Building Gene Networks by Information Extraction, Cleansing, & Integration

Limsoon Wong

Plan

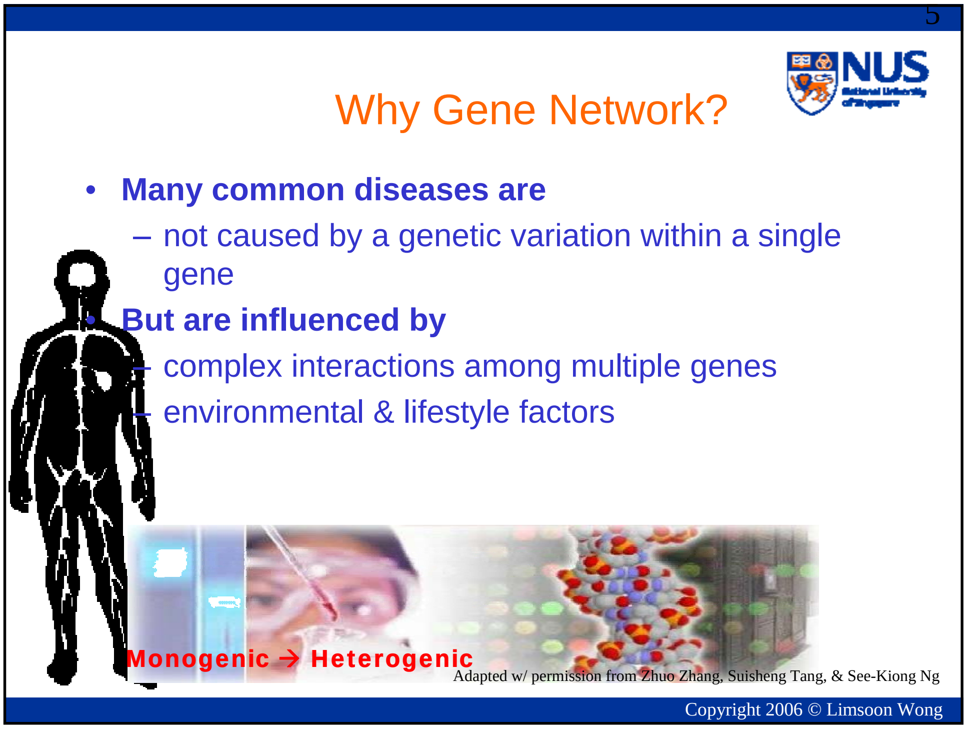
- **Motivation for study of gene network**
- **Example Efforts at I2R**
  - Disease Pathweaver
  - Dragon ERG Solution
- **Technical Challenges Involved**
  - Name entity recognition
  - Co-reference resolution
  - Protein interaction extraction
- **Discussion on issues**

Motivation

Some Useful Information Sources

- **Disease-centric resources**
  - OMIM [NCBI OMIM, 2004]
  - Gene2Disease [Perez-Iratxeta et al., *Nature*, 2002]
  - MedGene [Hu et al., *J. Proteome Res.*, 2003]
- **Emphasized direct gene-disease relationships**
  - Provide lists of disease-related genes
  - Do not provide info on gene-gene interactions & their networks
- **Related interaction resources**
  - KEGG [Kanehisa, *NAR*, 2000]
- **Manually constructed protein interaction networks**
- **Mostly metabolic pathways and few disease pathways**
  - Only 7 disease pathways

Adapted w/ permission from Zhou Zhang, Suisheng Tang, & See-Kiong Ng

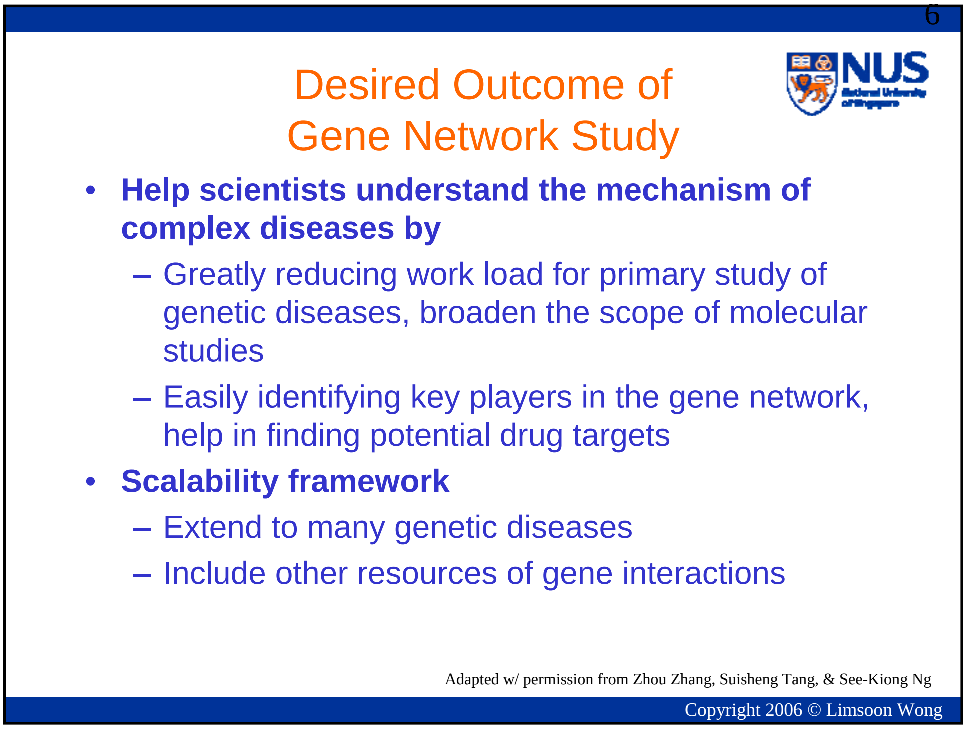


Why Gene Network?

- **Many common diseases are**
  - not caused by a genetic variation within a single gene
- **But are influenced by**
  - complex interactions among multiple genes
  - environmental & lifestyle factors

Monogenic → Heterogenic

Adapted w/ permission from Zhuo Zhang, Suisheng Tang, & See-Kiong Ng



Desired Outcome of Gene Network Study

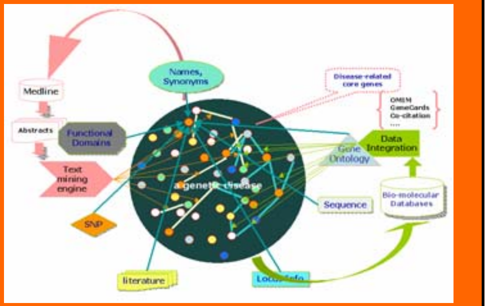
- **Help scientists understand the mechanism of complex diseases by**
  - Greatly reducing work load for primary study of genetic diseases, broaden the scope of molecular studies
  - Easily identifying key players in the gene network, help in finding potential drug targets
- **Scalability framework**
  - Extend to many genetic diseases
  - Include other resources of gene interactions

Adapted w/ permission from Zhou Zhang, Suisheng Tang, & See-Kiong Ng

Some Gene Network Study Efforts at I^{2}R

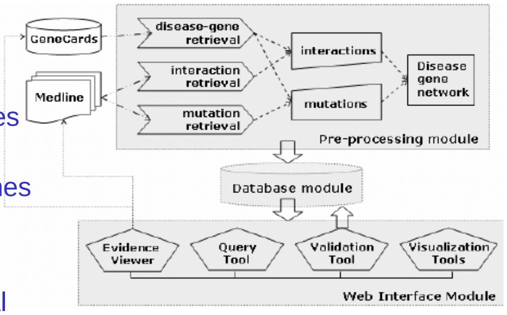
Disease Pathweaver

Dragon ERG Solution



Disease Pathweaver, Zhang et al. *APBC 2005*

- **Automatic constructing disease pathways**
  - Identify core genes
  - Mine info on core genes
  - Construct interaction network betw core genes
- **Data sources:**
  - Online literature
  - High-thru’put biological data



Adapted w/ permission from Zhou Zhang, Suisheng Tang, & See-Kiong Ng

Disease Pathweaver:

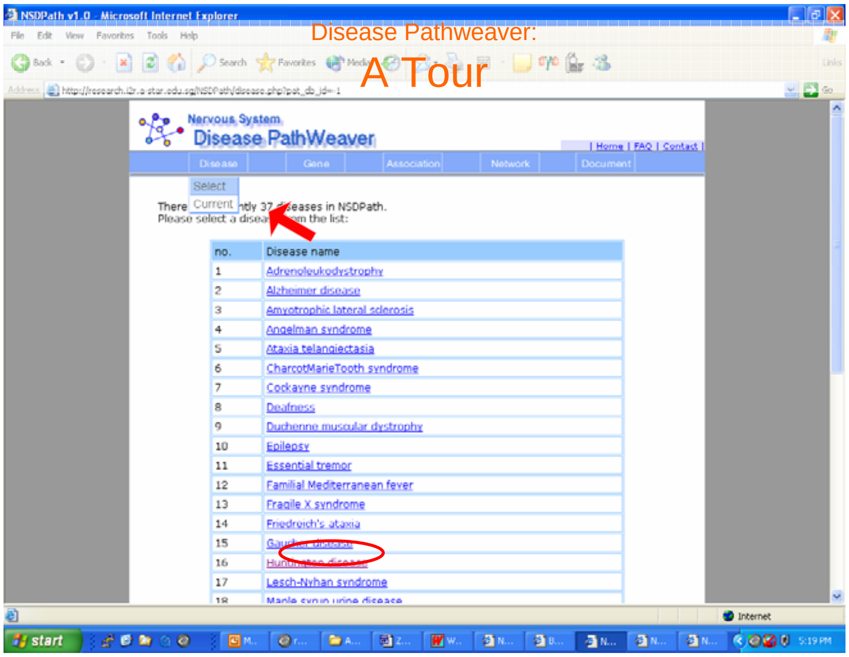
The Portal

- **Portal for human nervous diseases gene networks**
  - http://research.i2r.a-star.edu/NSDPath
- **Statistics**
  - 37 Human Nervous System Disorders
  - 7 ~ 60 core genes per disease
  - 2 ~ 320 core interactions per disease

Adapted w/ permission from Zhou Zhang, Suisheng Tang, & See-Kiong Ng

Disease Pathweaver:

