A knowledge-based T^2-statistic to perform pathway analysis for quantitative proteomic data

June En-Yu Lai

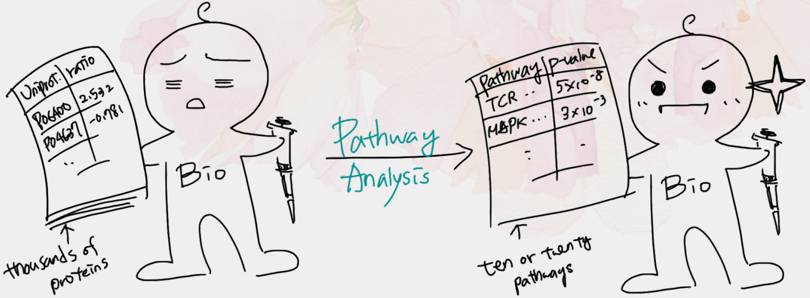
TIGP Bioinformatics Program, Academia Sinica, Taiwan

August 18, 2017

Functional Analysis for Quantitative Proteomic Data

- Proteomic data v.s. gene expression data
  - **Smaller sample size** (number of experiments).
  - Fewer identified entities.
  - The results are sensitive to experimental conditions and instruments.
- Functional analysis
  - **Pathway analysis** (PLoS Computational Biology, 2017).
  - Responsive subpathway locating (working manuscript).

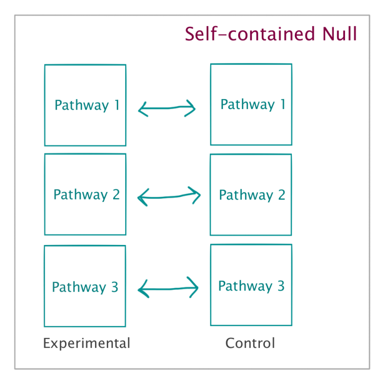
Pathway analysis — Data interpretation



Uniprot.	ratio
P06400	2.532
P04637	-0.781
..	

Pathway	p-value
TCR ..	5×10^{-8}
MAPK ...	3×10^{-3}
:	-

Null hypothesis: **Self-contained** v.s. Competitive



Self-contained Null

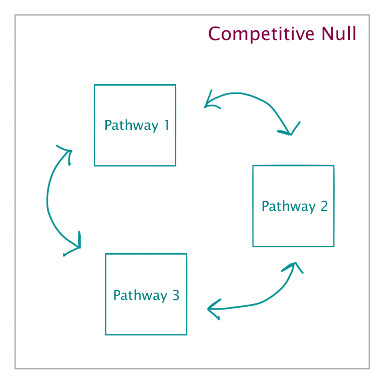
Pathway 1 ⟷ Pathway 1

Pathway 2 ⟷ Pathway 2

Pathway 3 ⟷ Pathway 3

Experimental

Control



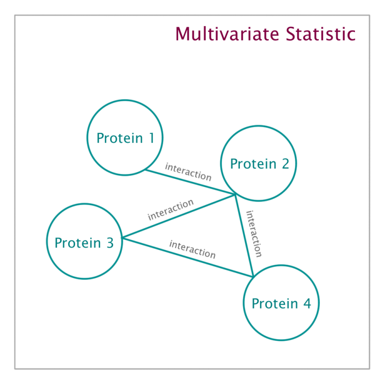
Competitive Null

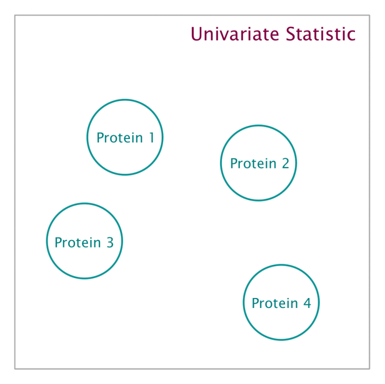
Pathway 1

Pathway 2

Pathway 3

Test statistic: **Multivariate** v.s. Univariate





The T^2-statistic for pathway analysis

The proposed T^2-statistic for a specific pathway $\mathcal{P}$ is then defined as,

$$T^2 = \boldsymbol{x}^T \boldsymbol{S}^{-1} \boldsymbol{x} \sim \chi_q^2$$

where

$\boldsymbol{x}$ is the vector of expression ratios,

$\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$,

$\boldsymbol{S}^{-1}$ is the inverse of **the covariance matrix $\boldsymbol{S}$**, and

q is the number of mapped proteins in $\mathcal{P}$.

