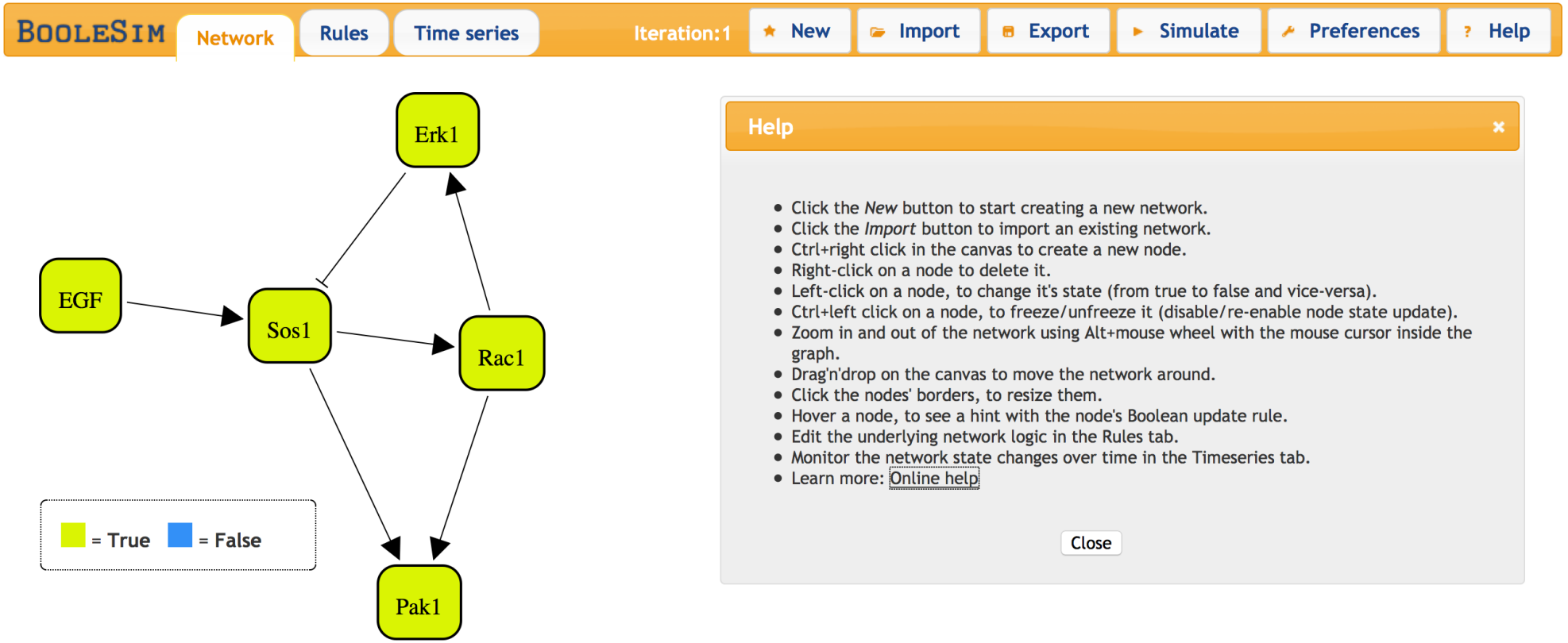
Genome-wide experimental validation

Experimental Source	Experimental Target	Model LSSA *in silico* simulation	Percentage (%)		
			Correct	Small error	Large error
U2OS cells with DNA damage	SaOS2 cells with DNA damage	p53 wild type DNA damage ON vs. p53 KO DNA damage ON	55	40	4.5
U2OS cells no DNA damage	SaOS2 cells no DNA damage	p53 wild type DNA damage OFF vs. p53 KO DNA damage OFF	58	37	5
U2OS cells no DNA damage	U2OS cells with DNA damage	p53 wild type DNA damage OFF vs. p53 wild type DNA damage ON	71	29	1
SaOS2 cells no DNA damage	SaOS2 cells with DNA damage	p53 KO DNA damage OFF vs. p53 KO DNA damage ON	68	31	1
HCT116 p53 +/+ no DNA damage	HCT116 p53 -/- no DNA damage	HCT116 p53 wild type vs. HCT116 p53 null	55	42	3

Hussain *et al.* (2015) *Genomics Comput Biol* **1**: e20

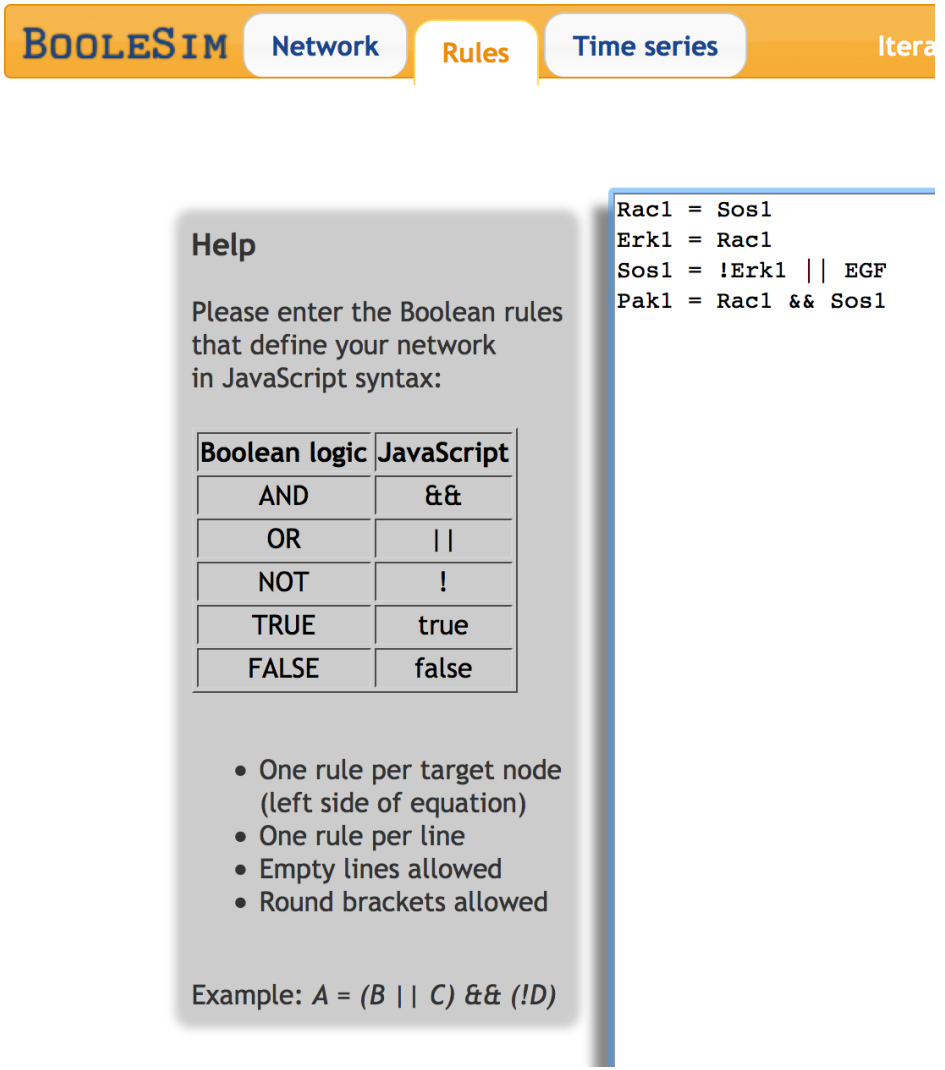
BooleSim

- Online interface available at: http://rumo.biologie.hu-berlin.de/boolesim/

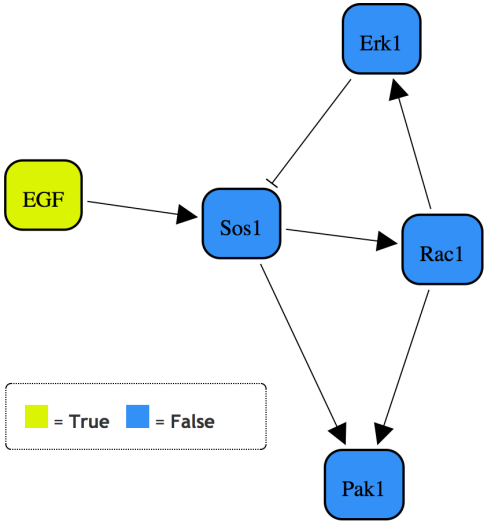


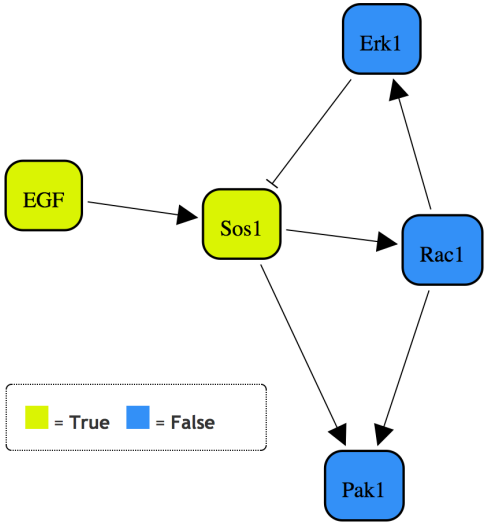
BooleSim

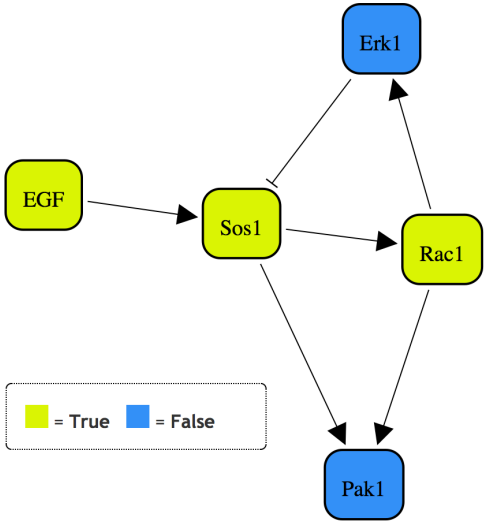
- EGF activates Sos1
- Sos1 activates Rac1
- Rac1 activates Erk1
- Erk1 inhibits Sos1
- Sos1 and Rac1 activate Pak1

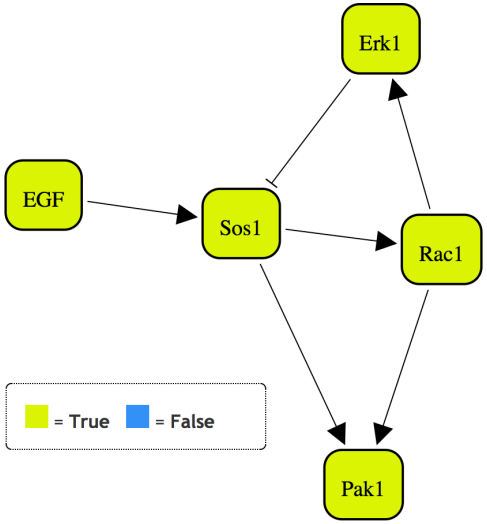


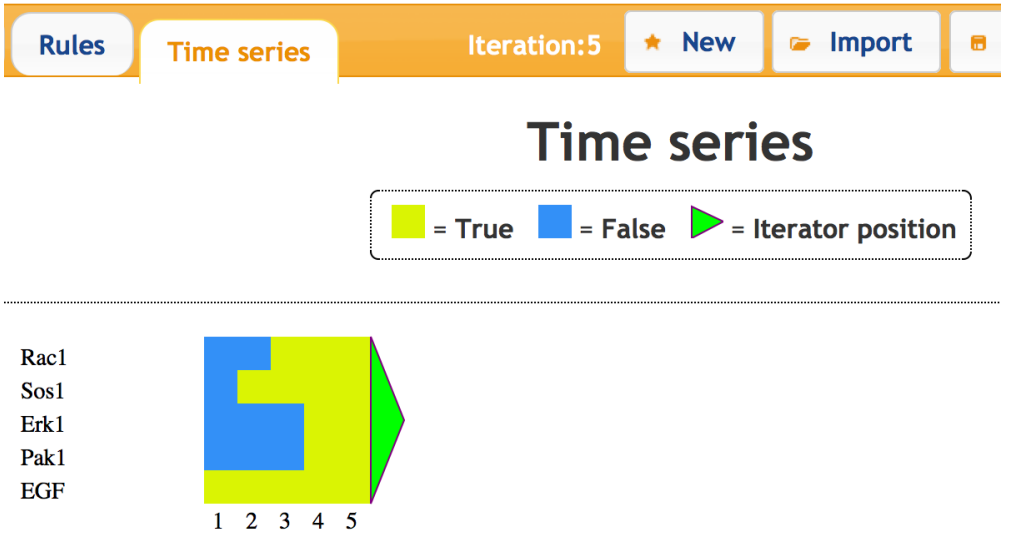
BooleSim


Erk1
EGF
Sos1
Rac1
= True
= False
Pak1


Erk1
EGF
Sos1
Rac1
= True
= False
Pak1

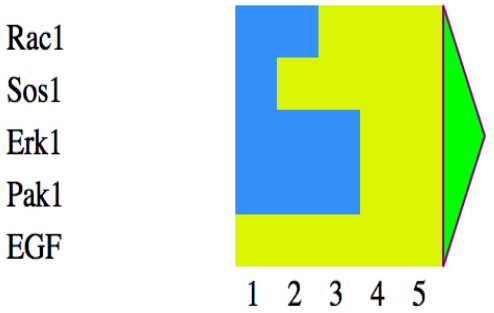

Erk1
EGF
Sos1
Rac1
= True
= False
Pak1


Erk1
EGF
Sos1
Rac1
= True
= False
Pak1

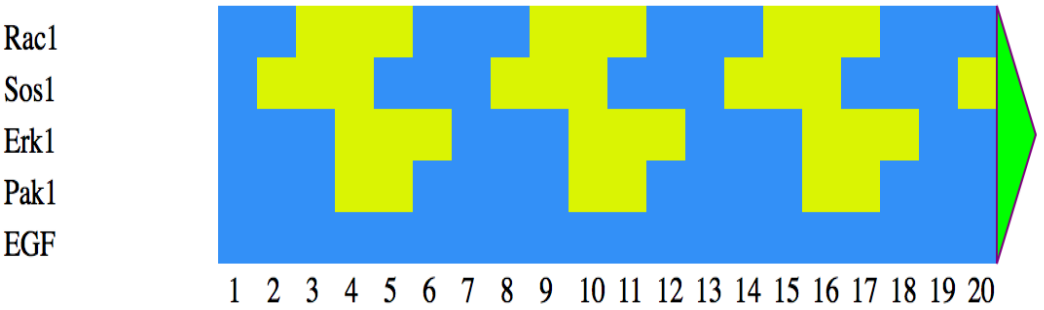

Rules
Time series
Iteration:5
New
Import
Time series
= True
= False
= Iterator position
Rac1
Sos1
Erk1
Pak1
EGF
1 2 3 4 5

BooleSim

- EGF is ON:
  the system reaches
  a steady state.



