

“Guilt by Association” as a Search Principle

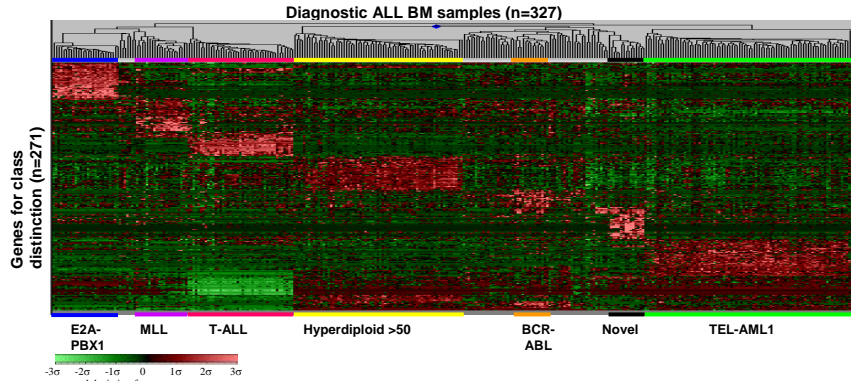
Limsoon Wong



A slight detour ...

Guilt by Association

- But do we know which ones are causal genes, which ones are surrogates, and which are noise?



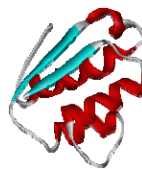
Biology 101

- Complete genomes are now available
- Knowing the **genes** is not enough to understand how biology **functions**
- **Proteins**, not genes, are responsible for many cellular activities
- Proteins function by **interacting** w/ other proteins and biomolecules

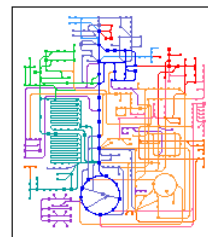
GENOME



PROTEOME



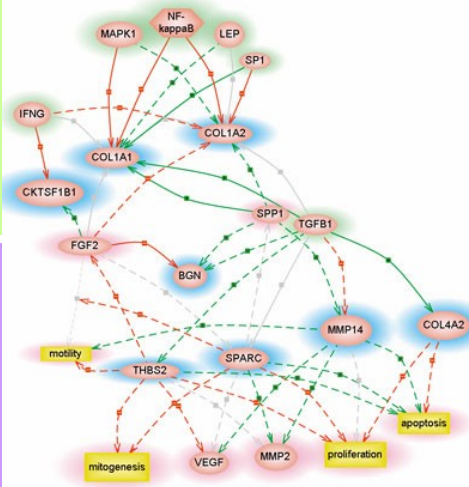
“INTERACTOME”



Slide credit: See-Kiong Ng

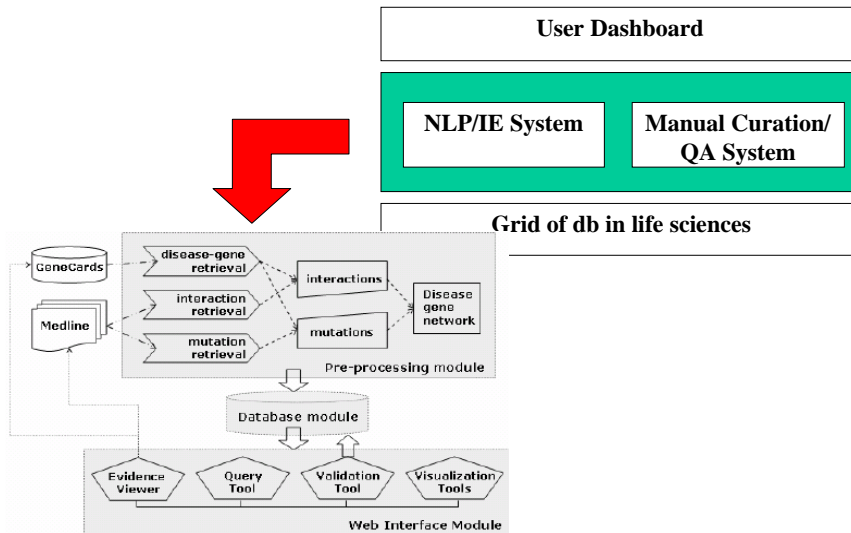
Gene Regulatory Circuits

- Extract functional annotations
- Extract relationships between genes, proteins, processes, diseases, & drugs
- Predict functional annotations
- Predict relationships between genes, proteins, processes, diseases, & drugs

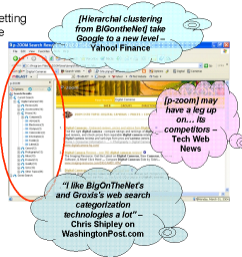


Information Extraction & Retrieval: Challenges in Context of Biomedicine

A Typical Architecture

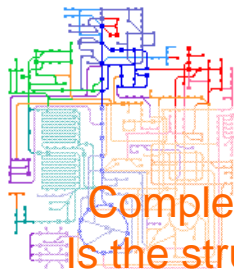


- Search engines are getting less useful than before
 - too many hits
 - not all relevant
 - not "organized"



- Take a search on p53. You get >300k hits or some # like that on MEDLINE
- It is not feasible for anyone to go thru all of that to find what he wants! And this problem is growing bigger as MEDLINE doubles every 1-2 year.
- Need to organize the db and/or the search results to make it easier for users to understand or to browse the results

Understandability:
Is our databases or search results understandable?



