Two Applications of Text Mining in Bioinformatics:

Enhancing Protein Function Prediction & Enhancing Drug Pathway Inference

Limsoon Wong

Plan

- **Protein function prediction**
  - Current approaches
  - Info fusion by Integrated Weighted Averaging
  - How can text mining help?
- **Drug pathway analysis**
  - Current approaches
  - Pathway consistency by Drug Pathway Decipherer
  - How can text mining help?

Protein Function Prediction:
Current Approaches

Protein Function Prediction

- **Protein function prediction is a key problem**
- **It is often solved using "guilt by association"**
  - Compare the target sequence T with sequences S_1, …, S_n of known function in a database
  - Determine which ones amongst S_1, …, S_n are the mostly likely homologs of T
  - Then assign to T the same function as these homologs
  - Finally, confirm with suitable wet experiments

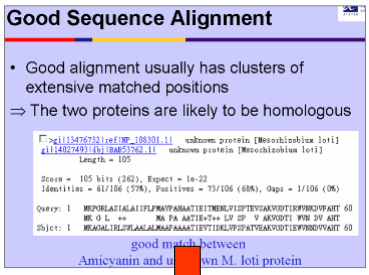
Guilt by Association of Seq Similarity

Compare *T* with seqs of known function in a db

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions

⇒ The two proteins are likely to be homologous

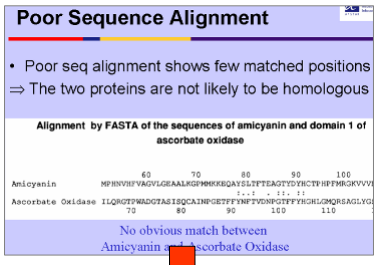


good match between Amicyanin and u[illegible]wn M. loti protein

Poor Sequence Alignment

- Poor seq alignment shows few matched positions

⇒ The two proteins are not likely to be homologous



No obvious match between Amicyanin a[illegible]scorbate Oxidase

Assign to *T* same function as homologs

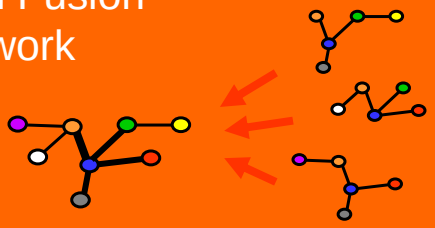
Discard this function as a candidate

Confirm with suitable wet experiments

Important Unsolved Challenges

- **What if there is no useful seq homolog?**
- **Guilt by other types of association!**
  - Domain modeling (e.g., HMMPFAM)
  - Similarity of dissimilarities (e.g., SVM-PAIRWISE)
  - Similarity of phylogenetic profiles
  - Similarity of subcellular co-localization & other physico-chemico properties(e.g., PROTFUN)
  - Similarity of gene expression profiles
  - Similarity of protein-protein interaction partners
- **Can text mining association help?**
- **Can fusion of multiple types of info help?**