A Tour

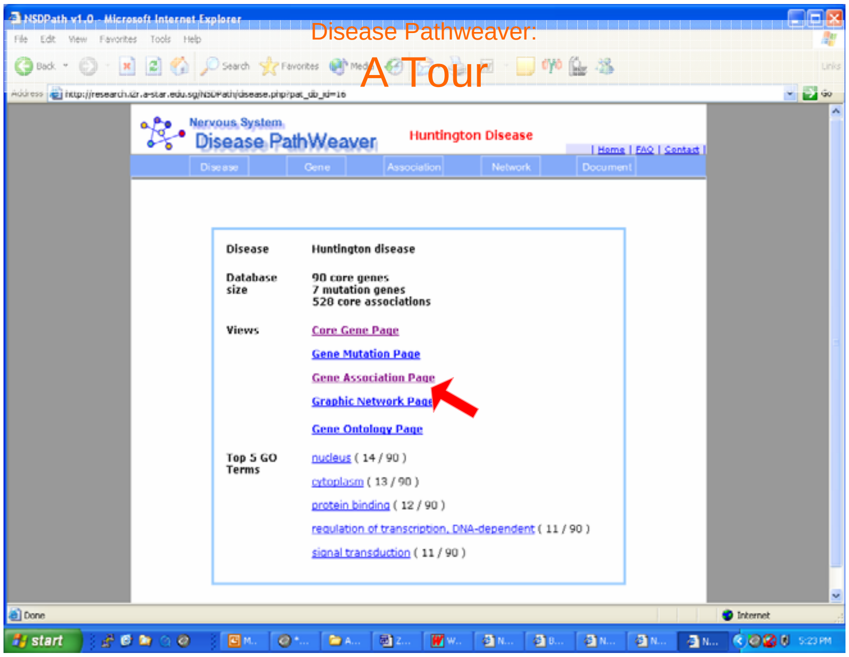


There Current ntly 37 di eases in NSDPath.
Please select a diseas om the list:

no.	Disease name
1	Adrenoleukodystrophy
2	Alzheimer disease
3	Amyotrophic lateral sclerosis
4	Angelman syndrome
5	Ataxia telangiectasia
6	CharcotMarieTooth syndrome
7	Cockayne syndrome
8	Deafness
9	Duchenne muscular dystrophy
10	Epilepsy
11	Essential tremor
12	Familial Mediterranean fever
13	Fragile X syndrome
14	Friedreich's ataxia
15	Gaucher disease
16	Huntington disease
17	Lesch-Nyhan syndrome
18	Maple syrup urine disease

Copyright © 2006 by Limsoon Wong. Adapted w/ permission from Zhuo Zhang, Suisheng Tang, & See-Kiong Ng

Disease Pathweaver: A Tour



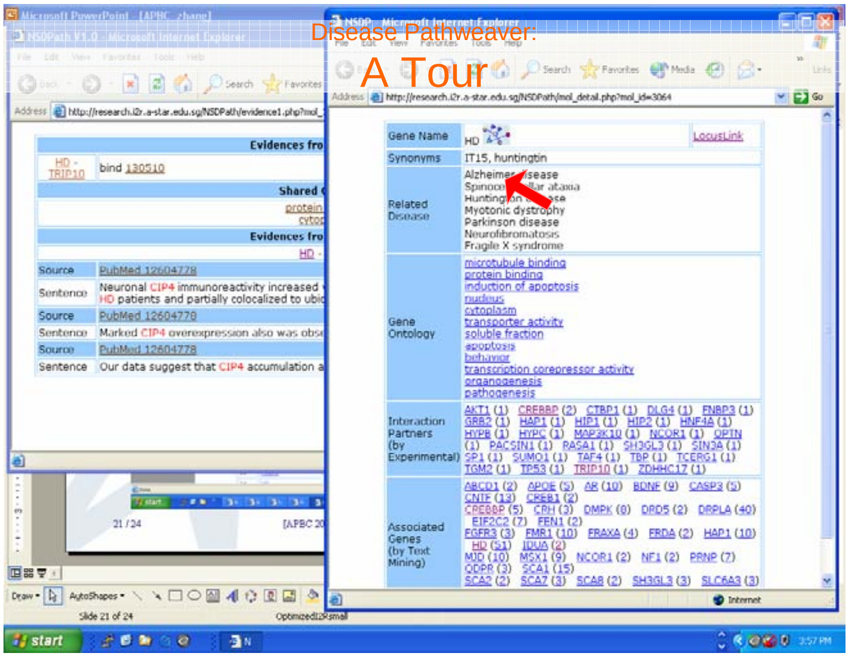
http://research.i2r.a-star.edu.sg/NSDPath/disease.php?pat_db_id=16

Nervous System Disease PathWeaver — Huntington Disease

| Home | FAQ | Contact |

Disease	Huntington disease
Database size	90 core genes 7 mutation genes 520 core associations
Views	Core Gene Page Gene Mutation Page Gene Association Page Graphic Network Page Gene Ontology Page
Top 5 GO Terms	nucleus (14 / 90) cytoplasm (13 / 90) protein binding (12 / 90) regulation of transcription, DNA-dependent (11 / 90) signal transduction (11 / 90)

Disease Pathweaver: A Tour



http://research.i2r.a-star.edu.sg/NSDPath/evidence1.php?mol_

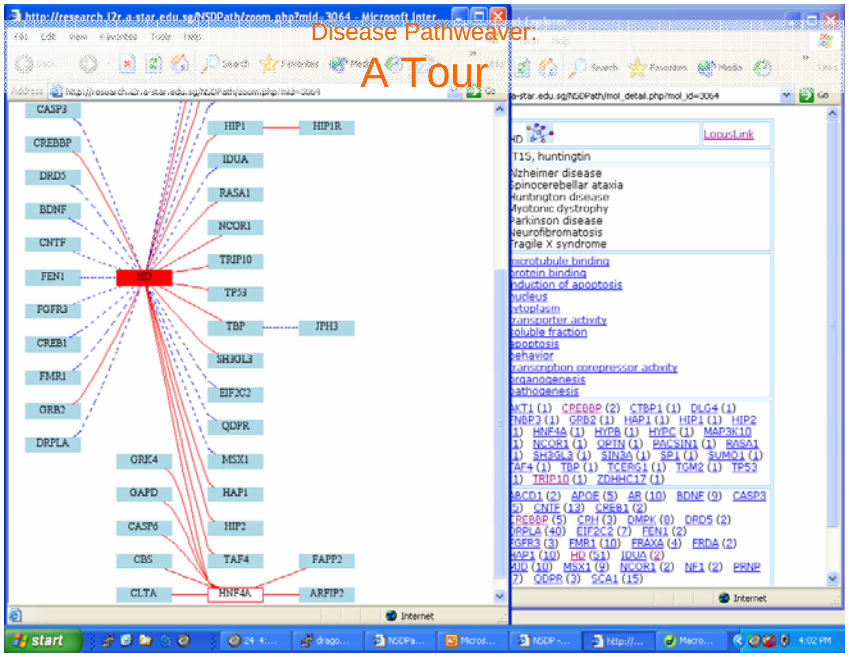
Evidences fro	
HD - TRIP10	bind 130510
Shared (	
	protein cytop
Evidences fro	
	HD -
Source	PubMed 12604778
Sentence	Neuronal CIP4 immunoreactivity increased HD patients and partially colocalized to ubiq
Source	PubMed 12604778
Sentence	Marked CIP4 overexpression also was obse
Source	PubMed 12604778
Sentence	Our data suggest that CIP4 accumulation a

http://research.i2r.a-star.edu.sg/NSDPath/mol_detail.php?mol_id=3064

Gene Name	HD — LocusLink
Synonyms	IT15, huntingtin
Related Disease	Alzheimer disease Spinocerebellar ataxia Huntington disease Myotonic dystrophy Parkinson disease Neurofibromatosis Fragile X syndrome
Gene Ontology	microtubule binding protein binding induction of apoptosis nucleus cytoplasm transporter activity soluble fraction apoptosis behavior transcription corepressor activity organogenesis pathogenesis
Interaction Partners (by Experimental)	AKT1 (1) CREBBP (2) CTBP1 (1) DLG4 (1) FNBP3 (1) GRB2 (1) HAP1 (1) HIP1 (1) HIP2 (1) HNF4A (1) HYPB (1) HYPC (1) MAP3K10 (1) NCOR1 (1) OPTN (1) PACSIN1 (1) RASA1 (1) SH3GL3 (1) SIN3A (1) SP1 (1) SUMO1 (1) TAF4 (1) TBP (1) TCERG1 (1) TGM2 (1) TP53 (1) TRIP10 (1) ZDHHC17 (1)
Associated Genes (by Text Mining)	ABCD1 (2) APOE (5) AR (10) BDNF (9) CASP3 (5) CNTF (13) CREB1 (2) CREBBP (5) CRH (3) DMPK (8) DRD5 (2) DRPLA (40) EIF2C2 (7) FEN1 (2) FGFR3 (3) FMR1 (10) FRAXA (4) FRDA (2) HAP1 (10) HD (51) IDUA (2) MJD (10) MSX1 (9) NCOR1 (2) NF1 (2) PRNP (7) QDPR (3) SCA1 (15) SCA2 (2) SCA7 (3) SCA8 (2) SH3GL3 (3) SLC6A3 (3)

Disease Pathweaver:

A Tour


http://research.i2r.a-star.edu.sg/NSDPath/zoom.php?mid=3064 - Microsoft Inter...
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://research.i2r.a-star.edu.sg/NSDPath/zoom.php?mid=3064
Go
CASP3
CREBBP
DRD5
BDNF
CNTF
FEN1
FGFR3
CREB1
FMR1
GRB2
DRPLA
HD
HIP1
HIP1R
IDUA
RASA1
NCOR1
TRIP10
TP53
TBP
JPH3
SH3GL3
EIF2C2
QDPR
MSX1
HAP1
HIP2
TAF4
FAPP2
HNF4A
ARFIP2
GRK4
GAPD
CASP6
CBS
CLTA
Internet
start
Search Favorites Media
a-star.edu.sg/NSDPath/mol_detail.php?mol_id=3064
Go
Internet

HD LocusLink

T15, huntingtin

lzheimer disease
Spinocerebellar ataxia
Huntington disease
Myotonic dystrophy
Parkinson disease
Neurofibromatosis
Fragile X syndrome

microtubule binding
protein binding
nduction of apoptosis
nucleus
ytoplasm
ransporter activity
soluble fraction
apoptosis
behavior
ranscription corepressor activity
organogenesis
pathogenesis

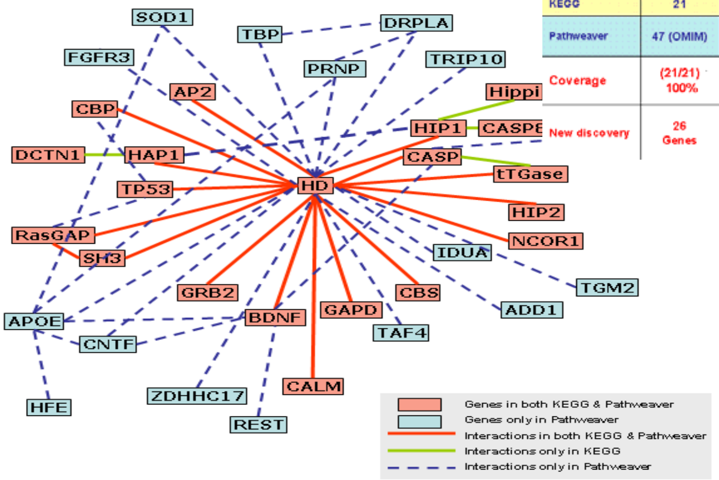
AKT1 (1) CREBBP (2) CTBP1 (1) DLG4 (1) NBP3 (1) GRB2 (1) HAP1 (1) HIP1 (1) HIP2 1) HNF4A (1) HYPB (1) HYPC (1) MAP3K10 1) NCOR1 (1) OPTN (1) PACSIN1 (1) RASA1 1) SH3GL3 (1) SIN3A (1) SP1 (1) SUMO1 (1) TAF4 (1) TBP (1) TCERG1 (1) TGM2 (1) TP53 1) TRIP10 (1) ZDHHC17 (1)

