Text Mining Special Interest Group

[Therese Vachon, NIBR]

Informatics and Knowledge Management/Information and Knowledge Engineering

Novartis Institute for Biomedical Research,
Cambridge, MA 4-6th October 2004

The PRISM Forum

At NIBR, Text Mining is considered a solution for

- Analyzing, tagging, annotating, exploring, structuring and classifying textual data and large document sets (internal and external)
- Identifying meaningful concepts from text and relationships between concepts
- Improving the quality of text retrieval methods
- Detecting novel patterns
- Detecting similarities across textual data
- Improving navigation across data sources and document sets

What are the Critical Business Needs that Text Mining Could Address?

- Discovery of unexpected relationship and relevant information
- Generation of new hypotheses
- Assisted annotation and curation
- Unified view of heterogeneous sources
- Analysis of trends and patterns
- Analysis of complex relationships between data elements
- Detection of unexpected or emerging information
- Knowledge inference
- Contextual and semantic navigation

What Questions do Users Want to Ask ?

- Find information about specific products, targets, genes, proteins, companies, etc.
- Link literature and experimental data
- Identify biological interactions
- Semantic text analysis of full text papers
- Dynamic data flow analysis and categorization
- Monitoring competitors' capabilities and activities
- Discovering networks of associations

Text Mining Applications/Components

Function	Current	Future
Text retrieval methods	Keyword searching, fuzzy search, semantic search	Case-based reasoning Query inference
Knowledge representation	Thesauri, taxonomies, ontologies	UIMA (unstructured information management architecture), …
Annotation	Manual annotation	Assisted annotation
Categorization & Clustering	K-means partitioning, hierarchical clustering, SOMs, rule-based categorization	SVM, Latent semantic analysis
NLP or OBIIE	Concept extraction IE 1st generation	IE 2nd generation Analysis of positive and negative associations among data objects
Full text access	Indexing , access and mining of full text internal documents	Indexing, access and mining of full text literature articles
Visualisation	Interactive network of co-occurences (XML SVG) Heat maps, factorial maps, etc. and graphlets technology	Network of typed relations

Current Text Mining Applications in Novartis/NIBR?

- Knowledge Space Portal
- Novartis Knowledge Miner (Ulix)
- Competitive Intelligence Analysis Platform
- News Analysis Platform
- Research EIS
- Text mining components

What Future Applications are Planned?