Completeness:
Is the structure of our databases expressive enough to capture critical information explicitly?

- Take a key paper such as the Kohn paper that summarises current knowledge on p53
- Is there a (semi)structured db that is able to capture all info in that paper explicitly?
- How well does this (semi-)structured database generalize to other similar type of papers?

Extract Entities & Annotations

Nomenclature loosely followed stimulates transcription of the **TAL1/SCL gene** and **phosphorylation** of its **protein products**.

Activation of the **TAL1 gene**, originally identified through its involvement by a **recurrent** **translocation**, is the most frequent **molecular lesion** recognized in **T-cell acute lymphoblastic leukemia**. Members of this gene contain the **basic-helix-loop-helix motif** characteristic of a **transcription factor** that bind to the **canonical DNA sequence CANN TG** as **protein products** by **hybrid cells** in vivo and in **chemical-induced erythroleukemia cell** lines. **In vivo** studies of the gene might regulate aspects of **erythroid differentiation**. Since the terminal events in **erythroid differentiation** are regulated by the **glycoprotein hormone erythropoietin (Epo)**, we investigated whether the expression or activity of the **TAL1 gene** and its **protein products** were affected by **Epo** in **erythroid cells** derived from mice infected with an **anemia-inducing strain of Friend virus (FVA cells)**. **Epo** elicited a rapid, dose-related increase in **TAL1 mRNA** by increasing **transcription** of the gene and **inducible TAL1 DNA binding activity** was identified in **FVA cell lines**. **TAL1 mRNA** levels rapidly decayed despite accumulating **mRNA** and **protein**. Induction of **DNA binding activity** was associated temporally with **Epo-induced phosphorylation** of nuclear **TAL1 protein**. These results indicate that **TAL1** acts at both **transcriptional and posttranscriptional levels** on the **TAL1 gene** and establish a link between **Epo signaling mechanisms** and **transcription factors** involved in the differentiation of **diverse cell lineages**.

Freq use of conjunction and disjunction in bio names with multiple bio-entity names sharing one head noun

Long descriptive names

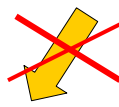
Names of genes & proteins used interchangeably

Zhou et al., *BioCreativE*, 2004

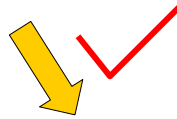
Extract Relationships

- Sentences describing relationships tend to be complicated
- Domain knowledge is often needed for interaction template filling


"c-Abl tyrosine kinase activity is blocked by pRb, which binds to the c-Abl kinase domain."



(pRb *inhibit* tyrosine kinase activity-of c-Abl)
(pRb *bind-to* c-Abl kinase domain)



(pRb *inhibit* tyrosine kinase activity-of c-Abl) *is-caused-by* (pRb *bind-to* c-Abl at kinase domain)



The screenshot shows the XTractor web interface. At the top, there is a navigation bar with buttons for Home, Create Queries, Queries, Explore, Search, Profile, My Contacts, and Invite Friend. Below this is a search bar and a list of categories: All, Protein, Disease, Drug, Process, Key Relationships, Categories, and Tag. The main content area displays a list of retrieved results, including: Adenoma, Pleomorphic; Adrenal Cortex Neoplasms; Adrenal Cortex Neoplasms; Adrenocortical Carcinoma; Bile Duct Neoplasms; Breast Neoplasms; Carcinoma, Hepatocellular; Carcinoma, Non-Small-Cell Lung; Carcinoma, Small Cell; Carcinoma, Squamous Cell; Carcinoma, Squamous Cell; Carcinoma, Transitional Cell; Cholangiocarcinoma; Colonic Neoplasms; Colorectal Neoplasms; Cushing Syndrome; Emphysema; Endometrial Neoplasms; Glioblastoma; Glioma; Head and Neck Neoplasms; Hepatitis B, Chronic; Laryngeal Neoplasms; Leukoplakia, Oral; Lung Neoplasms; Melanoma; Melanoma; Ovarian Neoplasms; Ovarian Neoplasms; Pancreatic Neoplasms; Pheochromocytoma. The footer of the interface credits 'Molecular Connections'.

The screenshot shows a web browser window with the URL `http://125.63.77.222:7070/xtractor/explore.do?tag=explore`. The page title is "Organizing Retrieval Results". The XTractor logo is visible with the tagline "Data mining simplified". A navigation menu includes "Home", "Create Queries", "Queries", "Explore", "Search", "Profile", "My Contacts", and "Invite Friend". Below the menu, there are tabs for "All", "Protein", "Disease", "Drug", "Process", "Key Relationships", "Categories", and "Tag". The "Categories" tab is active, displaying a list of relationship types: "Biomarker- Disease", "Drug- Disease", "Gene- Disease", "Gene- Drug", "Gene- Gene", "Gene- Knockout/Knockdown", "Gene- Mutation", "Gene- Pathways", and "Gene- Process". At the bottom, there are tabs for "My Sentences" and "Public Sentences". A credit line at the bottom right reads "Credit: Molecular Connections".