- EGF is OFF:
  the system reaches
  a cycle.

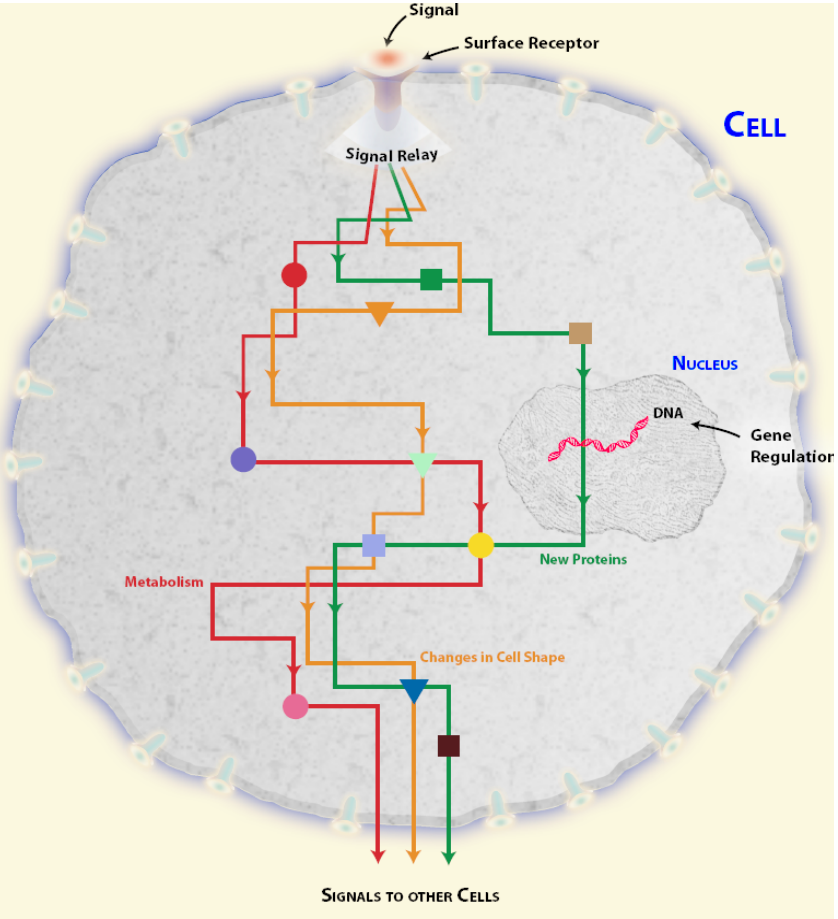


Challenges

- Manual curation of literature is long and cumbersome.
- Large systems are complex to represent and simulate.
- Can we generate models of diseases via large-scale text mining from literature?
- Can the integration of whole pathways as well as individual molecules add an extra level of understanding?

Text-mining biological pathways?

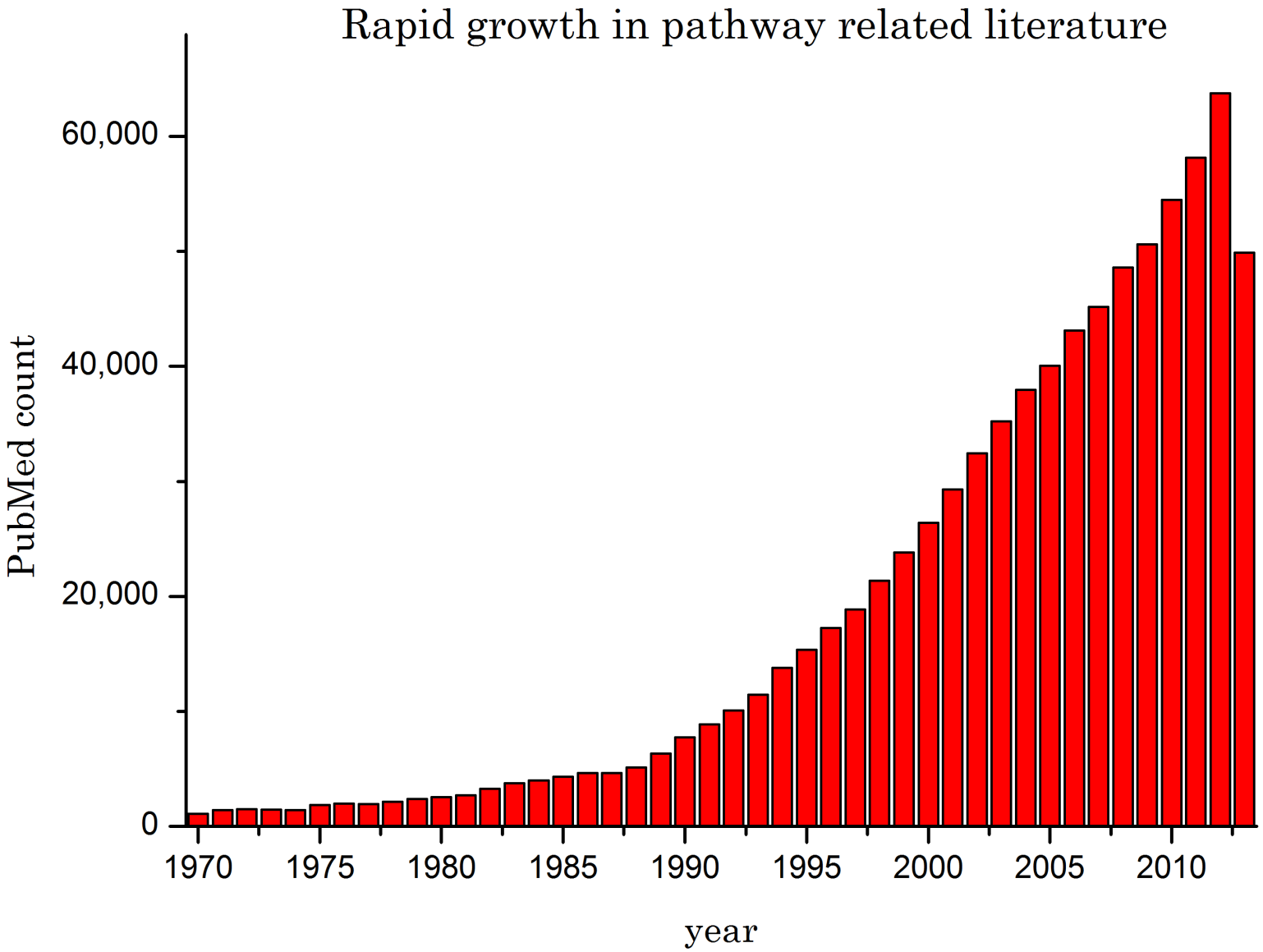
Recognising pathways names would help to prioritise curation.



- Set of interactions between molecules
- Leads to a certain product or change in the cell
- Involved in metabolism, gene regulation and signal transmission

Biological Pathways Fact Sheet. National Human Genome Research Institute

PubMed query: “pathway or signaling”



Pathway databases

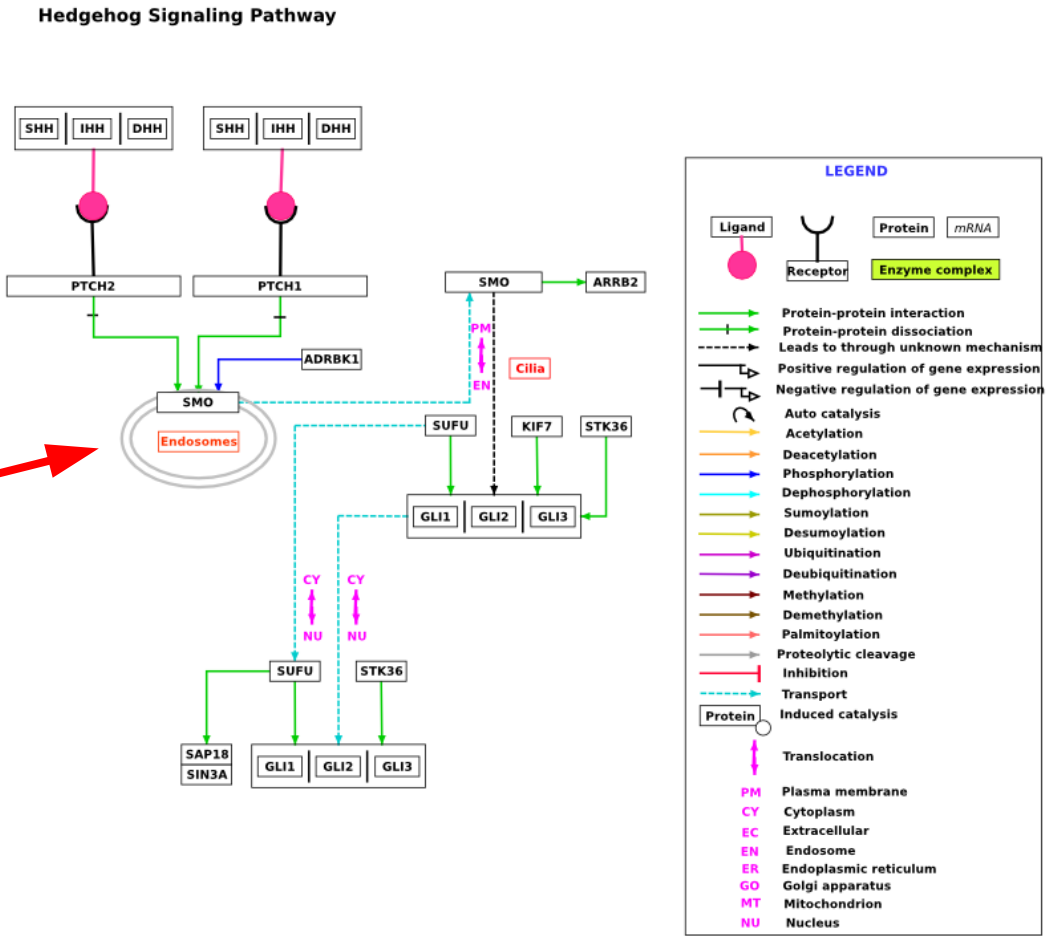
Curated databases

Integrative databases

And many more …

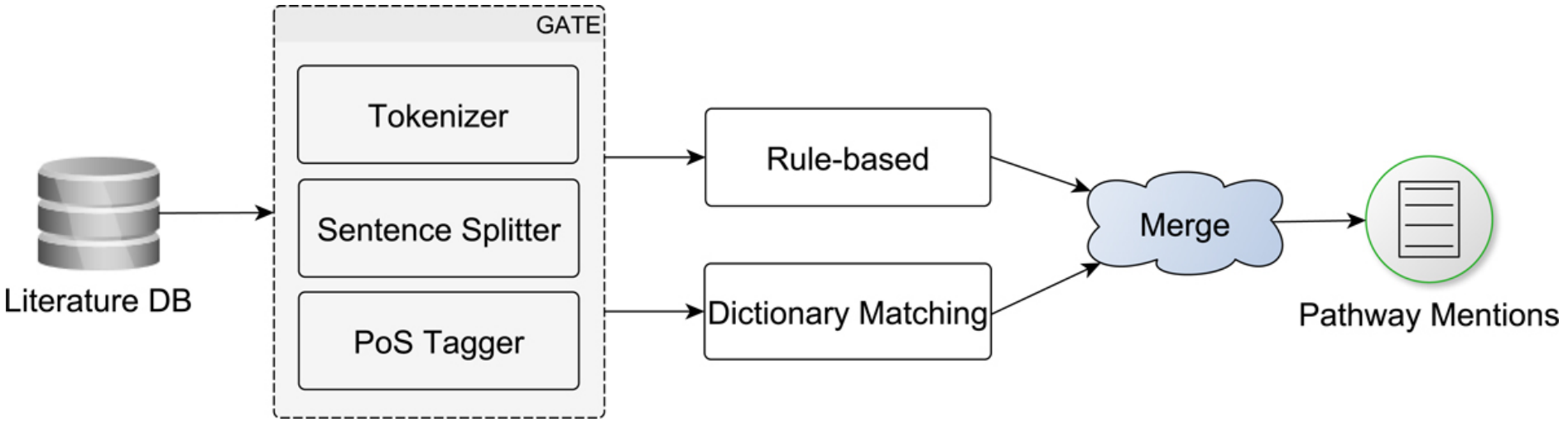
Pathway recognition

"HSulf-1 suppresses cell growth and down-regulates Hedgehog signaling in human gastric cancer cells."



Pathway Named Entity Recognition

- We developed a Named Entity Recognition (NER) tool for systematic identification of pathway mentions from literature: PathNER.



Wu *et al.* (2013) *BMC Syst Biol* **7**(Suppl 3): S2

Rule-based detection

- Backward rule
  - {GENE/PROTEIN} {KEYWORD}
    nf-kb signaling pathway
- Forward rule
  - {KEYWORD} [regulated/induced/...] [by/through..] {GENE/PROTEIN}
    Signaling Pathway Activated by IL-2

GENE/PROTEIN recognition are performed to filter out general mentions and non-biological mentions, for instance, “the diagnostic pathway”.

Soft dictionary matching

- Dictionary compiled from ***ConsensusPathDB*** and ***Pathway Ontology***

- To detect variants of pathway names
  - "calcium signaling", "calcium signa**ll**ing"
  - "p53 signaling pathway", "signaling pathway of p53"
  - "MAPK pathway", "MAPK signaling pathway"

Performance on gold corpus

Method	Strict			Lenient		
	Recall	Precision	F1-score	Recall	Precision	F1-score
Baseline	0.32	0.49	0.38	0.43	0.66	0.54
Soft dictionary	0.44	0.51	0.47	0.63	0.72	0.67
Rules	0.51	**0.86**	0.64	0.58	**0.97**	0.72
PathNER	**0.80**	0.74	**0.77**	**0.88**	0.81	**0.84**

- Gold corpus: manually annotated set of 726 pathway mentions.
- The baseline is exact dictionary matching implemented by LINNAEUS.

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = 2\frac{P \cdot R}{P + R}$$

Gerner et al. (2010) LINNAEUS: a species name identification system for biomedical literature. ***BMC Bioinformatics*** **11**: 85.

Application of PathNER

- AlzPathway: a map of signaling pathways for Alzheimer's disease
- AlzPathway is based on ~100 review articles
- Can we improve it using PathNER?

Ogishima et al. (2013) A map of Alzheimer's disease-signaling pathways: a hope for drug target discovery. *Clinical Pharmacology and Therapeutics* **93**: 399-401.

Mizuno et al. (2012) AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC Systems Biology* **6**: 52.

