Drug Pathway Decipherer and Some Exciting Results in Protein Complex Prediction

Difeng Dong, Guimei Liu, Limsoon Wong

Plan

- **Drug Pathway Decipherer**
  - Assumptions
  - Hypothesis
  - Approach
  - Preliminary implementation
- **Protein Complex Prediction**

Drug Pathway Decipherer: Assumptions, Hypothesis, Approach, & Preliminary Implementation

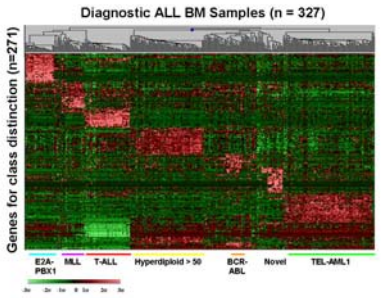
Gene Expression Analysis in Translational Medicine

- **Disease diagnosis**
- **Disease subtype discovery**
- **Treatment prognosis**

⇒ **Prediction accuracy is important**

- **Disease pathway inference**
- **Drug action pathway inference**
- **Drug escape pathway inference**

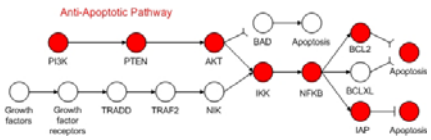
⇒ **Understanding cause and effect is important**



The patterns above tell us which patient has which ALL subtype. But they don't tell us why.

Assumptions

- **Circuit and state**
  - Two cells may have the same "molecular circuit"
  - But they may be in different "state" of the circuit



- **A drug acts thru an "unbroken" chain in the circuit --- the drug action pathway**
- **A drug fails to act when the chain is "broken" --- e.g., a gene in the chain is mutant or is under epigenetic effect**

Hypothesis

- **Chain X in the circuit is unbroken in drug responders**
- **Chain X in the circuit is broken in drug non-responders (possibly in different ways)**
- **Expression of genes on chain X in drug responders are consistent across drug responders**

Chain X is likely to be the drug's path of action

Approach

- Extract a chain from a known pathway
- Is the chain unbroken in responders and broken in non-responders?
- Is the state of this chain consistent in all responders?
- Candidate drug action pathway

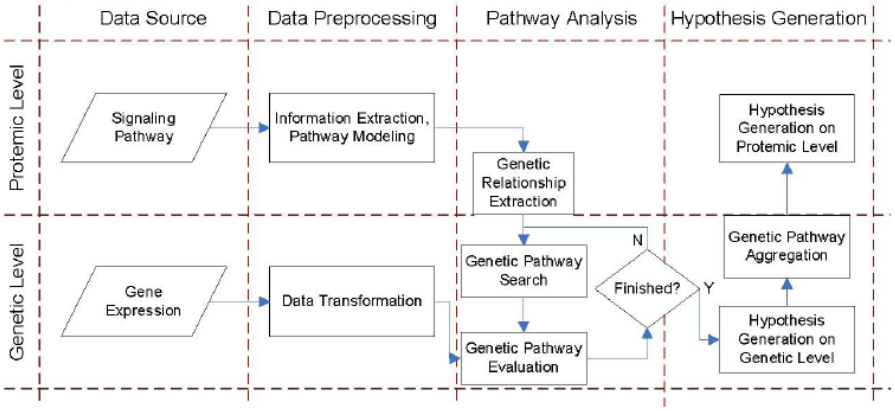
- Q: How to test a chain is broken?
- A: Test if a link is broken

- Q: How to test if a link (X → Y or X —| Y) is broken?
- A: Test if (X → Y or X —| Y) is behaving as expected

The idea is basic. But it sweeps many things under the carpet --- e.g., Y may be controlled by multiple genes. So how to implement this idea is quite tricky …