The slide features the NUS (National University of Singapore) logo in the top right corner. The main title is "Beyond IE & Retrieval: Predictions in Context of Biomedicine". At the bottom, there is a footer with the text "Invited keynote at SIGIR2008", the page number "14", and the copyright notice "Copyright 2008 © Limsoon Wong".

Plan

- Abductive Foundation of “Guilt by Association”
- Issue of Chance Association
- Novel Forms of Association
- Fusion of Multiple Evidence of Association
- Dichotomy of knowing two entities are in some relationship and yet not knowing what that relationship is

Abductive Foundation of “Guilt by Association”:

A Protein Function Prediction Perspective

```
PDGF-2 1 SLGSLTIAEPAMIAECKTREEVFCICRRL?DR?? 34  
p28sis 61 LARGKRSLGSLVAEPAMIAECKTRTEVFEISRRLIDRTN 100
```


Function Assignment to Protein Sequence

SPSTNRKYPPLPVDKLEEEINRRMADDNKLFREEFNALPACPIQATCEAASKEENKEKNR
YVNIILPYDHSRVHLTPVEGVPSDYINASFINGYQEKNFIAAQGPKEETVNDFWRMWE
QNTATIVMVTNLKERKECKCAQYWPDQGCWTYGNVRVSVEDVTVLVDYTVRKFCIQQVGD
VTNRKPQLITQFHFTSWPDFGVPTPIGMLKFLKVKACNPQYAGAIVVHCSAGVGRGTG
TFVVIDAMLDMMHSEKVDVYGFVSRIRAQRQMVQTDMQYVFIYQALLEHYLYGDTELE
VT

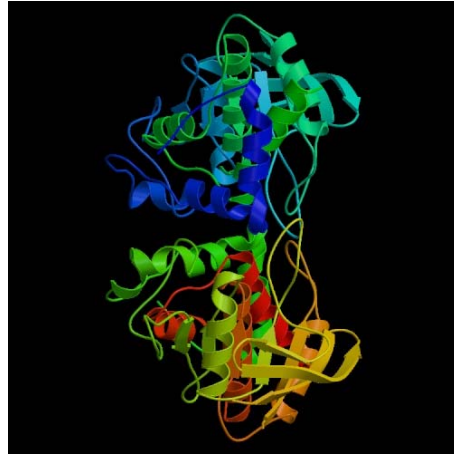
- How do we attempt to assign a function to a new protein sequence?

Parallels

- Given a protein, determine its functions
- Given a protein, find other proteins that share a common function with it
- Given a function, find all proteins having that function
- Given a document, determine what are the “things” it describes
- Given a document, find other documents that describe a common “thing” with it
- Given a “thing”, find all documents that describe that thing

A protein is a ...

- A protein is a large complex molecule made up of one or more chains of amino acids
- Protein performs a wide variety of activities in the cell

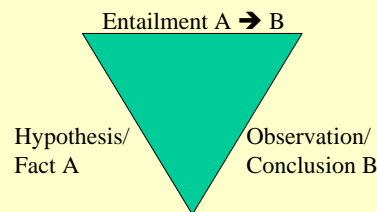


Invariant and Abductive Reasoning

- Function is determined by 3D struct of protein & environment protein is in
- Constraints imposed by 3D struct & environment give rise to “invariant” properties observed in proteins having the ancestor with that function

⇒ **Abductive reasoning**

- If those invariant properties are seen in a protein, then the protein is homolog of this protein



⇒ **“Guilt by association”**

What is the parallel of the above in IR?

“Guilt by Association”

- Compare the target sequence T with sequences S_1, \dots, S_n of known function in a database
- Determine which ones amongst S_1, \dots, 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

Sequence Alignment: Poor Example

- Poor seq alignment shows few matched positions
⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```

                60      70      80      90      100
Amicyanin      MPHNVHVFVAGVLGEAALKGPMKKEQAYSLTFTEAGTYDYHCTPHPPMRGKVVVE
                :..: . :. :.
Ascorbate Oxidase ILRGTFWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI
                70      80      90      100      110      120
    
```

No obvious match between
Amicyanin and Ascorbate Oxidase

Sequence Alignment: Good Example

- Good alignment usually has clusters of extensive matched positions
- ⇒ The two proteins are likely to be homologous

```

>gi113476732|ref|NP_108301.1| unknown protein [Mesorhizobium loti]
gi114027493|dbj|BAB53762.1| unknown protein [Mesorhizobium loti]
Length = 105

Score = 105 bits (262), Expect = 1e-22
Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1 MKPGRLASIALAIFLPMAVPAHAATIEITMENLVISPTESAKVGDITRWVVKDVFVHT 60
          MK G L ++ MA PA AATIE+T++ LV SP V AKVGDIT WVN DV AHT
Sbjct: 1 MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDITIEWVNDVVAHT 60
  
```

good match between
Amicyanin and unknown M. loti protein

Guilt by Association of Seq Similarity

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

Poor Sequence Alignment

- Poor seq alignment shows few matched positions
- ⇒ The two proteins are not likely to be homologous

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```

Amicyanin      60      70      80      90     100
MHNHVFVAVDVLGAAALGPHKRNQASLFFFAQDTVDICTEHPFRQKPVV
Ascorbate Oxidase  ILQKQTFWADGTASLDCAIMPGRFFVFVFVQVQFFFRHGLKQNRAGLVG
          70      80      90     100
  
