Information Retrieval and Text Mining Opportunities in Bioinformatics

Dr. N. JEYAKUMAR, M.Sc., Ph.D.,
Dept. of Bioinformatics
Bharathiar University
Coimbatore - 641046

Outline

- Introduction to IR and TM
- Biomedical Literature Resources
- Two basic tasks – Bio-Entity and Entity-Relation Identification
- Knowledge Discovery with text
- Text data integration
- Outlook

Part III: **Bio-Entity and Entity-Relation Extraction**

Text Mining:

Applications Areas in Biology

- Help to address the following problems:
  - Finding biological named entities (e.g. protein, gene, chemical names etc.) in context to particular study
  - Finding molecule interactions (e.g. protein-protein interactions, gene-gene relations etc.)
  - Finding relations between bio-concepts (e.g. relations between genes-disease, disease-drug)
  - Finding bio-chemical pathways
  - Finding sub-cellular localization information of proteins
  - Constructing biological vocabulary/ontology from text
  - Automatically Curating biological databases
  - Assisting gene expression data mining process
  - Knowledge-based information retrieval in context to biological repositories (e.g. MEDLINE etc.)

Text Mining:

Genetic Basics

- **Gene/Protein – Associate/interact – Gene/protein => pathway**
  (concept) (conceptual relation) (concept) =>(Biological process)

  (e.g) STAT3 interact BCL-X => apoptosis (cell death)

- **Gene/protein – symptom– disease**
  (concept) (function) (concept)

  (e.g.) p53 tumor suppressor cancer
  TNFRSF1B Insulin resistance diabetes

So, the main goal of any text mining/information extraction system in biomedical domain is identify the bio-entitles and their relationship

Information Extraction

Bio Entities and Relationships

Extract what?

- **Entities:** e.g., genes, proteins, diseases, chemical compounds, etc.
- **Relationships:** e.g., phosphorylation, activation of a gene by a transcription factor, etc.
- **Functions:** e.g., a protein is activated, a gene is transcribed, etc.

It is hard!

- **Entities** can have synonyms and be referred as anaphora (e.g., this gene, that protein, the former, etc.). **Unrelated entities** may share a name (polysyms), and one up to four words (e.g. p53, N-kappa beta protein)
- **Relationships** and **events** can be stated in various styles and indirect ways.

Information Extraction
Sample PubMed Record

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a 'mitotic cyclin', cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene.

Other recent studies have identified human cyclin D1 (PRAD1) as a putative G1 cyclin and candidate proto-oncogene.

However, the specific enzymatic activities and, hence, the precise biochemical mechanisms through which cyclins function to govern cell cycle progression remain unresolved.

In the present study we have investigated the coordinate interactions between these two potentially oncogenic cyclins, cyclin-dependent protein kinase subunits (cdks) and the Rb tumor-suppressor protein.

The distribution of cyclin D isoforms was modulated by serum factors in primary fetal rat lung epithelial cells.

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing's sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity.

Immobilized, recombinant cyclins A and D1 were found to associate with cellular proteins in complexes that contain the p105Rb protein.

Information Extraction

Sample PubMed Record with Named Entites

TI - Two potentially oncogenic cyclins, **cyclin A** and **cyclin D1**, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the **Rb** protein

AB - Originally identified as a 'mitotic cyclin', **cyclin A** exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an **S-phase-promoting factor** (**SPF**) as well as a candidate proto-oncogene.

Other recent studies have identified human **cyclin D1** (**PRAD1**) as a putative G1 cyclin and candidate proto-oncogene.

However, the specific enzymatic activities and, hence, the precise biochemical mechanisms through which cyclins function to govern cell cycle progression remain unresolved.

In the present study we have investigated the coordinate interactions between these two potentially oncogenic cyclins, cyclin-dependent protein kinase subunits (cdks) and the **Rb** tumor-suppressor protein.

The distribution of **cyclin D** isoforms was modulated by serum factors in primary fetal rat lung epithelial cells.

Moreover, **cyclin D1** was found to be phosphorylated on tyrosine residues in vivo and, like **cyclin A**, was readily phosphorylated by **pp60c-src** in vitro.

In synchronized human osteosarcoma cells, **cyclin D1** is induced in early G1 and becomes associated with **p9Ckshs1**, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing's sarcoma cells demonstrated that **cyclin D1** is associated with both **p34cdc2** and **p33cdk2**, and that **cyclin D1** immune complexes exhibit appreciable histone H1 kinase activity.

Immobilized, recombinant cyclins A and D1 were found to associate with cellular proteins in complexes that contain the **p105Rb** protein.

.

Information Extraction

Sample PubMed Record with NE Relations

TI - Two potentially oncogenic cyclins, **cyclin A** and **cyclin D1**, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the **Rb** protein

AB - Originally identified as a 'mitotic cyclin', **cyclin A** exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an **S-phase-promoting factor** (**SPF**) as well as a candidate proto-oncogene.

Other recent studies have identified human **cyclin D1** (**PRAD1**) as a putative G1 cyclin and candidate proto-oncogene.

However, the specific enzymatic activities and, hence, the precise biochemical mechanisms through which cyclins function to govern cell cycle progression remain unresolved.

In the present study we have investigated the coordinate interactions between these two potentially oncogenic cyclins, cyclin-dependent protein kinase subunits (cdks) and the **Rb** tumor-suppressor protein.

The distribution of **cyclin D** isoforms was modulated by serum factors in primary fetal rat lung epithelial cells.

Moreover, **cyclin D1** was found to be phosphorylated on tyrosine residues in vivo and, like **cyclin A**, was readily phosphorylated by **pp60c-src** in vitro.

In synchronized human osteosarcoma cells, **cyclin D1** is induced in early G1 and becomes associated with **p9Ckshs1**, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing's sarcoma cells demonstrated that **cyclin D1** is associated with both **p34cdc2** and **p33cdk2**, and that **cyclin D1** immune complexes exhibit appreciable histone H1 kinase activity.

Immobilized, recombinant cyclins A and D1 were found to associate with cellular proteins in complexes that contain the **p105Rb** protein.

Information Extraction

Named Entity Recognition (NER)

- NER involves identification of *proper names* in texts, and classification into a set of predefined categories of interest.
- Three universally accepted categories: person, location and organisation
- Other common tasks: recognition of date/time expressions, measures (percent, money, weight etc), email addresses etc.
- Other domain-specific entities: (e.g.) bio entites includes genes, proteins, names of drugs, etc.

Information Extraction:
Basic Problems in NER

- Variation of NEs – e.g. John Smith, Mr Smith, John.
- Ambiguity of NE types
  - John Smith (company vs. person)
  - May (person vs. month)
  - Washington (person vs. location)
  - 1945 (date vs. time)
- Ambiguity with common words, e.g. "may"

Information Extraction
Bio-NER

- Objective
  - Identify biological entities (proteins, genes) in articles and to link them to entries in biological databases.
- Methods
  - Rule-based
  - Dictionary based flexible matching,
  - Statistical and Machine Learning (naive Bayes, ME, SVM, CRF, HMM).

Information Extraction
Bio-NER - Challenges

- Authors often do not use the official gene symbols
- Genes have often synonyms.
- Use of full gene names and/or gene symbols/acronyms
- Gene names - medical terms ambiguity
- Gene names - common English words ambiguity (fly)
- Alternative typographical variants
- 14% of genes display inter-species ambiguity (Chen, 2005).
- Ambiguity between protein names and their protein family names
- Identification of new gene names (novel genes)

Information Extraction:
Bio-NER- Rule based approaches

- Pos tagger, trained on biological domain, chunking, semantic typing of chunks, identification of relations using pattern-matching rules

- Semantic typing of NPs: using combination of clue words, suffixes, acronyms etc (e.g. presence of Roman letters, Greek letters, ending with protein, gene names – nb-I, NF-beta, BC-Protein)

- Semantically typed sentences matched with rules

Information Extraction:
Bio-NER- Dictionary based Approaches

- Pos tagger, trained on biological domain, chunking, semantic typing of chunks, identification of relations using pattern-matching rules

- matching of NPs using specialized dictionary og genes and protein names

- The dictionary must be updated and up to date as new proteins and gene names are often discovered

Information Extraction

Bio-NER – Machine Learning based approaches

- Feature Set
  - Simple deterministic feature
  - Morphological feature
  - Part-of-Speech feature
  - Semantic trigger feature

Information Extraction

Bio-NER – Machine Learning based approaches

- Feature Set
  - Simple deterministic feature
  - Morphological feature
  - Part-of-Speech feature
  - Semantic trigger feature

Information Extraction

Bio-NER – Machine Learning based approches

Simple Deterministic Feature

- Word formations: capital letters, digits, …
- We used 29 simple deterministic features.