Protein Function Prediction:
Information Fusion Framework

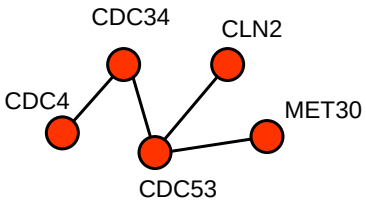


NUS National University of Singapore

NUS

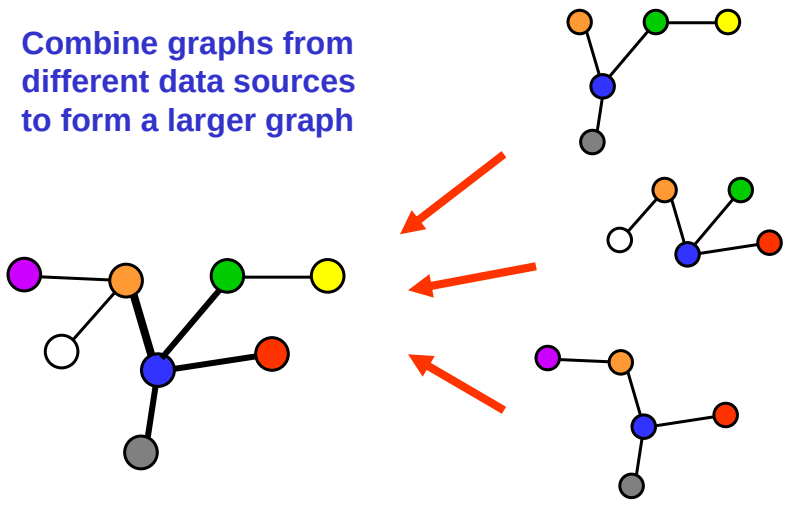
Strategy – Step 1

- **Model a data source as undirected graph G = ⟨V,E⟩**
  - V is a set of vertices; each vertex reps a protein
  - E is a set of edges; each edge (u , v) reps a relationship (e.g. seq similarity, interaction) betw proteins u and v



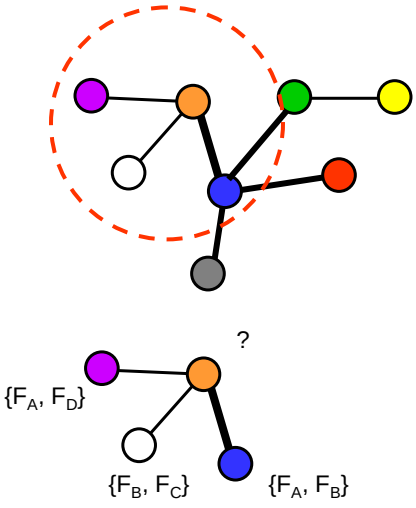
Strategy – Step 2

- **Combine graphs from different data sources to form a larger graph**



Strategy – Step 3

- **Estimate edge confidence from contributing data sources**
- **Predict function by observing which functions occur frequently in the high-confidence neighbours**



Unified Confidence Evaluation

- **Subdivide each data source into subtypes to improve precision (e.g., expt sources, sub-ranges of existing scores like E-scores)**
- **In general, estimate confidence of subtype k for sharing function f by:**

$$p(k, f) = \frac{\sum_{(u,v) \in E_{k,f}} S_f(u,v)}{|E_{k,f}| + 1}$$

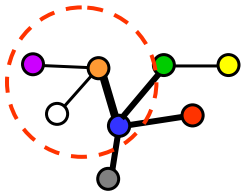
  - $E_{k,f}$ is subset of edges of subtype k where each edge has either one or both of its vertices annotated with function f
  - $S_f(u,v) = 1$ if u and v shares function f, 0 otherwise

Combination of Confidence

- **Combine confidence of data sources contributing to each edge:**

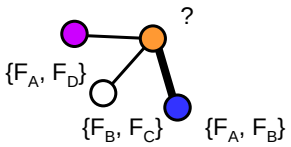
$$r_{u,v,f} = 1 - \prod_{k \in D_{u,v}} (1 - p(k, f))$$

  - p(k.f) is confidence of edges of subtype k sharing function f
  - $D_{u,v}$ is the set of subtypes of data sources which contains the edge (u,v)



Function Prediction

- **Integrated Weighted Average**



$$S_f(u) = \frac{\sum_{v \in N_u} \left(e_f(v) \times r_{u,v,f} \right)}{1 + \sum_{v \in N_u} r_{u,v,f}}$$

- $S_f(u)$ is score of function f for protein u
- $e_f(v)$ is 1 if protein v has function f, 0 otherwise
- N_u is set of neighbours of u
- $r_{u,v,f}$ is confidence of edge (u, v)

Protein Function Prediction:

Effect of Co-occurences of Protein Names in MEDLINE Abstracts

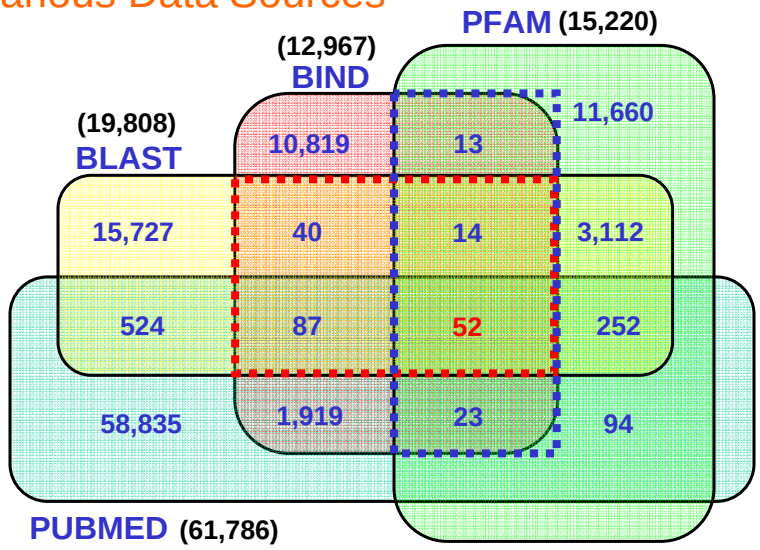
Data Sources

- **Protein Sequences**
  - Seqs from GO db
  - Each yeast seq is aligned w/ rest using BLAST (cutoff E-Score = 1)
  - -log(e-score) used as score
  - Top 5 results w/ known annotations
  - 19,808 unique pairs involving yeast proteins
- **Pfam Domains (SwissPfam)**
  - Precomputed Pfam domains for SwissProt and TrEMBL proteins w/ E-value threshold 0.01
  - No. of common domains as score
  - 15,220 unique pairs involving yeast proteins
- **PPI (BIND)**
  - 12,967 unique interactions betw yeast proteins
  - FS weight used as score
- **Pubmed Abstracts**
  - Pubmed abstracts obtained by searching protein's name and aliases on Pubmed
  - Limit to first 1000 abstracts returned
  - Fraction of abstracts w/ co-occurrence used as score

$$s(u,v) = \frac{|A_u \cap A_v|}{\sqrt{|A_u| \times |A_v|}}$$

  - 61,786 unique pairs involving yeast proteins

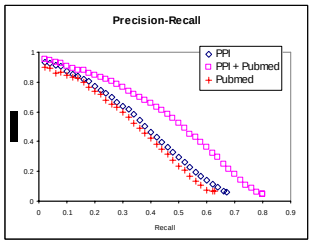
Pairs Involving Yeast Proteins in Various Data Sources

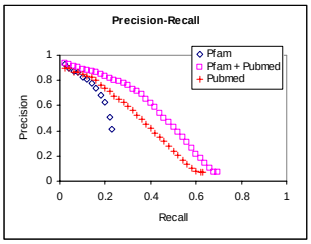


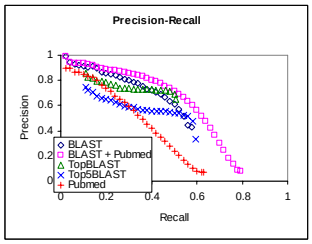
Can co-occurrence in abstracts help?

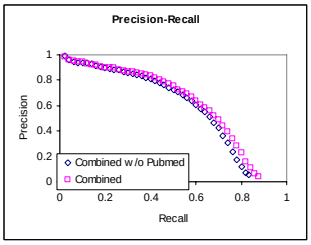
- **Need comparisons of**
  - PPI info w/ & w/o abstract occurrence info,
  - BLAST info w/ & w/o abstract occurrence info,
  - Pfam info w/ & w/o abstract occurrence info,
  - "Combined" w/ & w/o abstract occurrence info,
  - Top-blast info w/ & w/o abstract occurrence info