The T^2-statistic for pathway analysis

- We use the **confidence score** provided by protein-protein interaction databases to represent the strength of the covariance, and the **expression direction** provided by the testing dataset to indicate the sign of the covariance.
- On the basis of the confidence scores in $\mathfrak{J}$ and the protein expression ratios, each element s_{ij} of $\boldsymbol{S}$ is determined by the following four rules:

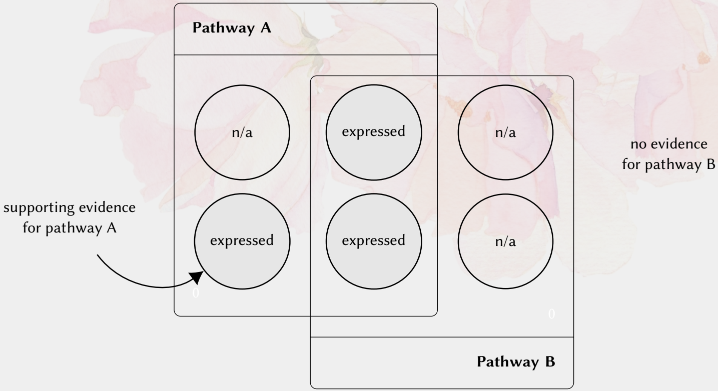
$$s_{ij} = \begin{cases} 0.4 & \text{if } i = j. \\ c_{p_i p_j} & \text{if } i \neq j, c_{p_i p_j} \in \mathfrak{J}, \text{ and } x_i \cdot x_j \geq 0. \\ -c_{p_i p_j} & \text{if } i \neq j, c_{p_i p_j} \in \mathfrak{J}, \text{ and } x_i \cdot x_j < 0. \\ 0.0 & \text{if } i \neq j \text{ and } c_{p_i p_j} \notin \mathfrak{J}. \end{cases}$$

The T^2-statistic for pathway analysis

$$T^2 = \boldsymbol{x}^T \boldsymbol{S}^{-1} \boldsymbol{x} \sim \chi^2_q$$

- If $\boldsymbol{S}$ is degenerate, we construct a Moore-Penrose pseudoinverse of $\boldsymbol{S}$ as a substitute, and q becomes the rank of $\boldsymbol{S}$.
- The p-value of the pathway $\mathcal{P}$ is derived from the χ^2_q distribution.

Pathway integration



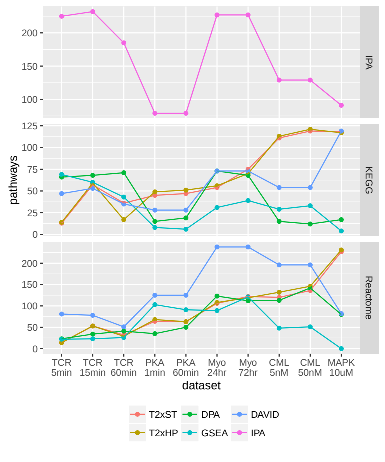
Performance evaluation — Really HARD!!!

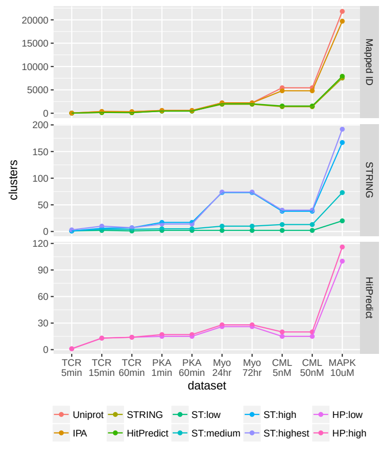
- No accepted gold-standards.
- We tried to match the results reported by these methods to the biological ideas provided by the original publication (true positives).
- A good statistic should be able to test if a pathway is significant:
  - if a statistic reports very little significant pathways, it might have the problem of false negatives;
  - if a statistic reports a large number of significant pathways, it might have the problem of false positive.
- We supposed that the number of significant pathways is one of the attribute to evaluate these methods.