Table 7 - Top 25 detected mentions in the ALZ_ARF_PUBMED corpus

No.	Freq	Detected Mention	Evidence	In AlzPathway?
1	1869	Alzheimer's disease	N/A	N/A
2	1121	Disease	N/A	N/A
3	201	Parkinson's disease	PMID: 12672864	FALSE
4	143	Amyotrophic lateral sclerosis	PMID: 1571856	FALSE
5	123	Metabolism	N/A	N/A
6	120	Apoptosis	PMID: 11227497	TRUE
7	99	Oxidative stress	PMID: 10681270	TRUE
8	99	Transcription	N/A	N/A
9	98	Long-term potentiation	PMID: 12399581	TRUE
10	94	Gene expression	N/A	TRUE
11	67	Proteasome	PMID: 10854289	FALSE
12	59	Huntington's disease	PMID: 15686606	FALSE
13	56	Cell cycle	PMID: 15936057	TRUE
14	35	Methylation	PMID: 19606065	FALSE
15	35	Translation	N/A	N/A
16	33	Acetylation	PMID: 19625751	TRUE
17	27	Endocytosis	PMID: 16442855	FALSE
18	21	Notch signaling	PMID: 19853579	FALSE
19	18	Glucose metabolism	PMID: 21971455	FALSE
20	17	Obesity	PMID: 19801534	FALSE
21	16	Long-term depression	PMID: 21854392	FALSE
22	15	Signal transduction	N/A	N/A
23	14	Glycolysis	PMID: 14718371	FALSE
24	14	Prion diseases	PMID: 15190676	FALSE
25	13	Creutzfeldt-Jakob disease	PMID: 7904883	FALSE

*N/A: Not applicable; TRUE: The pathway is present in AlzPathway; FALSE: the pathway is not in AlzPathways

Table 8 - Top 25 detected mentions in the ALZ_ARF_PMC corpus

No.	Freq	Detected Mention	Evidence	In AlzPathway?*
1	635	Disease	N/A	N/A
2	174	Alzheimer's disease	N/A	N/A
3	130	Amyotrophic lateral sclerosis	PMID: 1571856	FALSE
4	95	Methylation	PMID: 19606065	FALSE
5	78	Long-Term Potentiation	PMID: 12399581	TRUE
6	69	Oxidative Stress	PMID: 10681270	TRUE
7	65	Transcription	N/A	N/A
8	48	Parkinson's Disease	PMID: 12672864	FALSE
9	46	Cell cycle	PMID: 15936057	TRUE
10	44	Metabolism	N/A	N/A
11	32	Axon guidance	PMID: 17571925	TRUE
12	31	Gene expression	N/A	TRUE
13	23	Glucose metabolism	PMID: 21971455	FALSE
14	20	Calcium signaling	PMID: 21184278	TRUE
15	18	Acetylation	PMID: 19625751	TRUE
16	16	Apoptosis	PMID: 11227497	TRUE
17	15	Activation of the Rac signaling pathway	PMID: 10817927	FALSE
18	14	Notch signaling	PMID: 19853579	FALSE
19	12	Prion diseases	PMID: 15190676	FALSE
20	12	Proteasome	PMID: 10854289	FALSE
21	12	S phase	PMID: 19946466	TRUE
22	12	Translation	N/A	N/A
23	10	Endocytosis	PMID: 16442855	FALSE
24	10	Insulin/IGF-1 Signaling	PMID: 22817723	FALSE
25	9	Post-translational modifications	PMID: 21215781	TRUE

*N/A: Not applicable; TRUE: The pathway is present in AlzPathway;
FALSE: the pathway is not in AlzPathways

Biochem Biophys Res Commun. 2009 Dec 25;390(4):1093-7. doi: 10.1016/j.bbrc.2009.10.093. Epub 2009 Oct 22.

Alzheimer's disease and Notch signaling.

Woo HN, Park JS, Gwon AR, Arumugam TV, Jo DG.

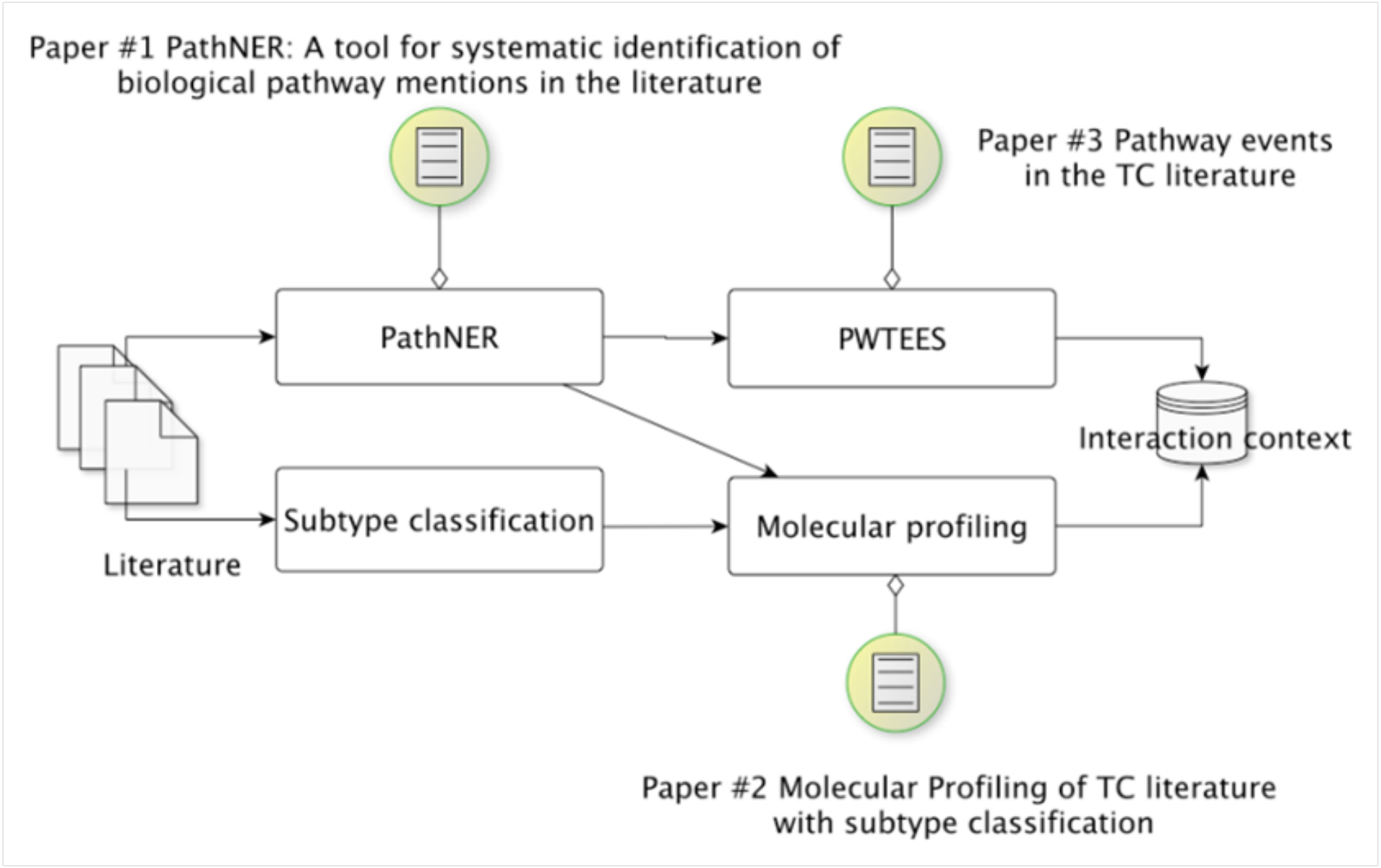
College of Pharmacy, Sungkyunkwan University, Suwon, Republic of Korea.

Abstract

Cleavage of the amyloid precursor protein (APP) by gamma-secretase generates a neurotoxic amyloid beta-peptide (Abeta) that is thought to be associated with the neurodegeneration observed in Alzheimer's disease (AD) patients. Presenilin is the catalytic member of the gamma-secretase proteolytic complex and mutations in presenilins are the major cause of early-onset familial Alzheimer's disease. In addition to APP, gamma-secretase substrates include Notch1 homologues, Notch ligands Delta and Jagged, and additional type I membrane proteins, raising concerns about mechanism-based toxicities that might arise as a consequence of inhibiting gamma-secretase. Notch signaling is involved in tumorigenesis as well as in determining the fates of neural and nonneural cells during development and in adults. Alterations in proteolysis of the Notch by gamma-secretase could be involved in the pathogenesis of AD. Inconsistently, several recent observations have indicated that enhanced Notch signaling and expression could be instrumental in neurodegeneration in AD. Therefore, detailed and precise study of Notch signaling in AD is important for elucidating diverse mechanisms of pathogenesis and potentially for treating and preventing Alzheimer's disease.

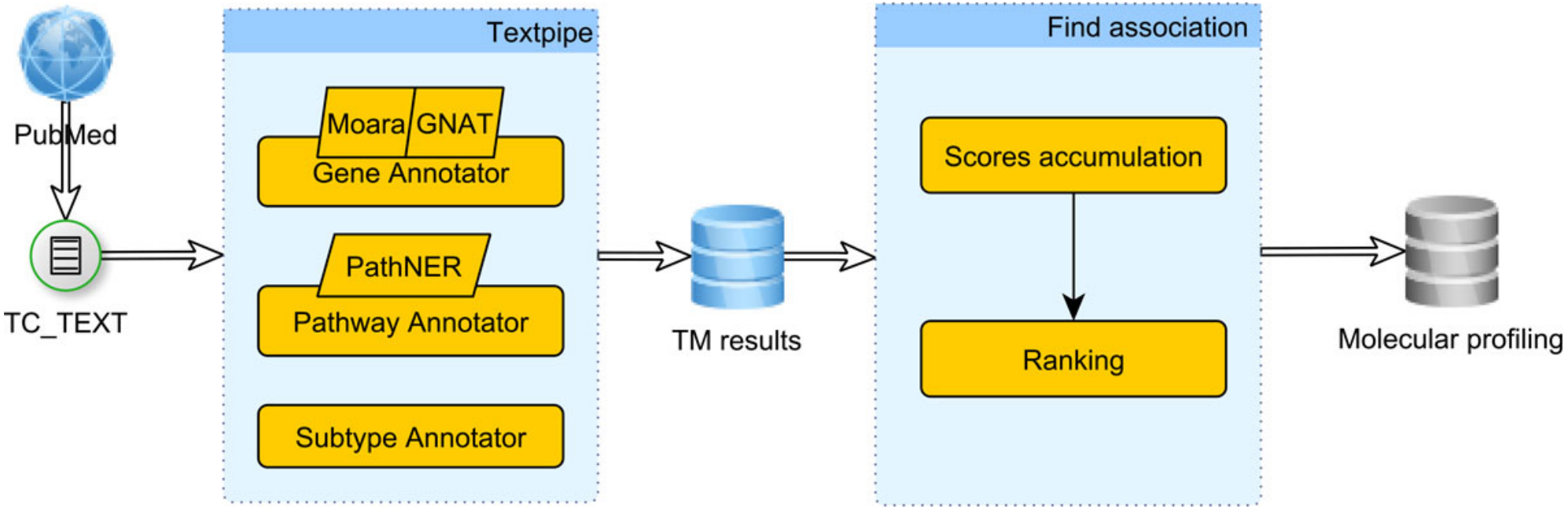
PMID: 19853579 [PubMed - indexed for MEDLINE]

Overview


Paper #1 PathNER: A tool for systematic identification of biological pathway mentions in the literature
Paper #3 Pathway events in the TC literature
PathNER
PWTEES
Interaction context
Subtype classification
Molecular profiling
Literature
Paper #2 Molecular Profiling of TC literature with subtype classification

Molecular profiling of TC subtypes

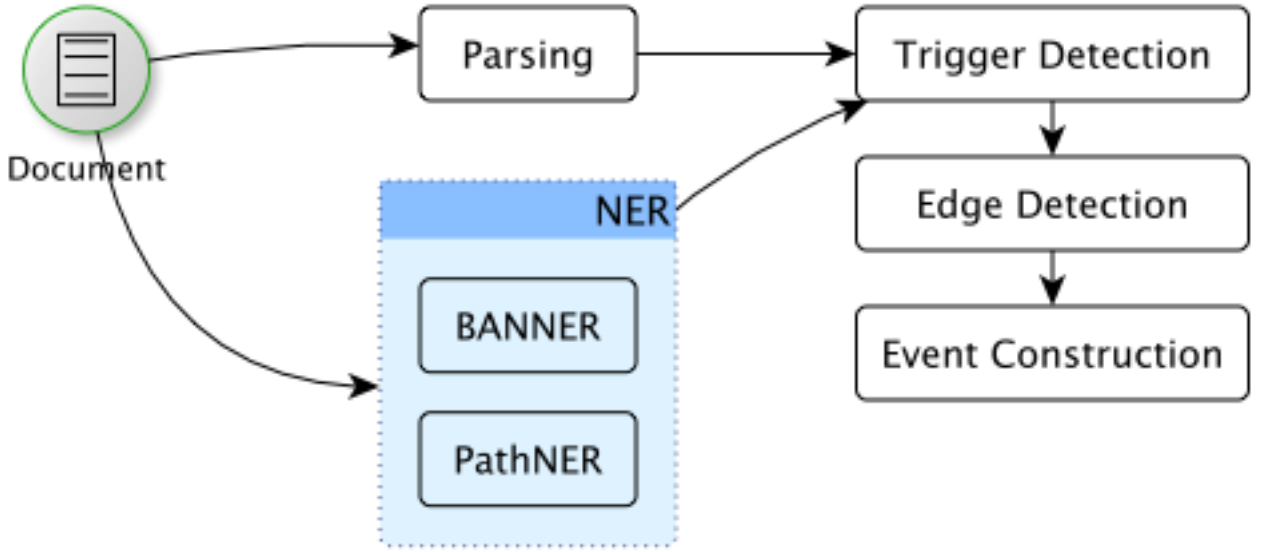
- A subtype classification method was developed for thyroid cancer to reconstruct specific molecular profiles.



Wu *et al.* (2014) *BMC Med Genomics* **7**(Suppl 3): S3

Network reconstruction of gene and pathway events

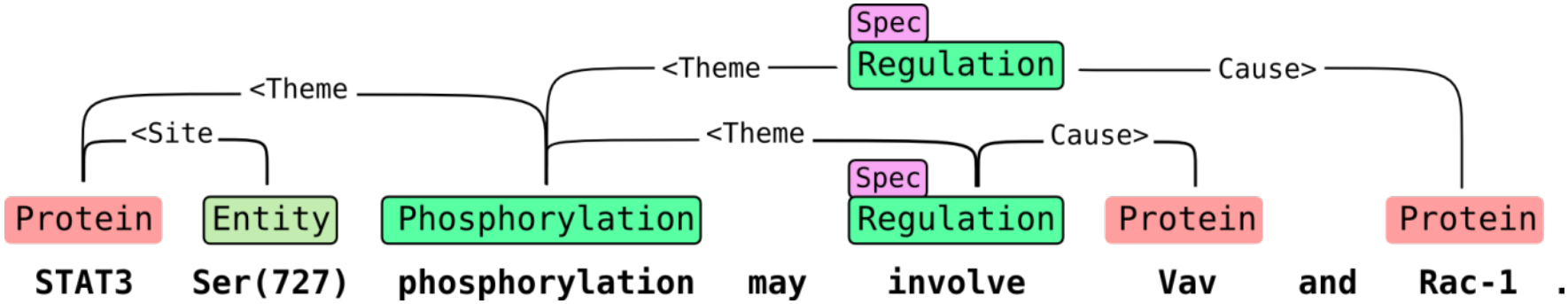
- We developed PWTEES, a system that can detect interactions between genes and pathways.
- BANNER is used for gene/protein name recognition, PathNER for pathway recognition.