```

No obvious match between Amicyanin and Ascorbate Oxidase

Discard this function as a candidate

Good Sequence Alignment

- Good alignment usually has clusters of extensive matched positions
- ⇒ The two proteins are likely to be homologous

```

>gi113476732|ref|NP_108301.1| unknown protein [Mesorhizobium loti]
gi114027493|dbj|BAB53762.1| unknown protein [Mesorhizobium loti]
Length = 105

Score = 105 bits (262), Expect = 1e-22
Identities = 61/106 (57%), Positives = 73/106 (68%), Gaps = 1/106 (0%)

Query: 1 MKPGRLASIALAIFLPMAVPAHAATIEITMENLVISPTESAKVGDITRWVVKDVFVHT 60
          MK G L ++ MA PA AATIE+T++ LV SP V AKVGDIT WVN DV AHT
Sbjct: 1 MKAGALIRLSWLAALALMAAPAAAATIEVTIDKLVFSPATVEAKVGDITIEWVNDVVAHT 60
  
```

good match between Amicyanin and unknown M. loti protein

Assign to *T* same function as homologs

Confirm with suitable wet experiments

Homologs obtained by BLAST

Sequences producing significant alignments:	Score (bits)	E Value
gi 14193729 cb AAK56109.1 AF332081.1 protein tyrosin phosph...	62.1	e-177
gi 126467 sp P18433 PTRA_HUMAN Protein-tyrosine phosphatase...	62.1	e-177
gi 4506303 ref NP_002827.1 protein tyrosine phosphatase, r...	62.1	e-176
gi 227294 prf I1701300A protein Tyr phosphatase	62.0	e-176
gi 18450369 ref NP_543030.1 protein tyrosine phosphatase, ...	62.1	e-176
gi 32067 emb CAA37447.1 tyrosine phosphatase precursor [Ho...	61.1	e-176
gi 285113 pir IJC1285 protein-tyrosine-phosphatase (EC 3.1....	61.9	e-176
gi 6981446 ref NP_036895.1 protein tyrosine phosphatase, r...	61.1	e-176
gi 2098414 pdb 1YFO A Chain A, Receptor Protein Tyrosine Ph...	61.1	e-174
gi 32313 emb CAA38662.1 protein-tyrosine phosphatase [Homo...	61.1	e-174
gi 450583 cb AAB04150.1 protein tyrosine phosphatase >gi 4...	60.5	e-172
gi 6679557 ref NP_033006.1 protein tyrosine phosphatase, r...	60.1	e-172
gi 483922 cb AAA17990.1 protein tyrosine phosphatase alpha	59.9	e-170

- Thus our example sequence could be a protein tyrosine phosphatase α (PTP α)

Issue of Chance Association:
A Twist in the Tale



Law of Large Numbers

- Suppose you are in a room with 365 other people
- Q: What is the prob that there is a person in the room having the same birthday as you?
- A: $1 - (364/365)^{365} = 63\%$
- Q: What is the prob that a specific person in the room has the same birthday as you?
- A: $1/365 = 0.3\%$
- Q: What is the prob that there are two persons in the room having the same birthday?
- A: 100%

Interpretation of P-value

- Seq. comparison progs, e.g. BLAST, often associate a P-value to each hit
 - P-value is interpreted as prob that a random seq has an equally good alignment
 - Suppose the P-value of an alignment is 10^{-6}
 - If database has 10^7 seqs, then you expect $10^7 * 10^{-6} = 10$ seqs in it that give an equally good alignment
- ⇒ Need to correct for database size if your seq comparison prog does not do that!

Note: $P = 1 - e^{-E}$

Lightning Does Strike Twice!

- **Roy Sullivan, a former park ranger from Virginia, was struck by lightning 7 times**
 - 1942 (lost big-toe nail)
 - 1969 (lost eyebrows)
 - 1970 (left shoulder seared)
 - 1972 (hair set on fire)
 - 1973 (hair set on fire & legs seared)
 - 1976 (ankle injured)
 - 1977 (chest & stomach burned)
- **September 1983, he committed suicide**



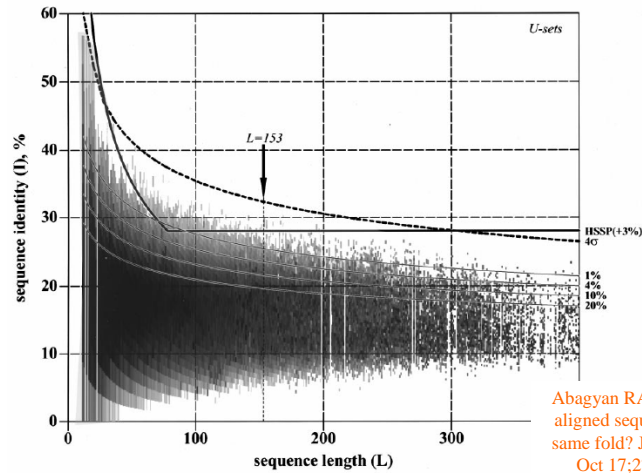
Cartoon: Ron Hipschman
Data: David Hand

Effect of Seq Compositional Bias

- **One fourth of all residues in protein seqs occur in regions with biased amino acid composition**
- **Alignments of two such regions achieves high score purely due to segment composition**
- ⇒ **While it is worth noting that two proteins contain similar low complexity regions, they are best excluded when constructing alignments**