General comparison

Tools	Significance requirement	Ranking statistic
T^2	raw p-value ≤ 0.05	number of mapped proteins
DPA	raw p-value ≤ 0.05	p-value
GSEA	FDR adjusted p-value (i.e. q-value) ≤ 0.25	NES
DAVID	EASE ≤ 0.1	p-value
IPA	Benjamini corrected p-value ≤ 0.05	p-value

General comparison: Consistency





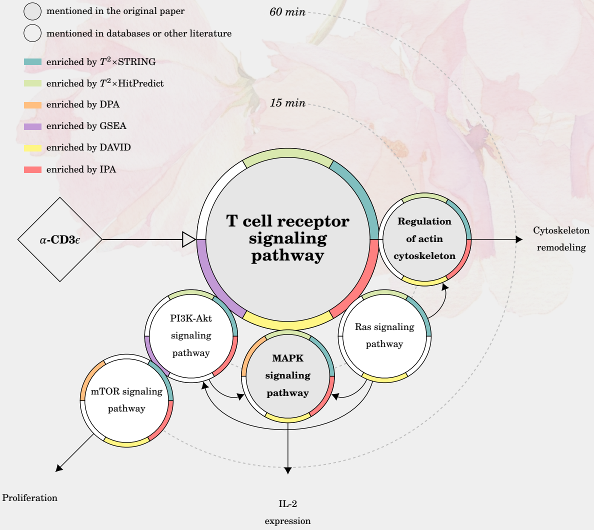
General comparison: Consistency

Database	KEGG					Reactome					IPA
Method	T^2+ST	T^2+HP	DPA	GSEA	DAVID	T^2+ST	T^2+HP	DPA	GSEA	DAVID	IPA
Uniprot	+	+	−	−	+	+	+	+	−	−	−
IPA	+	+	−	−	+	+	+	+	−	−	−
STRING	+	+	−	−	+	+	+	+	−	+	−
HitPredict	+	+	−	−	+	+	+	+	−	+	−
ST: low	+	+	−	−	+	+	+	+	−	−	−
ST: medium	+	+	−	−	+	+	+	+	−	−	−
ST: high	+	+	−	−	+	+	+	+	−	+	−
ST: highest	+	+	−	−	+	+	+	+	−	+	−
HP: low	+	+	−	−	+	+	+	+	−	−	−
HP: high	+	+	−	−	+	+	+	+	−	−	−

General comparison: Target pathways

Dataset	**TCR**			**PKA**		**Myogenesis**		**CML**		**MAPK**
Treatment	**α-CD3ϵ**			**PGE2**		**Serum-free**		**Dasatinib**		**U0126**
	5 min	*15 min*	*60 min*	*1 min*	*60 min*	*24 hr*	*72 hr*	*5 nM*	*50 nM*	*10 μM*
KEGG pathway	T cell receptor signaling pathway			cAMP signaling pathway		ECM-receptor interaction		Chronic myeloid leukemia		MAPK signaling pathway
$T^2\times$ST	1/13	1/57	2/36	16/45	17/47	15/54	16/75	20/111	23/119	3/118
p-value	< 0.0001	< 0.0001	0.0019	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
$T^2\times$HP	1/14	1/59	2/17	17/49	18/51	16/56	15/70	20/113	25/121	3/117
p-value	< 0.0001	0.0002	0.0011	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
DPA	-	8/68	20/71	11/15	7/19	18/73	-	-	-	2/17
p-value		0.0007	0.0024	0.0414	0.0178	0.0004				0.0007
GSEA	68/69	7/60	-	-	-	22/31	-	-	-	-
q-value	0.1786	0.1039				0.1416				
DAVID	1/47	1/53	1/35	-	-	40/73	40/73	8/54	8/54	8/119
p-value	< 0.0001	< 0.0001	< 0.0001			0.0041	0.0041	0.0004	0.0004	0.0002
IPA	1/225	2/232	89/185	60/79	60/79	-	-	52/129	52/129	-
p-value	< 0.0001	< 0.0001	0.0022	0.0186	0.0191			0.0013	0.0013	