Wu *et al.* (2015) *BMC Syst Biol*: in press

Event extraction

- We use TEES for event extraction: https://github.com/jbjorne/TEES/wiki/TEES-Overview
- Events include theme and cause:
  - A theme is the entity being regulated in the event.
  - A cause is the entity that regulates themes in the event.
- Event detection is triggered by keywords that give hints about an event's presence.



Björne *et al.* (2010) *Bioinformatics* **26**: i382-90

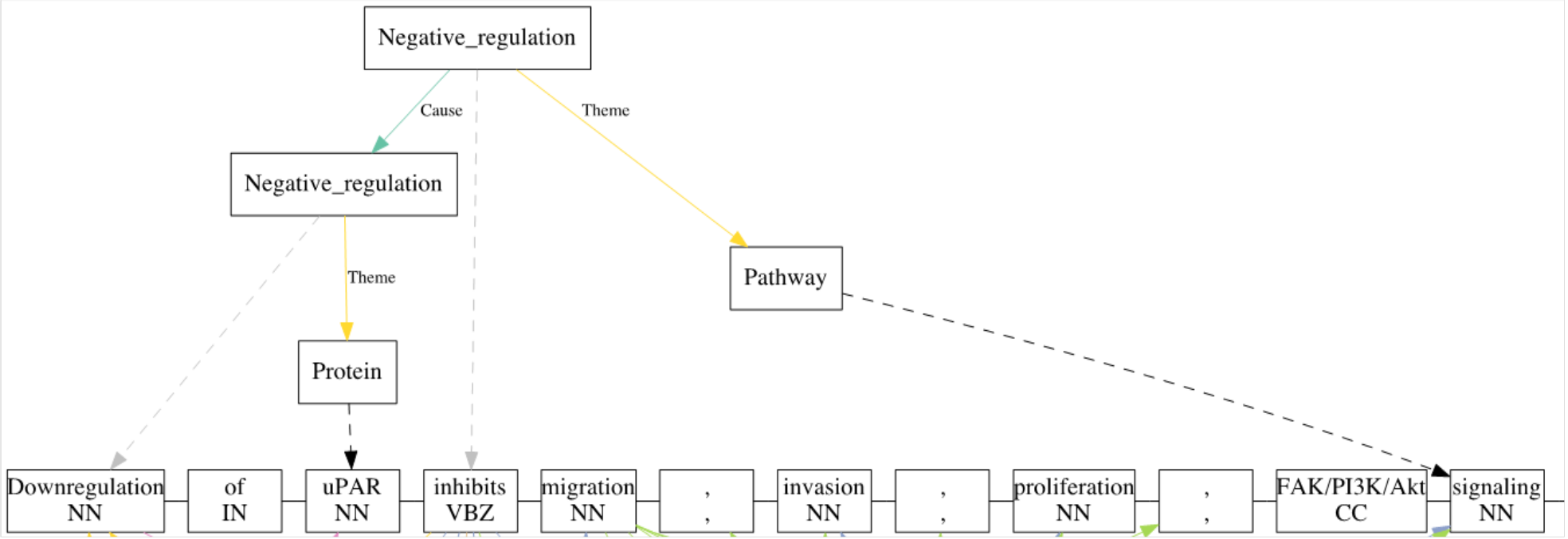
Event extraction

Type	Primary Args.	Second. Args.
Gene_expression	T(P)	
Transcription	T(P)	
Protein_catabolism	T(P)	
Phosphorylation	T(P)	Site
Localization	T(P)	AtLoc, ToLoc
Binding	T(P)+	Site+
Regulation	T(P/Ev), C(P/Ev)	Site, CSite
Positive_regulation	T(P/Ev), C(P/Ev)	Site, CSite
Negative_regulation	T(P/Ev), C(P/Ev)	Site, CSite

Event types and their arguments:
T = Theme, C = Cause, P = Protein, Ev = Event

Kim *et al.* (2009) *BioNLP '09*: 1-9

Interactions involving pathways



Example:
"Downregulation of uPAR inhibits migration, invasion, proliferation, FAK/PI3K/Akt signaling" (PMID: 21191179)

Pathway event detection

- Not all types of molecular events can involve pathways.
- We only consider regulation, positive regulation and negative regulation.
- We hypothesise that pathways appear in a similar context as genes/proteins when it comes to molecular events.
- We "disguise" pathway mentions recognised by PathNER as genes/proteins, with the same annotations as those produced by BANNER.

Nested events

Theme	**T_Theme**	*Gal-3*
	T_Cause	-
	T_Trigger	*Overexpression*
Cause	**C_Theme**	*HIPK2*
	C_Cause	-
	C_Trigger	*deficiency*
Trigger	*responsible*	

Example:
"HIPK2 deficiency might be responsible for such paradoxical Gal-3 overexpression in WDTC." (PMID: 21698151)

Large scale processing

- 38,572 abstracts.
- The whole PWTEES pipeline was implemented on the world's fastest supercomputer Tianhe-2 (Guangzhou).
- Time-consuming:
  - 125 hours on a single node.
  - 15 minutes with 500 parallel processes after optimisation.

Performance on pathway events

Dataset	TP	FP	TN	FN	P	PR
P_TEST	72	28	-	-	72%	-
PR_TEST	10	3	77	10	-	50%

P_TEST: evaluation against 100 reported pathway events for precision.

PR_TEST: evaluation against 100 <pathway, gene/protein> pairs for recall.

Comparison with EVEX

Example sentence	PMID	EVEX	PWTEES
Mutated BRAF, generates a constitutive activation of the mitogen-activated protein kinases (MAPK) signaling pathway	22863493	N/A	**T**: *Activation of MAPK pathway* **C**: *BRAF* **ET**: *Positive regulation*
PLD synergistically functions to activate the STAT3 signaling by interacting directly with the thyroid oncogenic kinase RET/PTC.	18498667	**T1**: *PLD* **T2**: *RET/PTC* **ET**: *Binding*	**T**: *PLD* **C**: *STAT3 signaling* **ET**: *Positive regulation*
CD40 stimulation inhibits cell growth and Fas-mediated apoptosis in a thyroid cancer cell line.	10223618	**Gene/Protein**: *CD40, Fas*	**T**: *Fas-medicated apoptosis* **C**: CD40 stimulation **ET**: *Negative regulation*
.. and that integration of the Ras/ERK1/2/ELK-1 and STAT3 pathways was required for up-regulation of the c-fos promoter by FMTC-RET	17209045	**T**: *c-fos promoter* **C**: *FMTC-RET* **ET**: *Positive regulation*	**T**: *up-regulation of the c-fos promoter by FMTC-RET* **C**: *Ras/ERK1/2/ELK-1 and STAT3 pathways* **ET**: *Positive regulation*

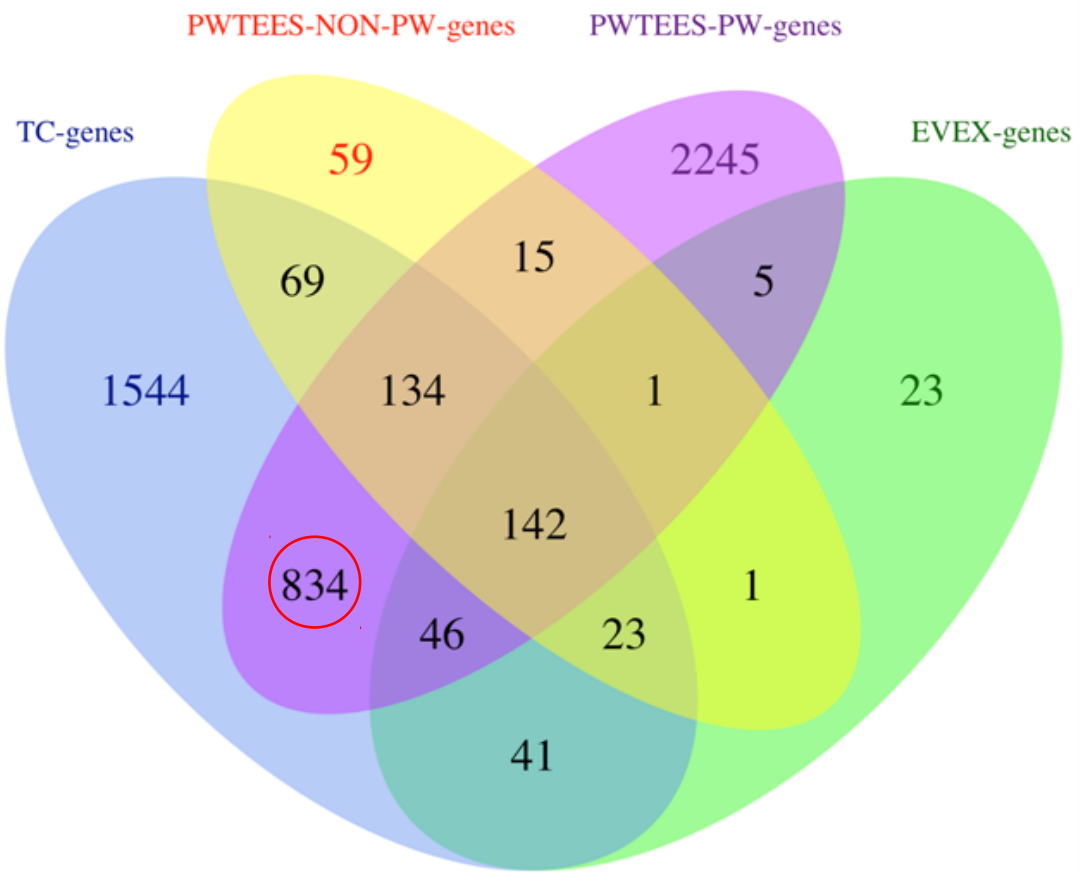
- EVEX is an event database created by applying TEES on 21.9 million PubMed abstracts and 460,000 PMC full-text articles.
- Introducing pathway entities significantly enhances event extraction.

Application on thyroid cancer

Type	Amount	Form
Genes/Proteins interactions	519	<Cause, Theme>
Binding pairs	145	<Theme1, Theme2>
Pathway interactions	313	<Cause, Theme>
EVEX-Human-TC-REL	599	<Source, Target>

Unique interactions detected on the whole thyroid cancer corpus.

Application on thyroid cancer

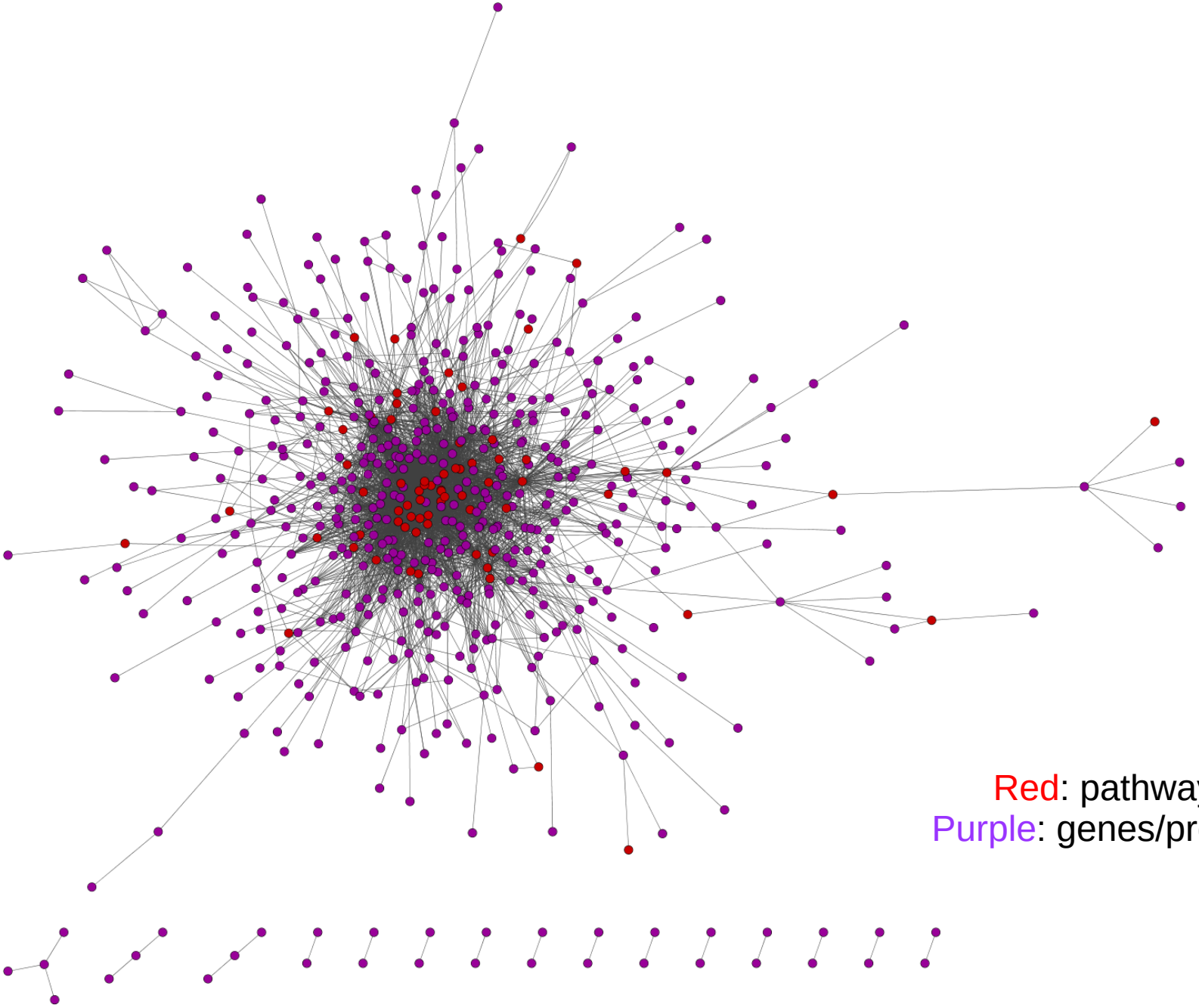


TC: 2833 thyroid cancer related genes.
PWTEES-NON-PW: genes involved in non-pathway interactions
PWTEES-PW: genes involved in pathway interactions
EVEX: genes involved in EVEX interactions

Network construction

- Nodes are genes/proteins and pathways.
- Edges:
  - All interactions detected by PWTEES are added to the edge set.
  - If a gene G is contained in a pathway P, then a new edge <G, P> is added.
- MERGE-PW network:
  - Takes both genes/proteins and pathways as nodes.
  - Contains 576 nodes and 3136 edges.
- NON-PW network
  - Based on gene/protein interactions.
  - Contains 444 nodes and 628 edges.