Feature	Example
Roman Digit	II, III, IV, …
Greek Letter	alpha, beta, …
CapNumCap	E1A, E2F, …
Caps1D	T4, CD4, …
allCaps	NFAT, MAZ, …
etc …	

Information Extraction

Bio-NER – Machine Learning based approaches

Morphological Feature

- Prefix and suffix
  - Important cue for terminology identification
  - Group prefixes/suffixes that have similar distribution over NE classes
  -

sOOC	~cin ~mide ~zole	actinomycin cycloheximide sulphamethoxazole
sLPD	~lipid ~rogen ~vitamin	phospholipids estrogen dihydroxyvitamin
etc …		

Information Extraction

Bio-NER – Machine Learning based approaches

Semantic Trigger Feature

- Head Nouns Triggers
  - Important clues for Bio NER!
- Example
  - PROTEIN: receptor, binding protein, …
  - CELL LINE: line, cell line, …
  - RNA: mRNA, messenger RNA, …
- Auto-generate top ranked unigram / bi-gram head noun list from training data for each class
- Very useful

Information Extraction

Bio-NER – Machine Learning based approaches

- model is learnt based on one of the following techniques:
  - Decision Trees, such as ID3
  - Support Vector Machines
  - Artificial Neural Network
  - HMM
  - Maximum Entropy
  - Conditional Random Fields
- Last two reported high precession and recall about 93% and 87% respectively

Information Extraction

Bio-NER – Machine Learning based approaches

- model is learnt based on one of the following techniques:
  - Decision Trees, such as ID3
  - Support Vector Machines
  - Artificial Neural Network
  - HMM
  - Maximum Entropy
  - Conditional Random Fields
- Last two reported high precession and recall about 93% and 87% respectively

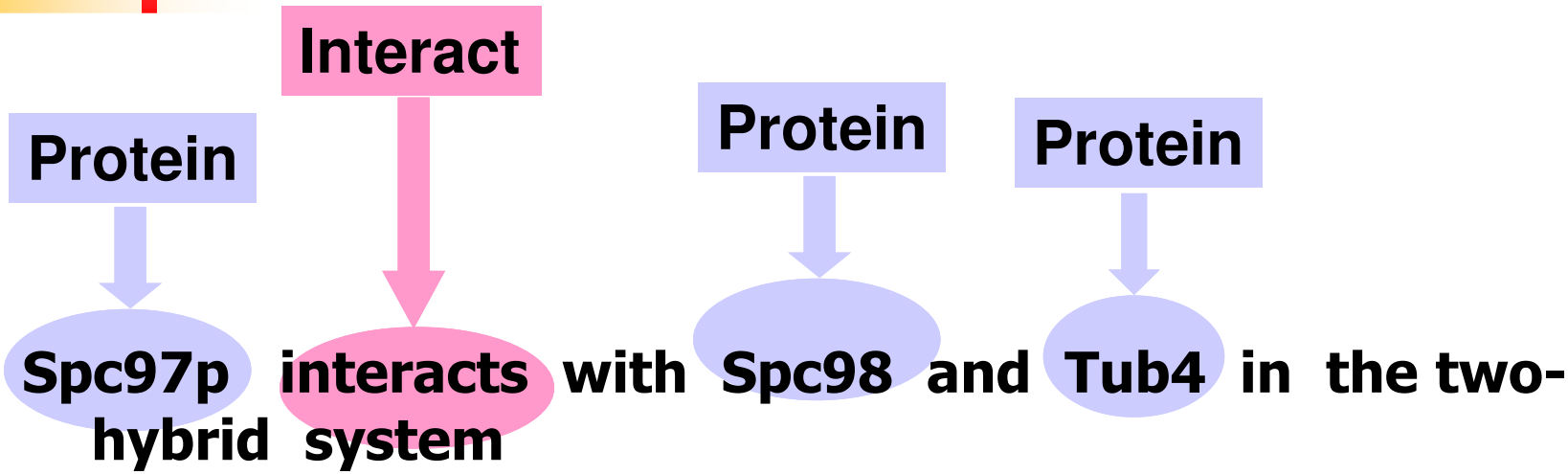
Information Extraction
Relation Extraction

- Objective
  - Extract interaction information between biological entities from literature. For example, protein-protein interaction, gene-gene relations etc.
- Methods
  - Co-occurrence of bioentities within close vicinity
  - Rule based
  - Machine learning based methods (Relationship extraction)
  - Linguistic methods (Dependency parsers, link parsers)

Information Extraction
Relation Extraction

- Objective
  - Extract interaction information between biological entities from literature. For example, protein-protein interaction, gene-gene relations etc.
- Methods
  - Co-occurrence of bioentities within close vicinity
  - Rule based
  - Machine learning based methods (Relationship extraction)
  - Linguistic methods (Dependency parsers, link parsers)

Information Extraction: Relation Extraction - example



Spc97p interact Spc98
Spc97p interact Tub4

Information Extraction: Relation Extraction - example

Co-occurrence statistics based approaches

Sentence level statistic

Relation	*Occurrence*
(Rb, IE2)	+1
(p53, IE2)	+1
(p53, Rb)	+1

Corpus level statistic

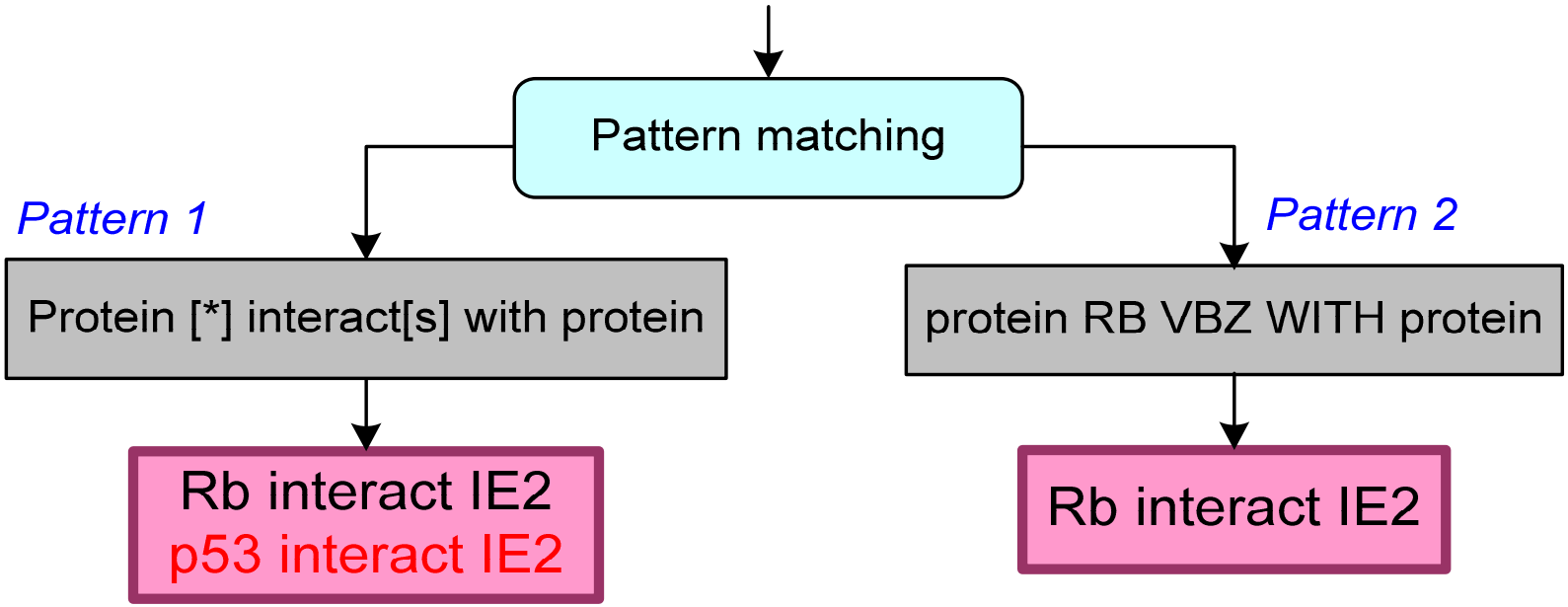
Relation	*Occurrence*
...	1
(p53, IE2)	8
...	6

Predefined threshold ***a = 7***

Relation	*Confidence*
...	...
(p53, IE2)	75%
...	...

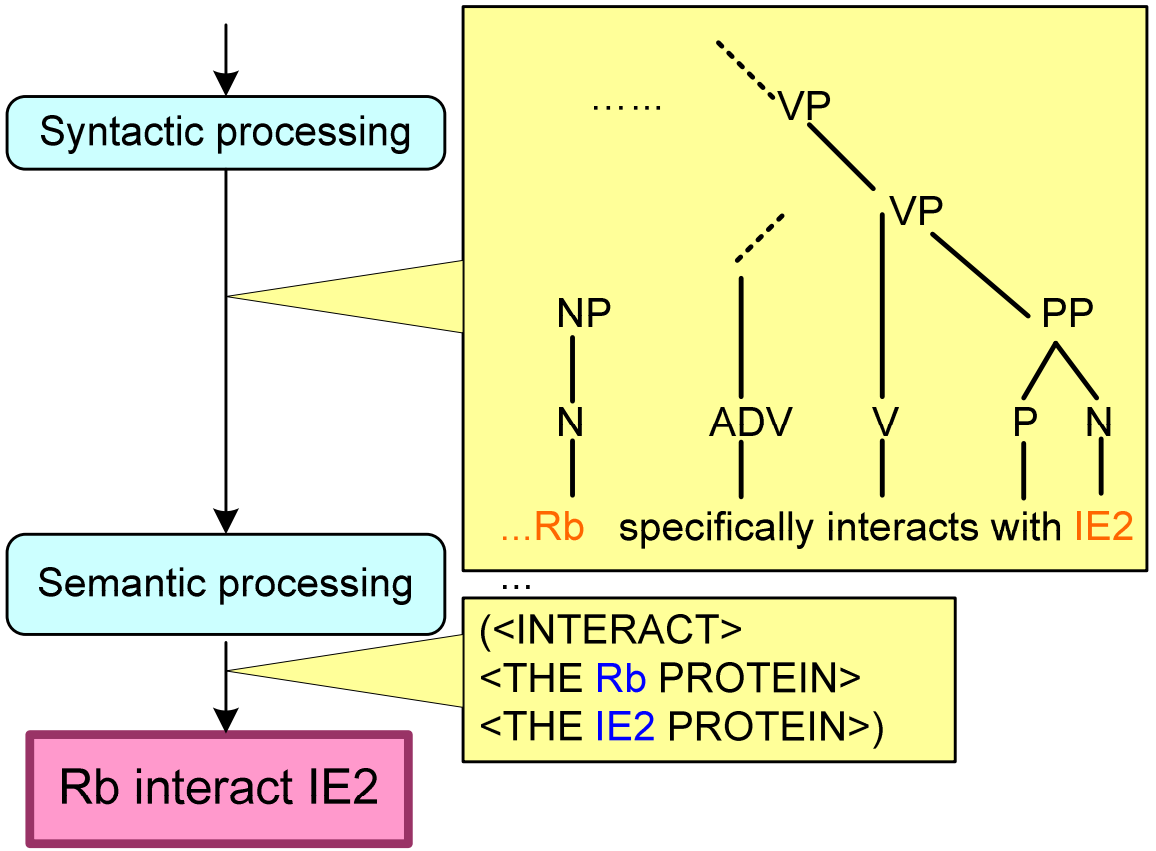
Information Extraction:
Relation Extraction

