

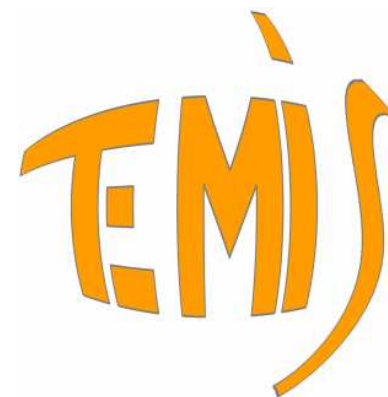
Beyond Search >>

Recognition of Chemical Entities in Scientific Literature



**Charles Huot, Chief Operating Officer
& Co-Founder**

**International Chemical Information Conference -
Nîmes, October 24th, 2006**



1

Drug Discovery Process

2

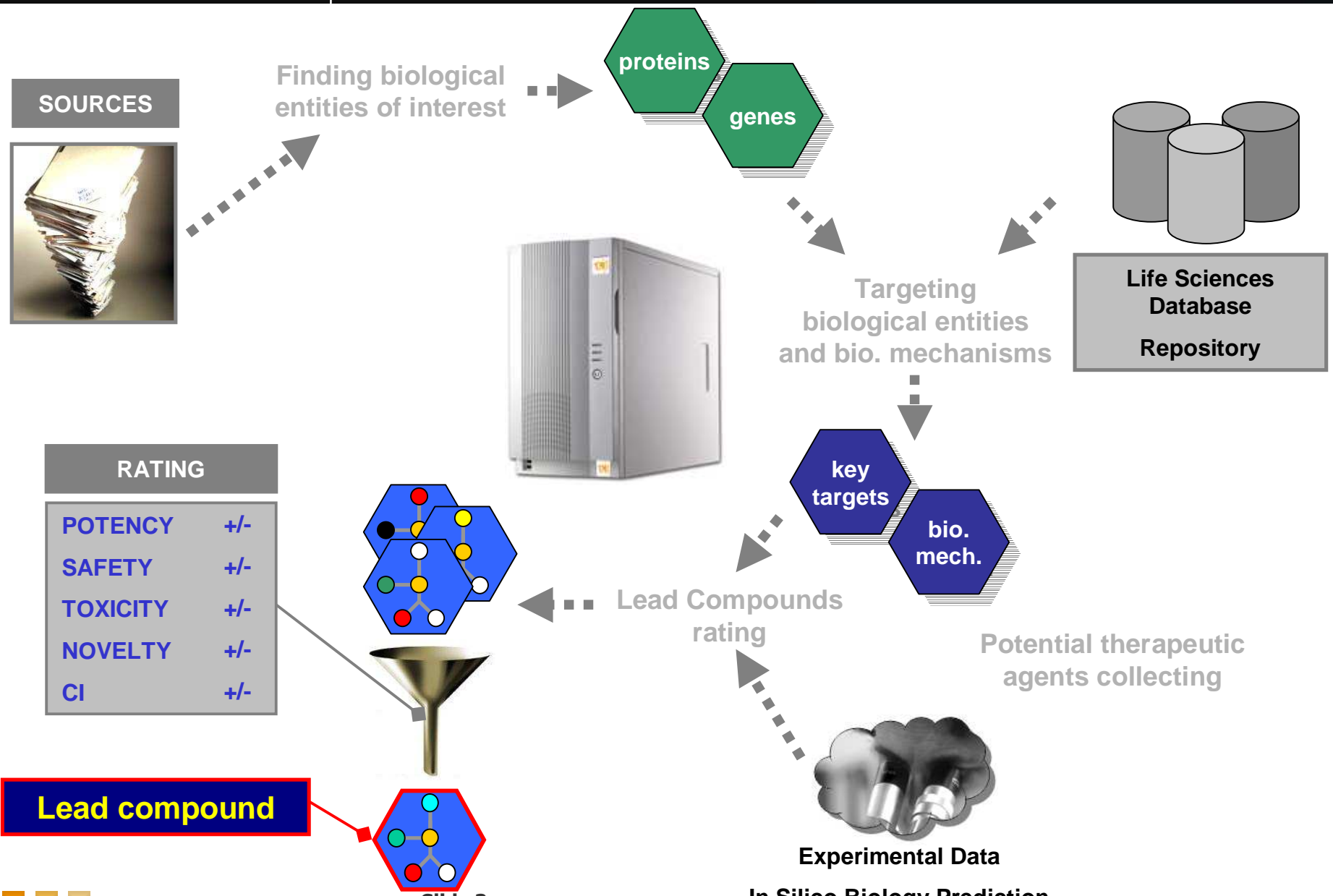
Recognition of Chemical Entities

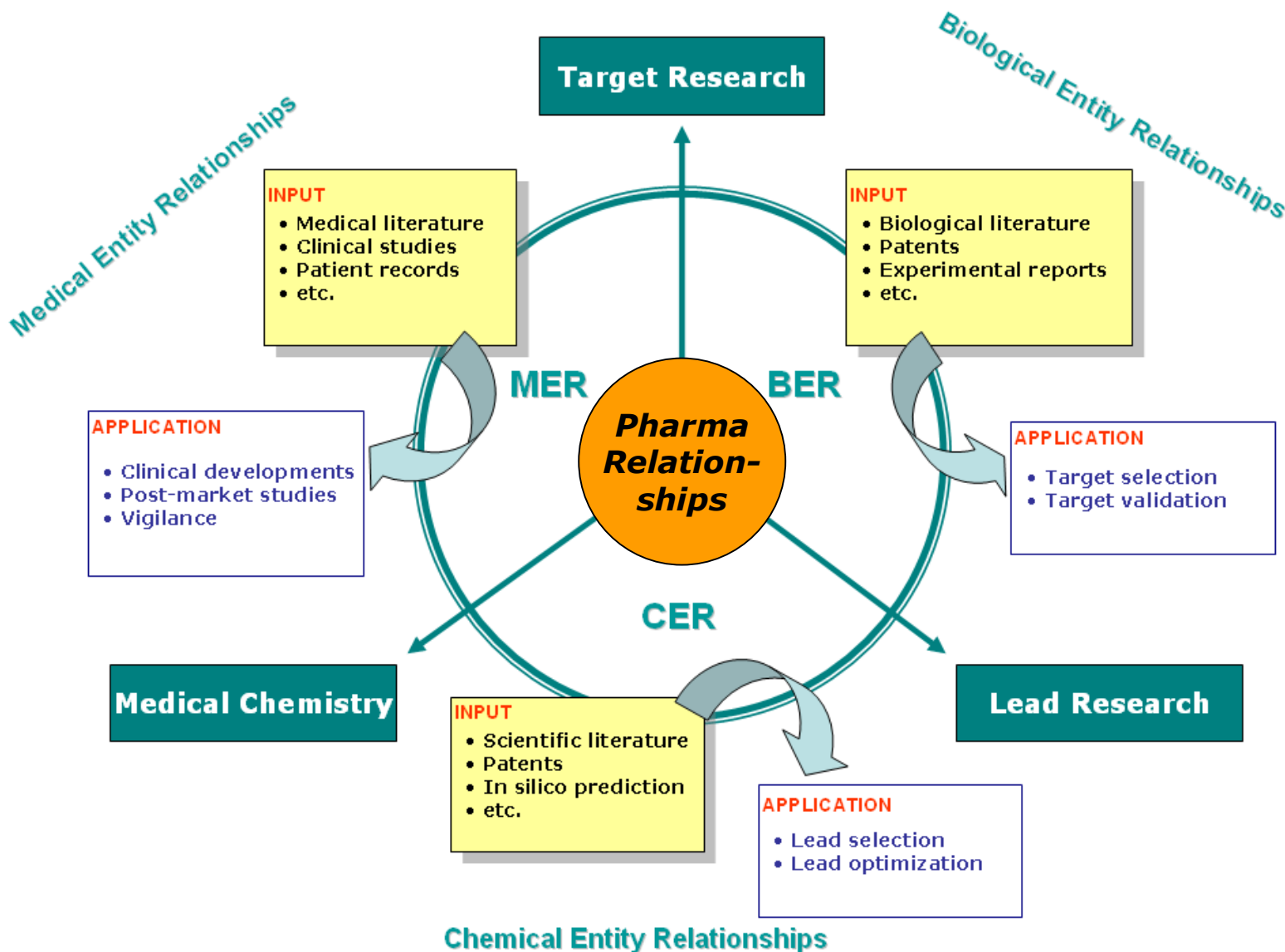
3

Case study - Scenario Highlight

4

TEMIS Presentation





1

Drug Discovery Process

2

Recognition of Chemical Entities

3

Case study - Scenario Highlight

4

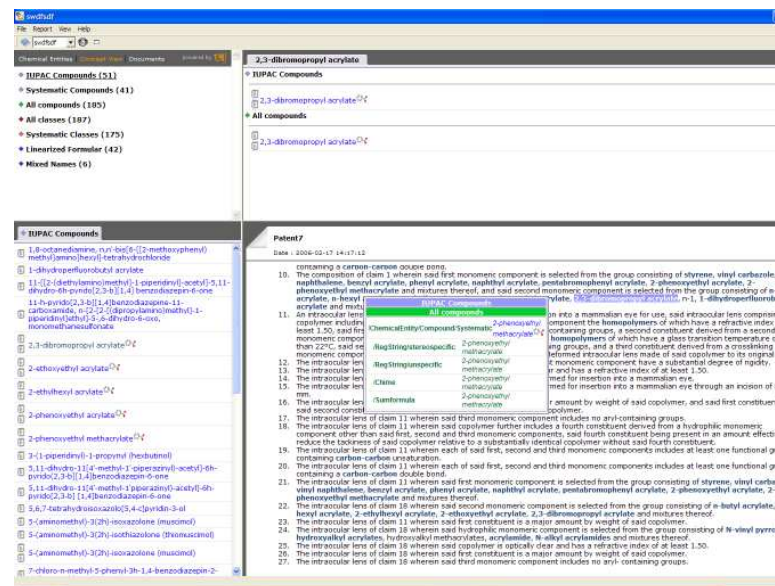
TEMIS Presentation

- A new information discovery weapon

- To analyze chemistry-related documents

- Key features

- Powerful recognition & extraction of chemical entities
- Highlighting within text for intelligent navigation
- Display graphical representation of the molecular structure