MERGE-PW network



Top 10 hubs in MERGE-PW network

No.	Hub (pathway name)	Degree
1	JAK/STAT3 pathway	276
2	MAPK/ERK pathway	245
3	TSHR signaling	159
4	PI3K/Akt pathway	141
5	Apoptosis	111
6	TSHR-induced G(q) signal transduction	92
7	EGFR signaling	83
8	TGFbeta transduction	81
9	epidermal growth factor receptor 1 signaling	78
10	T3/TR signaling	70

Top 10 gene/protein nodes

No.	NON-PW			MERGE-PW genes		
	Hub (Gene ID)	Symbol	Degree	Hub (Gene ID)	Symbol	Degree
1	5979	*RET*	58	207	*AKT1*	72
2	7157	*TP53*	24	5979	*RET*	63
3	207	*AKT1*	23	5594	*MAPK1*	57
4	7422	*VEGFA*	21	3265	*HRAS*	49
5	1950	*EGF*	19	4609	*MYC*	46
6	595	*CCND1*	18	5595	*MAPK3*	46
7	5594	*MAPK1*	15	6774	*STAT3*	45
8	673	*BRAF*	14	7157	*TP53*	45
9	6774	*STAT3*	14	1950	*EGF*	43
10	5727	*PTCH1*	14	595	*CCND1*	42

Top 10 bottlenecks in MERGE-PW network

Rank	Name	Betweeness Centrality
1	Apoptosis	110.0
2	5296 (*PIK3R2*)	45.0
3	5979 (*RET*)	44.0
4	Epidermal growth factor receptor 1 signaling	31.0
5	7157 (*TP53*)	27.0
6	207 (*AKT1*)	20.0
7	Cell cycle	20.0
8	1950 (*EGF*)	18.0
9	7124 (*TNF-alpha*)	18.0
10	5594 (*MAPK1*)	17.0

Conclusions (1)

- Text mining can assist in network reconstruction and enhance the coverage of interactions involving both genes/proteins and pathways.

- We demonstrated that integrating curated information about pathways provides an extra layer of understanding, highlighting key genes and functions.

- Availability: source code files are available at https://github.com/chengkun-wu/PWTEES

Pleiotropy

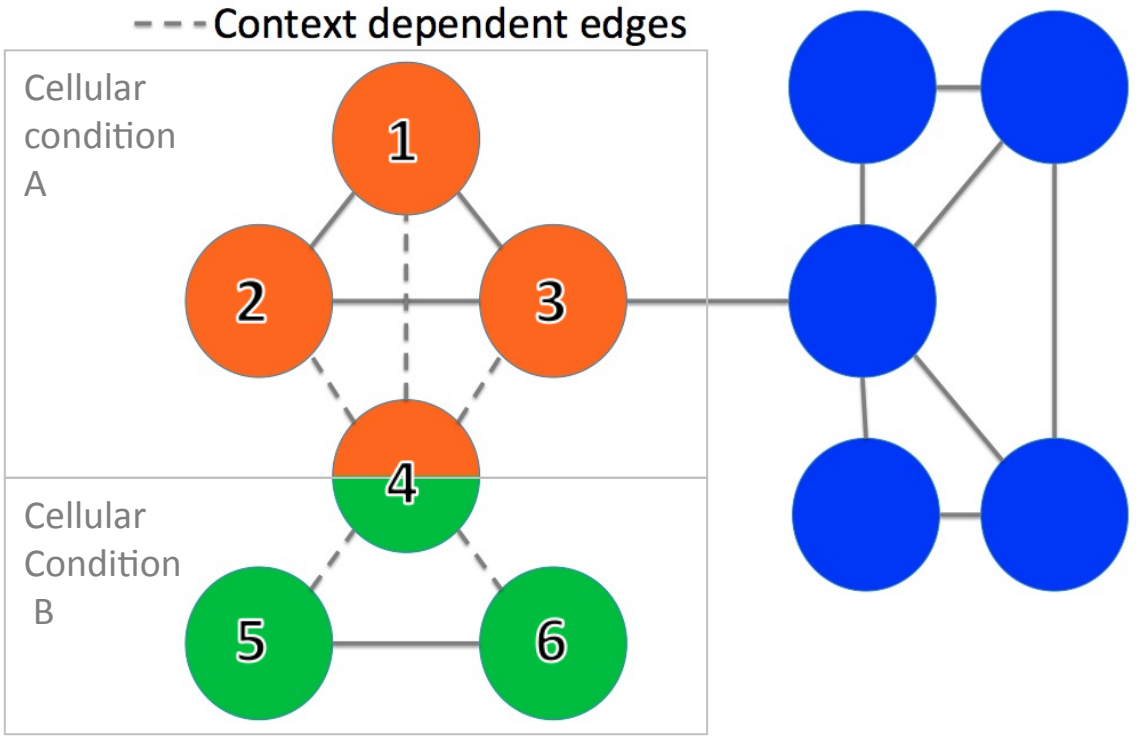
- Pleiotropy is the ability of genes to affect multiple seemingly unrelated traits.

- Genes in multifunctional pathways.

- Genes in multiple pathways with different functions.

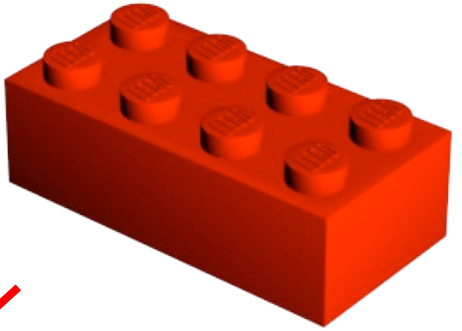
Pleiotropy

- Simply because two genes can interact does not mean they will interact in every context.



Gene 4 interacts with genes 1, 2 & 3 in condition A

And with genes 5 & 6 in condition B



Pathways

- Modules are sometimes used to define functional units in networks, but they are based on network topology.
- Biological pathways are experimentally validated sets of interacting proteins, contributing to a particular cellular function.
- Assign function to pathways.
- Pathways provide cellular context for genes.

PPI vs. pathway networks

PPI networks are good for examining the sum of each genes functions – i.e. effect of knockouts.

Pathway networks are good for examining the relationship / flow of data between functions.

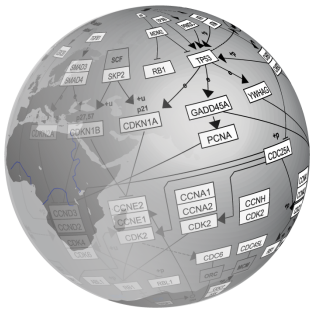
Aims

- Develop a method to map function, incorporating context dependent cellular function.
- Present a functional network of *Saccharomyces cerevisiae* pathways.
- Examine the organisation and physical implementation of function within the yeast cell.
- Network validation.

Pathways

- Pathway data: ConsensusPathDB
  - Pathway names and gene sets.
  - Collects data from major databases.

- Remove duplicated pathway data.



WIKIPATHWAYS
Pathways for the People

CPDB

KEGG
Kyoto Encyclopedia of Genes and Genomes

NATIONAL CANCER INSTITUTE
nature PathwayInteractionDatabase

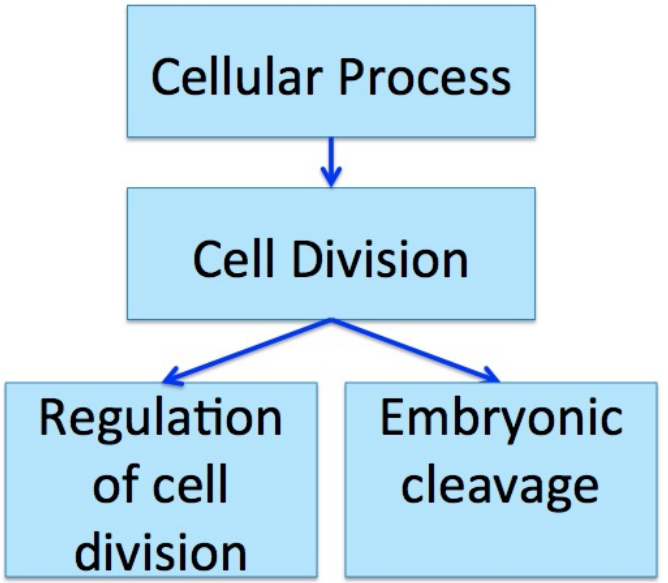
REACTOME

http://cpdb.molgen.mpg.de/

Gene Ontology annotations

- The Gene Ontology provides hierarchical functional annotations for genes.

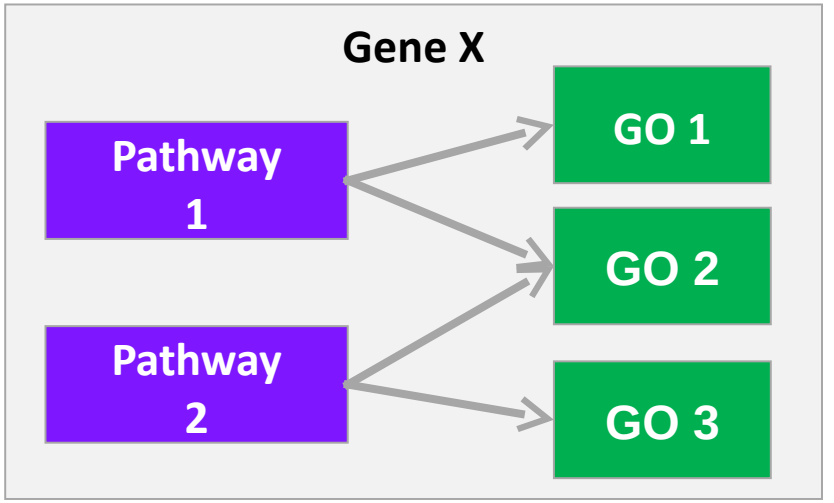
- Used experimentally validated experimental process annotations.

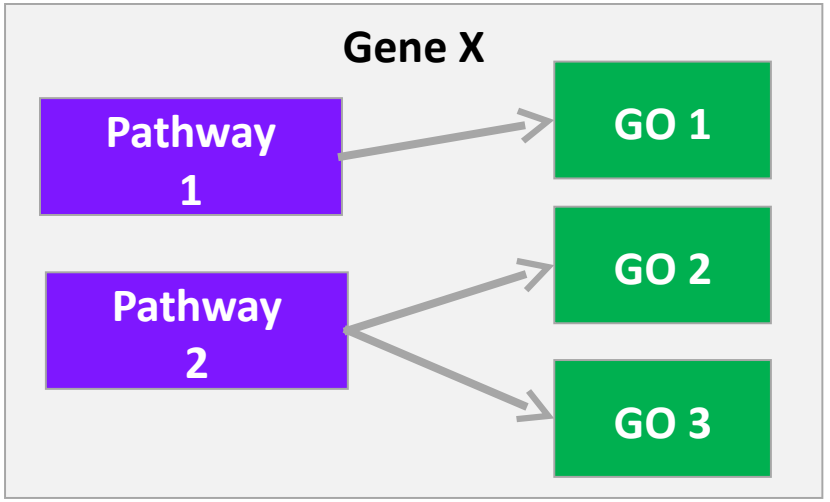


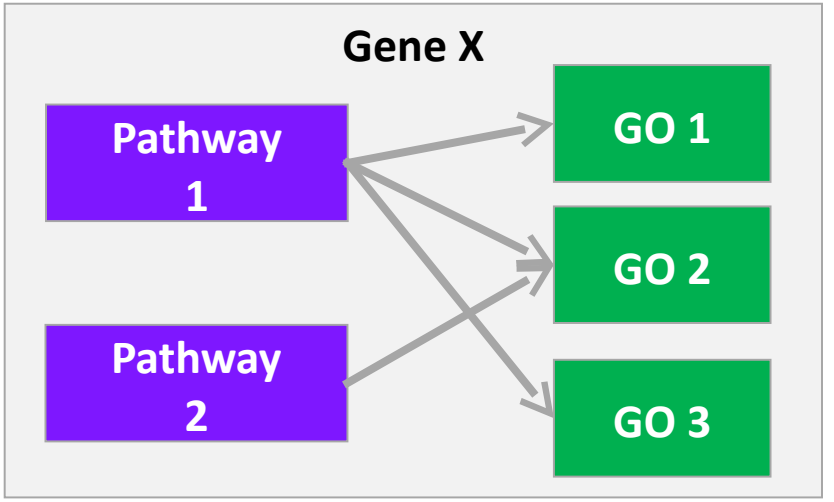
Mapping multiple functions onto multiple pathways

Suppose Gene X has 3 functions and is found in 2 pathways.

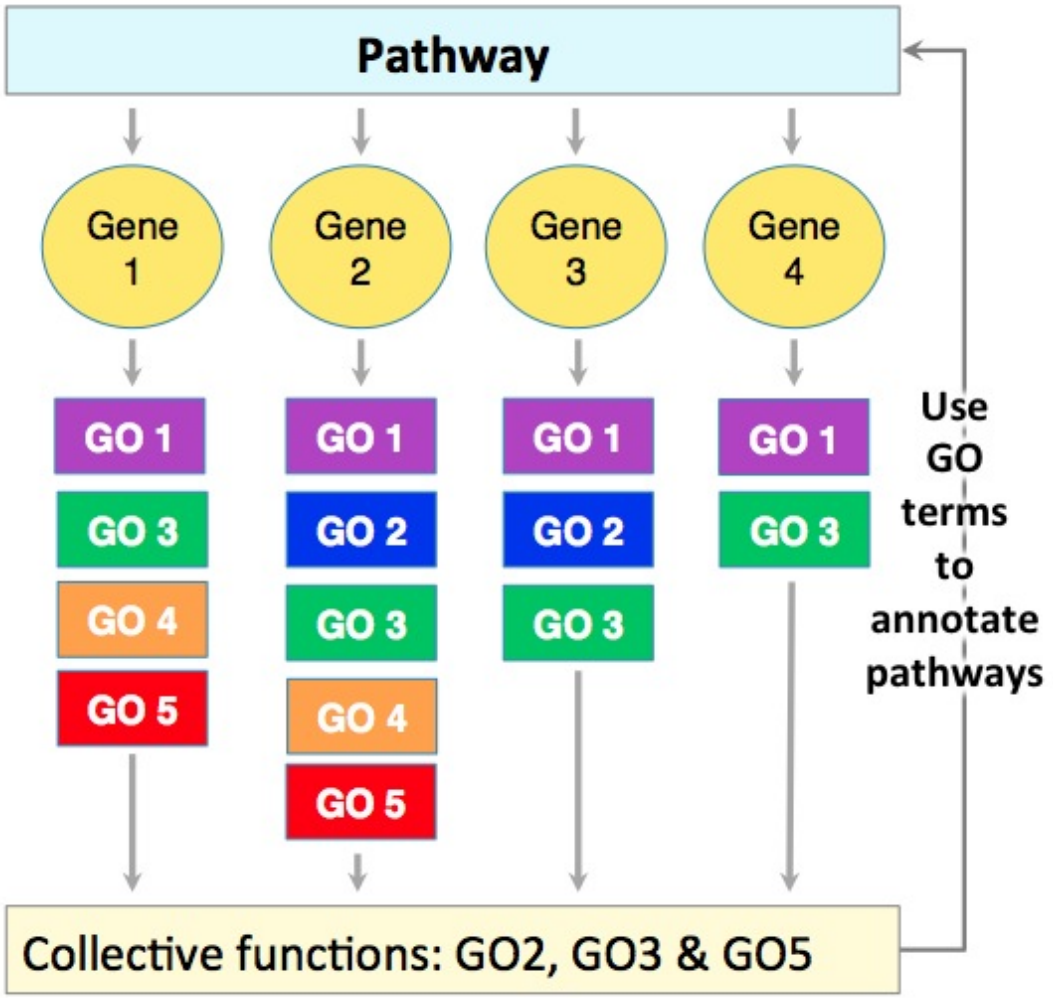
Which functions correspond to its role in which pathway?