Rule based pattern matching approaches



Information Extraction:
Relation Extraction

Parsing-Based Approaches



Information Extraction
Relation Extraction

- Events/Relations in Life Science Text
  - Protein Protein Interaction
  - Gene Regulation
  - Ligand Protein Interaction
  - Drug Disease Association
  - Drug Side-effects Association
  - Gene Disease Association

Text Mining:

BioMedical Text Mining Systems - Examples

- iHOP
  - http://www.ihop-net.org/UniPub/iHOP/
  - Gene centric search Engine
- EBIMed
  - http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp
  - Concept based search linked to Uniprot
- GoPubMed
  - http://www.gopubmed.org/
  - Clusters documents based on Gene/MesH Ontology
- BioMinT
  - http://biomint.pharmadm.com/
  - An easy to use information retrieval and extraction tool
- Textpresso
  - http://www.textpresso.org/
  - Text categorization genome search engine

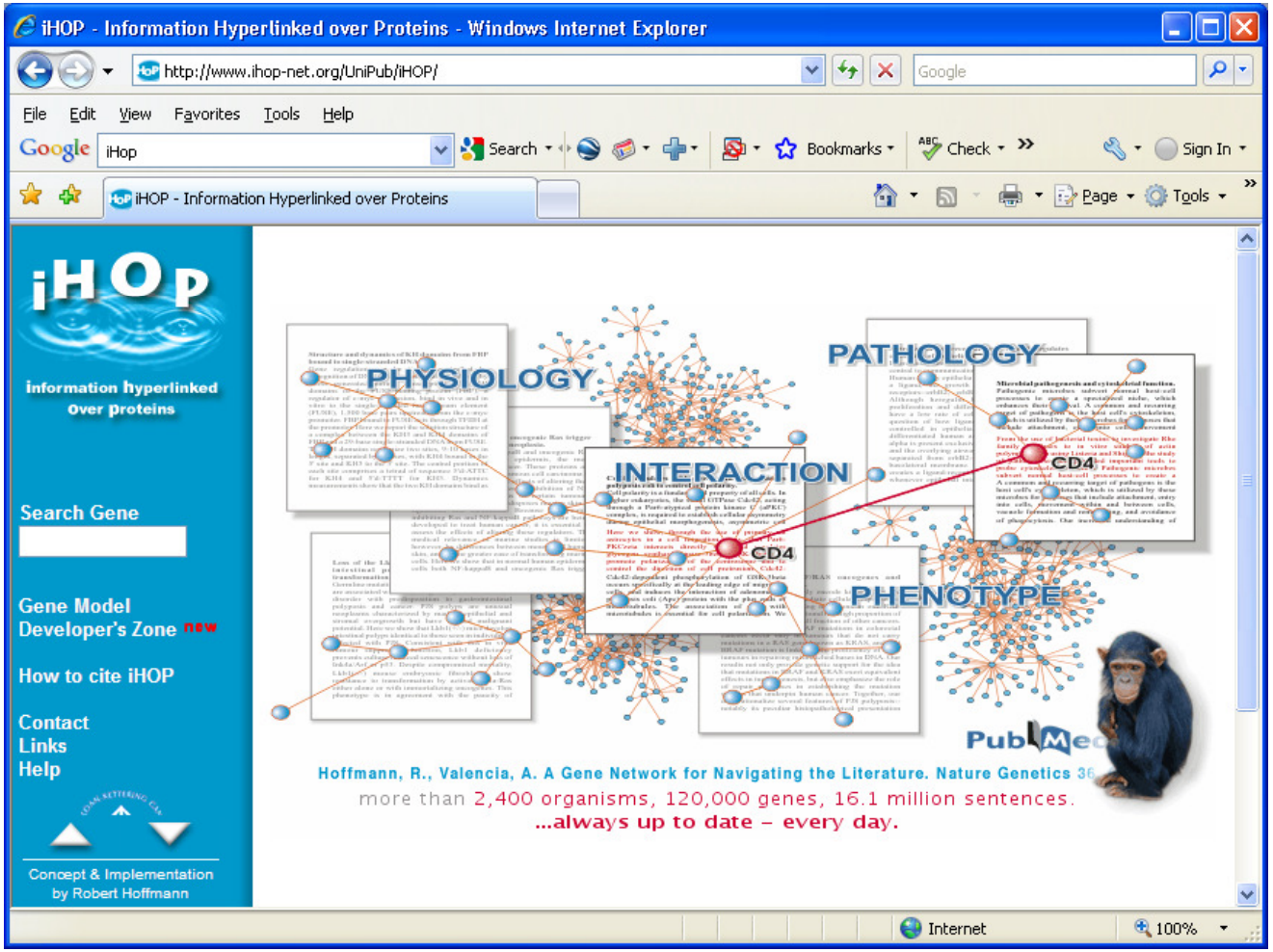
Text Mining:

BioMedical Text Mining Systems - Examples

- iHOP
  - http://www.ihop-net.org/UniPub/iHOP/
  - Gene centric search Engine
- EBIMed
  - http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp
  - Concept based search linked to Uniprot
- GoPubMed
  - http://www.gopubmed.org/
  - Clusters documents based on Gene/MesH Ontology
- BioMinT
  - http://biomint.pharmadm.com/
  - An easy to use information retrieval and extraction tool
- Textpresso
  - http://www.textpresso.org/
  - Text categorization genome search engine

Text Mining

iHOP – Web Page



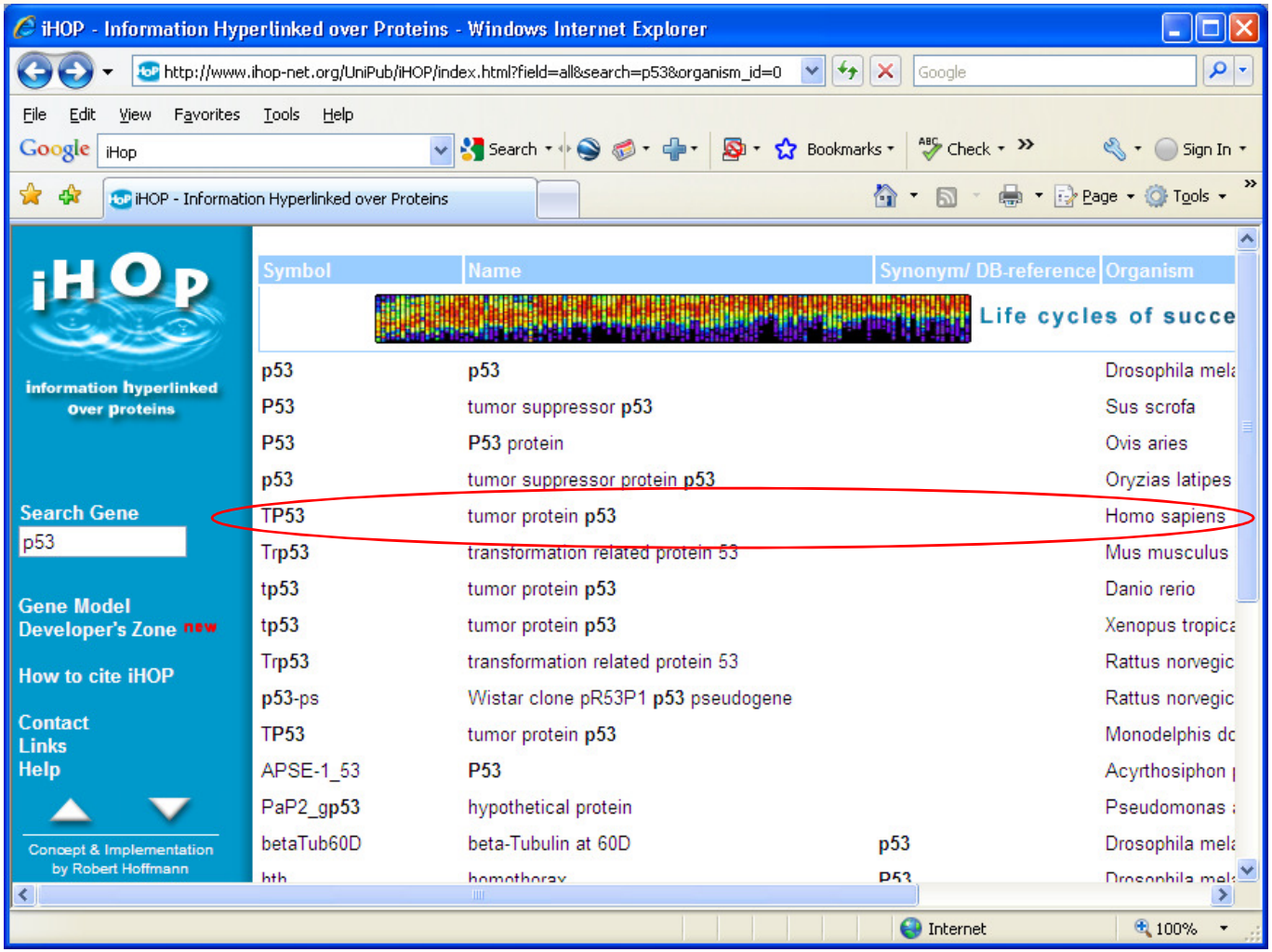
Text Mining:

iHOP

- iHOP - Information Hyperlinked over Proteins
  - A network of concurring genes and proteins extends through the scientific literature touching on phenotypes, pathologies and gene function.
  - iHOP provides this network as a natural way of accessing millions of PubMed abstracts. By using genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed can be converted into one navigable resource, bringing all advantages of the internet to scientific literature research.
- Reference
  - A Gene Network for Navigating the Literature. Hoffmann, R., Valencia, A. Nature Genetics 36, 664 (2004)

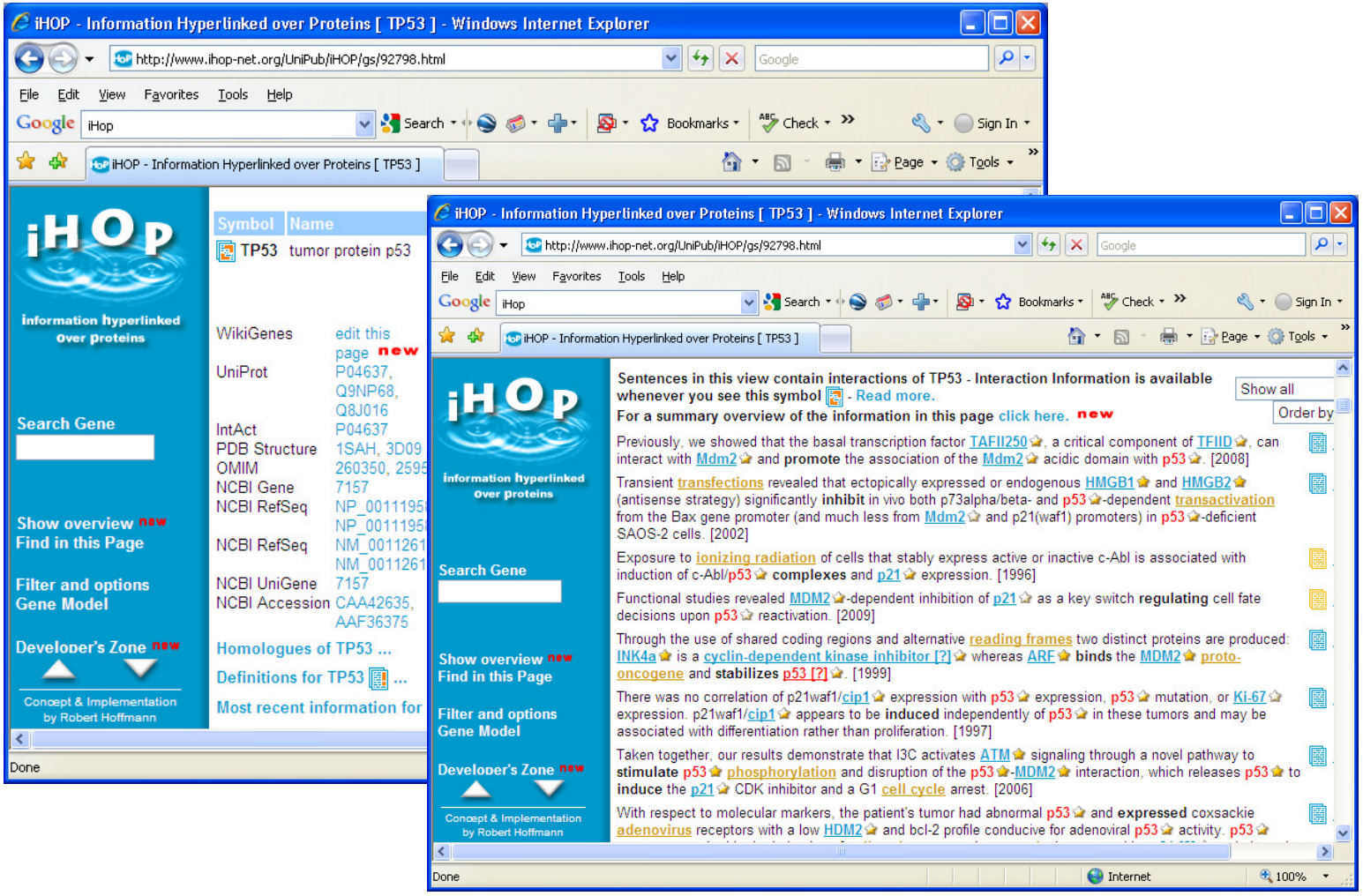
Text Mining

iHOP – Gene Searching



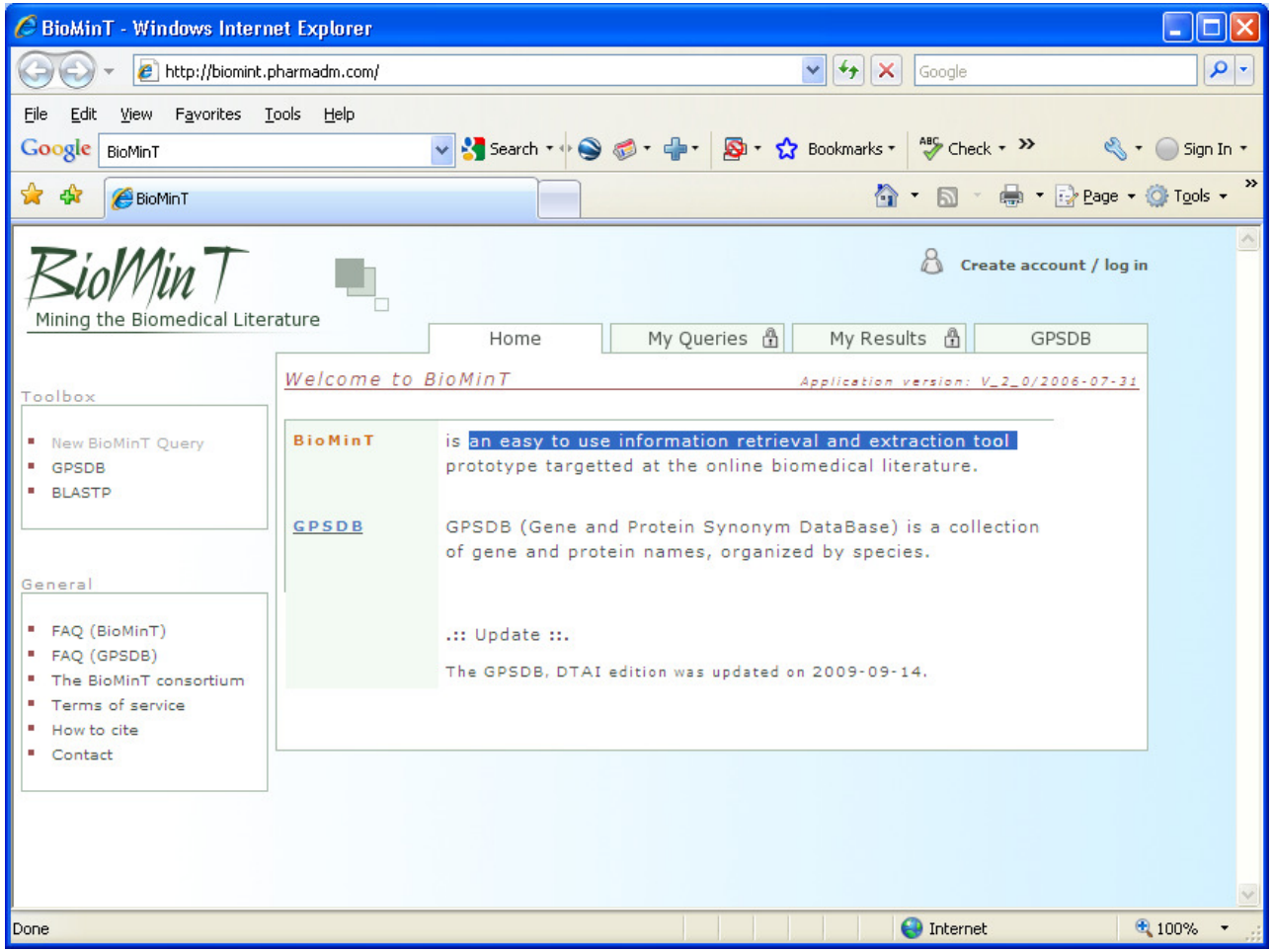
Text Mining

iHOP – Gene Searching



Text Mining

BioMinT



Text Mining:

MedMiner

- One of earlier initiative for biomedical text mining
- Searches and integrates information from text and data resources such as PubMed and GeneCards
- Later organizes the complied information around topics relevant to user query
- Reference
  - Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN: MedMiner: an internet text-mining tool for biomedical information with application to gene expression profiling. BioTechniques 1999, 27:1210-1217

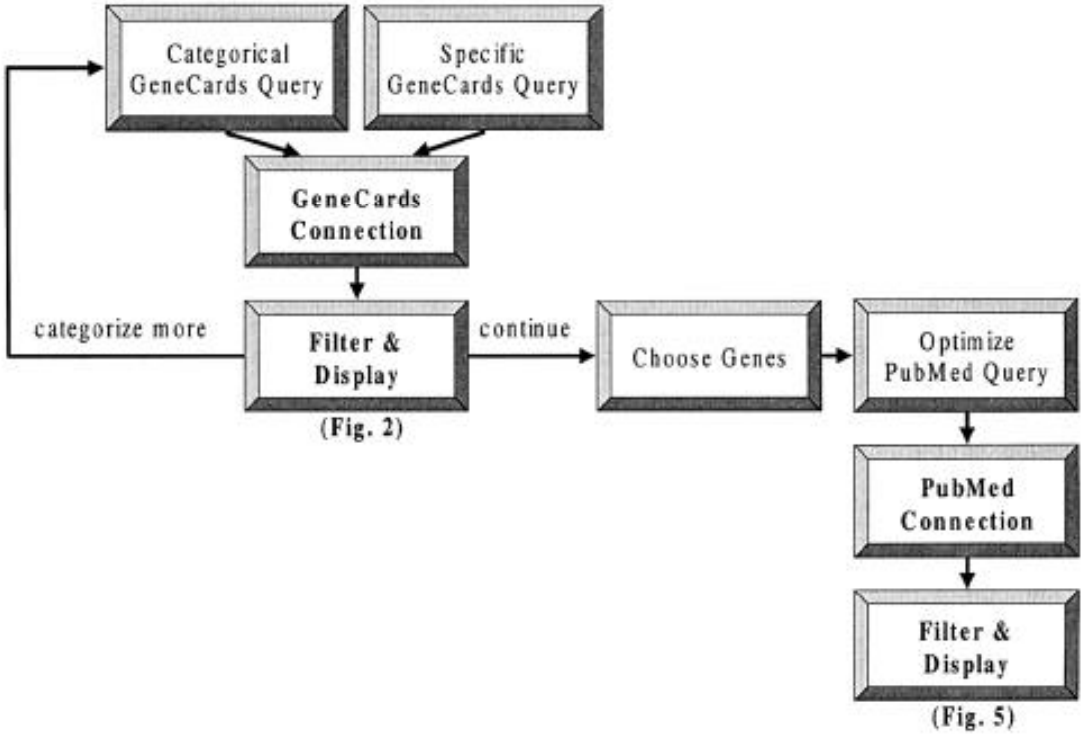
Text Mining:

MedMiner – Processing steps

- User inputs query genes (i.e. list of genes) or biological concept (e.g. apoptosis)
- MedMiner collects the relevant information about query genes from GeneCards database as gene-gene relations with keywords
- The new query is now searched for PubMed database

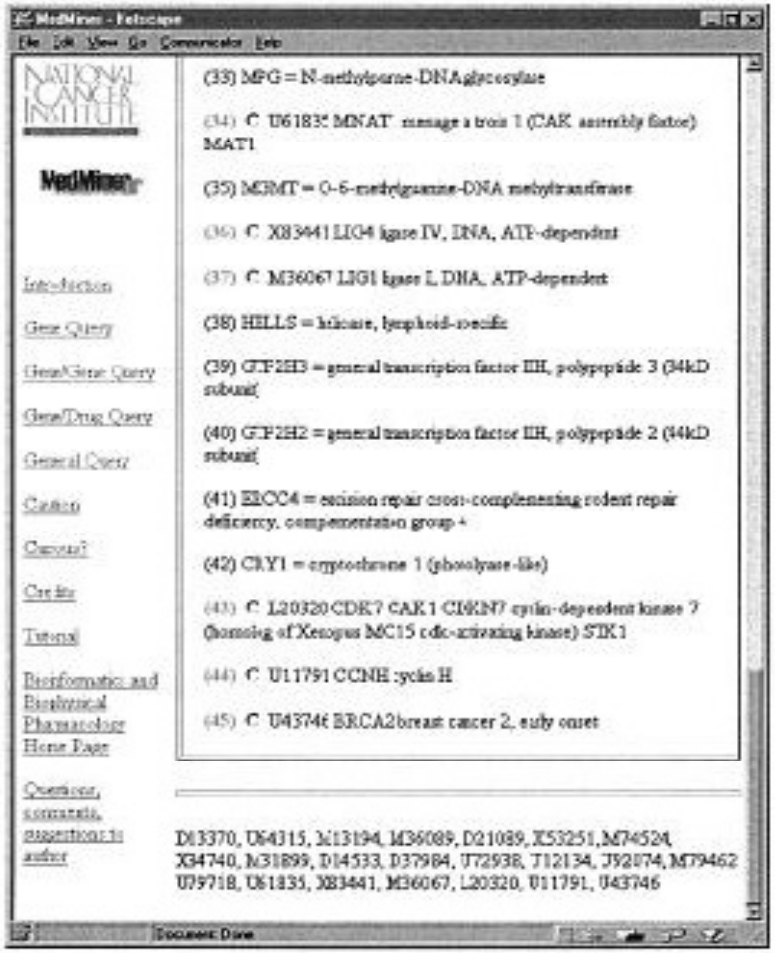
Text Mining

MedMiner – Work Flow



Text Mining

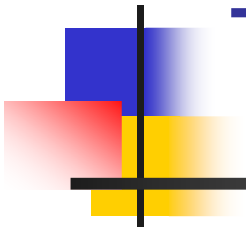
MedMiner – Outputs



Text Mining

MedMiner – Outputs

	PubMed only	MedMiner
Query	p53 AND mdm2 AND inhibit*	p53,mdm2
Cutoff date	past 60 days	past 60 days
Number of abstracts returned as relevant	12	11
Initial results format	Abstract titles	Abstract sentences
Number of initial results	12	17
Number of initial results with "inhibit" relationship	2/12	17/17
Number of sentences returned as relevant	104	17



Text Data Integration

Text Mining and Microarray Gene Expression Analysis

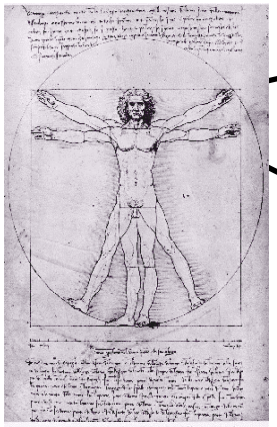
Text Mining and Microarrays

Introduction - Gene Expression

- Cells are different because of differential gene expression.
- About 40% of human genes are expressed at one time.
- Gene is expressed by transcribing DNA exons into single-stranded mRNA
- mRNA is later translated into a protein
- Microarrays measure the level of mRNA expression

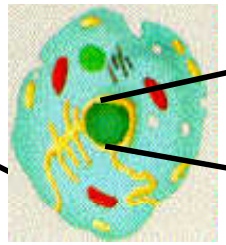
Text Mining and Microarrays

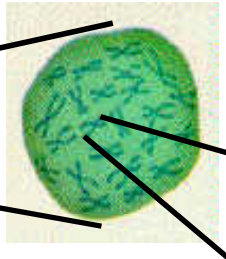
Gene Expression – The Big Picture



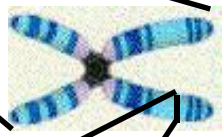
Cell

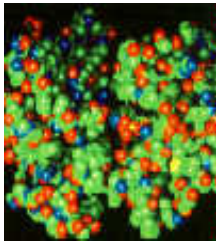
Nucleus



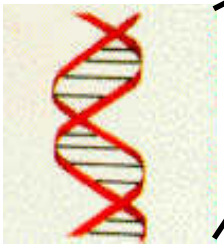


Chromosome





Gene expression



Protein

Gene (mRNA), single strand

Gene (DNA)

Graphics courtesy of the National Human Genome Research Institute

Text Mining and Microarrays

Gene expression data analysis

- Gene expression microarrays have tremendous potential in biology and medicine
- Microarray data analysis is difficult and poses unique challenges
- Text mining of microarray data analysis process is critical for good, reliable results

Text Mining and Microarrays

An example application to microarray data analysis

- The platelet-derived lipid mediator sphingosine-1-phosphate (S1P) is an endogenous ligand of the endothelial differentiation gene (EDG) family of G protein-coupled receptors. S1P is involved in various cellular responses such as apoptosis, proliferation, and cell migration. S1P also involved in tumor cell invasion (Invasion – spread of cancer cells into healthy tissues adjutant to the tumor)

- To date, the impact of S1P on human glioblastoma (pediatric brain tumor) is not fully understood. The gene expression analysis to investigate the response of a glioblastoma cell line (U373MG) to S1P administration reveals seventy-two genes were found to be differentially expressed.