Diff in Recall-Precision by Literature Co-Occurrence

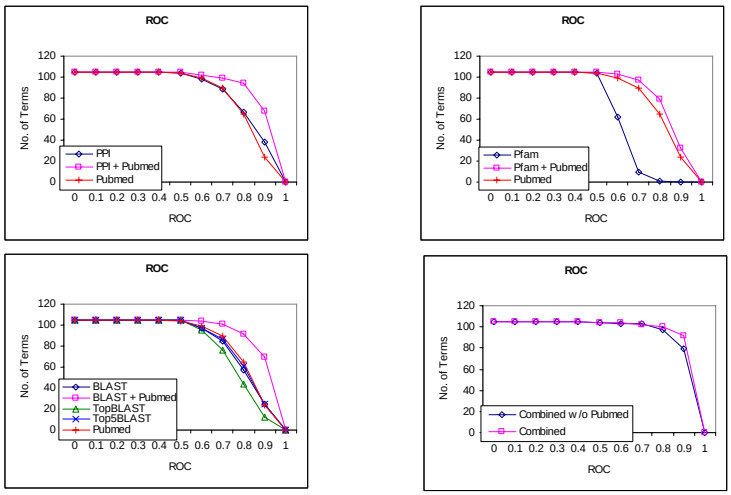


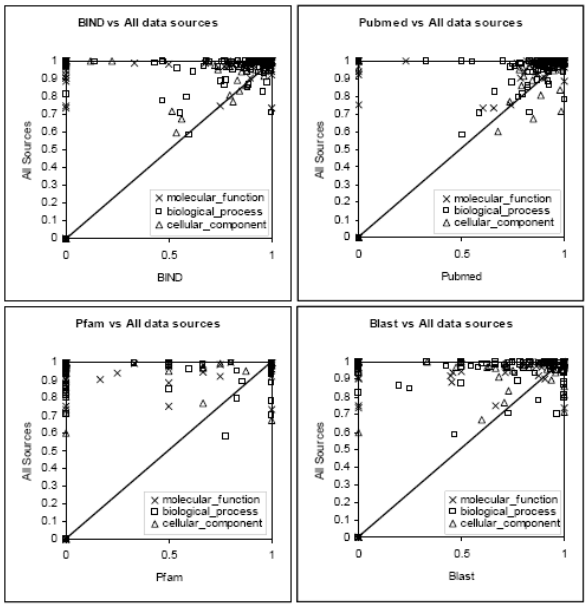






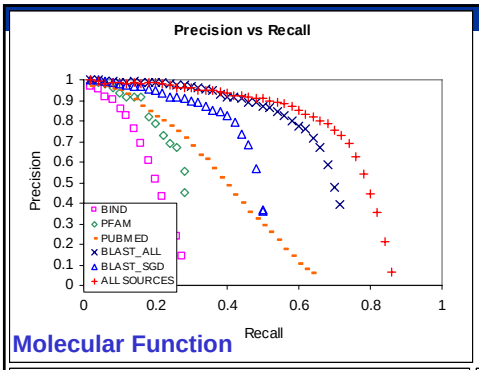
Diff in No. of Terms w/ Better ROC by Literature Co-Occurrence





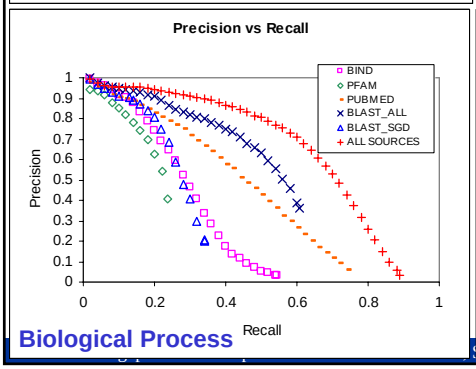
Literature co-occurrence seems to contribute especially well to cellular component & biological processes