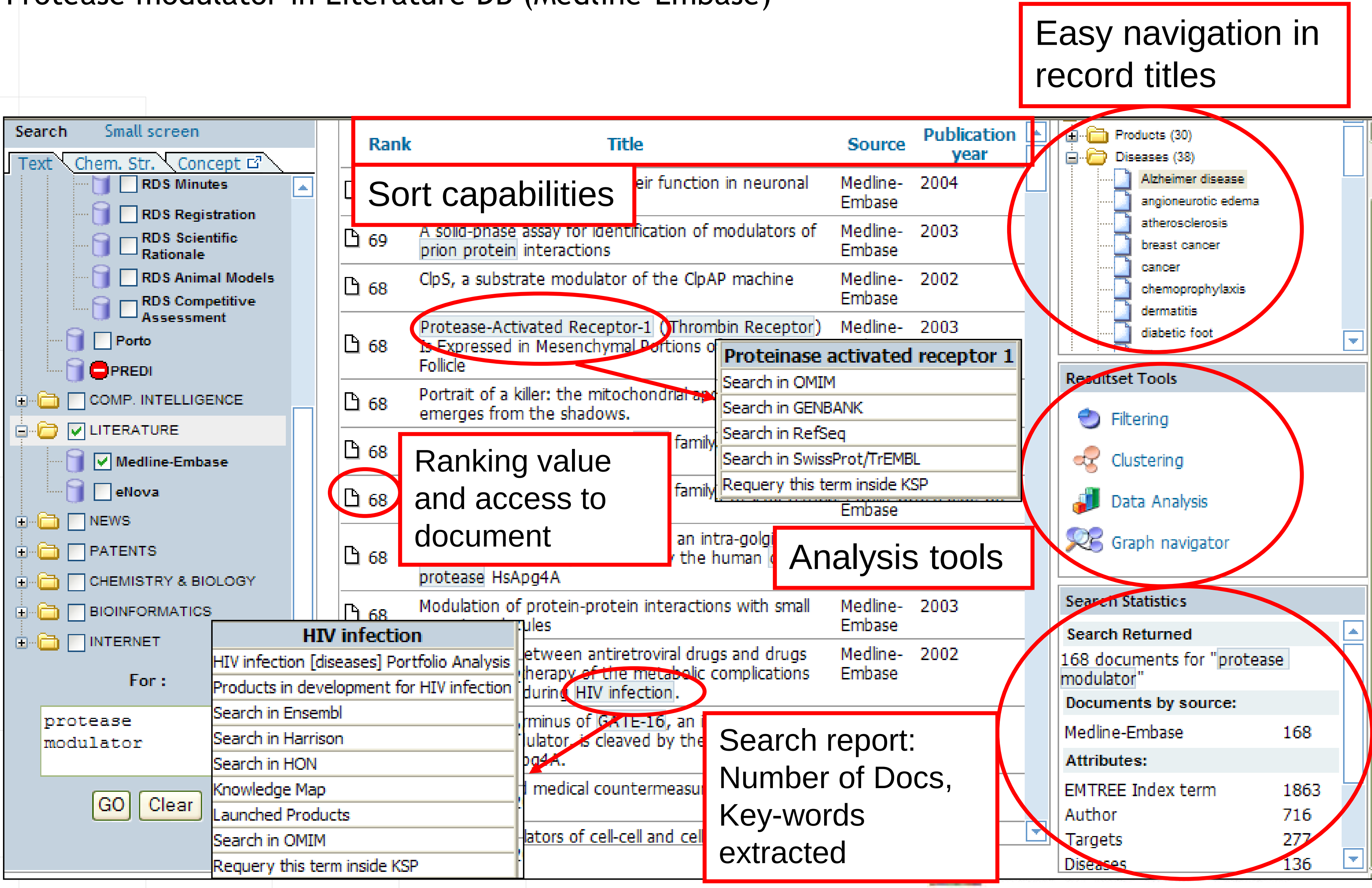
- Text mining in genomics
- Text mining in chemoinformatics
- Text mining of full text publications
- Text mining for marketing and sales
- Mining of medical enquiries
- Expert location systems
- Generic text analysis platform
- Generic categorical variable analysis platform

Knowledge Space Portal

- Provide key elements for efficiently accessing Novartis-internal and external information relevant to daily decision in the drug discovery and development process:
  - Data integration across heterogeneous data sources and applications (internal and external)
  - Consistent user interface for data retrieval, exploration and analysis across all data types
  - Contextual (ultralink), tree-based (static or dynamic taxonomies) and semantic (knowledge map) navigation
  - Data exploration and analysis methods
  - Personalized views
  - Collaborative, annotation and information sharing tools
  - Alerting

Display-Navigation-Ultralink

Protease modulator in Literature DB (Medline-Embase)



Easy navigation in record titles

Sort capabilities

Ranking value and access to document

Analysis tools

Search report: Number of Docs, Key-words extracted

Rank	Title	Source	Publication year
	...eir function in neuronal	Medline-Embase	2004
69	A solid-phase assay for identification of modulators of prion protein interactions	Medline-Embase	2003
68	ClpS, a substrate modulator of the ClpAP machine	Medline-Embase	2002
68	Protease-Activated Receptor-1 (Thrombin Receptor) Is Expressed in Mesenchymal Portions o... Follicle	Medline-	2003
68	Portrait of a killer: the mitochondrial ap... emerges from the shadows.		
68	... family		
68	... family	Embase	
68	... an intra-golgi ... the human ... protease HsApg4A		
68	Modulation of protein-protein interactions with small ...ules	Medline-Embase	2003
	...etween antiretroviral drugs and drugs ...herapy of the metabolic complications ...during HIV infection.	Medline-Embase	2002
	...rminus of GATE-16, an i... ...ulator, is cleaved by the ...pg4A.		
	... medical countermeasu...		
	...lators of cell-cell and cel...		