Case Study: TCR downstream phosphoproteome



● mentioned in the original paper

○ mentioned in databases or other literature

🟦 enriched by T^2×STRING

🟩 enriched by T^2×HitPredict

🟧 enriched by DPA

🟪 enriched by GSEA

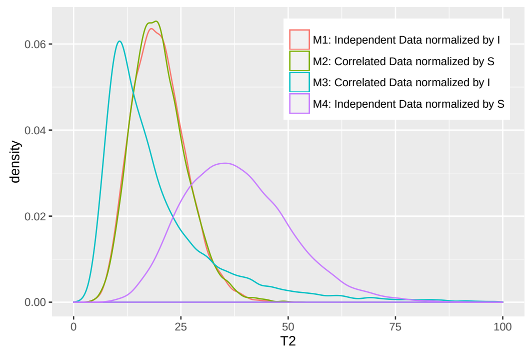
🟨 enriched by DAVID

🟥 enriched by IPA

Case Study: TCR downstream phosphoproteome

Dataset	Pathway Title	$\mathbf{T}^2\times$ST	$\mathbf{T}^2\times$HT	DPA	GSEA	DAVID	IPA
5 min	T cell receptor signaling pathway	1/13	1/14	-	68/69	1/47	1/225
15 min	Ras signaling pathway	2/57	2/59	-	-	10/53	-
	Regulation of actin cytoskeleton	3/57	3/59	-	-	9/53	15/232
	MAPK signaling pathway	4/57	4/59	65/68	-	41/53	4/232
	PI3K-Akt signaling pathway	24/57	25/59	-	6/60	-	169/232
60 min	mTOR signaling pathway	25/36	-	67/71	-	24/35	68/185

Accurate and inaccurate estimation: Toy example



	Correlated Data (generated by $\boldsymbol{S}$)	Independent Data (generated byï¡đI)
Correlated Null (normalized by $\boldsymbol{S}$)	M2: accurate estimation	M4: inaccurate due to false positive PPI scores
Independent Null (normalized by I)	M3: inaccurate due to incomplete knowledge	M1: accurate estimation

Accurate and inaccurate estimation: Real cases

Dataset	Experiment	Pathway Title	original *p*-value	30% permuted	60% permuted	30% purged	60% purged
TCR	5 min	T cell receptor signaling pathway	< 0.0001	100%	100%	100%	100%
	15 min	Ras signaling pathway	0.0017	100%	100%	100%	100%
		Regulation of actin cytoskeleton	< 0.0001	100%	100%	100%	100%
		MAPK signaling pathway	< 0.0001	100%	100%	100%	100%
		PI3K-Akt signaling pathway	< 0.0001	100%	100%	100%	100%
	60 min	mTOR signaling pathway	0.0002	97%	98%	100%	100%
PKA	1 min	Regulation of actin cytoskeleton	< 0.0001	100%	100%	100%	100%
		PI3K-Akt signaling pathway	< 0.0001	92%	81%	86%	84%
		MAPK signaling pathway	< 0.0001	100%	100%	100%	100%
		Rap1 signaling pathway	< 0.0001	100%	100%	100%	100%
		cAMP signaling pathway	< 0.0001	100%	100%	100%	100%
		Glycolysis / Gluconeogenesis	1	0%	0%	0%	0%
	60 min	Cell cycle	< 0.0001	100%	100%	100%	100%
		mTOR signaling pathway	0.0024	100%	100%	100%	100%
		Base excision repair	< 0.0001	100%	100%	100%	100%

Thank You ♥