21

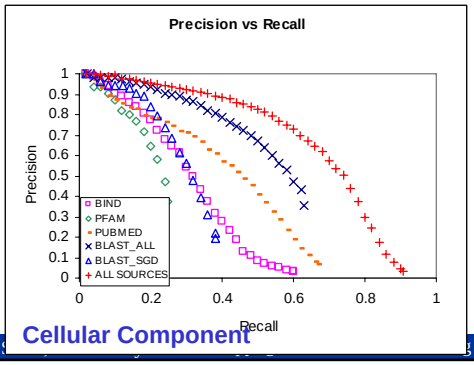


Molecular Function

Combining all data sources outperforms any individual data source

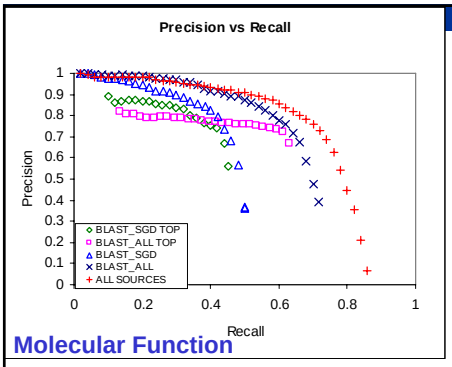


Biological Process



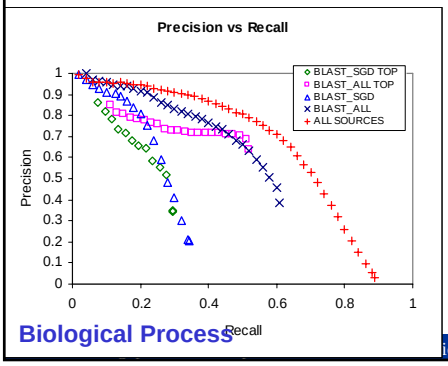
Cellular Component

22

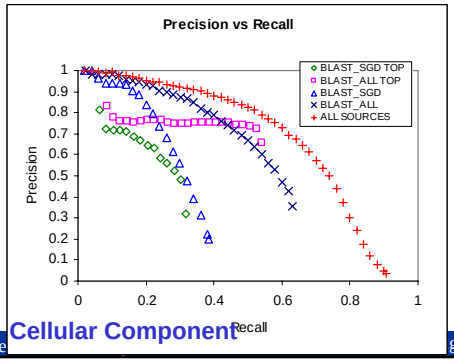


Molecular Function

- **Weighted Averaging predicts w/ better precision than transferring function from top blast hit**
- **Using all data sources outperforms topblast in both sensitivity and precision**



Biological Process



Cellular Component

Conclusions & Ongoing Works

- A simple graph-based method that combines multiple sources of data sources for function prediction
- Even simple co-occurrence count can give reasonable sensitivity & precision for function prediction
- Combining multiple info sources outperforms any single info source

- Can we improve on this formula?

$$s(u,v) = \frac{|A_u \cap A_v|}{\sqrt{|A_u| \times |A_v|}}$$

- Can we identify "good" abstracts?
- Can we use co-occurrence at sentence level? Can we use richer sentence level analysis?
- Can we identify "good" sentences?

Drug Pathway Inference:

Current Approaches

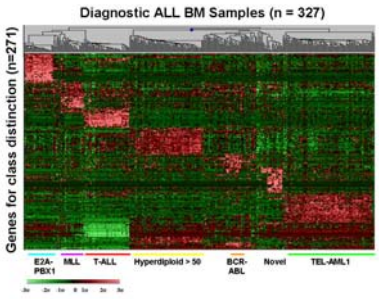
Gene Expression Analysis in Translational Medicine

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

⇒ **Prediction accuracy is important**

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

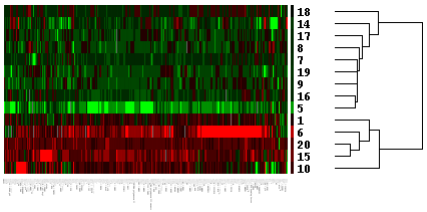
⇒ **Understanding cause and effect is important**



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

Understanding Drug Response

- **Key objectives**
  - Identify significantly diff regulated genetic pathways correlating well to treatment response
  - Identify drug-resistant genetic phenotypes for treatment non-responders



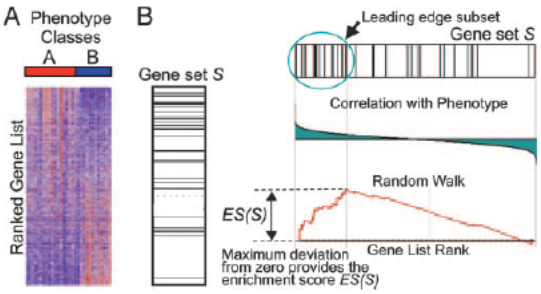
The patterns tell us which NPC patients respond to CYC202. But They don't tell us why.

Approaches

- **Intersection Analysis**
  - Genes are ranked wrt correlation w/ phenotypes
  - Selected genes are evaluated on pre-defined gene groups, based on pathways for significance of intersection
  - Issues
    - **Which test statistics to use to rank genes?**
    - **Which cut-off to use?**
- **T-Profiler, PAGE**
  - Average fold change of genes is first computed
  - T-test/Z-test is then performed on each pre-defined gene sets and the remaining genes
  - Issues
    - **All genes with a gene group are considered**
    - **Which test statistics to use?**
    - **Expression variance within each group is ignored**

Approaches

- **GSEA**
  - Genes are ranked according to their correlation with phenotypes
  - Go across the ranked gene list, whenever meets a gene in the specified gene set, increase the score, otherwise, decrease it
  - Issue
    - **Gene-gene relationship ignored**