For : protease modulator

Search Statistics

Search Returned

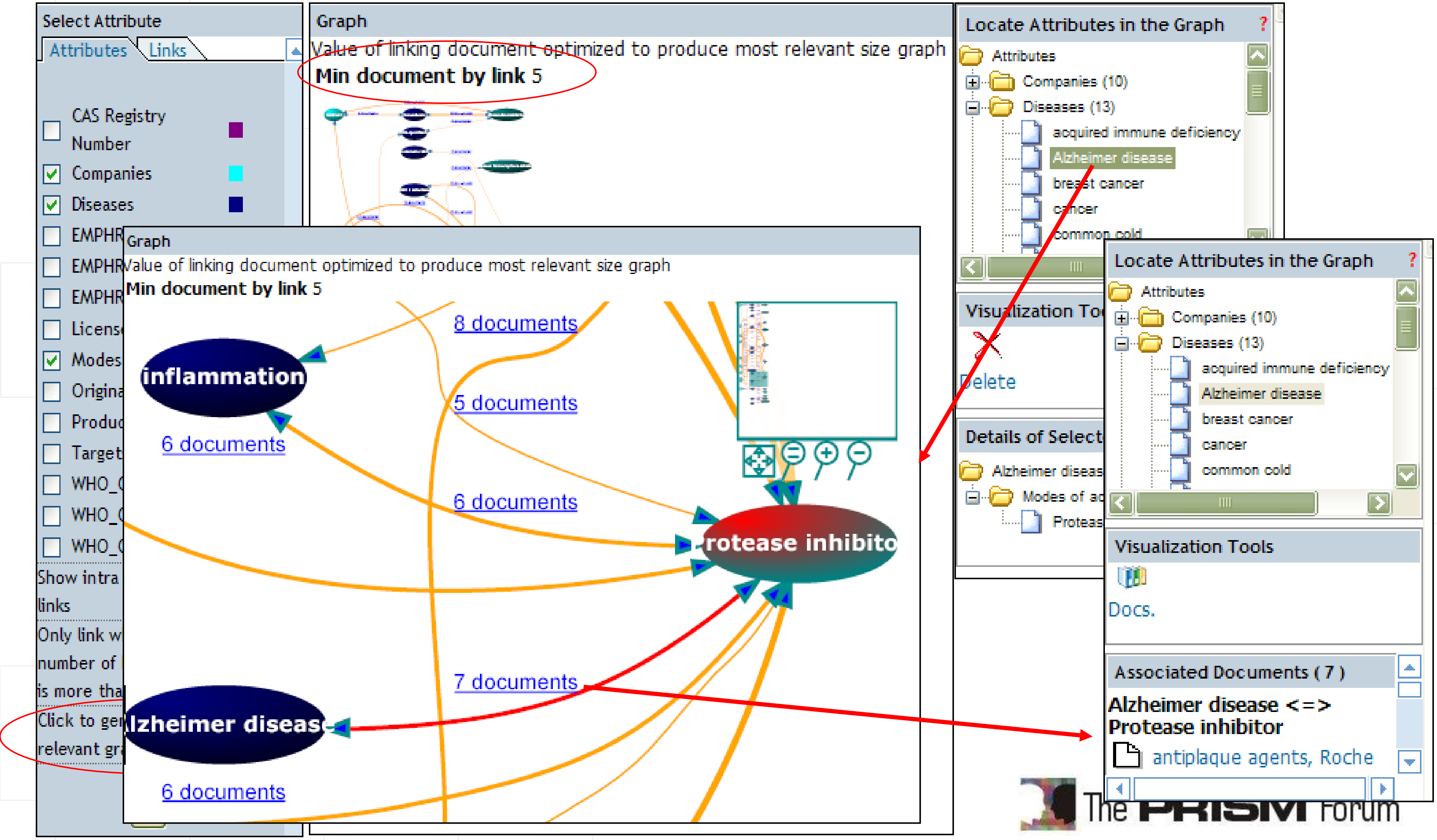
168 documents for "protease modulator"

Documents by source:	
Medline-Embase	168
Attributes:	
EMTREE Index term	1863
Author	716
Targets	277
Diseases	136