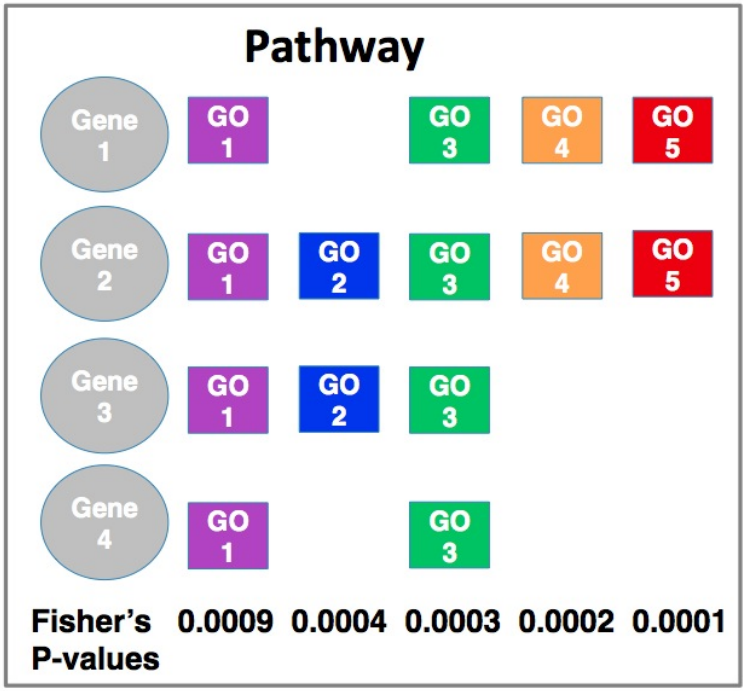
Functionally annotating pathways



1. Create **enrichment profiles** of over-represented GO terms
2. Create **functional profiles** of the most over represented terms
3. Add pleiotropic functions

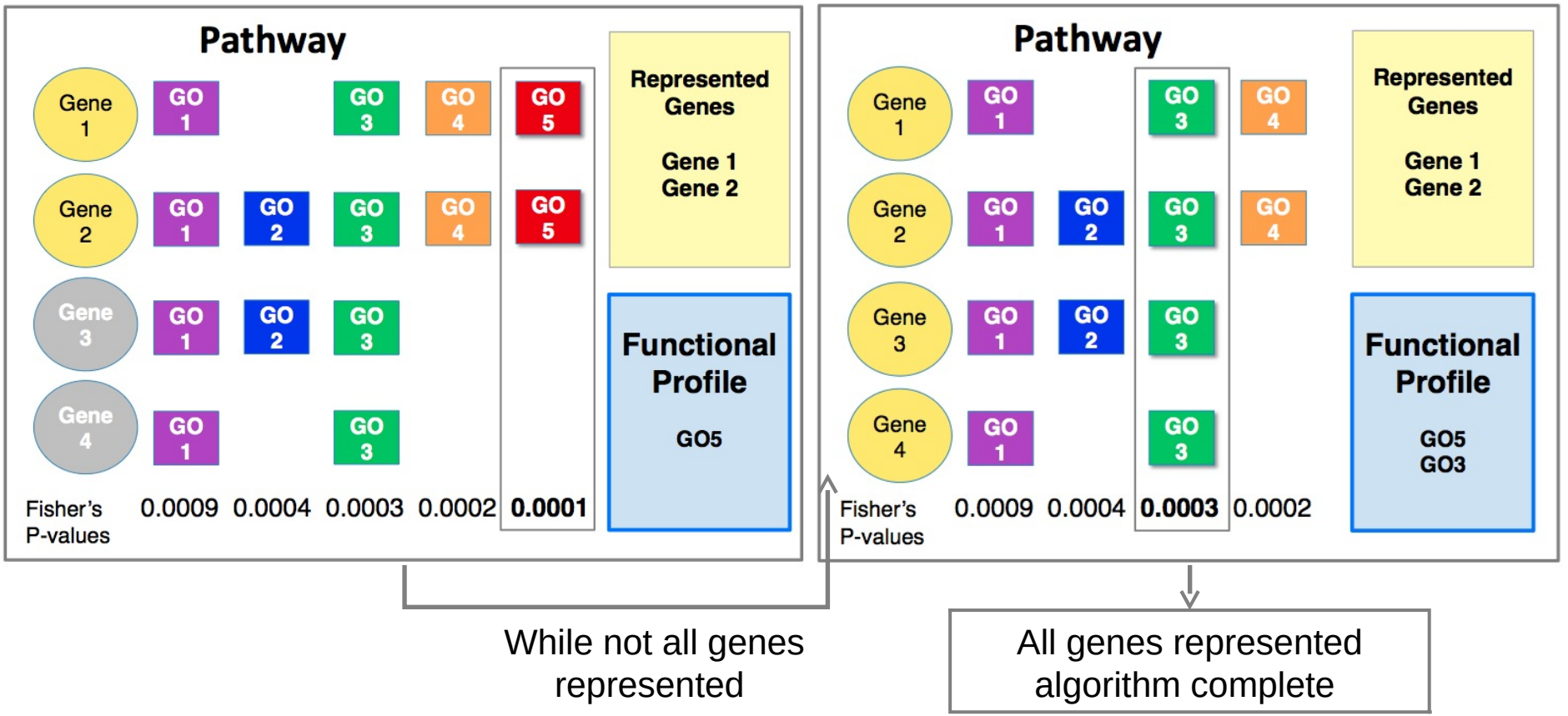
Enrichment profiles

- Select enriched genes Fisher's exact test
  - $p<0.01$ for high sensitivity
- Pathways had a median of 26 GO terms
  - Hierarchical nature of GO
  - Multifunctional pathways

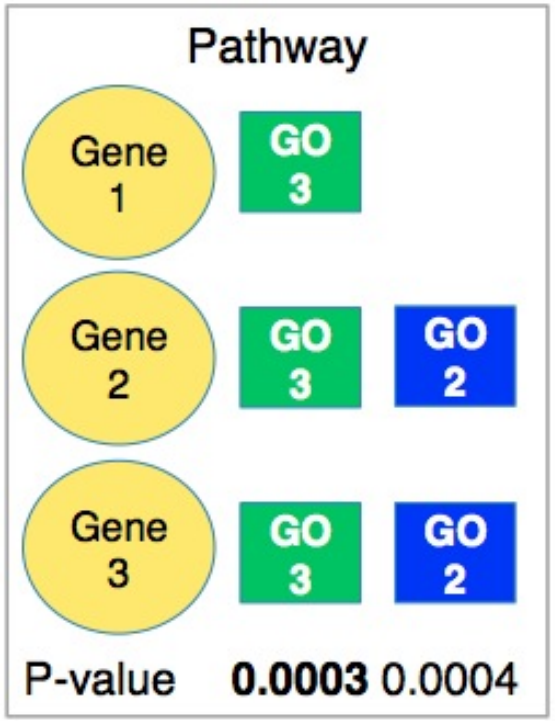


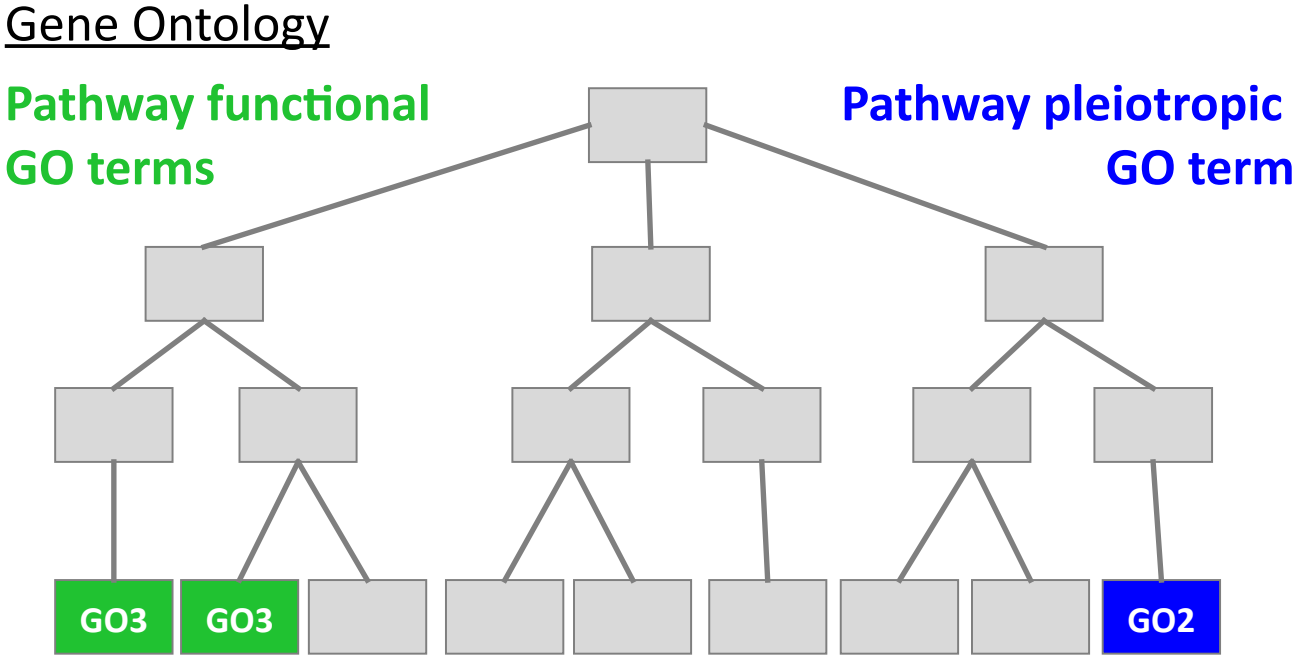
Functional profiles

Finding the most enriched GO terms that are sufficient to represent the genes within the pathway.



Multifunctional pathways

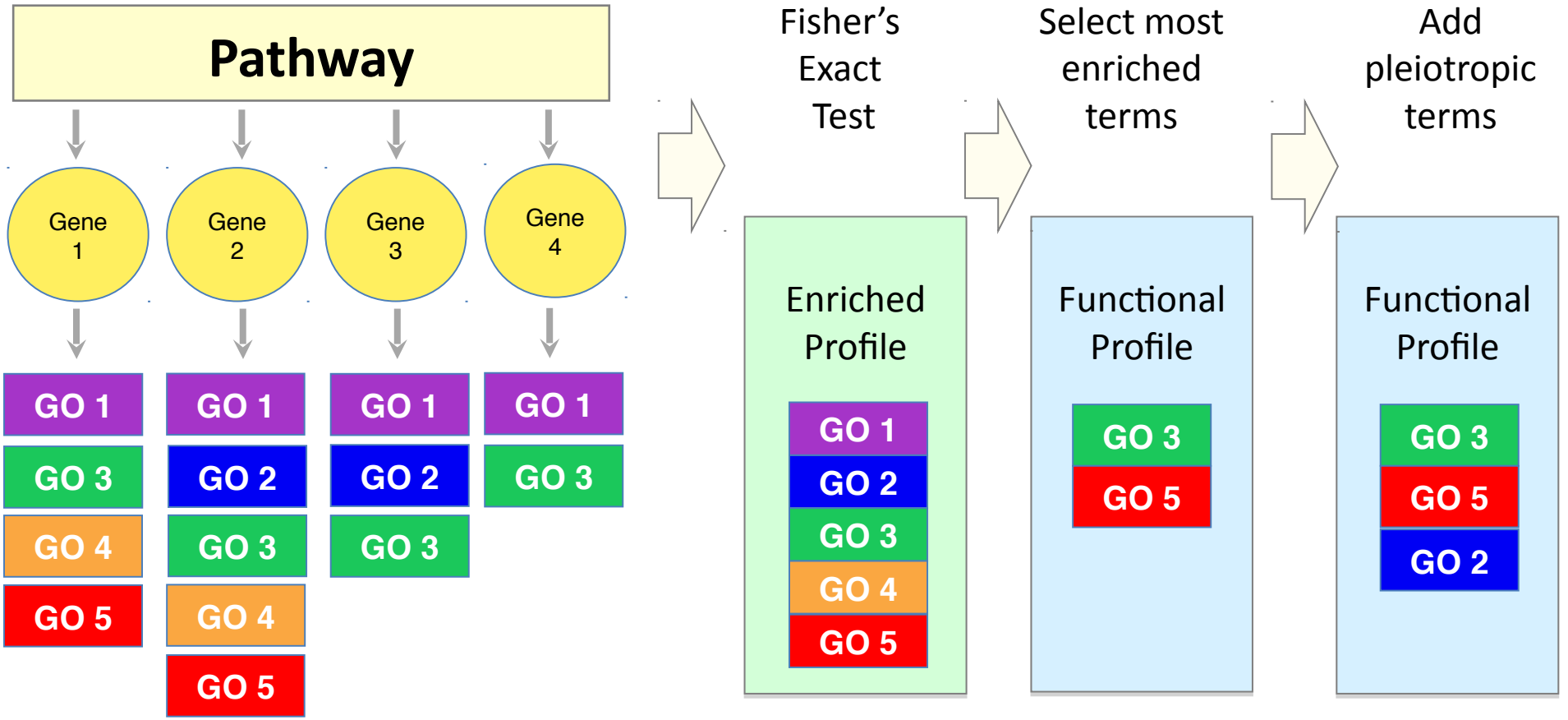




- Pathways can participate in >1 function.

- Pathways with enriched GO terms highly semantically different to their functional annotations had pleiotropic terms added.

Functionally annotated pathways



Network generation

- Link functionally connected pathways by shared GO terms.