We gave this approach a try. Here is the first attempt…

DPD Workflow

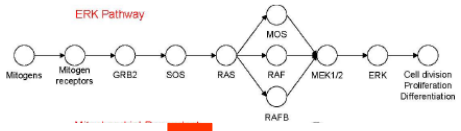


Idea: Look for pathways whose expected genetic /signaling interactions most consistent with those observed in the samples

DPD Details

- **Extract genetic relationship from signaling pathways**
- **Connect genetic relationship into genetic pathways**

A signaling pathway



Extracted genetic relationship

GRB2➔SOS2, SOS2➔HRAS, …

Connected genetic pathway

GRB2➔SOS2➔HRAS➔RAF1➔MAP2K1➔MAPK1➔MAPK Pathway;

DPD Details (cont.)

3. **Compute gene expression change correlation for each edge q in a genetic pathway for each sample**
4. **Derive z-score z(q) of correlation above wrt background**
5. **Compute pathway score for genetic pathway ϑ:**

$$z(\vartheta) = \frac{1}{\sqrt{k}} \sum_{q \in \vartheta} (-1)^{\alpha} z(q),$$

score(ϑ) = p-value of z(ϑ)
conf(ϑ) = 1 – score(ϑ)

6. **Apply p-value and FDR control to obtain significant hypothesized genetic pathways**
7. **Compute signaling pathway score & conf for signaling pathway γ**

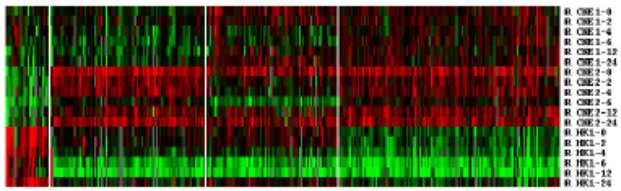
$$Z_i^{\gamma} = \sum_{\vartheta \sim \gamma} \sum_{g \in G_{\vartheta}} \left(\frac{1}{|G_{\vartheta}|} \times impact(g) \times r_{gi} \times \frac{conf(\vartheta)}{\sum_{\vartheta' \sim \gamma} conf(\vartheta')} \right)$$

$$conf(Z^{\gamma}) = \sum_{\vartheta \sim \gamma} \left(conf(\vartheta) \times \frac{conf(\vartheta)}{\sum_{\vartheta' \sim \gamma} conf(\vartheta')} \right)$$

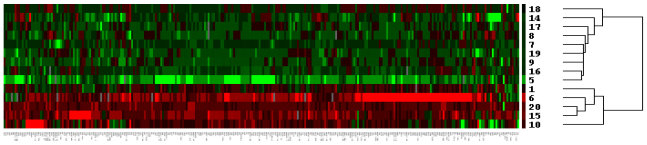
α= 0 if q is +ve relationship.
α = 1 if q is -ve relationship

Example: CYC202 Response in NPC

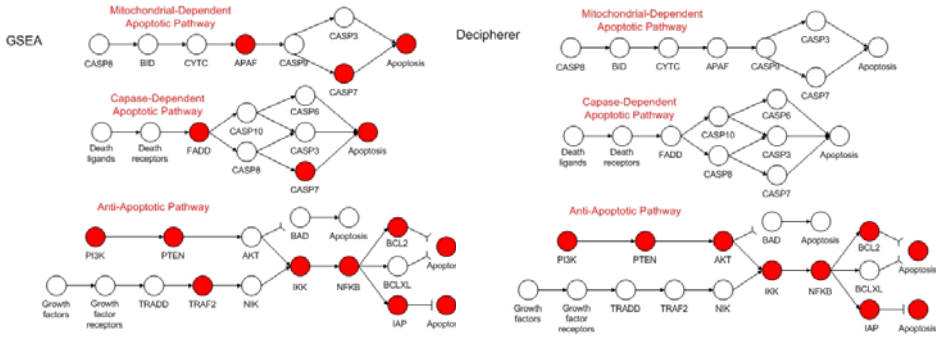
In vitro: 3 cell lines, expression measured at 6 time points. CNE1 resistant to treatment; CNE2 partial response; HK1 full response