Graph Navigator –

Protease modulators in CI DBs July 2004 - ADIS & Pharmaprojects

Graph navigator



Select Attribute

Attributes Links

- □ CAS Registry Number
- ☑ Companies
- ☑ Diseases
- □ EMPHR
- □ EMPHR
- □ EMPHR
- □ License
- ☑ Modes
- □ Origina
- □ Produc
- □ Target
- □ WHO_(
- □ WHO_(
- □ WHO_(

Show intra links

Only link w number of is more tha

Click to gen relevant gra

Graph

Value of linking document optimized to produce most relevant size graph

Min document by link 5

Graph

Value of linking document optimized to produce most relevant size graph

Min document by link 5

Locate Attributes in the Graph

- Attributes
  - Companies (10)
  - Diseases (13)
    - acquired immune deficiency
    - Alzheimer disease
    - breast cancer
    - cancer
    - common cold

Visualization To

Delete

Details of Select

- Alzheimer diseas
- Modes of ac
  - Proteas

Locate Attributes in the Graph

- Attributes
  - Companies (10)
  - Diseases (13)
    - acquired immune deficiency
    - Alzheimer disease
    - breast cancer
    - cancer
    - common cold

Visualization Tools

Docs.

Associated Documents (7)

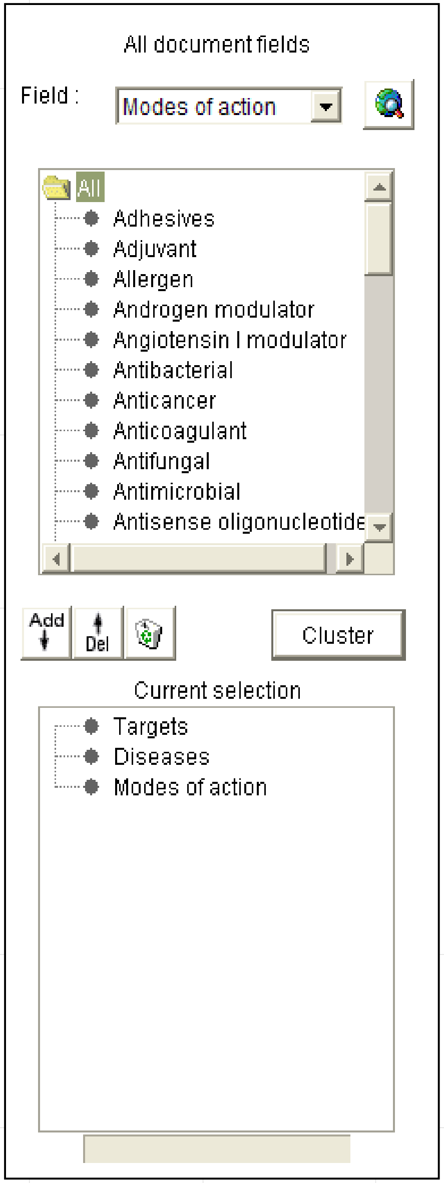
Alzheimer disease <=> Protease inhibitor