Source: NCBI

Effect of Sequence Length



Parallels

- **P-value and E-value**
- **Ranking measures?**
 - Different concepts
 - Not necessarily same effect
- **Compositionally biased regions**
- **Stop words?**
- **Length, conserved site, transitive assignment, and other caveats**
- **???**

Novel Forms of Associations



Important Unsolved Challenge



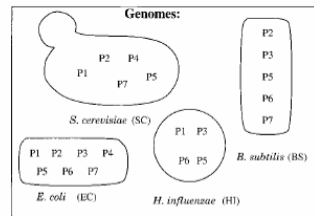
- **What if there is no useful seq homolog?**
- **Guilt by other types of association!**
 - Domain modeling (e.g., HMMPFAM)
 - ✓ Similarity of phylogenetic profiles
 - ✓ Similarity of dissimilarities (e.g., SVM-PAIRWISE)
 - Similarity of subcellular co-localization & other physico-chemico properties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

Guilt by Association of Phylogenetic Profile



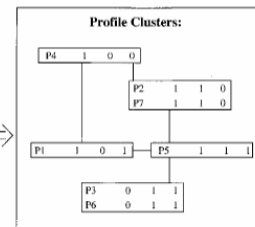
- A protein is not alone when performing its biological function

⇒ Gene (and hence proteins) with identical patterns of occurrence across phyla tend to function together



Phylogenetic Profile:

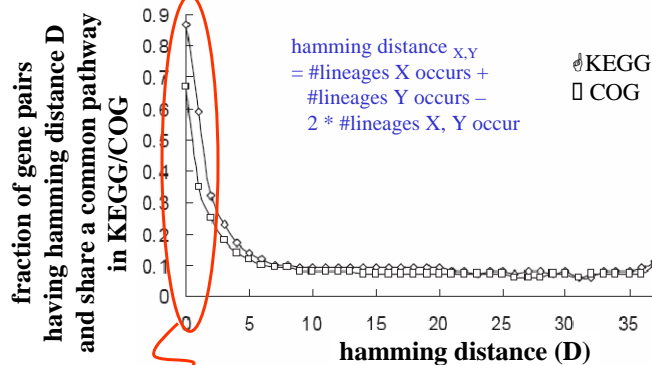
	EC	SC	BS	HI
P1	1	0	1	0
P2	1	1	0	0
P3	0	1	1	0
P4	1	0	0	0
P5	1	1	1	1
P6	0	1	1	1
P7	1	1	0	0



Conclusion: P2 and P7 are functionally linked.
P3 and P6 are functionally linked.

Phylogenetic Profiling: Evidence

Wu et al., *Bioinformatics*, 19:1524--1530, 2003



- Proteins having low hamming distance (thus highly similar phylogenetic profiles) tend to share common pathways
- Exercise: Why do proteins having high hamming distance also have this behaviour?

Guilt by Association of Dissimilarities



Differences of "unknown" to other fruits are same as "apple" to other fruits



"unknown" is an "apple"!

	Orange ₁	Banana ₁	...
Apple ₁	Color = red vs orange Skin = smooth vs rough Size = small vs small Shape = round vs round	Color = red vs yellow Skin = smooth vs smooth Size = small vs small Shape = round vs oblong	...
Orange ₂	Color = orange vs orange Skin = rough vs rough Size = small vs small Shape = round vs round	Color = orange vs yellow Skin = rough vs smooth Size = small vs small Shape = round vs oblong	...
Unknown ₁	Color = red vs orange Skin = smooth vs rough Size = small vs small Shape = round vs round	Color = red vs yellow Skin = smooth vs smooth Size = small vs small Shape = round vs oblong	...
...

SVM-Pairwise Framework

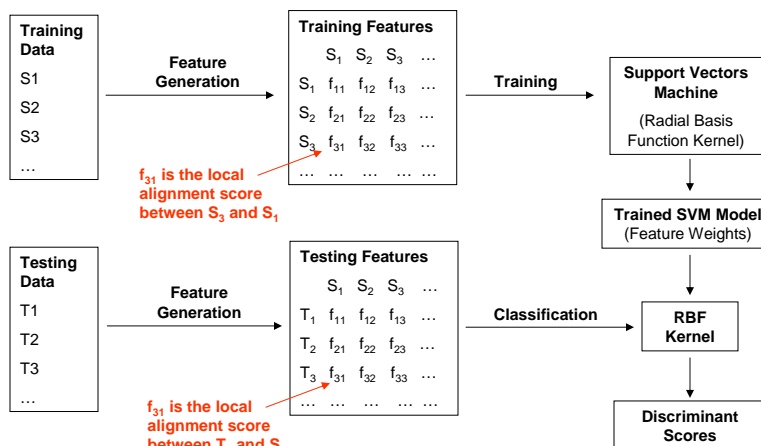
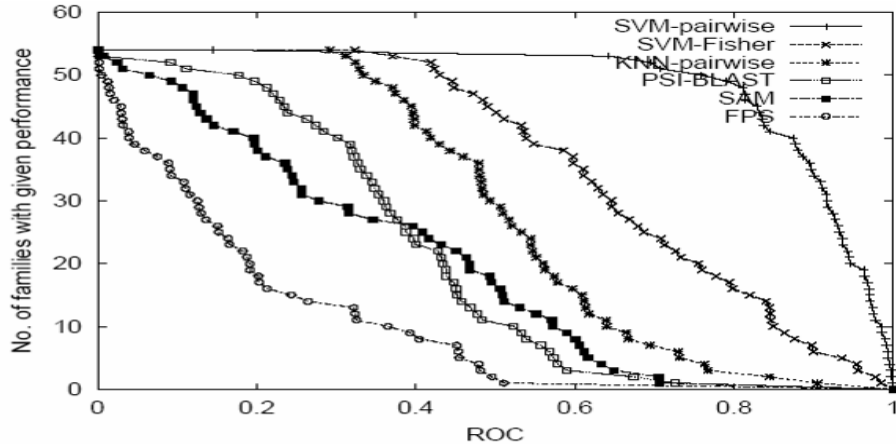