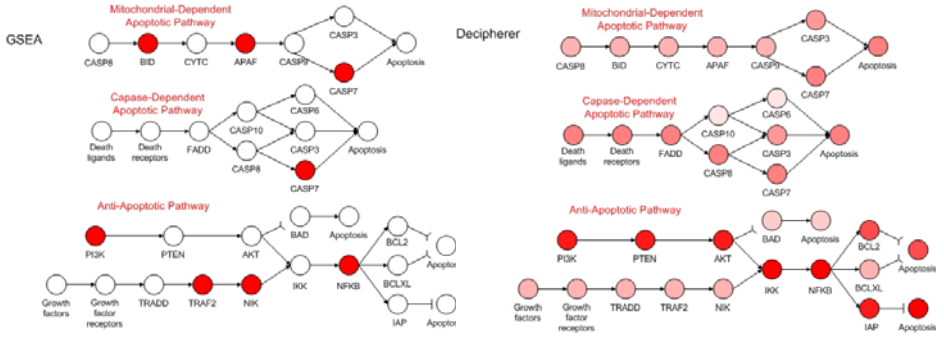
In vivo:12 patients, expression measured before and after treatment. Patients are classified into two responding groups wrt their genetic responding phenotype



GSEA vs DPD: In vitro



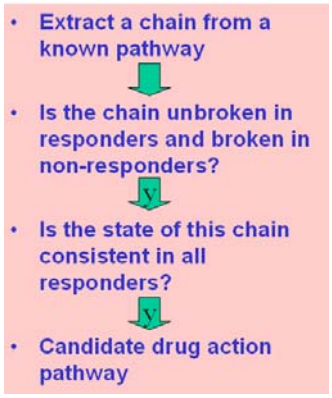
GSEA vs DPD: In vivo



Comments

- **Pros**
  - Pathway structures are considered, so more specific hypotheses are generated
  - Gene co-expression are considered, so higher significance level are reached
- **Cons**
  - Limited pathway structures available to evaluate
  - Evaluation procedure is too complicated

- **However, do you think we have actually implemented this approach?!**



Some Exciting Results in Protein Complex Prediction

Motivation

- **Can a protein interaction with so many proteins simultaneously?**

⇒ **A big "hub" and its "spokes" should probably be decomposed into subclusters**

- Each subcluster is a set proteins that interact in the same space and time
- Viz., a protein complex

Some Protein Interaction Data Sets

Sprinzak et al., *JMB*, 327:919-923, 2003

Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results)	516	53	33
A2: GY2H Ito *et al.* (core)	798	64	40
A3: GY2H Ito *et al.* (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, *in vitro*	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

Large disagreement betw methods

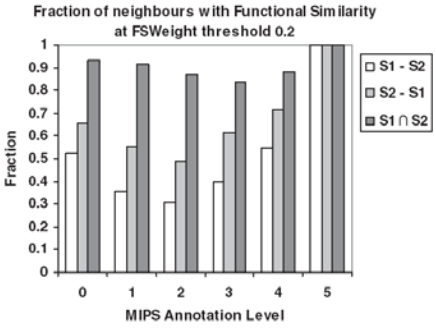
- **High level of noise**

⇒ **Need to clean up before protein complex prediction**

Guilt by Association of Common Interaction Partners

- **Two proteins participating in same biological process are likely to interact**
- **Two proteins having a large proportion of their interaction partners in common are likely to participate in same biological process**

⇒ **Two proteins having a large proportion of their interaction partners in common are likely to directly interact also**

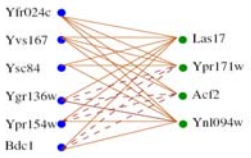


Chua et al, Bioinformatics, 22:1623--1630, July 2006

Combining Local and Global Measures

- **Local measure**

$$w_L^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} w_L^{k-1}(x,u) + \sum_{x \in N_u \cap N_v} w_L^{k-1}(x,v)}{\sum_{x \in N_u} w_L^{k-1}(x,u) + \sum_{x \in N_v} w_L^{k-1}(x,v) + \lambda_u^k + \lambda_v^k}$$