ABCD1 (2) APOE (5) AR (10) BDNF (9) CASP3 5) CNTF (13) CREB1 (2) CREBBP (5) CRH (3) DMPK (8) DRD5 (2) DRPLA (40) EIF2C2 (7) FEN1 (2) FGFR3 (3) FMR1 (10) FRAXA (4) FRDA (2) HAP1 (10) HD (51) IDUA (2) MJD (10) MSX1 (9) NCOR1 (2) NF1 (2) PRNP 7) QDPR (3) SCA1 (15)

Disease Pathweaver:

DPW vs KEGG on Huntington Disease

	Core Genes	Gene Associations
KEGG	21	19
Pathweaver	47 (OMIM)	59
Coverage	(21/21) 100%	(12/19) 63.1%
New discovery	26 Genes	47 (37 True Positive) 78.7% TP rate

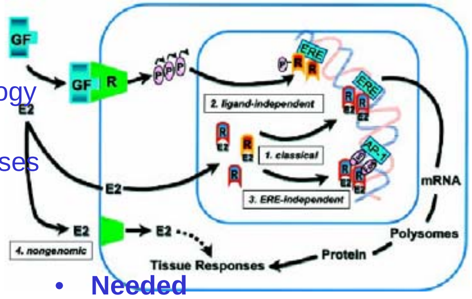


🟥 Genes in both KEGG & Pathweaver
🟦 Genes only in Pathweaver
— Interactions in both KEGG & Pathweaver
— Interactions only in KEGG
- - - Interactions only in Pathweaver

Adapted w/ permission from Zhuo Zhang, Suisheng Tang, & See-Kiong Ng

Estrogen-Responsive Genes

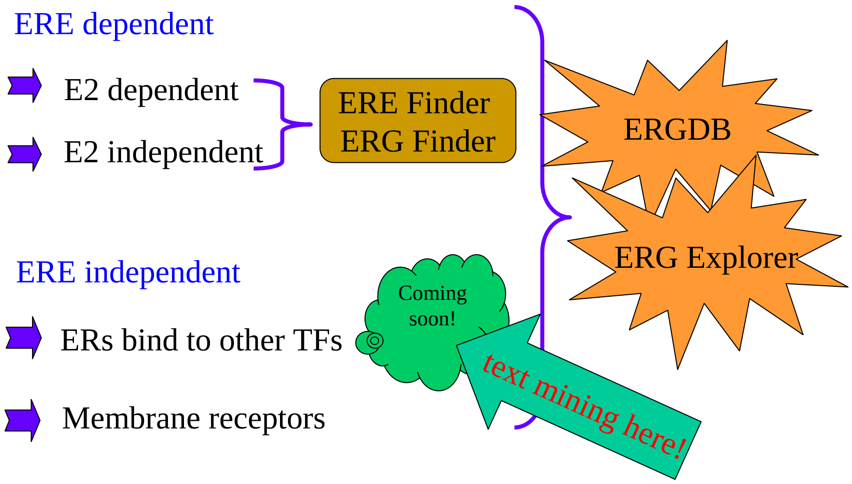
- **Why**
  - Affects human physiology in many aspects
  - Related to many diseases
  - Widely used in clinic
- **Challenges**
  - Multiple pathways
  - Difficult to predict ERE
  - Many estrogen-responsive genes but only a few are well-studied
  - Difficult to keep up w/ speed of knowledge accumulation



- **Needed**
  - Tools to predict ERE & estrogen-responsive genes
  - Database of useful info
  - Systems to predict impt regulatory units, associate gene functions, & generate global view of gene network

Adapted w/ permission from Suisheng Tang & Vlad Bajic

Dragon ERG Solution



ERE dependent

- E2 dependent
- E2 independent

ERE Finder
ERG Finder

ERGDB

ERG Explorer

ERE independent

- ERs bind to other TFs
- Membrane receptors

Coming soon!

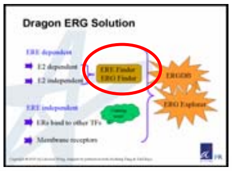
text mining here!

Adapted w/ permission from Suisheng Tang & Vlad Bajic

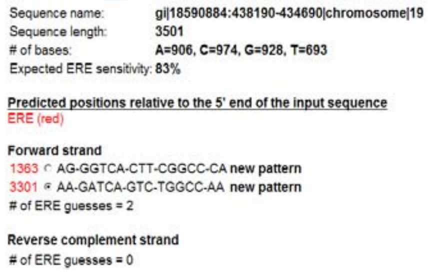
Dragon ERG Solution:

Dragon ERE Finder,

Bajic et al, *NAR*, 2003



- ➢Predict functional ERE in genomic DNA
- ➢One prediction in 13.3k bp
- ➢Allow further analysis

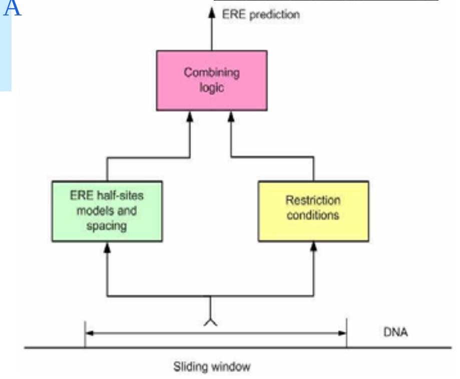


```
Sequence name:          gi|18590884:438190-434690|chromosome|19
Sequence length:        3501
# of bases:             A=906, C=974, G=928, T=693
Expected ERE sensitivity: 83%

Predicted positions relative to the 5' end of the input sequence
ERE (red)

Forward strand
1363 ○ AG-GGTCA-CTT-CGGCC-CA new pattern
3301 ● AA-GATCA-GTC-TGGCC-AA new pattern
# of ERE guesses = 2

Reverse complement strand
# of ERE guesses = 0
```

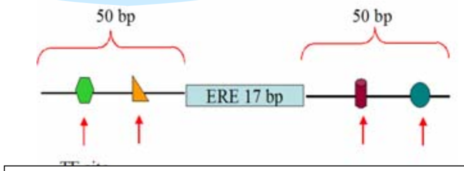


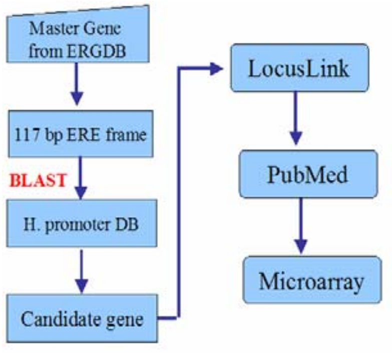
http://sdmc.i2r.a-star.edu.sg/ERE-V2/index

Dragon ERG Solution:

Dragon Estrogen-Responsive Gene Finder, Tang et al, *NAR*, 2004a

- ➢Only for human genome
- ➢Using 117 bp ERE frame
- ➢Evaluated by PubMed & microarray data



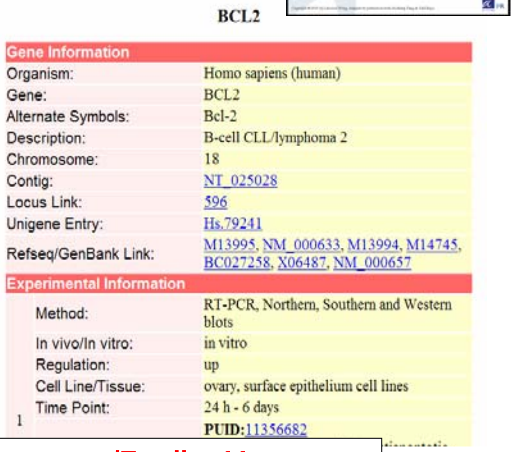


http://sdmc.i2r.a-star.edu.sg/DRAGON/ERGP1_0

Dragon ERG Solution:

Dragon Estrogen-Responsive Genes Database, Tang et al, *NAR*, 2004b

- ➢Contains >1000 genes
- ➢Manually curated
- ➢Basic gene info
- ➢Experimental evidence
- ➢Full set of references
- ➢ERE sites annotated



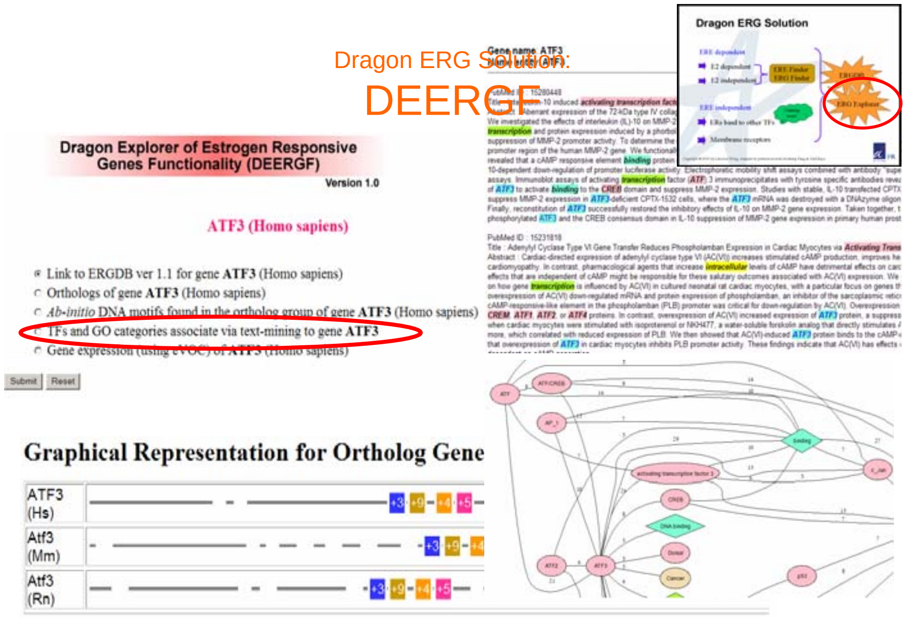
BCL2

Gene Information	
Organism:	Homo sapiens (human)
Gene:	BCL2
Alternate Symbols:	Bcl-2
Description:	B-cell CLL/lymphoma 2
Chromosome:	18
Contig:	NT_025028
Locus Link:	596
Unigene Entry:	Hs.79241
Refseq/GenBank Link:	M13995, NM_000633, M13994, M14745, BC027258, X06487, NM_000657
Experimental Information	
Method:	RT-PCR, Northern, Southern and Western blots
In vivo/In vitro:	in vitro
Regulation:	up
Cell Line/Tissue:	ovary, surface epithelium cell lines
1 Time Point:	24 h - 6 days
	PUID:11356682

http://sdmc.i2r.a-star.edu.sg/promoter/Ergdb-v11

Dragon ERG Solution:

DEERGF



Dragon Explorer of Estrogen Responsive Genes Functionality (DEERGF)

Version 1.0

ATF3 (Homo sapiens)

- ● Link to ERGDB ver 1.1 for gene **ATF3** (Homo sapiens)
- ○ Orthologs of gene **ATF3** (Homo sapiens)
- ○ *Ab-initio* DNA motifs found in the ortholog group of gene **ATF3** (Homo sapiens)
- ○ TFs and GO categories associate via text-mining to gene **ATF3**
- ○ Gene expression (using eVOC) of **ATF3** (Homo sapiens)

Graphical Representation for Ortholog Gene

Gene name: ATF3

PubMed ID : 15280448

Title : Interleukin-10 induced activating transcription fact

Abstract : Aberrant expression of the 72-kDa type IV collag

We investigated the effects of interleukin (IL)-10 on MMP-2

transcription and protein expression induced by a phorbol

suppression of MMP-2 promoter activity. To determine the

promoter region of the human MMP-2 gene. We functional

revealed that a cAMP responsive element binding protein

10-dependent down-regulation of promoter luciferase activity. Electrophoretic mobility shift assays combined with antibody "supe

assays. Immunoblot assays of activating transcription factor (ATF) 3 immunoprecipitates with tyrosine specific antibodies revea

of ATF3 to activate binding to the CREB domain and suppress MMP-2 expression. Studies with stable, IL-10 transfected CPTX

suppress MMP-2 expression in ATF3-deficient CPTX-1532 cells, where the ATF3 mRNA was destroyed with a DNAzyme oligon

Finally, reconstitution of ATF3 successfully restored the inhibitory effects of IL-10 on MMP-2 gene expression. Taken together, t

phosphorylated ATF3 and the CREB consensus domain in IL-10 suppression of MMP-2 gene expression in primary human prost

PubMed ID : 15231818

Title : Adenylyl Cyclase Type VI Gene Transfer Reduces Phospholamban Expression in Cardiac Myocytes via Activating Trans

Abstract : Cardiac-directed expression of adenylyl cyclase type VI (AC(VI)) increases stimulated cAMP production, improves he

cardiomyopathy. In contrast, pharmacological agents that increase intracellular levels of cAMP have detrimental effects on card

effects that are independent of cAMP might be responsible for these salutary outcomes associated with AC(VI) expression. We

on how gene transcription is influenced by AC(VI) in cultured neonatal rat cardiac myocytes, with a particular focus on genes th

overexpression of AC(VI) down-regulated mRNA and protein expression of phospholamban, an inhibitor of the sarcoplasmic retic

cAMP-responsive-like element in the phospholamban (PLB) promoter was critical for down-regulation by AC(VI). Overexpression

CREM, ATF1, ATF2, or ATF4 proteins. In contrast, overexpression of AC(VI) increased expression of ATF3 protein, a suppress

when cardiac myocytes were stimulated with isoproterenol or NKH477, a water-soluble forskolin analog that directly stimulates A

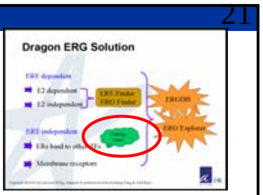
more, which correlated with reduced expression of PLB. We then showed that AC(VI)-induced ATF3 protein binds to the cAMP-

that overexpression of ATF3 in cardiac myocytes inhibits PLB promoter activity. These findings indicate that AC(VI) has effects

dependent on cAMP generation.

Dragon ERG Solution:

Case-Specific TF relation networks, Pan et al, *NAR*, 2004



Dragon ERG Solution

ERE dependent

- E2 dependent
- E2 independent

ERE Finder
ERG Finder

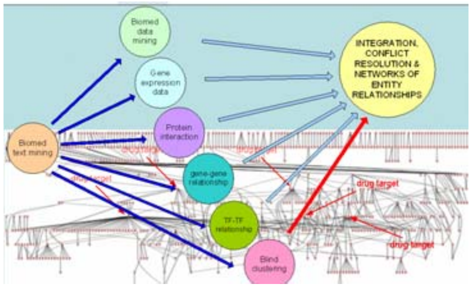
ERGDB

ERE independent

- ERs bind to other TFs
- Membrane receptors

ERG Explorer

- Analyse abstracts
- Stemming, POS tagging
- Use ANNs, SVM, discriminant analysis
- Simplified rules for sentence analysis
- Constraints on the forms of sentences
- Sensitivity ~75%
- Precision ~82%



- Produce reports & direct links to PubMed docs, & graphical presentations of entity links

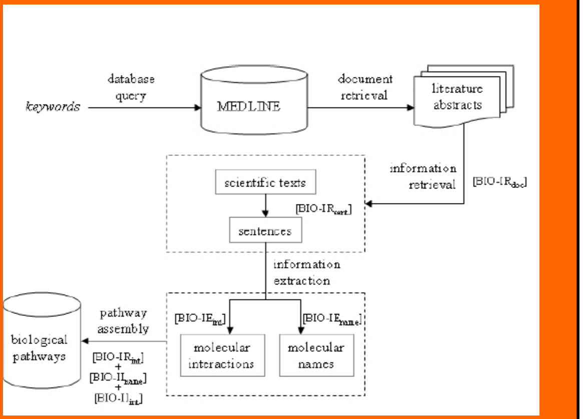
Adapted w/ permission from Vlad Bajic

Technical Challenges

Named entity recognition

Co-reference resolution

Data cleansing



keywords

database query

MEDLINE

document retrieval

literature abstracts

information retrieval

[BIO-IR$_{doc}$]

scientific texts

[BIO-IR$_{sent}$]

sentences

information extraction

[BIO-IE$_{int}$]

[BIO-IE$_{name}$]

molecular interactions

molecular names

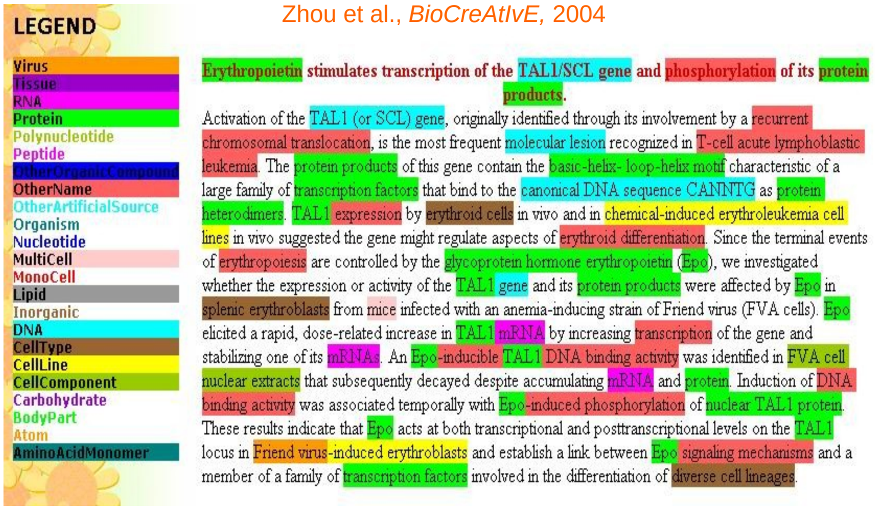
pathway assembly

biological pathways

[BIO-IR$_{int}$] + [BIO-II$_{name}$] + [BIO-II$_{int}$]

Bio Entity Name Recognition,

Zhou et al., *BioCreAtIvE,* 2004



LEGEND

- Virus
- Tissue
- RNA
- Protein
- Polynucleotide
- Peptide
- OtherOrganicCompound
- OtherName
- OtherArtificialSource
- Organism
- Nucleotide
- MultiCell
- MonoCell
- Lipid
- Inorganic
- DNA
- CellType
- CellLine
- CellComponent
- Carbohydrate
- BodyPart
- Atom
- AminoAcidMonomer

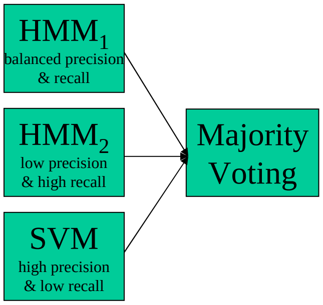
Erythropoietin stimulates transcription of the TAL1/SCL gene and phosphorylation of its protein products.

Activation of the TAL1 (or SCL) gene, originally identified through its involvement by a recurrent chromosomal translocation, is the most frequent molecular lesion recognized in T-cell acute lymphoblastic leukemia. The protein products of this gene contain the basic-helix- loop-helix motif characteristic of a large family of transcription factors that bind to the canonical DNA sequence CANNTG as protein heterodimers. TAL1 expression by erythroid cells in vivo and in chemical-induced erythroleukemia cell lines in vivo suggested the gene might regulate aspects of erythroid differentiation. Since the terminal events of erythropoiesis are controlled 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 splenic erythroblasts 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 stabilizing one of its mRNAs. An Epo-inducible TAL1 DNA binding activity was identified in FVA cell nuclear extracts that subsequently 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 Epo acts at both transcriptional and posttranscriptional levels on the TAL1 locus in Friend virus-induced erythroblasts and establish a link between Epo signaling mechanisms and a member of a family of transcription factors involved in the differentiation of diverse cell lineages.