- Result of close cooperation between MDL & TEMIS

- Extraction & structure recognition performed thru MDL technology embedded within a TEMIS *Skill Cartridge*TM
- Architecture, workflow & display built by TEMIS

1. Recognition of different name representations
2. Name-To-Structure layer
 - Integration of a 3rd party tools (CambridgeSoft,...) and database look-up (internal, Beistein, ACD,...)
 - Customizable dictionary incl. pattern recognition
3. Structure-based normalization via Registration String

- 95% of chemical compounds extracted
 - Systematical Chemical Names according to IUPAC
 - Inverted Chemical Names according to CAS for specific stems
 - Trivial names, element names, element symbols
 - Half-systematic chemical names (with trivial fragments)
 - Explicitly allowed abbreviations
 - Chemical formula-like constructions, linearized structural formulas
 - Compound names from systematic and/or trivial fragments
 - Amino acids and nucleic acids
 - Chemical compound classes
 - Enzyme names according to IUPAC

- 95% of chemical components extracted
- Highlighting of chemical entities within text
 - Enable instant & rapid visualization of most-valuable information within large documents

Highlighting

On-demand explanations

Patent7
Date: 2006-02-17 14:17:12

containing a carbon-carbon double bond.

10. The composition of claim 1 wherein said first monomeric component is selected from the group consisting of styrene, vinyl carbazole, vinyl naphthalene, benzyl acrylate, phenyl acrylate, naphthyl acrylate, pentabromophenyl acrylate, 2-phenoxyethyl acrylate, 2-phenoxyethyl methacrylate and mixtures thereof, and said second monomeric component is selected from the group consisting of n-butyl acrylate, n-hexyl acrylate, 2-ethylhexyl acrylate, 2-ethoxyethyl acrylate, 2,3-dibromopropyl acrylate, n-1, 1-dihydroperfluorobutyl acrylate and mixtures thereof.