Text Mining and Microarrays

List of Differentially Expressed Genes

AKAP2	CORO1C	FLJ23476	JMJD3	PBEF1	STMN1
BCL6	CTBP1	FOSB	KIAA0092	PDE4C	THBS1
BTG1	DOC1	FOSL2	KIAA1718	PIM1	TMSNB
BTG2	DSCR1	FOXG1B	KLF5	PLAU	TNFA1P
C9ORF3	DUSP14	FZD7	LBH	RBMS2	TOP1
CALDI	EHD1	GADD45	MAP2K3	RGS3	TPM1
CASKIN2	EHD4	GBP1	MIG2	SACS	TPM4
CCL2	EPOR	GLIPR1	NAB1	SDC4	TRIPBR2
CDKN1A	ETS2	HRB2	NFKB1A	SERD2	TWIST1
CEBPD	F3	IL6	NR4A1	SFRS3	TXNIP
CITED2	FLJ13448	IL8	NRG1	SOCS5	UBE2E3
COPED	FLJ23231	JAG1	PALM2	STK17A	WDR1

Text Mining and Microarrays

Data

- Abstracts related to Brain Tumors are downloaded from PubMed/MEDLINE
- Full-text articles are downloaded form 20 journals related to Cancer (Table 1)

Text Mining and Microarrays

Table 1 – List of Full-text Journals (1999-2004)

Biochemistry	Cell	Jr. of Biological Chemistry	Neurology
BBRC	EMBO Journal	Jr. of Cell Biology	Nucleic Acid Research
Brain Research	FEBS Letters	Jr. of Neuroscience	Oncogene
Cancer	Genes and Development	Nature	PNAS
Cancer Research	International Jr. of Cancer	Neuron	Science

Text Mining and Microarrays

Methodology

- **Gene/Protein name and synonym dictionary creation**
  - This uses *Entrez Gene* as central resource for creation of gene/protein name and synonym dictionary of all the known kinases
- **Gene-name normalization:**
  - This process replaces all the known protein/gene names in the abstract with its unique canonical identifier (Entrez gene ID) using the gene-synonym dictionary specially constructed for this study.
- **Sentence parsing and relation filtering**:
  - Various biomedical based NLP tools such as Brill tagger to ENG parser with user defined rules will be used for the accurate extraction of protein/gene relations (e.g. Table 2)

Text Mining and Microarrays

Methodology (contd)

- **Data Warehouse and Web service Development**
  - The database will contain all human protein and their relationships with other proteins/genes and pathway maps (e.g. Table 3)

- **Visualization of protein kinase pathways**
  - the extracted protein kinase relationships will be visualized as kinase pathway maps using publicly available tools or using JAVA programming language

Text Mining and Microarrays

Table 2 - List of Extraction Rules

Type: Pattern: Sentence: Output:	**Nouns describing agents** ($gene (is) (the\|an\|a) @{0,2}$action of @{0,2} $gene) IL6, a known mediator of STAT3 response Interleukin 6 mediates STAT3
Type: Pattern: Sentence: Output:	**Passive verbs** ($gene @{0.6} (is|was|be|are) @{0,1} $action $(by|via) @{0,3} $gene) Protein kinase c (PKC) has been shown to be activated by parathyroid hormone Parathyroid hormone activates pkc
Type: Pattern: Sentence: Output:	**Active verbs** ($gene $sub-action @{0,1} $action @{0,2} $gene) Insulin mediated inhibition of hormone sensitivity lipase activity Insulin inhibits lipase
Type: Pattern: Sentence: Output:	**Nouns describing actions** ($gene @{0,6} $action (of|with) @{0,1} $gene) abi5 domains required for interaction with abi3 abi5 interacts abi3

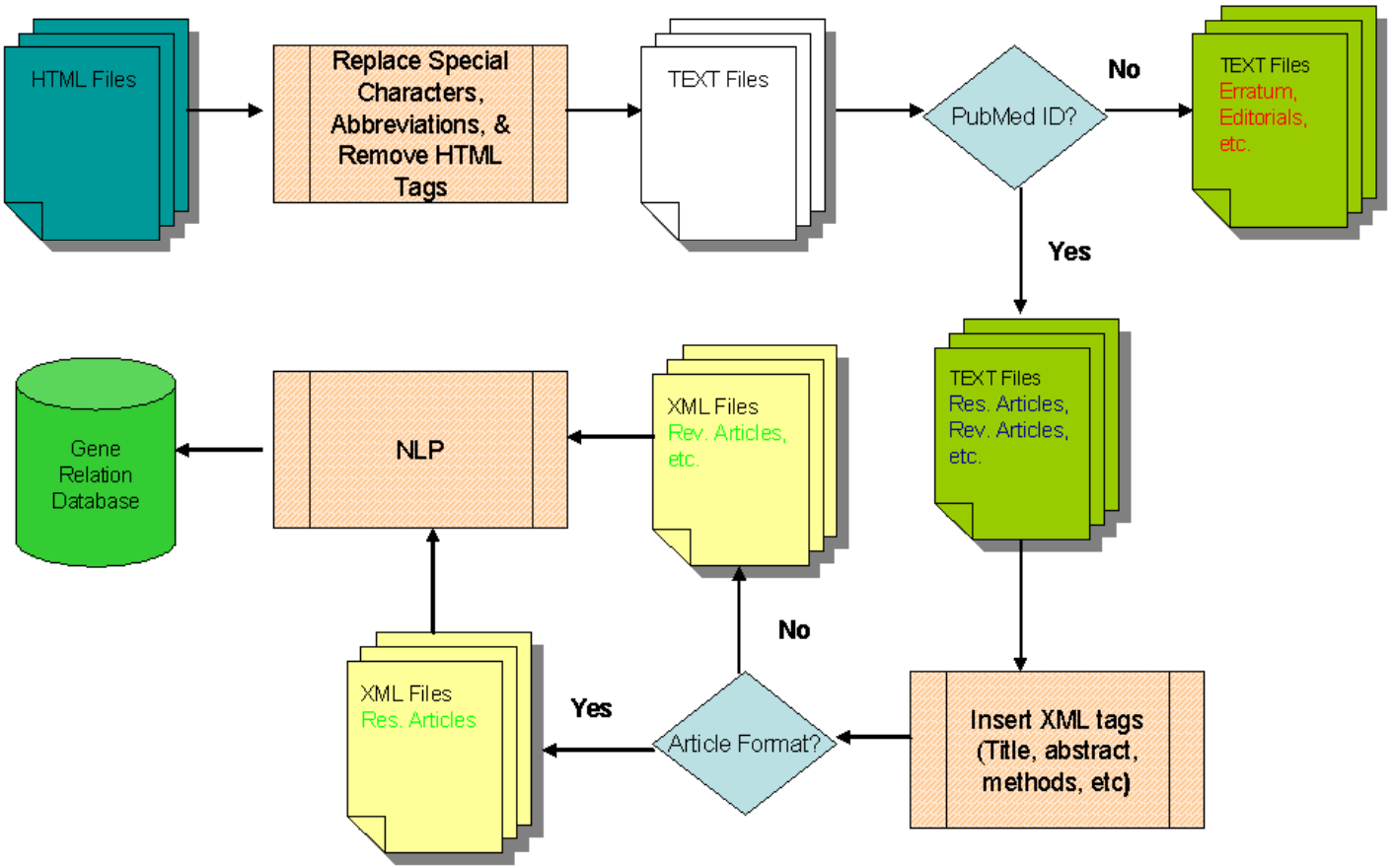
Text Mining and Microarrays

Table 3 – Data warehouse of gene relations

PubMed ID	Gene 1	Gene 2	Relation	#Source Sentence
12881431	APOBEC2	AICDA	mediates	Corresponding sentence
12101418	CTPB1	P53	Inhibits	-do-
15131130	DOC1	nf-kappa b	activates	-do-
12154096	ETHD-1	Pkb	activates	-do-