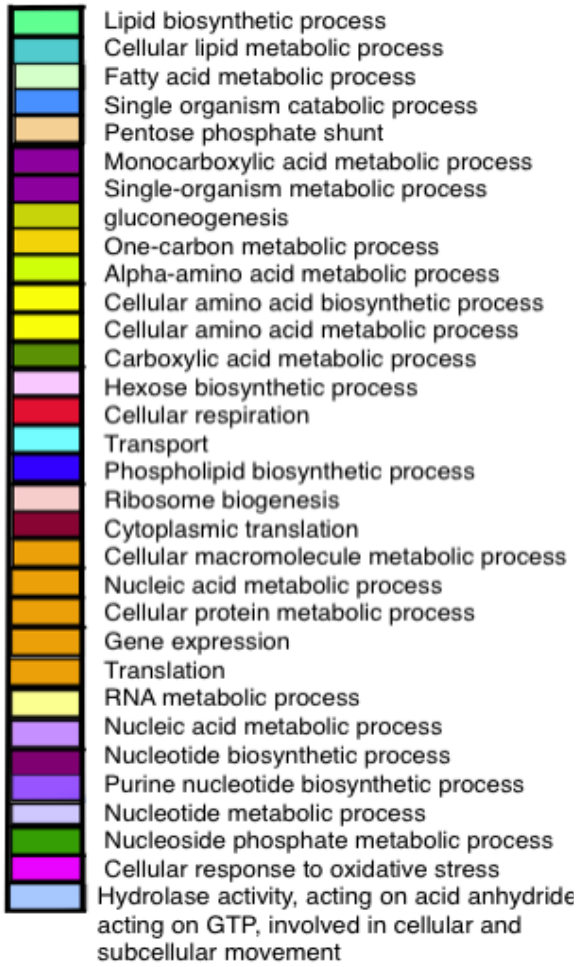
Jaccard coefficient:

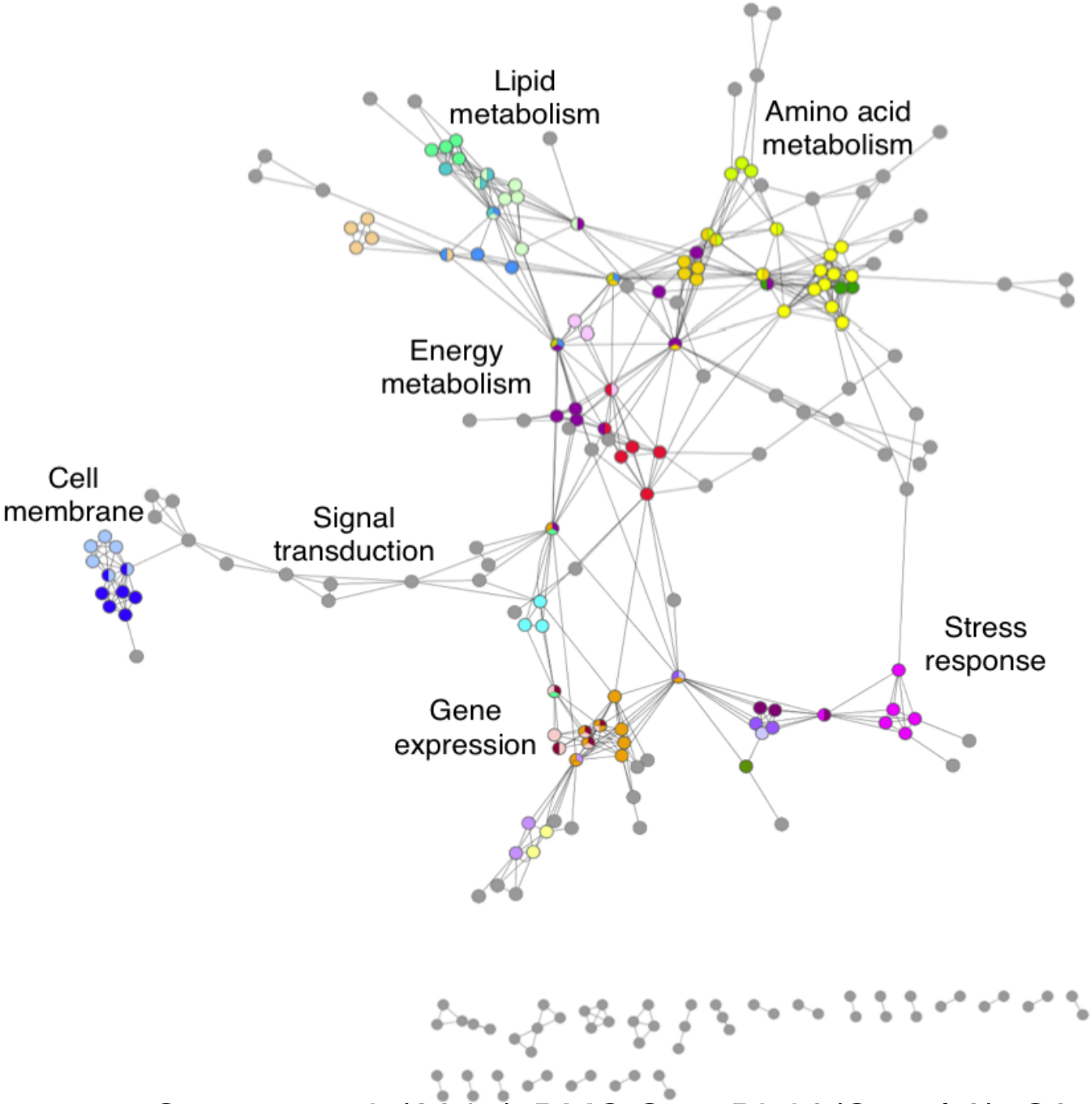
$$\text{Jaccard (A, B)} = \frac{\text{Number of shared GO terms}}{\text{Sum of both pathways' GO terms}}$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Scores between 0 and 1 weight edges.

Functionally linked pathways

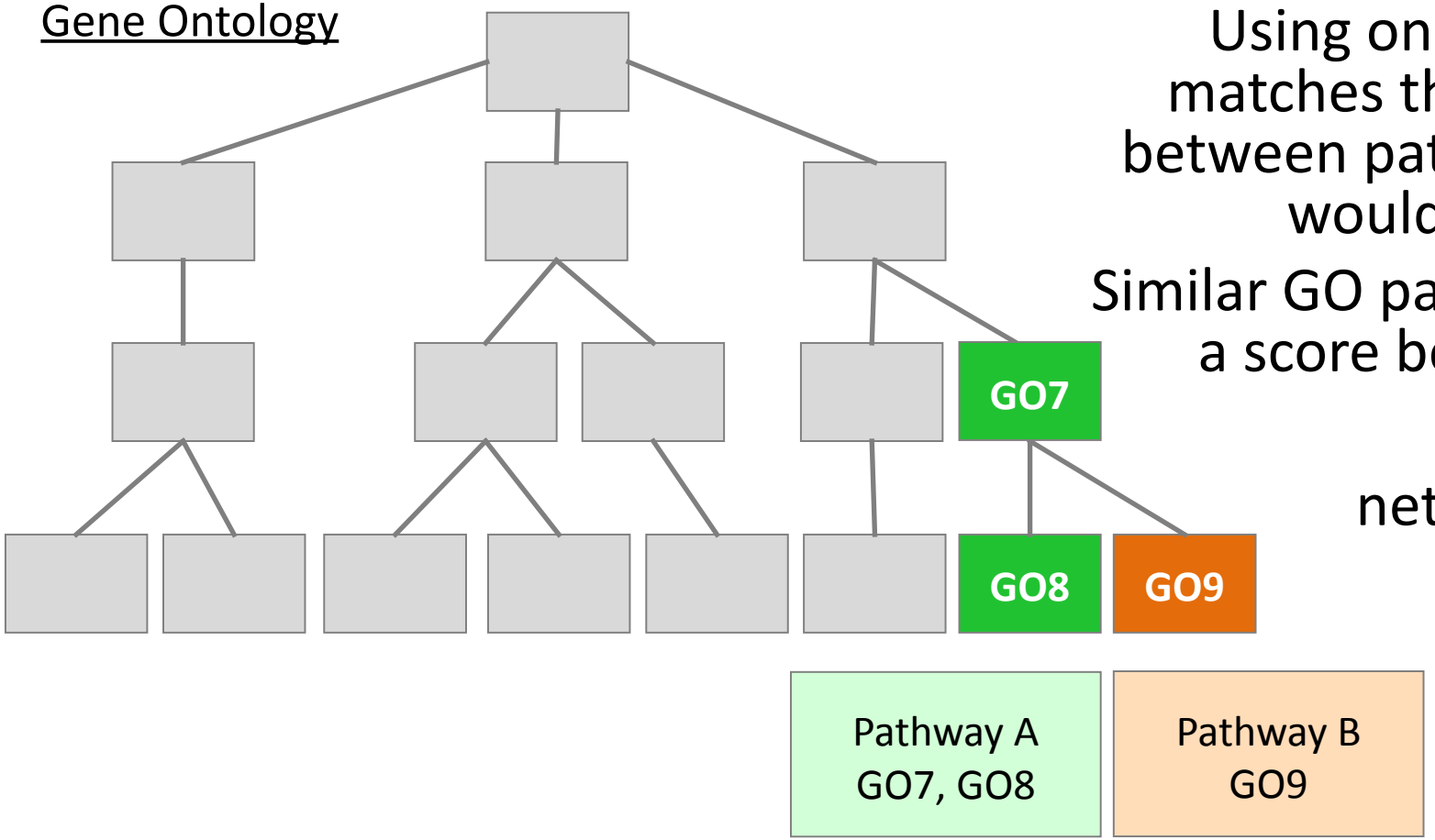




Stoney *et al.* (2015) *BMC Syst Biol* **9**(Suppl 6): S3

Network generation

Link functionally similar GO terms



Using only direct GO matches the similarity between pathways A&B would be missed.

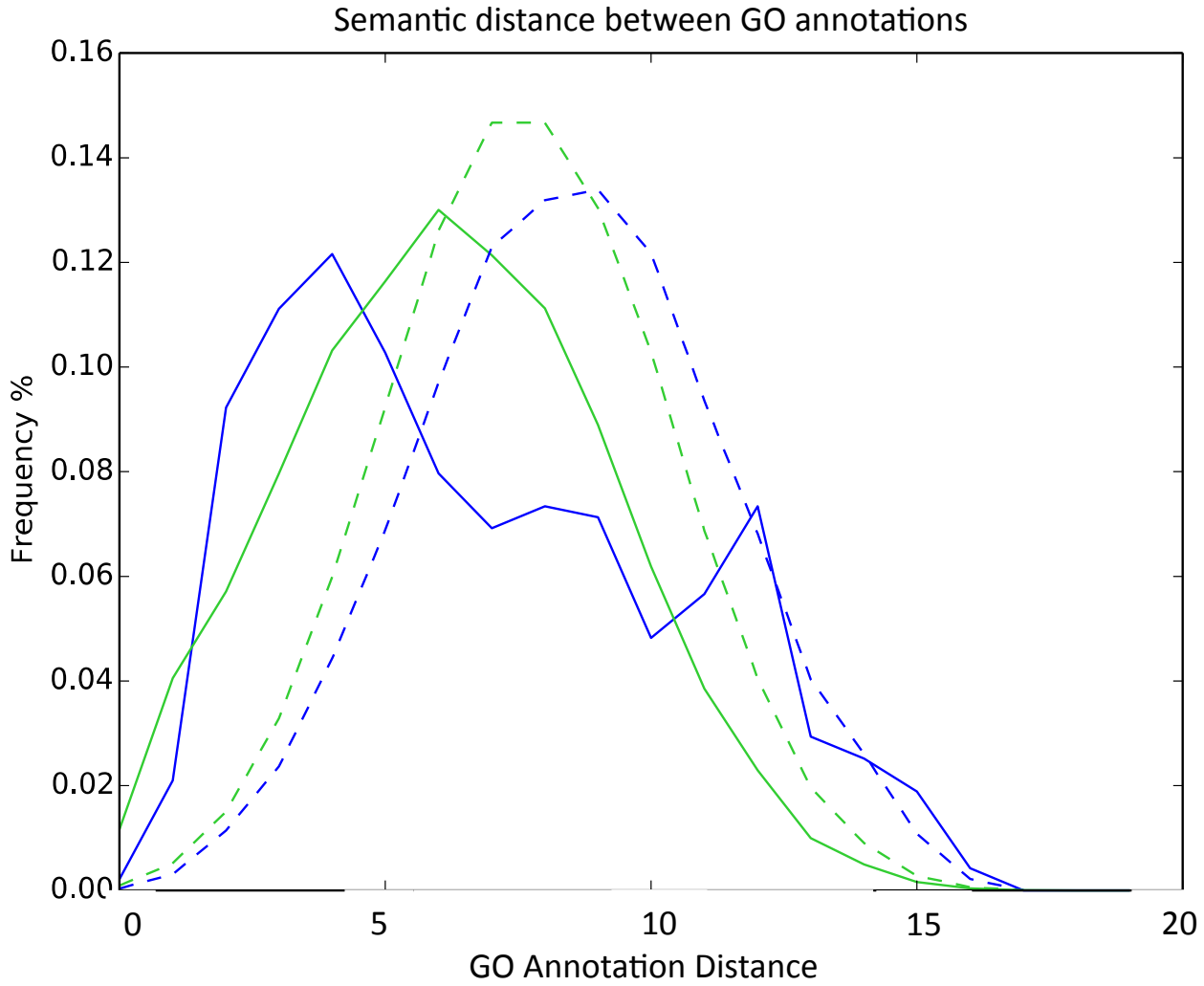
Similar GO pairs received a score between 0-1,

added to

network edges

Semantic similarity within profiles

Functional Profiles have a median of 2 GO terms



Blue line

Semantic distance within functional profile

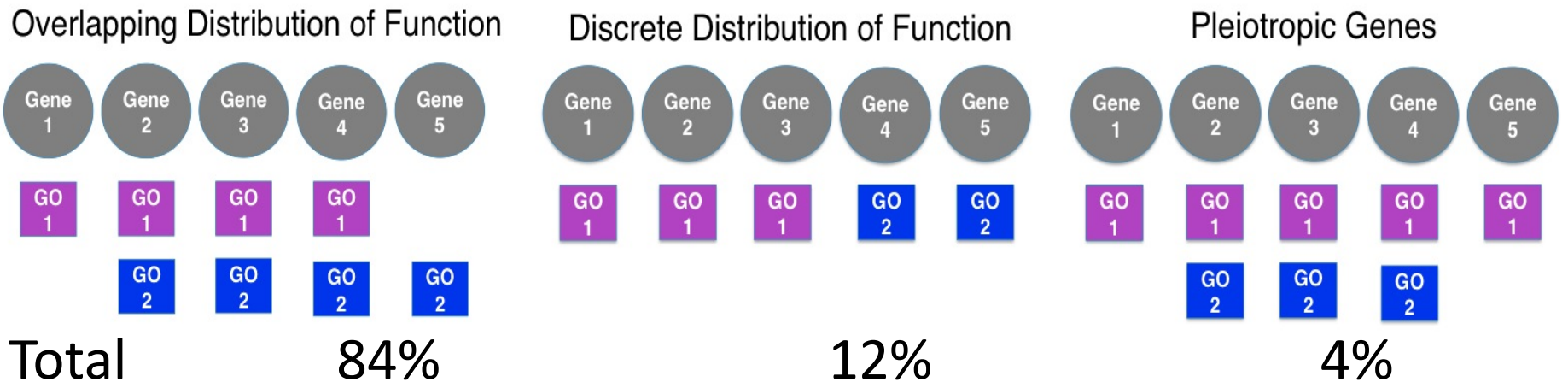
Green line

Semantic distance within enriched profile

Dashed lines

Chance semantic distributions

Functional variance within pathways



- Overlapping functional distributions show the flow of information between functions
- Discrete functional distributions have a median semantic distance of 10.0
- Discrete distributions may form ‘functional bridges’
- Pleiotropic distributions show genes participating in multiple functions