11. An intraocular lens size of at least 1.50, said first monomeric component other than 22°C, said second monomeric component in a copolymer including a first monomeric component other than 22°C, said second monomeric component in a copolymer including a first monomeric component other than 22°C.

12. The intraocular lens of claim 11 wherein said first monomeric component includes no aryl-containing groups.

13. The intraocular lens of claim 11 wherein said copolymer further includes a fourth constituent derived from a hydrophilic monomeric component other than said first, second and third monomeric components, said fourth constituent being present in an amount effective to reduce the tackiness of said copolymer relative to a substantially identical copolymer without said fourth constituent.

14. The intraocular lens of claim 11 wherein each of said first, second and third monomeric components includes at least one functional group containing carbon-carbon unsaturation.

15. The intraocular lens of claim 11 wherein each of said first, second and third monomeric components includes at least one functional group containing a carbon-carbon double bond.

16. The intraocular lens of claim 11 wherein said first monomeric component is selected from the group consisting of styrene, vinyl carbazole, vinyl naphthalene, benzyl acrylate, phenyl acrylate, naphthyl acrylate, pentabromophenyl acrylate, 2-phenoxyethyl acrylate, 2-phenoxyethyl methacrylate and mixtures thereof.

17. The intraocular lens of claim 11 wherein said second monomeric component is selected from the group consisting of n-butyl acrylate, n-hexyl acrylate, 2-ethylhexyl acrylate, 2-ethoxyethyl acrylate, 2,3-dibromopropyl acrylate and mixtures thereof.

18. The intraocular lens of claim 11 wherein said first constituent is a major amount by weight of said copolymer.

19. The intraocular lens of claim 18 wherein said hydrophilic monomeric component is selected from the group consisting of N-vinyl pyrrolidone, hydroxyalkyl acrylates, hydroxyalkyl methacrylates, acrylamide, N-alkyl acrylamides and mixtures thereof.

20. The intraocular lens of claim 18 wherein said copolymer is optically clear and has a refractive index of at least 1.50.

21. The intraocular lens of claim 18 wherein said first constituent is a major amount by weight of said copolymer.

22. The intraocular lens of claim 18 wherein said third monomeric component includes no aryl-containing groups.

23. The intraocular lens of claim 18 wherein said first constituent is a major amount by weight of said copolymer.

24. The intraocular lens of claim 18 wherein said hydrophilic monomeric component is selected from the group consisting of N-vinyl pyrrolidone, hydroxyalkyl acrylates, hydroxyalkyl methacrylates, acrylamide, N-alkyl acrylamides and mixtures thereof.

25. The intraocular lens of claim 18 wherein said copolymer is optically clear and has a refractive index of at least 1.50.