Adapted w/ permission from Jian Su

Bio Entity Name Recognition: Ensemble Classification Approach

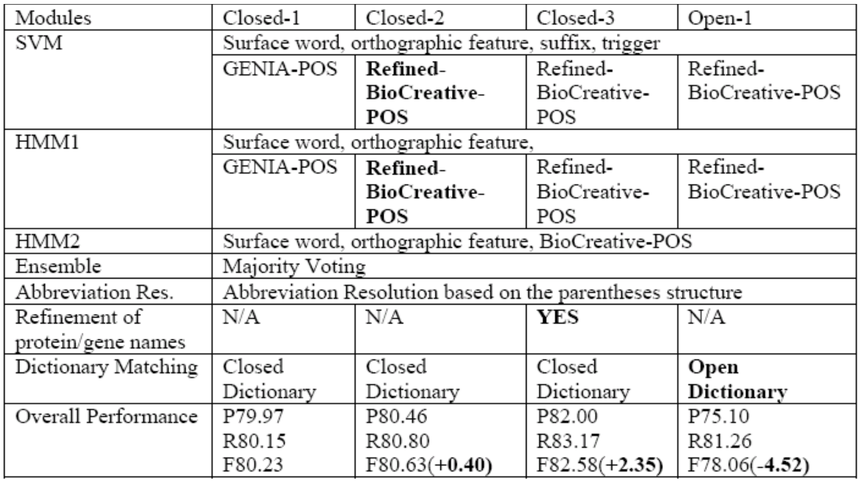
- **Features considered**
  - orthographic, POS, morphologic, surface word, trigger words (TW$_1$: receptor, enhancer, etc. TW$_2$: activation, stimulation, etc.)
- **SVM**
  - Context of 7 words
  - Each word gives 5 features, plus its position
- **HMMs**
  - 3 features used (orthographic, POS, surface word)
  - HMM$_1$ & HMM$_2$ use POS taggers trained on diff corpora



HMM$_1$
balanced precision & recall
HMM$_2$
low precision & high recall
SVM
high precision & low recall
Majority Voting

Bio Entity Name Recognition:

Performance at *BioCreAtIvE* 2004

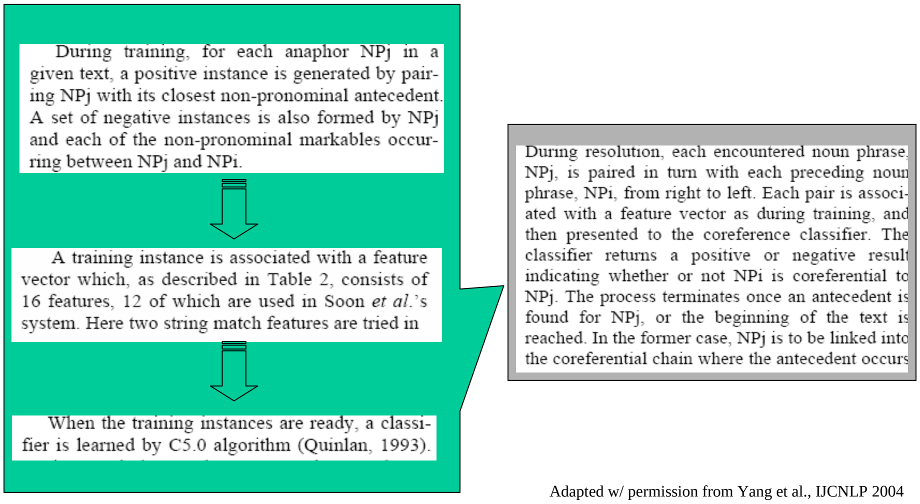


Modules	Closed-1	Closed-2	Closed-3	Open-1
SVM	Surface word, orthographic feature, suffix, trigger			
	GENIA-POS	**Refined-BioCreative-POS**	Refined-BioCreative-POS	Refined-BioCreative-POS
HMM1	Surface word, orthographic feature,			
	GENIA-POS	**Refined-BioCreative-POS**	Refined-BioCreative-POS	Refined-BioCreative-POS
HMM2	Surface word, orthographic feature, BioCreative-POS			
Ensemble	Majority Voting			
Abbreviation Res.	Abbreviation Resolution based on the parentheses structure			
Refinement of protein/gene names	N/A	N/A	**YES**	N/A
Dictionary Matching	Closed Dictionary	Closed Dictionary	Closed Dictionary	**Open Dictionary**
Overall Performance	P79.97 R80.15 F80.23	P80.46 R80.80 F80.63(**+0.40**)	P82.00 R83.17 F82.58(**+2.35**)	P75.10 R81.26 F78.06(**-4.52**)

Adapted w/ permission from Zhou et al., BioCreAtIve 2004

Co-Reference Resolution,

Yang et al., *IJCNLP*, 2004



During training, for each anaphor NPj in a given text, a positive instance is generated by pairing NPj with its closest non-pronominal antecedent. A set of negative instances is also formed by NPj and each of the non-pronominal markables occurring between NPj and NPi.

A training instance is associated with a feature vector which, as described in Table 2, consists of 16 features, 12 of which are used in Soon *et al.*'s system. Here two string match features are tried in

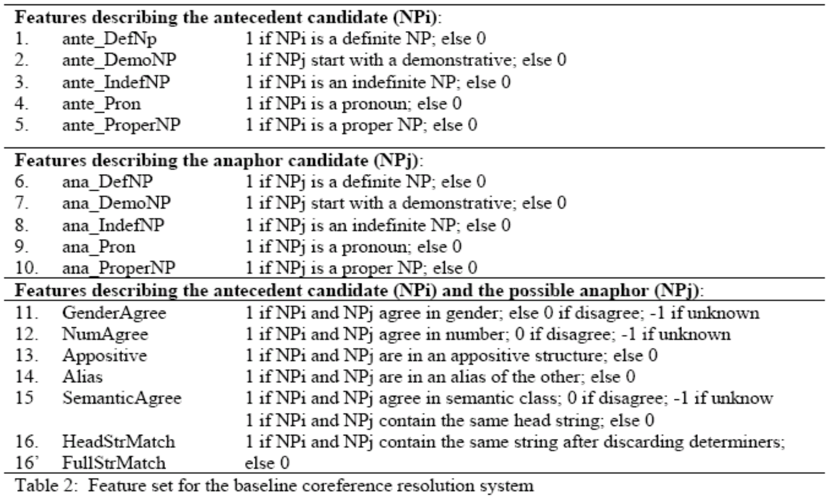
When the training instances are ready, a classifier is learned by C5.0 algorithm (Quinlan, 1993).

During resolution, each encountered noun phrase, NPj, is paired in turn with each preceding noun phrase, NPi, from right to left. Each pair is associated with a feature vector as during training, and then presented to the coreference classifier. The classifier returns a positive or negative result indicating whether or not NPi is coreferential to NPj. The process terminates once an antecedent is found for NPj, or the beginning of the text is reached. In the former case, NPj is to be linked into the coreferential chain where the antecedent occurs

Adapted w/ permission from Yang et al., IJCNLP 2004

Co-Reference Resolution:

Baseline Features Used



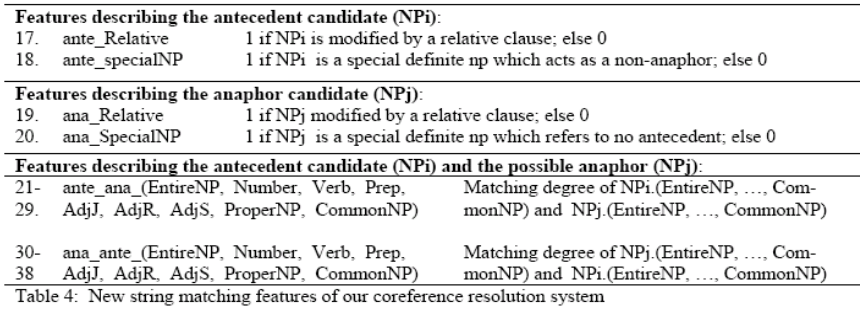
Features describing the antecedent candidate (NPi):		
1.	ante_DefNp	1 if NPi is a definite NP; else 0
2.	ante_DemoNP	1 if NPj start with a demonstrative; else 0
3.	ante_IndefNP	1 if NPi is an indefinite NP; else 0
4.	ante_Pron	1 if NPi is a pronoun; else 0
5.	ante_ProperNP	1 if NPi is a proper NP; else 0
Features describing the anaphor candidate (NPj):		
6.	ana_DefNP	1 if NPj is a definite NP; else 0
7.	ana_DemoNP	1 if NPj start with a demonstrative; else 0
8.	ana_IndefNP	1 if NPj is an indefinite NP; else 0
9.	ana_Pron	1 if NPj is a pronoun; else 0
10.	ana_ProperNP	1 if NPj is a proper NP; else 0
Features describing the antecedent candidate (NPi) and the possible anaphor (NPj):		
11.	GenderAgree	1 if NPi and NPj agree in gender; else 0 if disagree; -1 if unknown
12.	NumAgree	1 if NPi and NPj agree in number; 0 if disagree; -1 if unknown
13.	Appositive	1 if NPi and NPj are in an appositive structure; else 0
14.	Alias	1 if NPi and NPj are in an alias of the other; else 0
15	SemanticAgree	1 if NPi and NPj agree in semantic class; 0 if disagree; -1 if unknow
		1 if NPi and NPj contain the same head string; else 0
16.	HeadStrMatch	1 if NPi and NPj contain the same string after discarding determiners;
16’	FullStrMatch	else 0

Table 2: Feature set for the baseline coreference resolution system

Adapted w/ permission from Yang et al., IJCNLP 2004

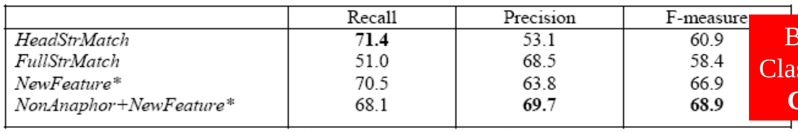
Co-Reference Resolution:

New Features Used & Performance



Features describing the antecedent candidate (NPi):		
17.	ante_Relative	1 if NPi is modified by a relative clause; else 0
18.	ante_specialNP	1 if NPi is a special definite np which acts as a non-anaphor; else 0
Features describing the anaphor candidate (NPj):		
19.	ana_Relative	1 if NPj modified by a relative clause; else 0
20.	ana_SpecialNP	1 if NPj is a special definite np which refers to no antecedent; else 0
Features describing the antecedent candidate (NPi) and the possible anaphor (NPj):		
21- 29.	ante_ana_(EntireNP, Number, Verb, Prep, AdjJ, AdjR, AdjS, ProperNP, CommonNP)	Matching degree of NPi.(EntireNP, …, CommonNP) and NPj.(EntireNP, …, CommonNP)
30- 38	ana_ante_(EntireNP, Number, Verb, Prep, AdjJ, AdjR, AdjS, ProperNP, CommonNP)	Matching degree of NPj.(EntireNP, …, CommonNP) and NPi.(EntireNP, …, CommonNP)