Pathway multi-functionality

Green squares: discrete functional annotations

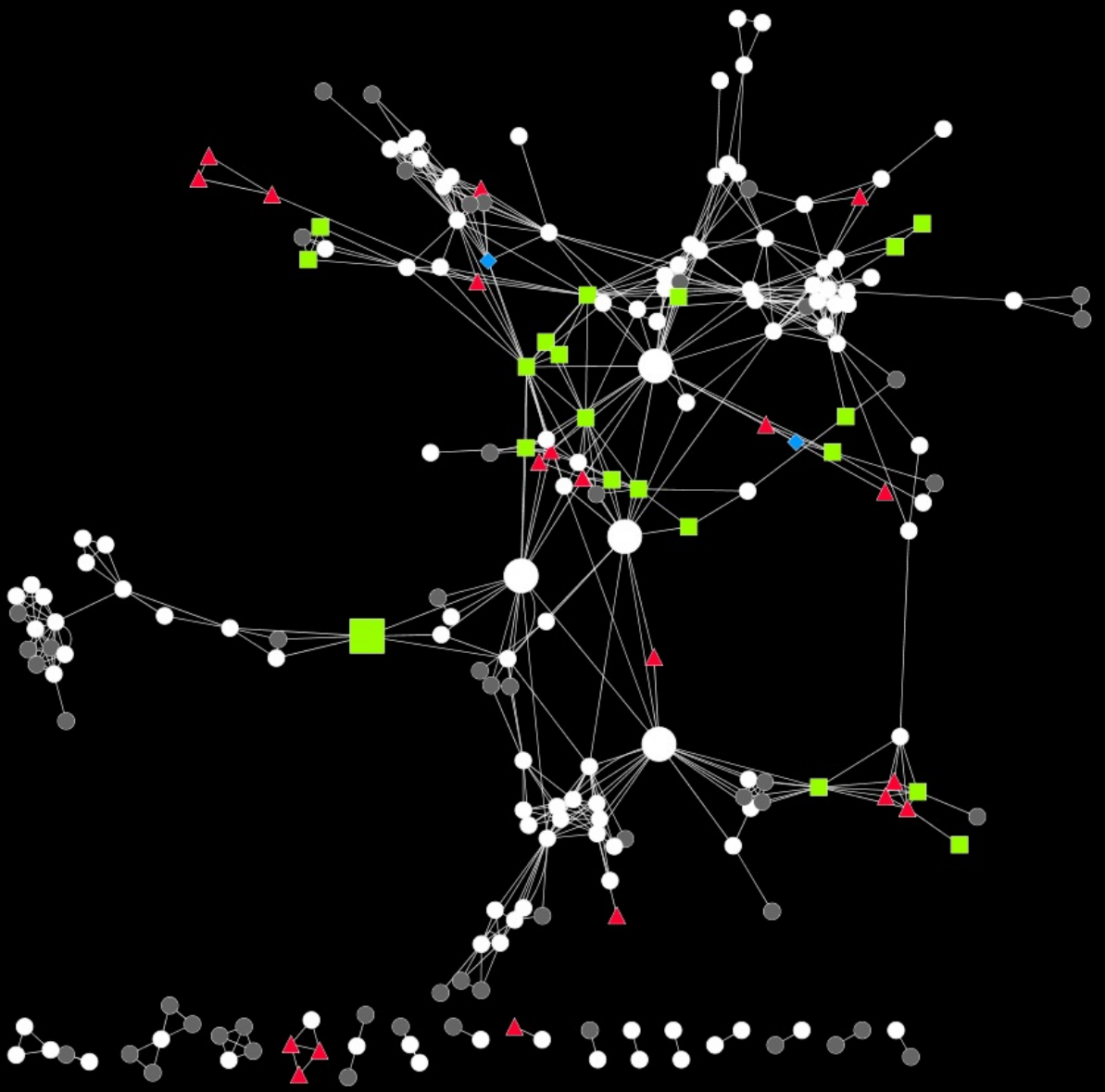
Red triangles: pleiotropic genes

Blue diamonds: discrete functionality and pleotropic genes

White circles: overlapping multifunctional distributions

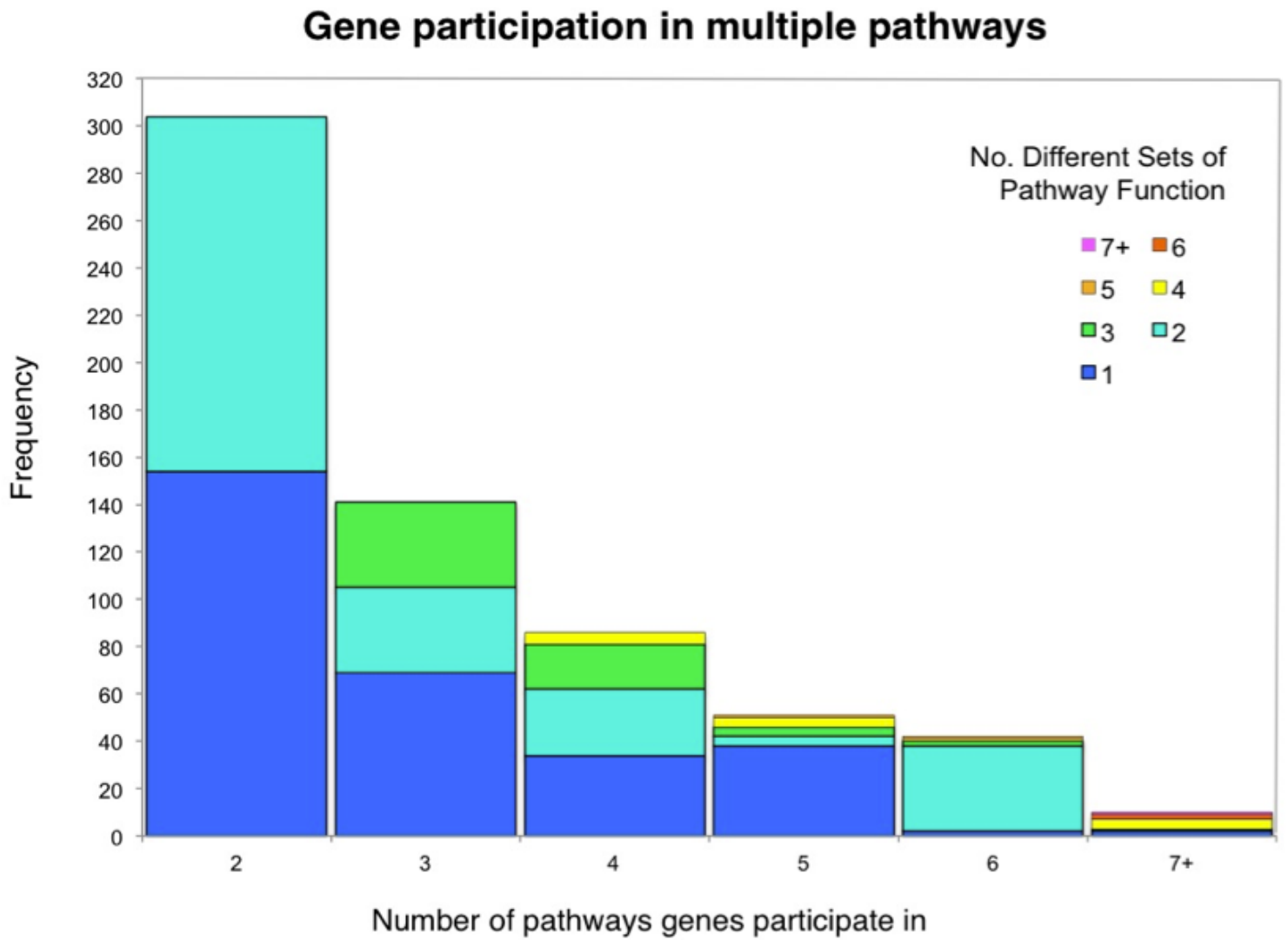
Light grey circles: single pathway function

Large nodes: high betweeness centralities

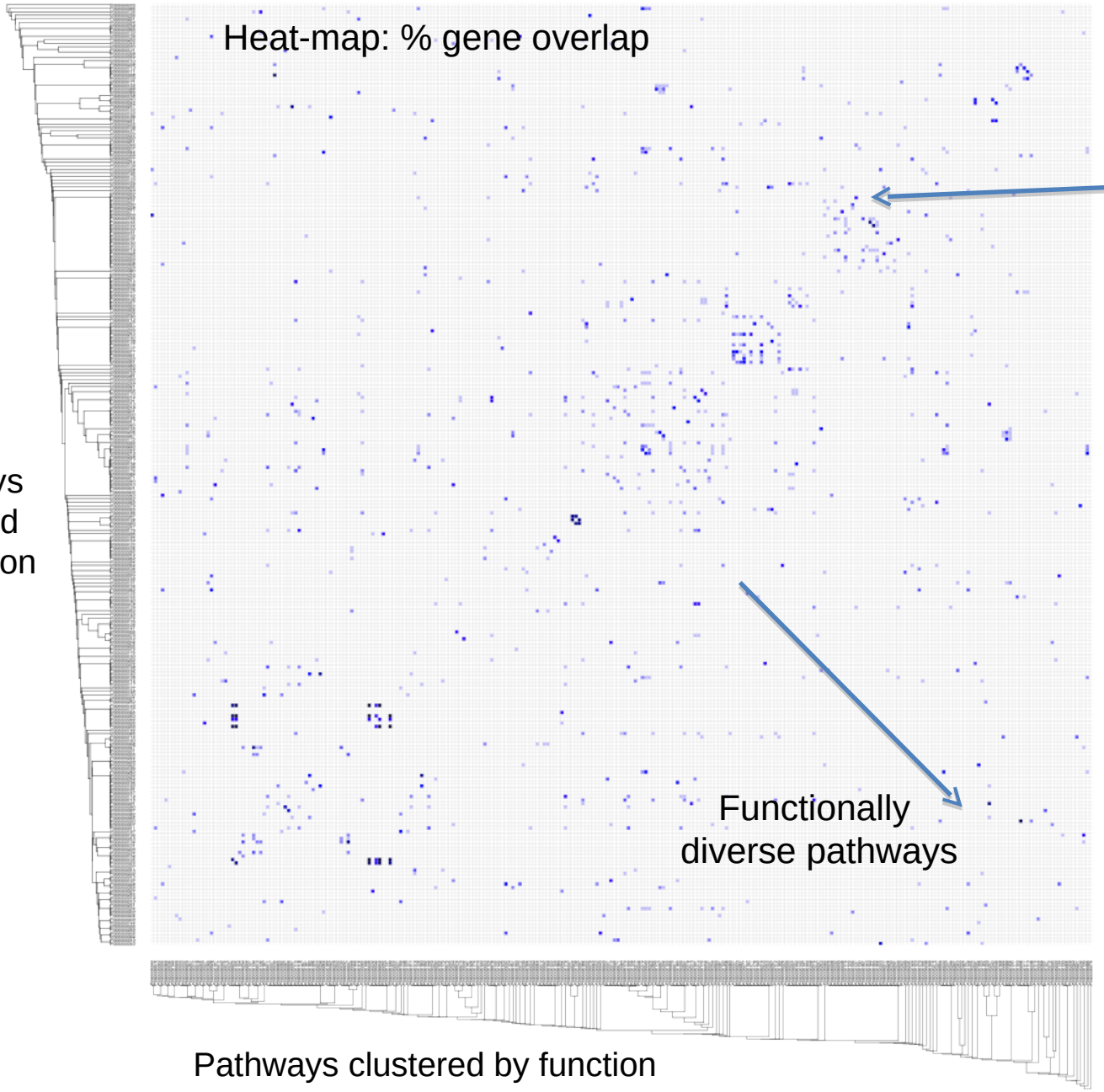


Gene participation in multiple pathways

- 44 % of genes appeared in multiple pathways.

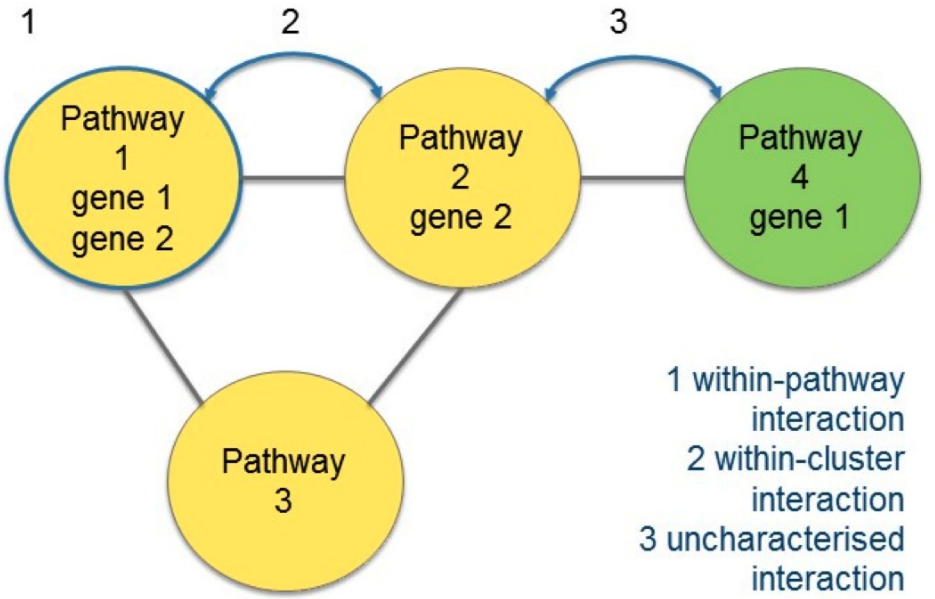


Gene overlap in functionally diverse pathways



Genetic interactions

- Genetic interactions (GI):
  Double mutants that show a significant deviation in fitness compared to the expected multiplicative effect of combining two single mutants.
- Increased GIs within network clusters suggest biological significance.
- GIs were enriched by a factor
  - 6.5 within pathways
  - 5.5 within clusters



Conclusions (2)

- Gene functions depend on the context, which can be inferred from pathway composition.
- Biological function should become the focus of models, instead of molecular function.
- The use of pathways as network entities can speed up model construction and enhance biological interpretability.