Image credit: Kenny Chua

Performance of SVM-Pairwise

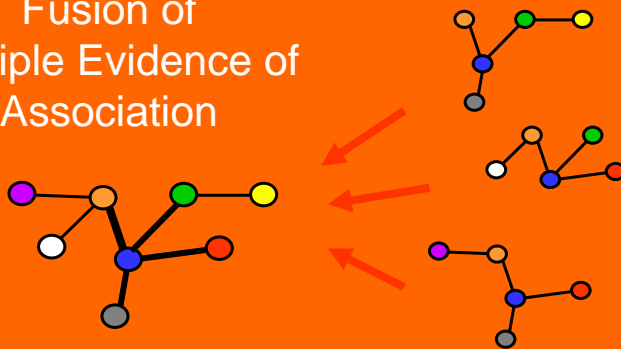


- **ROC: The area under the curve derived from plotting true positives as a function of false positives for various thresholds**

Parallels

- **Guilt by association of genome phylogenetic profile**
- **Guilt by association of dissimilarities**
- **Two “things” are “equivalent”**
 - If they occur in same documents
 - If they are mutually exclusive in documents they occur in?
- **Is this a new association concept in IR?**

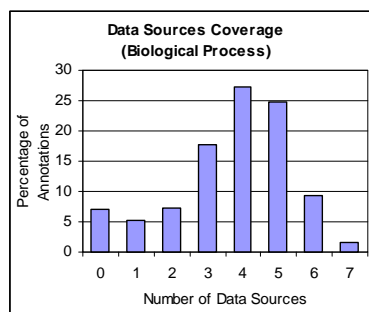
Fusion of Multiple Evidence of Association



Coverage of Data Sources



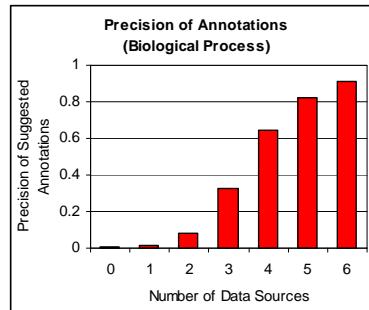
- **> 90% of known protein annotations are suggested from at least one data source**
- **A large percentage (80%) of known protein annotations are suggested by 3 or more data sources**



% of known annotations suggested by different number of data sources

Confidence of Overlapping Evidence

- **Protein annotations suggested by more data sources are more likely to be correct**
- **Protein annotations suggested by 4 or more data sources are correct > 60% of the time**



Fraction of annotations suggested from different number of datasources that are correct

Difficulties w/ Information Fusion

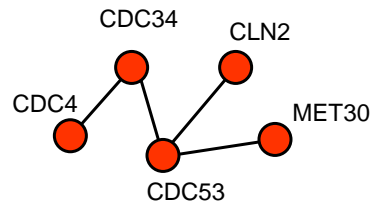
- **Differences in nature**
 - E.g., sequence homology vs PPI are very different relationships
- **Differences in reliability**
 - E.g., noisy datasets such as Y2H PPI and gene expression
- **Differences in scoring metrics**
 - E.g., E-Score from BLAST vs Pearson correlation between expression profiles

Integrated Weighted Averaging – Step 1

- Model a data source as undirected graph $G = \langle V, E \rangle$

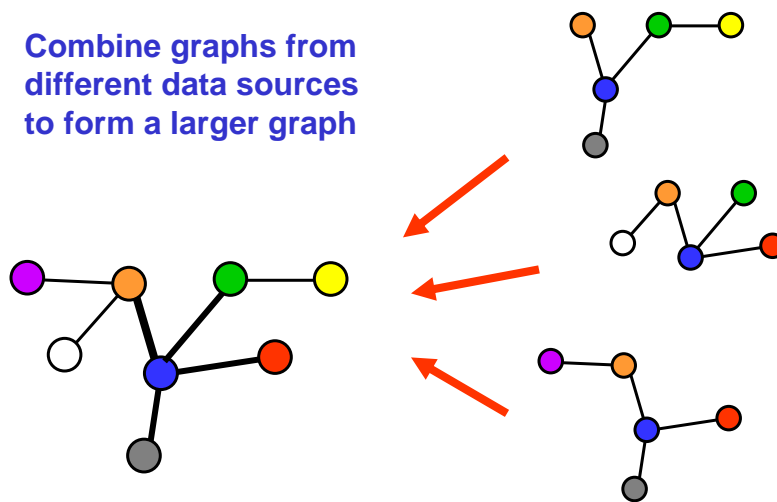
- V is a set of vertices; each vertex reps a protein