Table 4: New string matching features of our coreference resolution system

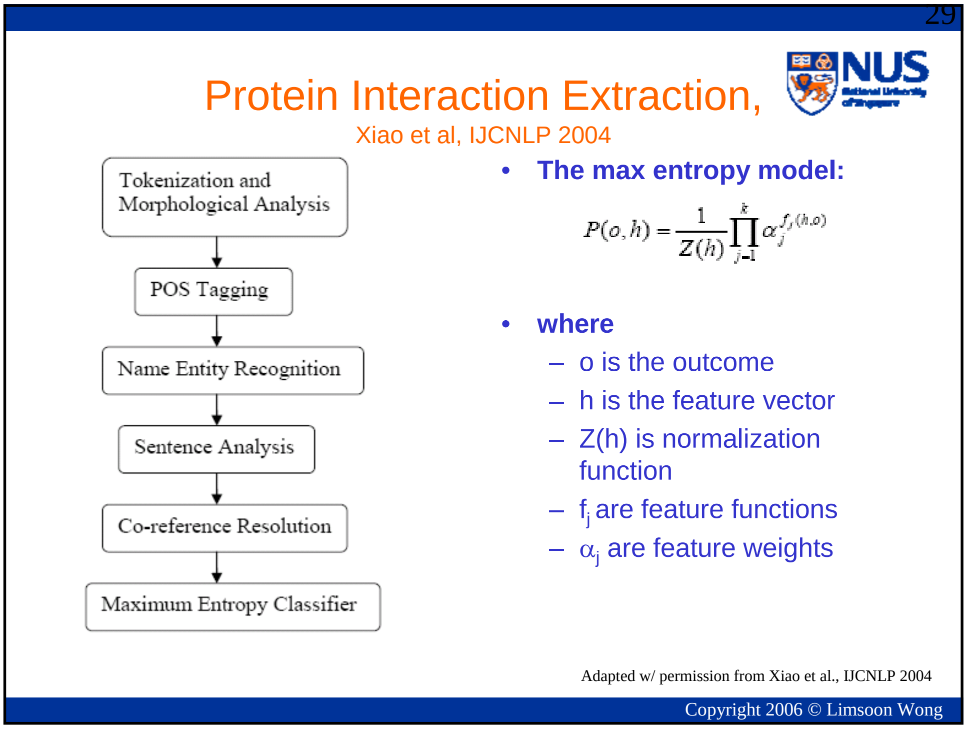


	Recall	Precision	F-measure
HeadStrMatch	**71.4**	53.1	60.9
FullStrMatch	51.0	68.5	58.4
*NewFeature**	70.5	63.8	66.9
*NonAnaphor+NewFeature**	68.1	**69.7**	**68.9**

Table 5: Experimental results on the Medline data set using C5.0 (the *ed systems use *ContainRation* metric with *Binary* weighting scheme)

Base Classifier: **C5.0**

Adapted w/ permission from Yang et al., IJCNLP 2004



Protein Interaction Extraction,

Xiao et al, IJCNLP 2004

Tokenization and Morphological Analysis

POS Tagging

Name Entity Recognition

Sentence Analysis

Co-reference Resolution

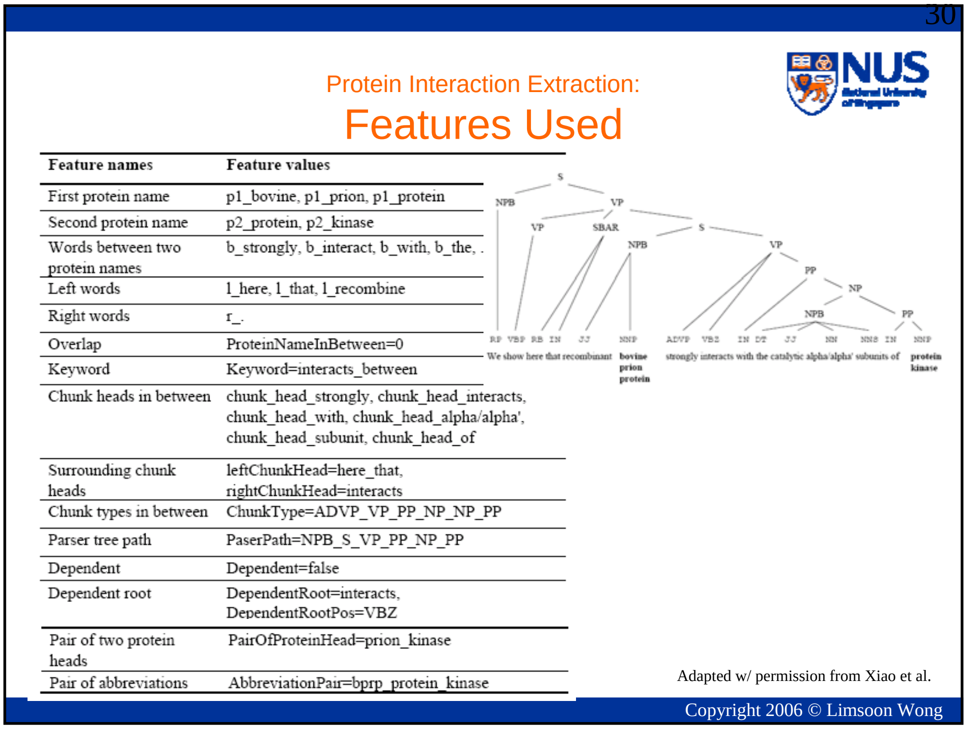
Maximum Entropy Classifier

- **The max entropy model:**

$$P(o,h) = \frac{1}{Z(h)} \prod_{j=1}^{k} \alpha_j^{f_j(h,o)}$$

- **where**
  - o is the outcome
  - h is the feature vector
  - Z(h) is normalization function
  - f_j are feature functions
  - α_j are feature weights

Adapted w/ permission from Xiao et al., IJCNLP 2004



Protein Interaction Extraction:

Features Used

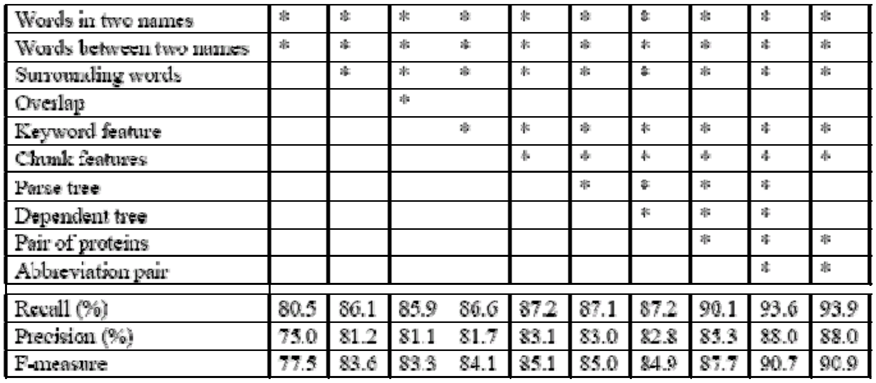
Feature names	Feature values
First protein name	p1_bovine, p1_prion, p1_protein
Second protein name	p2_protein, p2_kinase
Words between two protein names	b_strongly, b_interact, b_with, b_the, .
Left words	l_here, l_that, l_recombine
Right words	r_.
Overlap	ProteinNameInBetween=0
Keyword	Keyword=interacts_between
Chunk heads in between	chunk_head_strongly, chunk_head_interacts, chunk_head_with, chunk_head_alpha/alpha', chunk_head_subunit, chunk_head_of
Surrounding chunk heads	leftChunkHead=here_that, rightChunkHead=interacts
Chunk types in between	ChunkType=ADVP_VP_PP_NP_NP_PP
Parser tree path	PaserPath=NPB_S_VP_PP_NP_PP
Dependent	Dependent=false
Dependent root	DependentRoot=interacts, DependentRootPos=VBZ
Pair of two protein heads	PairOfProteinHead=prion_kinase
Pair of abbreviations	AbbreviationPair=bprp_protein_kinase

S
NPB
VP
VP
SBAR
S
NPB
VP
PP
NP
NPB
PP
RP VBP RB IN JJ NNP ADVP VBZ IN DT JJ NN NNS IN NNP
We show here that recombinant
bovine prion protein
strongly interacts with the catalytic alpha/alpha' subunits of
protein kinase

Adapted w/ permission from Xiao et al.

Protein Interaction Extraction:

Performance on IEPA Corpus



Words in two names	*	*	*	*	*	*	*	*	*	*
Words between two names	*	*	*	*	*	*	*	*	*	*
Surrounding words		*	*	*	*	*	*	*	*	*
Overlap			*							
Keyword feature				*	*	*	*	*	*	*
Chunk features					*	*	*	*	*	*
Parse tree						*	*	*	*	
Dependent tree							*	*	*	
Pair of proteins								*	*	*
Abbreviation pair									*	*
Recall (%)	80.5	86.1	85.9	86.6	87.2	87.1	87.2	90.1	93.6	93.9
Precision (%)	75.0	81.2	81.1	81.7	83.1	83.0	82.8	85.3	88.0	88.0
F-measure	77.5	83.6	83.3	84.1	85.1	85.0	84.9	87.7	90.7	90.9

IEPA = Interaction Extraction Performance Assessment

Adapted w/ permission from Xiao et al.

Some Interesting Issues in Constructing Gene/Protein Networks

33

Issues

- **Sound:**
  - Is the contents of our databases correct?
- **Complete:**
  - Is the structure of our databases expressive enough to capture critical information explicitly?
- **Understandable:**
  - Is our databases or search results understandable?
- **Other issues relating to NLP/IE**

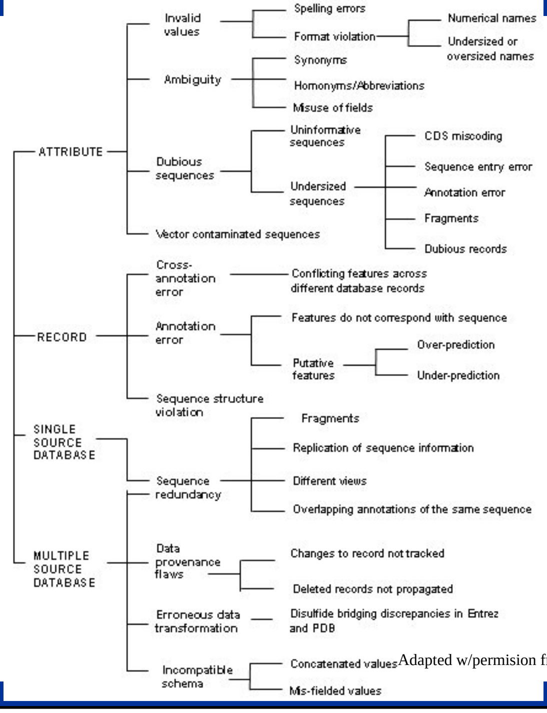
Soundness: Is the content of our databases correct?

Sources of Errors, Koh et al, *DBiDB*, 2005

- **11 types & 28 subtypes of data artifacts**
  - Critical artifacts (vector contaminated sequences, duplicates, sequence structure violations)
  - Non-critical artifacts (misspellings, synonyms)
- **> 20,000 seq records in public contain artifacts**
- **Identification of these artifacts are impt for accurate knowledge discovery**

- **Sources of artifacts**
  - Diverse sources of data
    - **Repeated submissions of seqs to db's**
    - **Cross-updating of db's**
  - Data Annotation
    - **Db's have diff ways for data annotation**
    - **Data entry errors can be introduced**
    - **Different interpretations**
  - Lack of standardized nomenclature
    - **Variations in naming**
    - **Synonyms, homonyms, & abbrevn**
  - Inadequacy of data quality control mechanisms
    - **Systematic approaches to data cleaning are lacking**

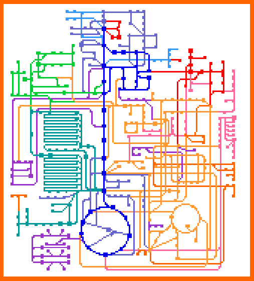
Adapted w/ permission from Judice Koh, Mong Li Lee, & Vladimir Brusic



Classification of Errors

Adapted w/permision from Judice Koh, Mong Li Lee, & Vladimir Brusic

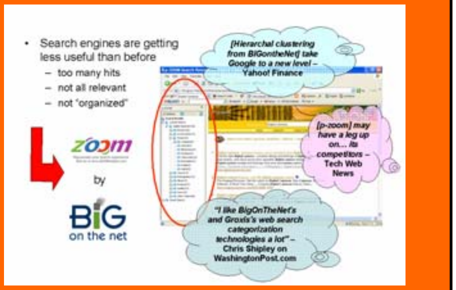
Completeness: Is the structure of our databases expressive enough to capture critical information explicitly?