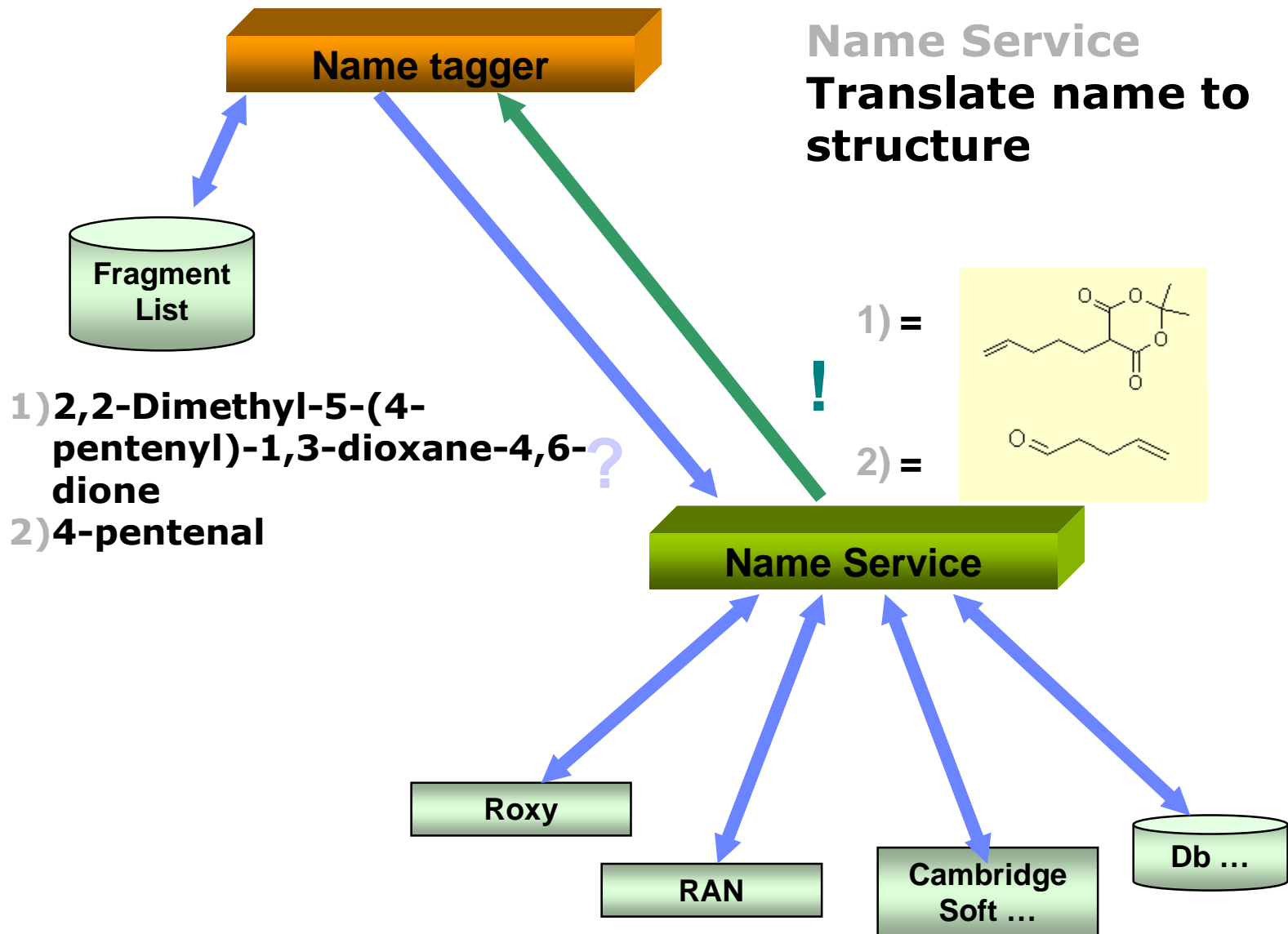
26. The intraocular lens of claim 18 wherein said first constituent is a major amount by weight of said copolymer.

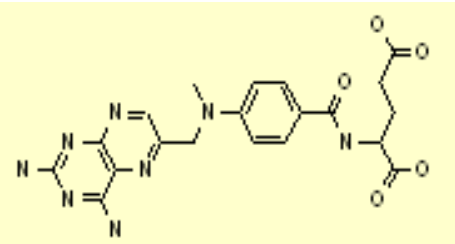
27. The intraocular lens of claim 18 wherein said third monomeric component includes no aryl-containing groups.