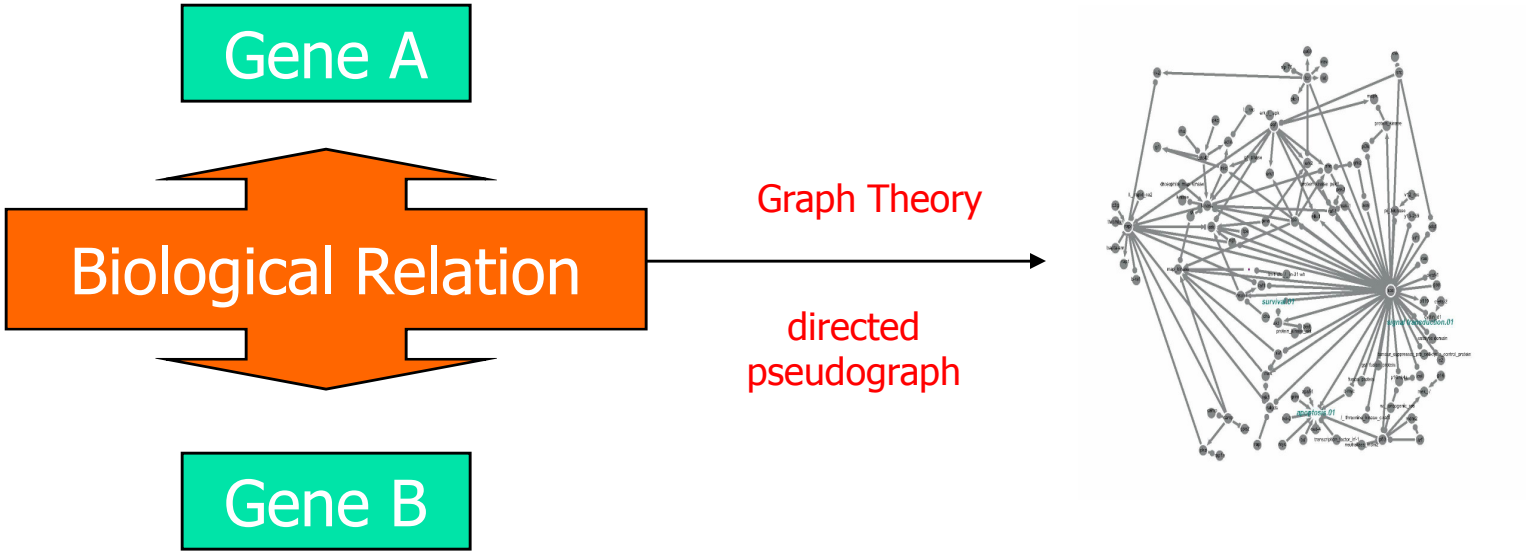
Text Mining and Microarrays

Data and Analysis - Overview



Text Mining and Microarrays

Post–processing: Network Construction



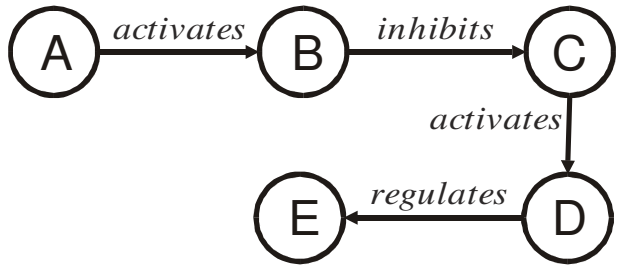
Text Mining and Microarrays

Definition: pseudograph, directed pseudograph

- Informally speaking, a graph is a set of nodes (or vertices) that are connected by links (or edges).
- A multigraph is defined as a set V of vertices, a set E of edges, and a function $f: E \rightarrow \{\{u, v\} | \{u, v\} \in V, u \neq v\}$, specifying which vertices are connected by which edge. If u = v, then the graph is considered a pseudograph, i.e. it contains a loop connecting a vertex with itself.
- If the edges have a direction then the graph is referred to as directed graph or digraph. The network is a directed pseudograph, if it contain multiple edges and loops between the same vertices.
- In the network structure in the present study the genes/proteins are represented as vertices and the relationships as directed edges.

Text Mining and Microarrays

Network construction – Transitive dependencies

1. Gene *A* activates gene *B*. 2. Gene *B* inhibits gene *C*. 3. Gene *C* activates gene *D*. 4. Gene *D* regulates gene *E*	

In this example, to the interaction $A \rightarrow B$ as *direct interaction*, whereas $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$ represents a *transitive dependency of degree 4*, because this dependency involves a path length of 4. In this study transitive dependency of up to degree 3 was used to construct the network

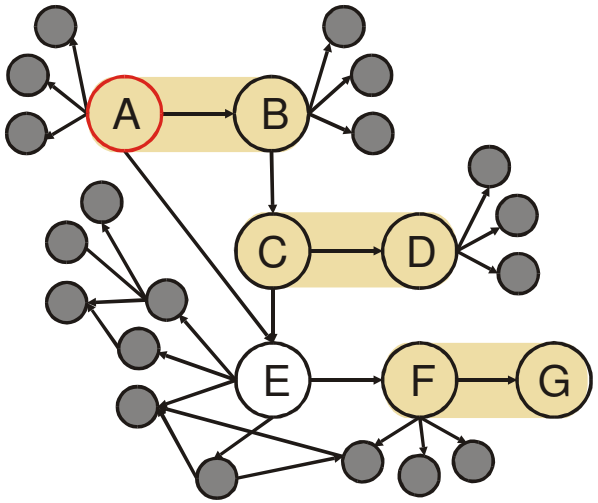
Text Mining and Microarrays

Network Construction - Pruning

- For the set of differentially expressed genes, retrieve all relations that specify transitive dependencies of degree up to 3.
- Based on relation sentences, identify all interactions that meet a specific inclusion criterion (S1P and Invasion)
- Retain only those patterns that meet the inclusion criterion.
- Each pattern contains a pair of entities (i.e., canonical gene/protein names). Use each entity as seed vertex in the network.
- For each seed vertex, find all transitive dependencies of degree 1, 2, and 3 that lead back to a differentially expressed gene and connect the vertices that are involved in the path.
- Find and display all interactions between the vertices

Text Mining and Microarrays

Network Construction - Pruning



- The vertices $A \rightarrow B$, $C \rightarrow D$, and $F \rightarrow G$, are known as seed vertices as these relations contains either one of the keywords 'S1P' or 'invasion'.

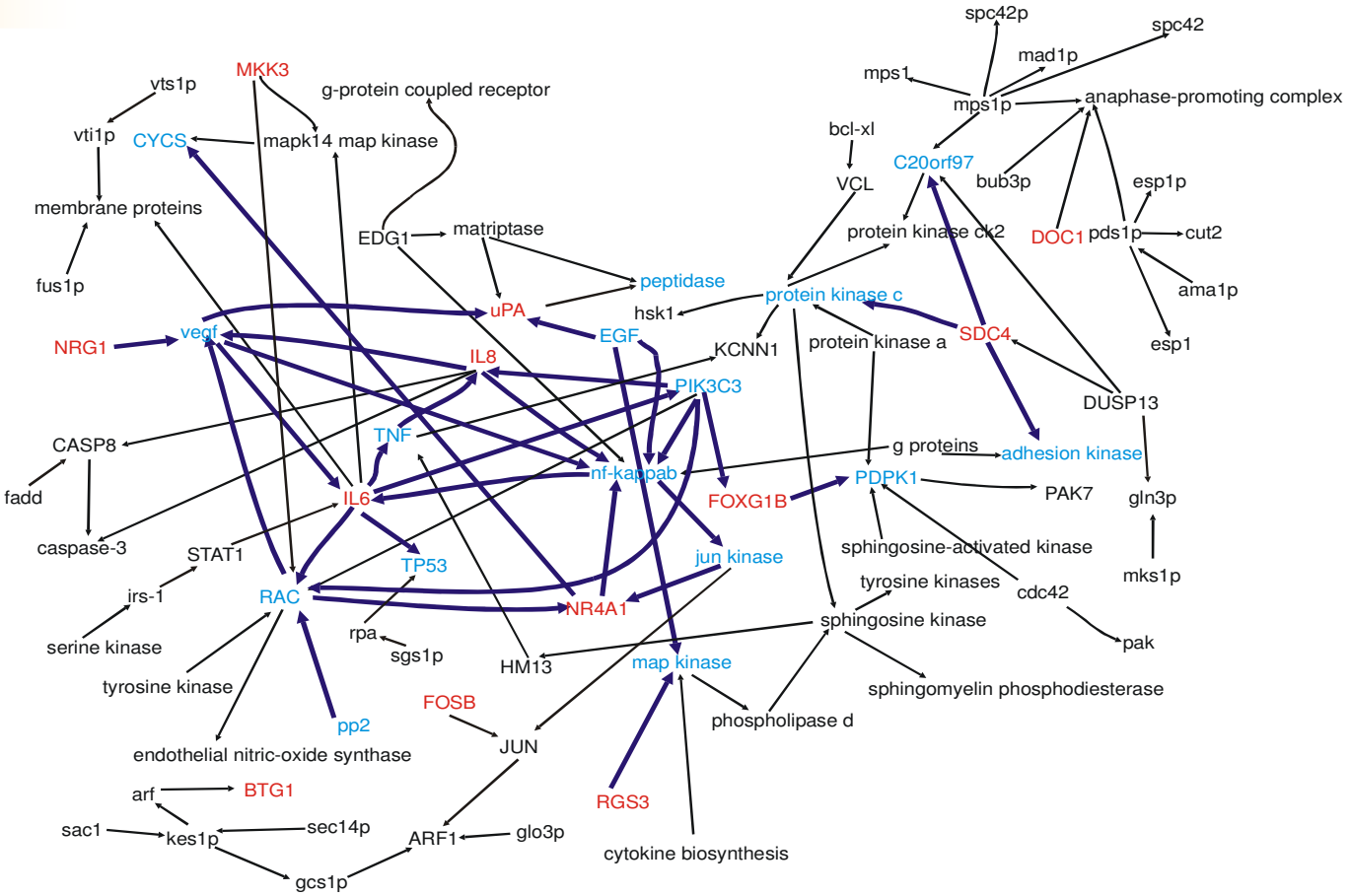
Text Mining and Microarrays

Post–processing: Network Construction

- Using the graph theory and gene-gene relations data warehouse, two types of networks were constructed
- The network that links the differentially expressed genes to S1P (Figure 1)
- the network that links the genes to tumor invasivity (Figure 2)
- Gene interaction network derived from an intersection of the S1P- and invasion-network (Figure 3)
- The resultant network is manually curated and analyzed