Expressive Power

- **Take a key paper such as the Kohn paper that summarises current knowledge on p53 regulation.**
- **Is there a structured database that is able to capture all info in that paper explicitly?**
- **Is there a semi-structured database 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?**

Understandability: Is our databases or search results understandable?



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

zoom by BIG on the net

[Hierarchal clustering from BIGontheNet] take Google to a new level – Yahoo! Finance

[p-zoom] may have a leg up on… its competitors – Tech Web News

"I like BigOnTheNet's and Groxis's web search categorization technologies a lot" – Chris Shipley on WashingtonPost.com

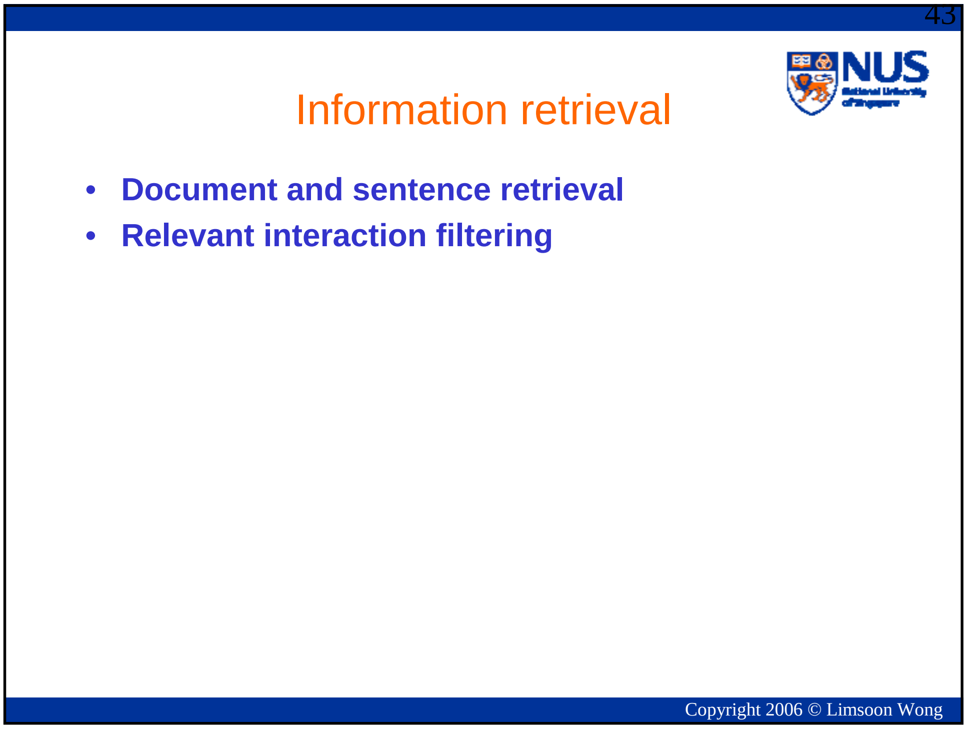
Self-Organization

- Take a search on p53. You will get >300k hits or some number 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 database and/or the search results into hierarchy or "semantic" net to make it easier for users to understand or to browse the results
- How do we define this hierarchy/net?
- Can this hierarchy/net be self-organized?

Problems relating to NLP/IE

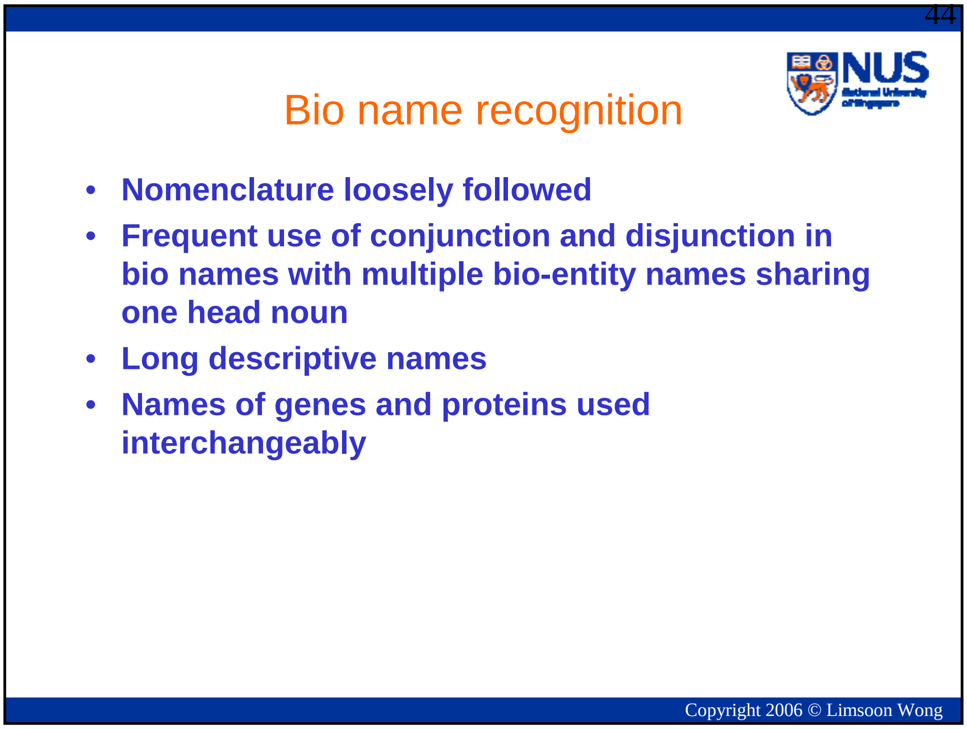
Handling full-length papers

- **Source document structure parsing**
- **Hyper-linked file tracking**
- **Figure and table processing**
- **Special symbol handling**



Information retrieval

- Document and sentence retrieval
- Relevant interaction filtering

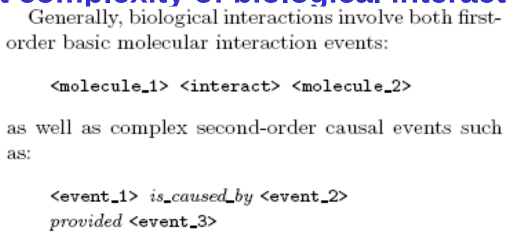


Bio name recognition

- Nomenclature loosely followed
- Frequent use of conjunction and disjunction in bio names with multiple bio-entity names sharing one head noun
- Long descriptive names
- Names of genes and proteins used interchangeably

Bio-interaction extraction

- **Inherent complexity of biological interactions**



Generally, biological interactions involve both first-order basic molecular interaction events:

<molecule_1> <interact> <molecule_2>

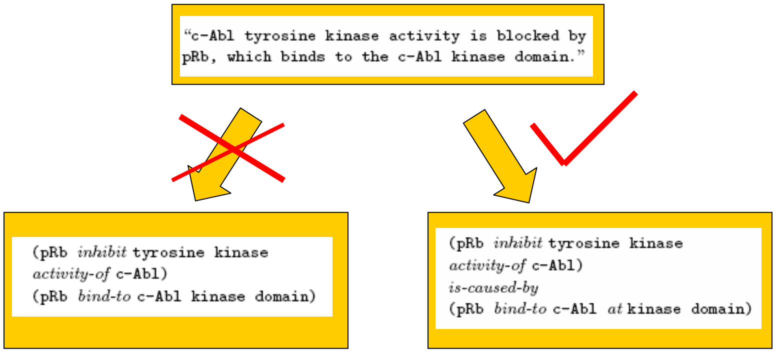
as well as complex second-order causal events such as:

<event_1> *is_caused_by* <event_2>
provided <event_3>

⇒ **Sentences describing them also tend to be complicated**

Bio-interaction extraction

- **Domain knowledge is often needed for interaction template filling**



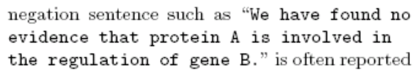
"c-Abl tyrosine kinase activity is blocked by pRb, which binds to the c-Abl kinase domain."

(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)

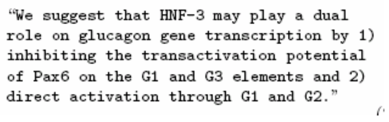
Extraction of other relevant info

- **Contextual information**
  - Species, cell type, cellular localisation, etc
- **Negative information**



negation sentence such as "We have found no evidence that protein A is involved in the regulation of gene B." is often reported

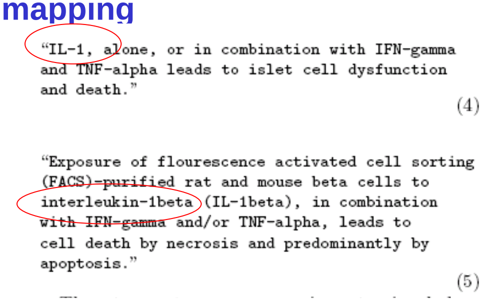
- **Speculative & incomplete facts**



"We suggest that HNF-3 may play a dual role on glucagon gene transcription by 1) inhibiting the transactivation potential of Pax6 on the G1 and G3 elements and 2) direct activation through G1 and G2."