- E is a set of edges; each edge (u, v) reps a relationship (e.g. seq similarity, interaction) betw proteins u and v



Integrated Weighted Averaging – Step 2

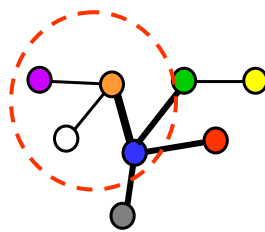
- Combine graphs from different data sources to form a larger graph



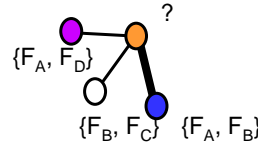
Integrated Weighted Averaging – Step 3

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

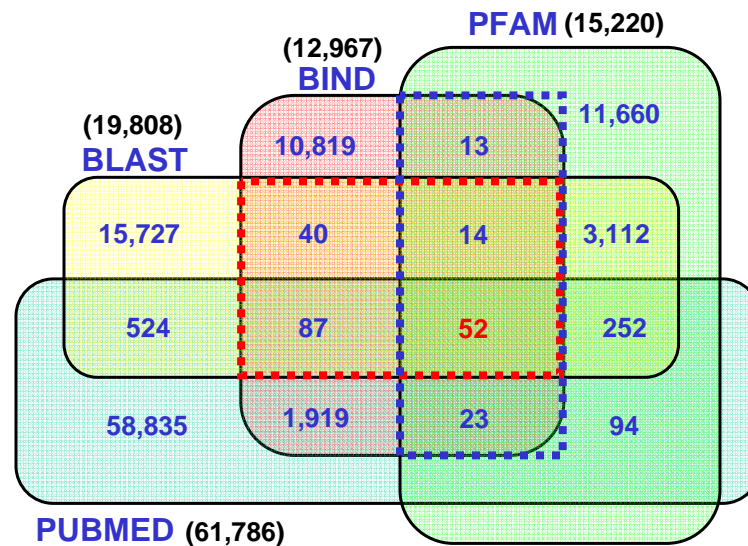
$$r_{u,v,f} = 1 - \prod_{k \in D_{u,v}} (1 - p(k, f))$$

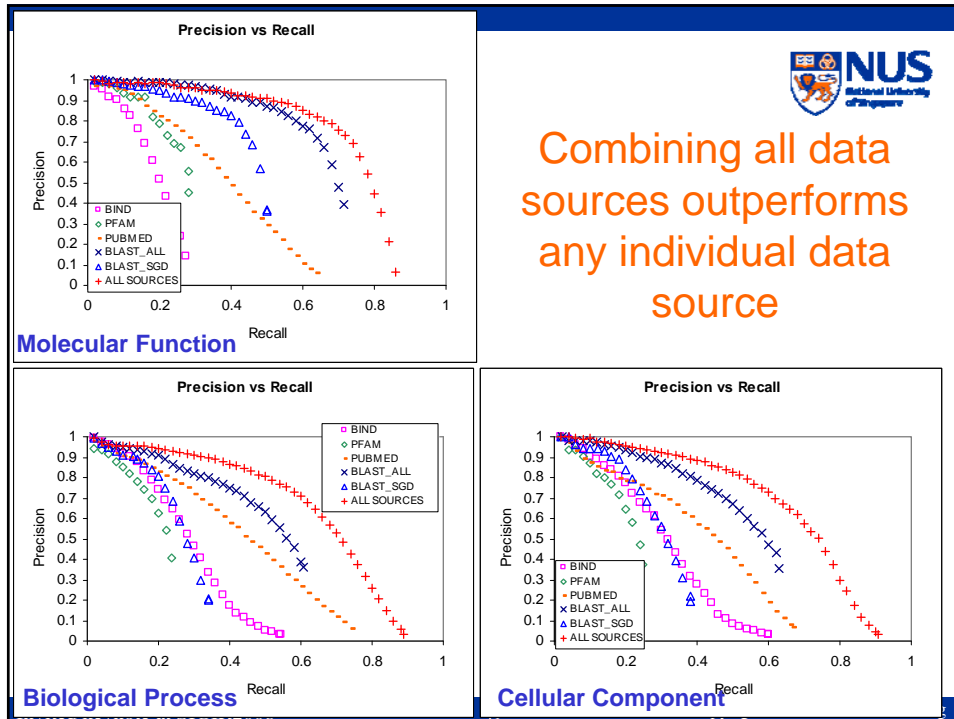


$$S_f(u) = \frac{\sum_{v \in N_u} (e_f(v) \times r_{u,v,f})}{1 + \sum_{v \in N_u} r_{u,v,f}}$$