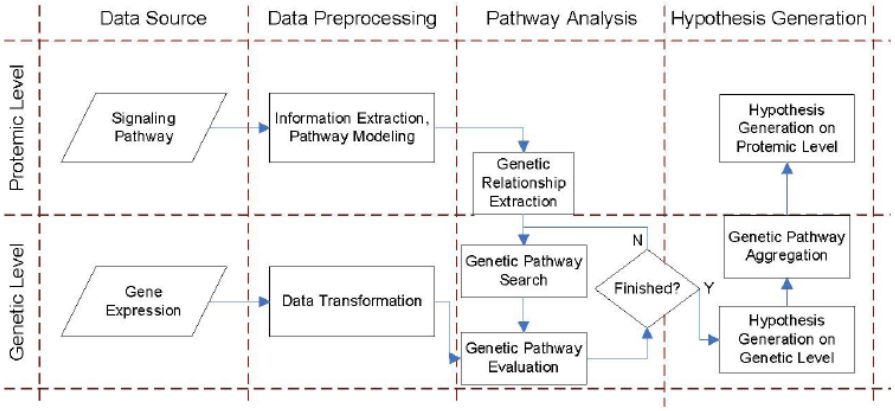
Gaps

- **Intersection analysis**
  - Which test to use to rank genes?
  - Which cut-off to use?
- **T-Profiler and PAGE**
  - All genes with a gene group are considered
  - Which test to use?
  - Expression variance within each group is ignored
- **GSEA**
  - Gene-gene relationship is ignored
- **Obvious next stage: consider gene-gene relationship**
  - How to get this info?
  - How to use them?
  - Can text mining help?

Drug Pathway Inference:

Drug Pathway Decipherer

DPD Workflow

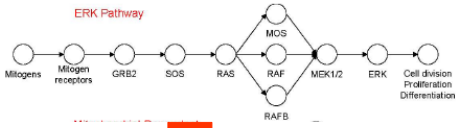


Idea: Look for pathways whose expected genetic /signaling interactions most consistent with those observed in the samples

DPD Details

1. **Compute for each gene its relative expression change betw pre- and post-treatment**
2. **Extract genetic relationship from signaling pathways**
3. **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.)

4. **Compute gene expression correlation for each edge q in a genetic pathway for each sample**
5. **Derive z-score z(q) of correlation above wrt background**
6. **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)

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

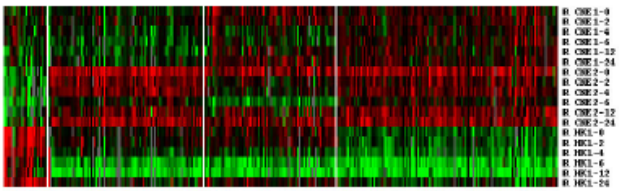
7. **Apply p-value and FDR control to obtain significant hypothesized genetic pathways**
8. **Compute signaling pathway score & conf for signaling pathway γ**

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

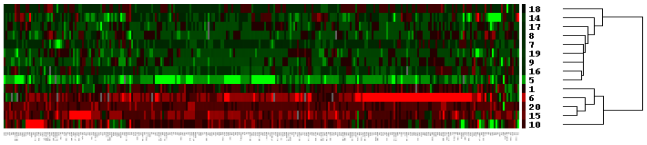
$$conf(Z^{\gamma}) = \sum_{\vartheta \sim \gamma} \left(conf(\vartheta) \times \frac{conf(\vartheta)}{\sum_{\vartheta' \sim \gamma} conf(\vartheta')} \right)$$

Example: CYC202 Response in NPC

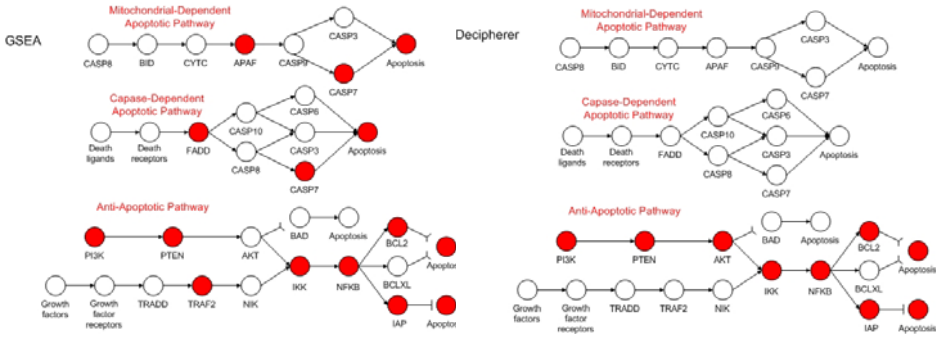
In vitro: 3 cell lines, expression measured at 6 time points. CNE1 resistant to treatment; CNE2 partial response; HK1 full response