Information integration

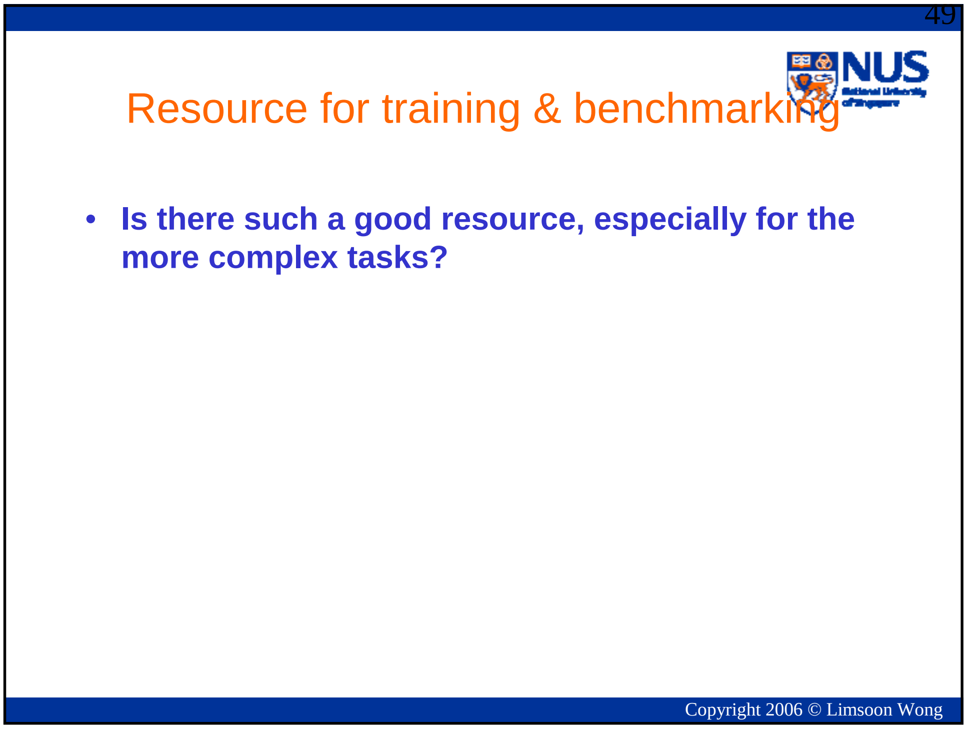
- **Bio-name mapping**



"IL-1, alone, or in combination with IFN-gamma and TNF-alpha leads to islet cell dysfunction and death." (4)

"Exposure of flourescence activated cell sorting (FACS)-purified rat and mouse beta cells to interleukin-1beta (IL-1beta), in combination with IFN-gamma and/or TNF-alpha, leads to cell death by necrosis and predominantly by apoptosis." (5)

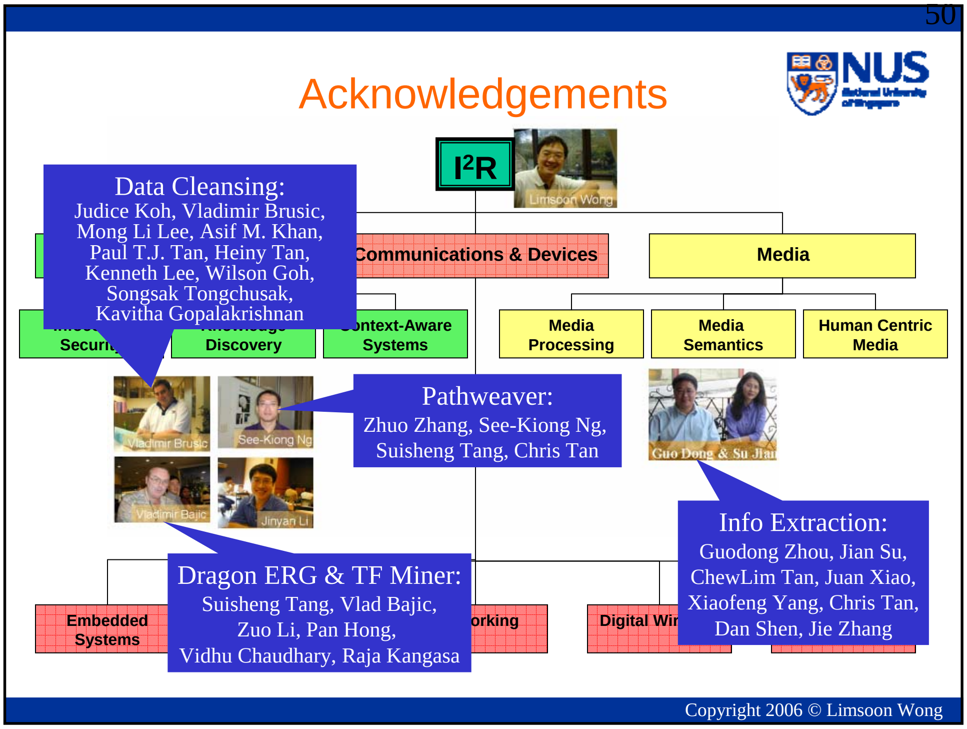
- **Bio-interaction mapping**
  - how do you know two complex sentences are talking about the same interaction?



49

Resource for training & benchmarking

- **Is there such a good resource, especially for the more complex tasks?**



50

Acknowledgements

I²R

Communications & Devices

Media

Securi

Discovery

ontext-Aware Systems

Media Processing

Media Semantics

Human Centric Media

Embedded Systems

orking

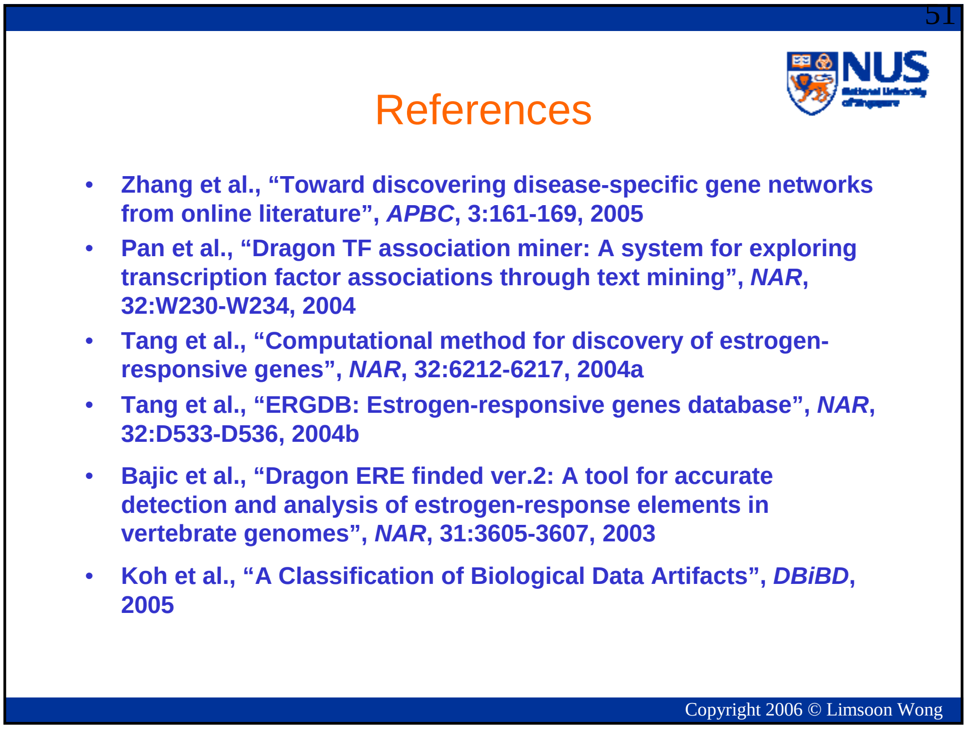
Digital Wir

Data Cleansing:
Judice Koh, Vladimir Brusic, Mong Li Lee, Asif M. Khan, Paul T.J. Tan, Heiny Tan, Kenneth Lee, Wilson Goh, Songsak Tongchusak, Kavitha Gopalakrishnan

Pathweaver:
Zhuo Zhang, See-Kiong Ng, Suisheng Tang, Chris Tan

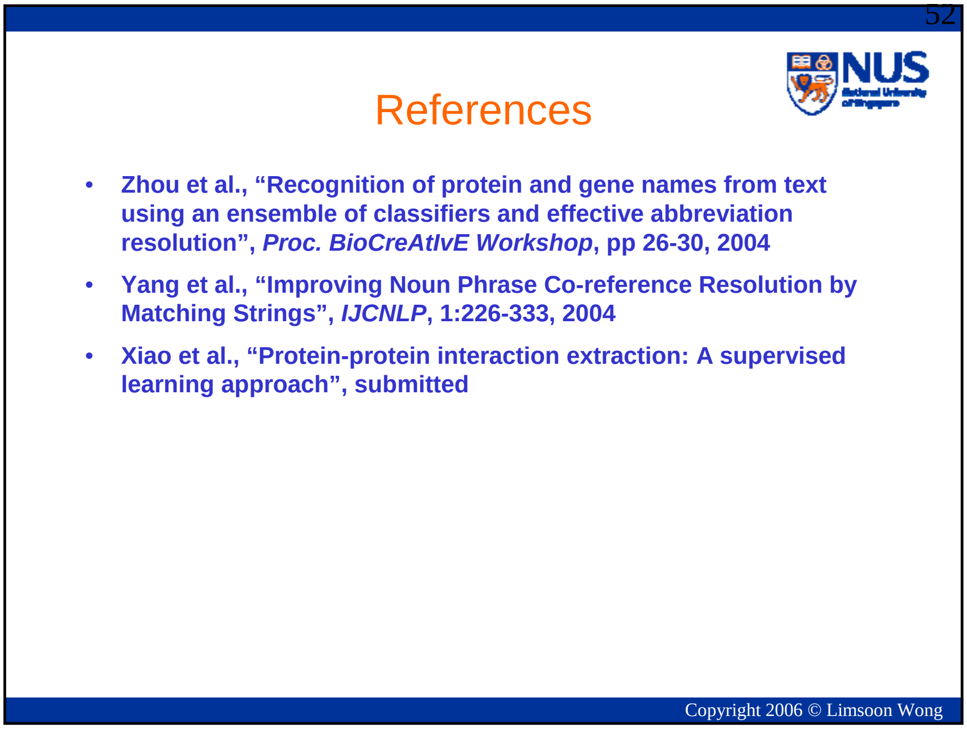
Info Extraction:
Guodong Zhou, Jian Su, ChewLim Tan, Juan Xiao, Xiaofeng Yang, Chris Tan, Dan Shen, Jie Zhang

Dragon ERG & TF Miner:
Suisheng Tang, Vlad Bajic, Zuo Li, Pan Hong, Vidhu Chaudhary, Raja Kangasa



References

- Zhang et al., "Toward discovering disease-specific gene networks from online literature", *APBC*, 3:161-169, 2005
- Pan et al., "Dragon TF association miner: A system for exploring transcription factor associations through text mining", *NAR*, 32:W230-W234, 2004
- Tang et al., "Computational method for discovery of estrogen-responsive genes", *NAR*, 32:6212-6217, 2004a
- Tang et al., "ERGDB: Estrogen-responsive genes database", *NAR*, 32:D533-D536, 2004b
- Bajic et al., "Dragon ERE finded ver.2: A tool for accurate detection and analysis of estrogen-response elements in vertebrate genomes", *NAR*, 31:3605-3607, 2003
- Koh et al., "A Classification of Biological Data Artifacts", *DBiBD*, 2005



References

- Zhou et al., "Recognition of protein and gene names from text using an ensemble of classifiers and effective abbreviation resolution", *Proc. BioCreAtIvE Workshop*, pp 26-30, 2004
- Yang et al., "Improving Noun Phrase Co-reference Resolution by Matching Strings", *IJCNLP*, 1:226-333, 2004
- Xiao et al., "Protein-protein interaction extraction: A supervised learning approach", submitted