antiplaque agents, Roche

The PRISM Forum

Clustering

Clustering



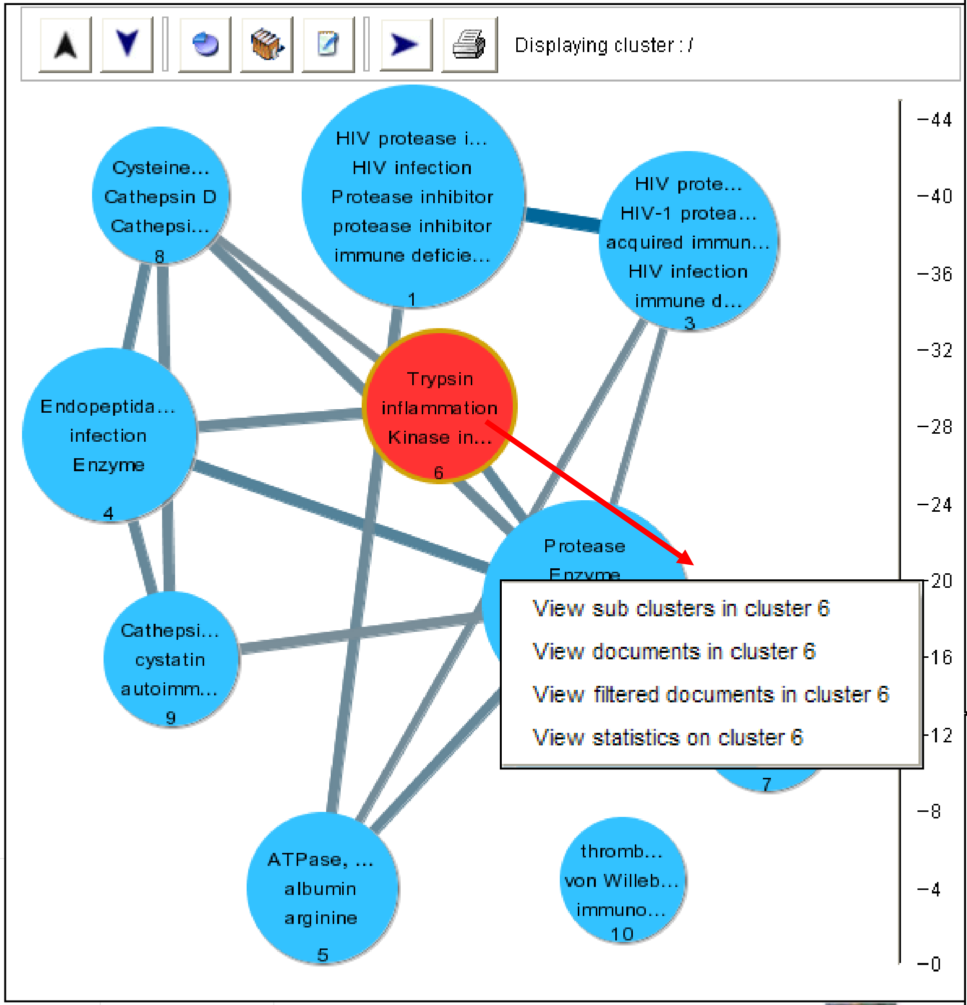
All document fields

Field : Modes of action

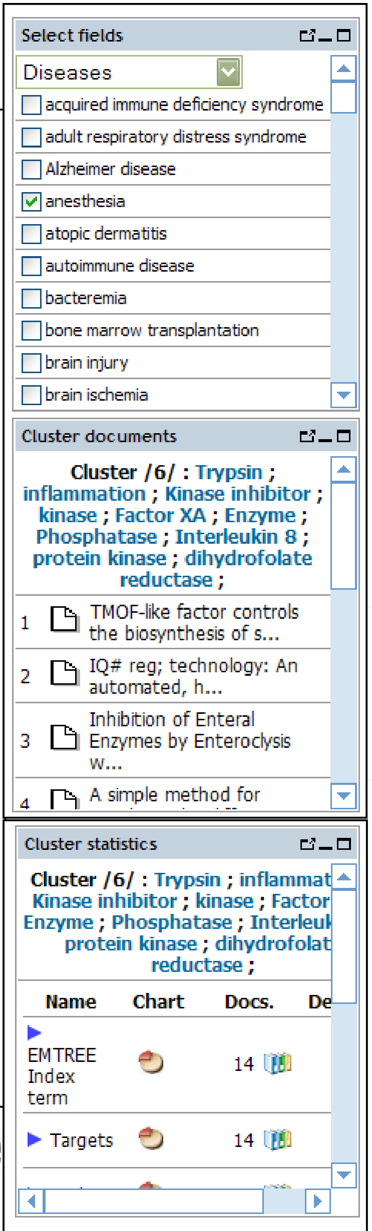
- All
  - Adhesives
  - Adjuvant
  - Allergen
  - Androgen modulator
  - Angiotensin I modulator
  - Antibacterial
  - Anticancer
  - Anticoagulant
  - Antifungal
  - Antimicrobial
  - Antisense oligonucleotide

Current selection

- Targets
- Diseases
- Modes of action



Displaying cluster : /



Select fields

Diseases

- ☐ acquired immune deficiency syndrome
- ☐ adult respiratory distress syndrome
- ☐ Alzheimer disease
- ☑ anesthesia
- ☐ atopic dermatitis
- ☐ autoimmune disease
- ☐ bacteremia
- ☐ bone marrow transplantation
- ☐ brain injury
- ☐ brain ischemia