Associations from Multiple Data Sources





Parallels

- **Given a protein, determine its function via “guilt by association of multiple evidence types”**
 - Sensitivity is improved
 - Precision is improved
- **Given a document, determine the “thing” it talks about by multimodal IR**
 - The more media types talking about the same “thing” that this document talks about, the higher chance to determine accurately what this “thing” is
- **Is this a new problem concept for multimodal IR?**

Invited keynote at SIGIR2008 50 Copyright 2008 © Limsoon Wong

Dichotomy of knowing two entities are in some relationship and yet not knowing what that relationship is



Good Source for Evidence of “Guilt”



- **IWA and other methods perform better when**
 - Graph has fewer nodes with no annotated neighbours
 - Unannotated nodes in graph are connected to a greater number of annotated nodes
- ⇒ **A data source that is able to contribute a large # of association edges connecting to annotated proteins should provide the greatest gain in prediction accuracy**

Gabow et al, BMC Bioinformatics, 9:198, 2008

	Source	Yeast MIPS
Annotation Terms		
Percentage Unknown Nodes	PPI	85
	GENETIC	23
	COLIT	14
Percentage Connected to ≥ 1 Unknown	PPI	2
	GENETIC	31
	COLIT	14
Percentage Only Surrounded by Unknowns	PPI	4
	GENETIC	0.9
	COLIT	0.08
Percent Edges Connecting Nodes Sharing Function	PPI	37
	GENETIC	48
	COLIT	80

But only 2% of nodes derived from Medline are unannotated

And more disturbingly...

So Medline abstracts seem like a good source of association info

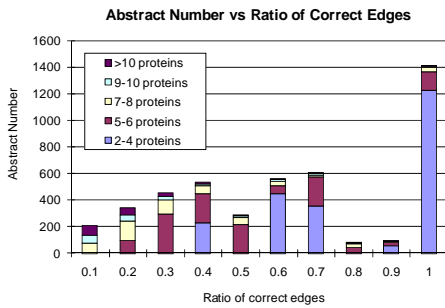
Invited keynote at SIGIR2008 53 Copyright 2008 © Limsoon Wong

A Dichotomy

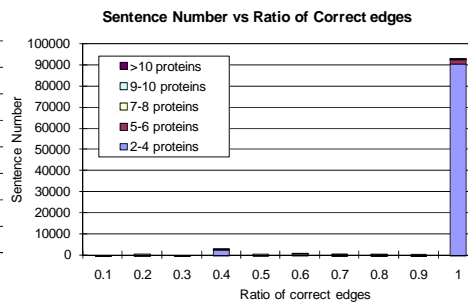
- 80% of protein pairs that co-occurred in “enough” Medline abstracts have 1 function in common
- Yet precision-recall curve is far below that expectation!

Invited keynote at SIGIR2008 54 Copyright 2008 © Limsoon Wong

Co-occurrence at too coarse a level?



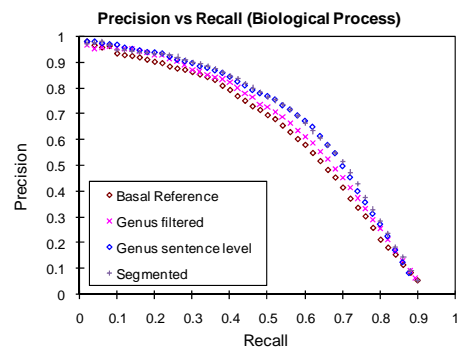
Half of the abstracts has less than 50% of their co-mentioned protein pairs sharing some function



Nearly all sentences have all their co-mentioned protein pairs sharing some function

Dichotomy Remains

- Nearly all sentences have all their co-mentioned protein pairs sharing some function
- Yet the precision-recall curve is far below expectation!



Dichotomy Explained?



- An edge in the graph simply means an association between two proteins
- ⇒ The two proteins have a function in common

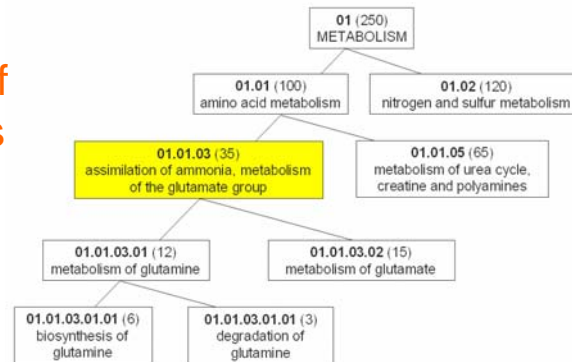
- But each protein may have several functions
- Don't know which one is the function they have in common

Parallel in IR: We know two documents are related, but we don't know what that relationship is

Closing Remarks



Shades of Meanings



- **GO Functional Annotation**
 - Hierarchical
 - 3 Namespaces (molecular function, biological process, cellular component)

“Guilt by Association” as a Search Principle



- **Founded on abduction based on invariants of biology & physics**
- **Many forms of associations**
 - Sequence
 - Phylogenetic profile
 - Dissimilarities
- **Fusion of Multiple Evidence of Association**
- **Issue of Chance Association**
- **Dichotomy of knowing two entities are in some relationship and yet not knowing what that relationship is**
- **Shades of meanings**

Acknowledgements



- See-Kiong Ng
- Molecular Connections Pvt Ltd

- Hon Nian CHUA
- Zhihui LI