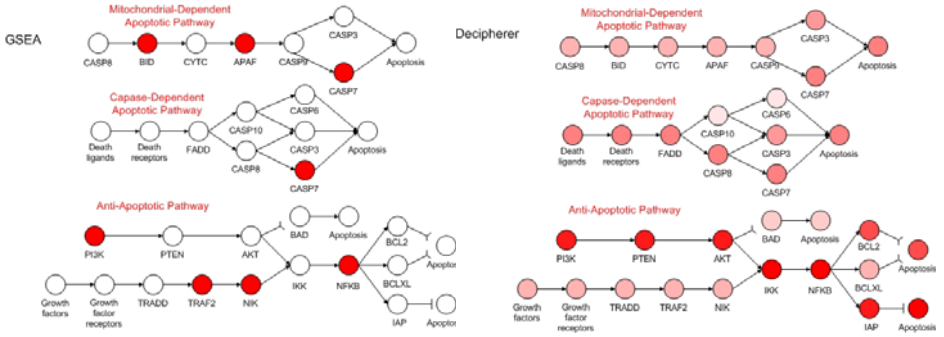
In vivo:12 patients, expression measured before and after treatment. Patients are classified into two responding groups wrt their genetic responding phenotype



GSEA vs DPD: In vitro



GSEA vs DPD: In vivo



Differentiation Status of Pathways

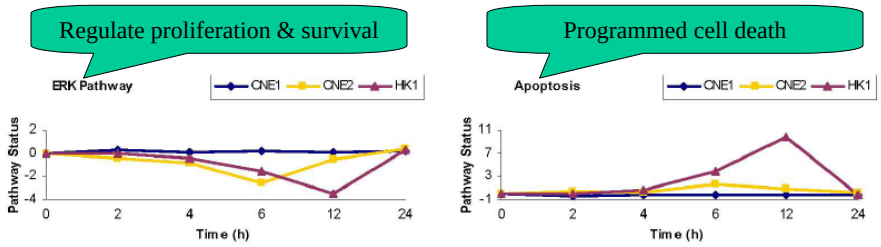


Table 1. *p*-values for the differentiations of status of signaling pathways.

Comparison Group	ERK	Apoptosis	JNK/p38	G1/S
CNE1 *vs.* CNE2	< 0.0001	0.0028	0.2921	-
CNE1 *vs.* HK1	< 0.0001	0.0006	-	0.4992
CNE2 *vs.* HK1	0.0004	0.0022	-	-

Identification of Genetic Pathways

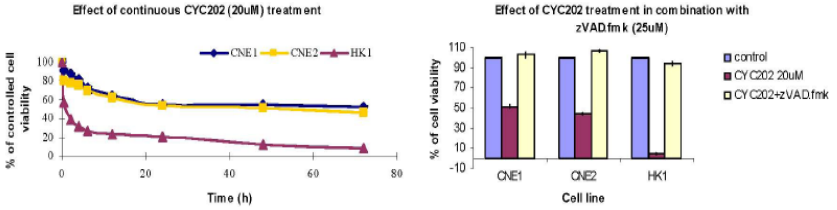


Table 2. List of the identified genetic pathways: Genes for replacement are separated by "/".