Cluster documents

Cluster /6/ : Trypsin ; inflammation ; Kinase inhibitor ; kinase ; Factor XA ; Enzyme ; Phosphatase ; Interleukin 8 ; protein kinase ; dihydrofolate reductase ;

1 TMOF-like factor controls the biosynthesis of s...

2 IQ# reg; technology: An automated, h...

3 Inhibition of Enteral Enzymes by Enteroclysis W...

4 A simple method for

Cluster statistics

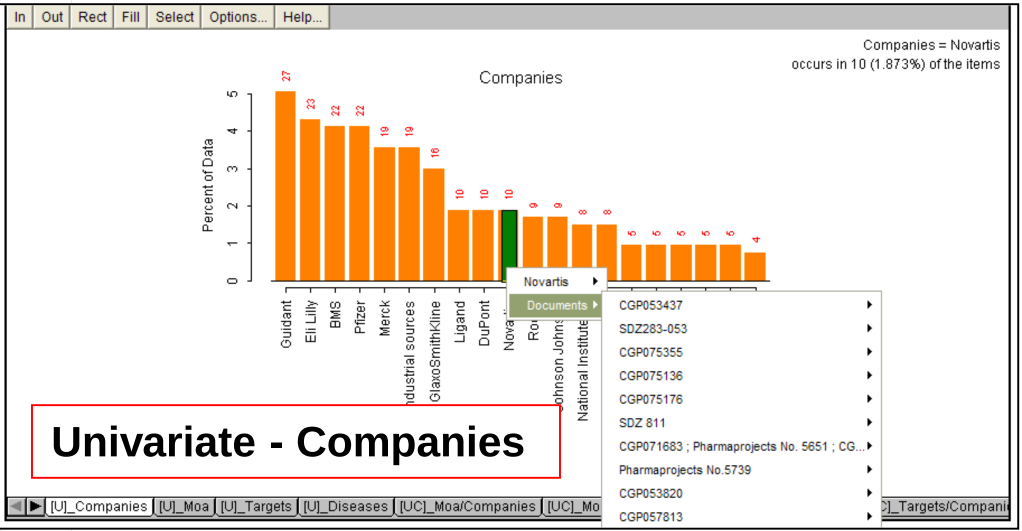
Cluster /6/ : Trypsin ; inflammat Kinase inhibitor ; kinase ; Factor Enzyme ; Phosphatase ; Interleuk protein kinase ; dihydrofolat reductase ;