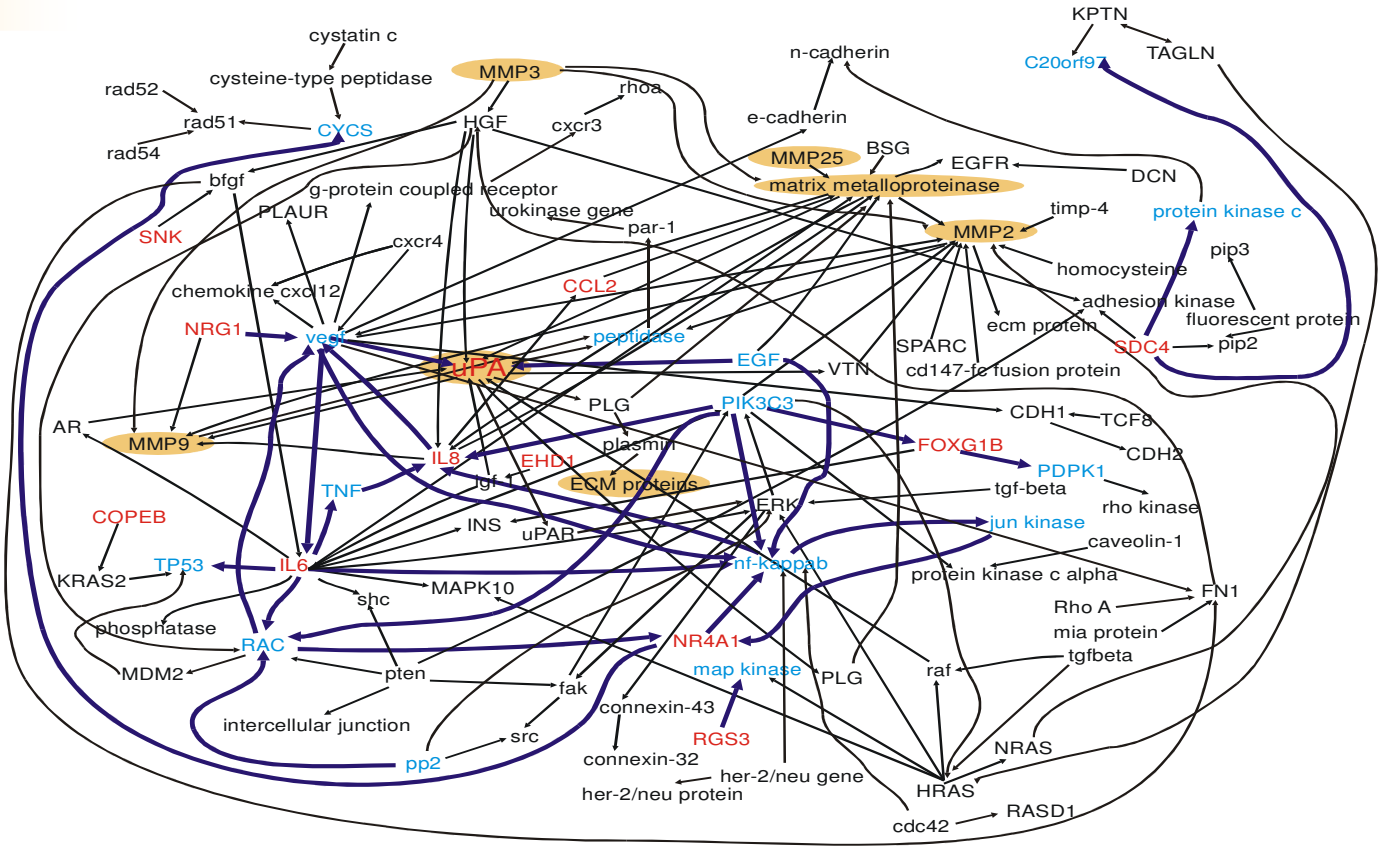
Text Mining and Microarrays

Figure 1: S1P Network



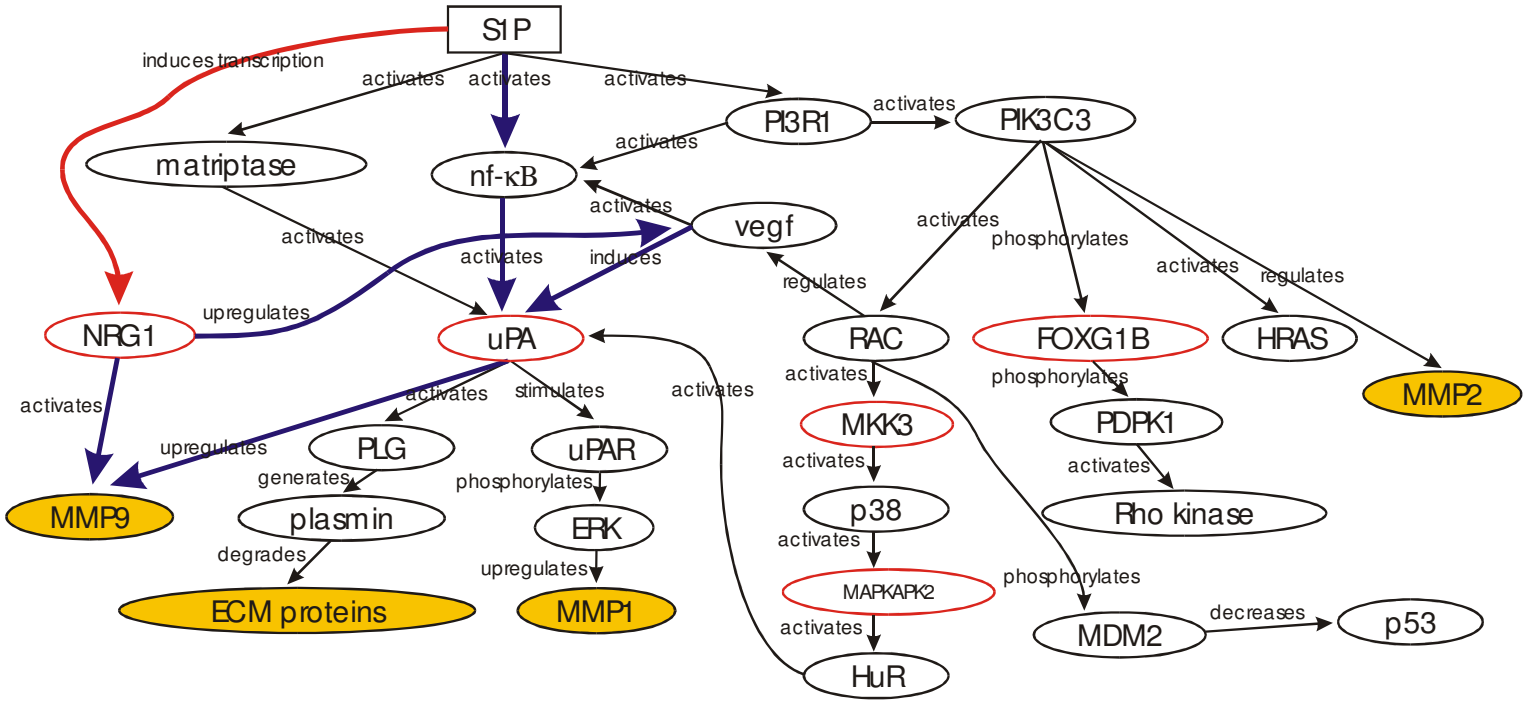
Text Mining and Microarrays

Figure 2: Invasion Network



Text Mining and Microarrays

Figure 3: Intersection Network



Gene interaction network derived from an intersection of the S1P- and invasion-Relations

Text Mining and Microarrays

Results

- Analysis of the network reveals a interesting relation – "*regulation of uPA, NRG-1 and MMP-9 by S1P*" could be a key player in the invasion of glioblastoma cells (J. Natarajan et al., BMC Bioinformatics, 7(1):373, 2006.)
- Better than other existing text mining systems such as PubGene and iHoP (abstract based)
- Most of our network information came from full-text literature that were not mentioned in abstracts

Text Mining:

Related Publications

- Text mining of full text articles and creation of a knowledge base for analysis of microarray data", Proc. Intl. Symposium on Knowledge Exploration in Life Sciences Informatics, Milan, Italy, 84-95, 2004

- Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell", BMC Bioinformatics, Aug 10;7(1):373, 2006.

Conclusions

- R&D in biology & biotechnology (B&B) are generating unprecedented volumes of literature information (abstracts and full-text)
- **Text Mining** ≡ Application & development of IT to analyze & model biological information
- Bioinformatics is not only concerned with biological sequences and structures data alone.
- IT techniques such as text mining will play dominant role in future biomedical knowledge exploration studies

Reference

- Shatkay H., "Hairpins in bookstacks: Information retrieval from biomedical text", *Briefings in Bioinformatics,* Vol. 6(3), 222-238, (2005).
- Natarajan J., Berrar D., Hack C.J., Dubitzky W., "Knowledge discovery in biology and biotechnology texts: A review of techniques, evaluation strategies, and applications", *Critical Reviews in Biotechnology*, Vol. 25, 31-52, (2005).
- Krallinger M., Valencia A., "Text-Mining and Information-Retrieval Services for Molecular Biology", *Genome Biology*, Vol 6, 224 (2005).

Acknowledgement

- Prof. Werner Dubitzky – Univeristy of Ulster
- Dr. Daniel Berrar – Unveristy of Ulster
- Martin Krallinger and Ashish V Tendulkar – APBIO Text Mining Tools in Biology
- Dr. Hagit Shatkay http://www.shatkay.org/

Thank You

Contact:

N. JEYAKUMAR: n.jeyakumar@yahoo.co.in