- **Global measure**

$$w_G(u,v) = \max\{conf(V_1,V_2) \cdot \frac{2|N_u \cap V_2|}{|V_2|+|N_u|} \cdot \frac{2|N_v \cap V_1|}{|V_1|+|N_v|} \mid u \in V_1, v \in V_2\}$$

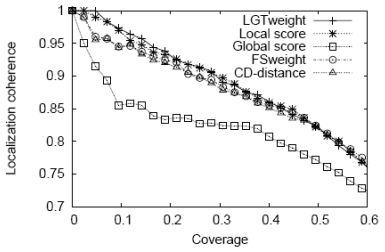
Where conf(V_1, V_2) is the ratio of # of interactions betw V_1 and V_2 to # of distinct protein pairs contained in (V_1, V_2)

- **Combined measure**

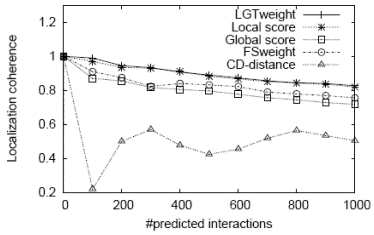
$$LGTweight(u,v) = w_L^2(u,v) + w_G(u,v).$$

Liu & Wong, GIW 2008

High functional homogeneity and localization coherence are observed in PPIs that are ranked high by LGTweight



Localization coherence of PPIs reported in DIP



Localization coherence of PPIs predicted

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%

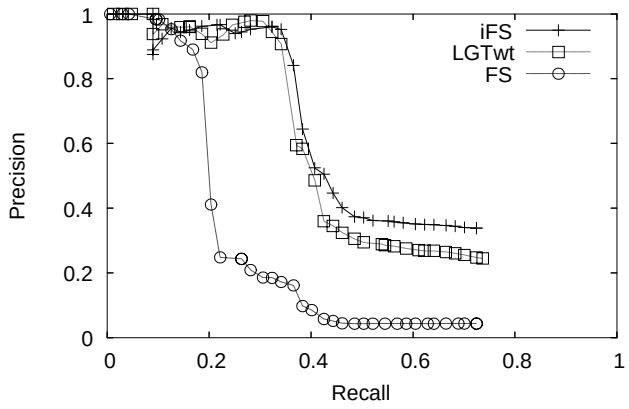
Now we can make protein complex prediction as follows…

- **Remove noise edges in the input PPI network by discarding edges having low LGTweight**
- **Augment the input PPI network by addition of missing edges having high LGTweight**
- **Predict protein complex by simple clique finding or other techniques**

Preliminary Experiments

- **Matching a predicted complex S with a true complex C**
  - Vs: set of proteins in S
  - Vc: set of proteins in C
  - Overlap(S, C) = $|Vs \cap Vc|^2 / |Vs||Vc|$
  - Overlap(S, C) ≥ 0.25
- **Evaluation**
  - Precision = matched predictions / total predictions
  - Recall = matched complexes / total complexes
- **Datasets: BioGrid yeast**
  - #interactions: 38555
  - #interactions with >0 common neighbor: 27940

Results



20000 new interactions added

What have we learned?

Drug Pathway Decipherer

- Extract a chain from a known pathway
- ↓
- Is the chain unbroken in responders and broken in non-responders?
- ↓ y
- Is the state of this chain consistent in all responders?
- ↓ y
- Candidate drug action pathway

Protein complex prediction

- **Guilt by association of common interaction partners is useful for predicting**
  - PPI cellular localization
  - Missing PPIs
  - Protein complexes

- **Acknowledgement**
  - Difeng Dong

- **Acknowledgement**
  - Guimei Liu

Any Question?