Name	Chart	Docs.	De
EMTREE Index term		14	
Targets		14	

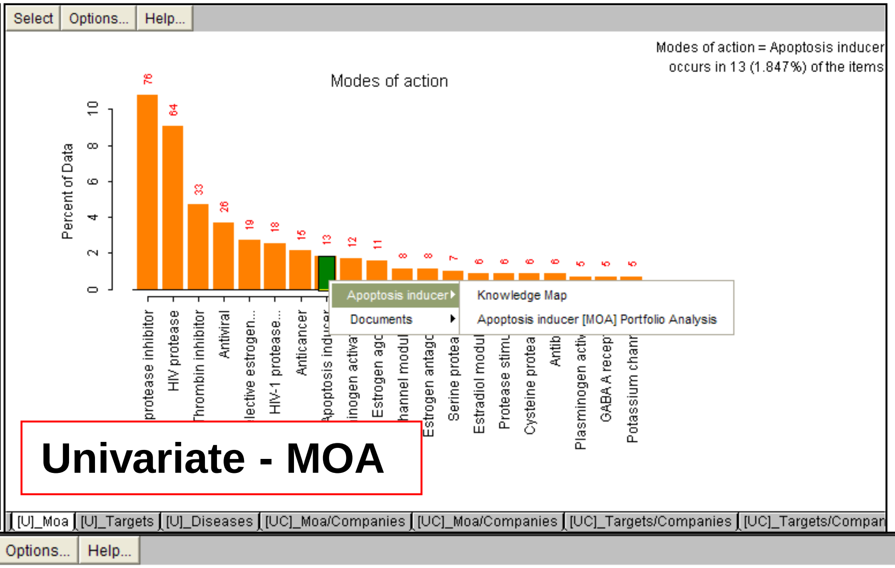
Data Analysis –

Protease modulators in CI DBs July 2004 - ADIS & Pharmaprojects

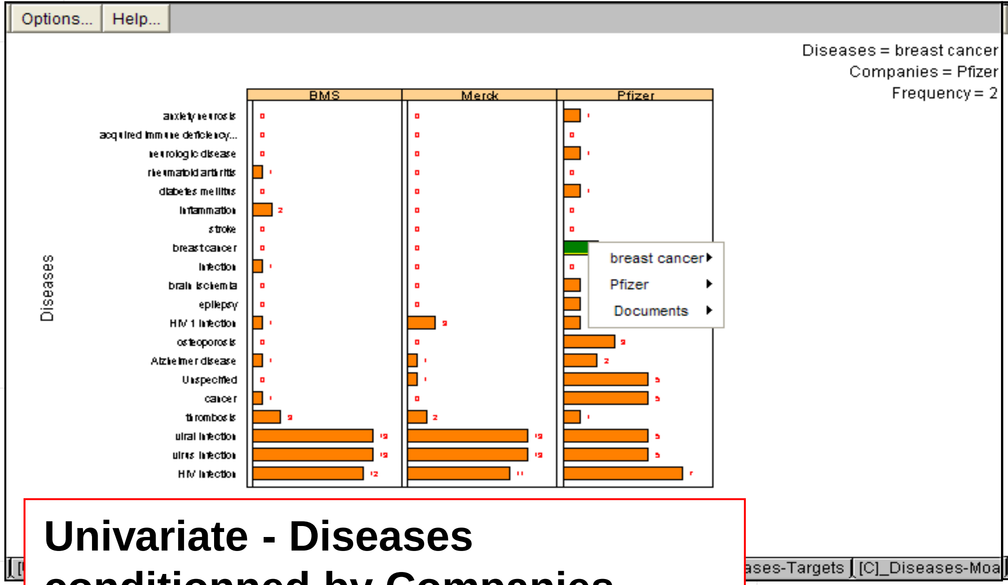
Data Analysis



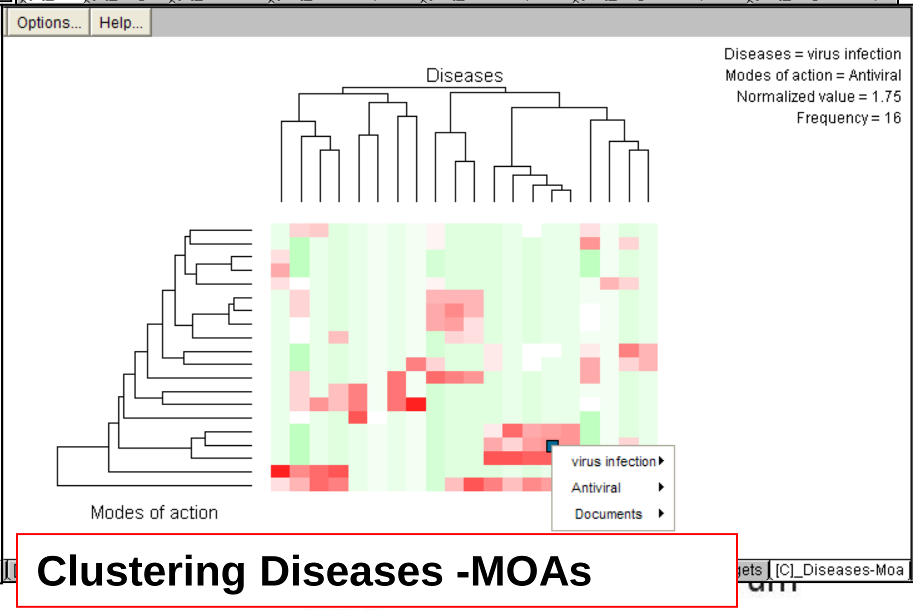
Univariate - Companies



Univariate - MOA



Univariate - Diseases conditionned by Companies



Clustering Diseases -MOAs

Conclusions

- Text mining techniques have already been implemented in relevant areas
- Additional techniques need to be developed/tested/implemented especially in the field of information extraction/NLP and in the field of categorization
- Collaboration with external partners/research institutes needed
- Text mining components to be applied to all applications dealing with text