Signaling Pathway	Genetic Pathway	Confidence
CNE1		
ERK	GRB2→SOS2→HRAS→RAF1→MAP2K1→ MAPK1/MAPK3	$\geqslant 0.999$
Apoptosis	PIK3CB→PTEN→AKT2/AKT3→CHUK/IKBKB/IKBKG→NFKB2→BIRC2/BIRC5	$\geqslant 0.9998$
JNK/p38	MAP3K12→MAP2K7→MAPK9	0.9665
G1/S	CCND1→CDK4⊣RB1→E2F2/E2F3	$\geqslant 0.9906$
CNE2		
ERK	GRB2→SOS1→MRAS/KRAS/NRAS/RRAS→BRAF→MAP2K1→MAPK1	$\geqslant 0.9885$
Apoptosis	PIK3CA/PIK3CB→PTEN→AKT1→IKBKB→RELA→BIRC2/BIRC5	$\geqslant 0.9949$
JNK/p38	MAP4K3/TRAF2→MAP3K1→MAP2K4→MAPK8/MAPK10	$\geqslant 0.9658$
HK1		
ERK	GRB2→SOS1→HRAS→BRAF→MAP2K1/MAP2K2→MAPK1/MAPK3	$\geqslant 0.9646$
Apoptosis	PIK3R1→PTEN→AKT2/AKT3→IKBKB→NFKB2/RELA→BCL2/BIRC2	$\geqslant 0.9663$
G1/S	CUL1→SKP2→CDKN1A⊣CDK6⊣RB1⊣E2F2/E2F3	$\geqslant 0.9645$

Differentiation Status of Pathways

Table 3. The results of signaling pathway status estimation for the *in vivo* dataset: The "response" column shows the molecular response to treatment for patients. The "status" column shows the estimated post-treatment pathway status.

Patient	Response	ERK		JNK/p38		G1/S		Apoptosis	
		Status	Conf.	Status	Conf.	Status	Conf.	Status	Conf.
Pt5	P(ositive)	-2.25	0.98	-3.08	0.99	-	-	1.34	0.99
Pt8	P	-	-	-1.01	0.99	-	-	0.82	0.98
Pt9	P	-0.97	0.98	-	-	0.76	0.95	-	-
Pt14	P	-	-	-	-	-0.61	0.99	-0.86	0.99
Pt16	P	-0.20	0.99	-0.20	0.95	0.29	0.99	1.42	0.97
Pt17	P	-1.02	0.99	-1.02	0.99	-0.33	0.96	1.01	0.99
Pt19	P	-	-	-0.86	0.98	-	-	0.91	0.98
Pt18	No Tumor	**-0.15**	0.99	-	-	0.28	0.99	**0.13**	0.99
Pt1	N(egative)	0.21	0.95	0.52	0.99	1.06	0.97	-1.00	0.98
Pt7	N	-0.10	0.97	-0.68	0.96	0.28	0.98	0.11	0.98
Pt10	N	1.02	0.99	1.16	0.99	-	-	-1.57	0.97
Pt15	N	-	-	-	-	-	-	-1.01	0.98
Pt20	N	1.30	0.98	-	-	-0.93	0.96	-1.68	0.99

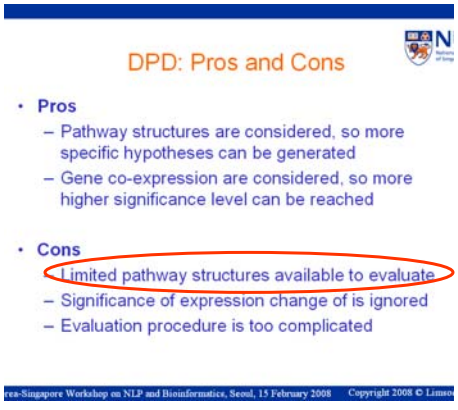
DPD: Pros and Cons

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

Drug Pathway Inference:

Where Can Text Mining Contribute?

Where can text mining contribute?



DPD: Pros and Cons

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

- **Extract signaling pathway details (i.e., inhibit, activate)**
- **Link proteins (i.e., signaling pathway level) to genes (i.e., genetic pathway level)**
- **Link pathway to phenotypes**

References & Acknowledgements

This talk is based on work done in …

- Hon Nian Chua, Wing-Kin Sung, Limsoon Wong. **An Efficient Strategy for Extensive Integration of Diverse Biological Data for Protein Function Prediction**, *Bioinformatics*, 23:3364-3373, 2007
- Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments.** *ACM SIGKDD Explorations*, 9:3-14, 2007
- Difeng Dong, Chun-Ying Cui, Benjamin Mow, Limsoon Wong**. Deciphering Drug Action and Escape Pathways: An Example on Nasopharyngeal Carcinoma.** In preparation.

Any Question?