IUPAC Compounds

All compounds

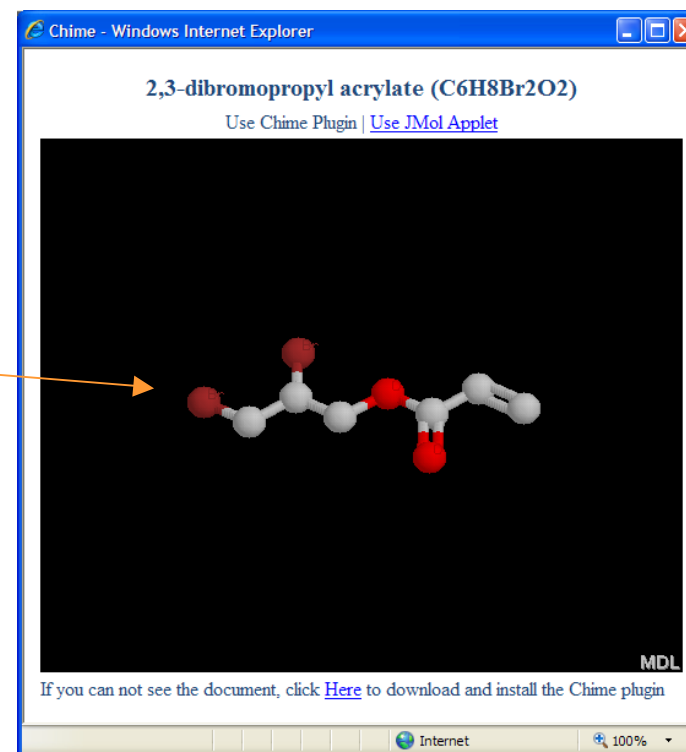
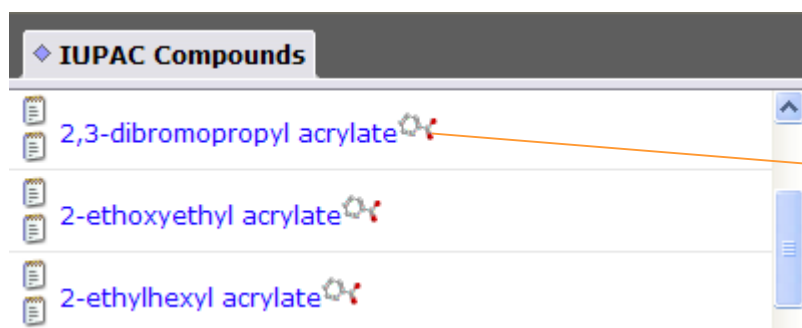
/ChemicalEntity/Compound/Systematic	2-phenoxyethyl methacrylate
/RegString/stereospecific	2-phenoxyethyl methacrylate
/RegString/unspecific	2-phenoxyethyl methacrylate
/Chime	2-phenoxyethyl methacrylate
/Sumformula	2-phenoxyethyl methacrylate



Chemical name	structure
<p>Amethopterin 4-amino-N10-methylpteroylglutamic acid MEXATE METHOTREXATE 2-{4-[(2,4-Diamino-pteridin-6-ylmethyl)-methyl-amino]-benzoylamino}-pentanedioic acid</p>	 <p style="text-align: center;">↓</p>

- **REGSTRING**
 - Unique fingerprint
 - Easy duplicate check

- 95% of chemical components extracted
- Highlighting of chemical entities within text
- Link chemical name to associated structure
 - Thru structure graphical display



Enables the user to search, navigate, analyze the scientific literature ...

- ... interactively ...
 - From full-text documents, abstracts and database text fields
- ... with a Life Sciences specific view ...
 - Detection of Medical, Biological and Chemical entities (e.g. diseases, proteins, chemical names)
 - Translation of the entities to their native representations (e.g. chemical structures & bio-sequences)
 - Recognition of entity relationships (e.g. drug-disease or protein-ligand)
- & Support decisions with facts
 - Scientists, information professionals, patent officers are supported to backup their intuition with factual information in a project.

1

Drug Discovery Process

2

Recognition of Chemical Entities

3

Case study - Scenario Highlight

4

TEMIS Presentation

- You have 2 hours to produce a synthetic report for your manager and give your colleagues first elements in an crisis management meeting
- The topic? atherosclerosis
 - which potential targets are mentioned ?
 - what are the latest studied ones
 - which associated diseases fall in our therapeutic areas ?
 - which treatments were successfully used on these associated diseases
 - can we find potential new compounds similar to these known treatments ?)

First, perform a search on « atherosclerosis » to locate relevant documents

The screenshot displays the Luxid information intelligence search interface. At the top, the Luxid logo and navigation menu are visible. The search bar contains the term "atherosclerosis". Below the search bar, there are options for "Max items" (set to "No limit") and "Results per page" (set to "10"). A "Search now" button is present. The search criteria are displayed as "Search atherosclerosis into text, title having Source contain, Disorder".

Below the search results area, there are two filter panels. The left panel shows filters for "PublishingDate" and "Source", with "Source" checked. The right panel shows filters for "Disorder" and "Enzyme", with "Disorder" checked. Two callout boxes with arrows point to these filters: "Browse all sources" points to the "Source" filter, and "Focus on disorders" points to the "Disorder" filter.

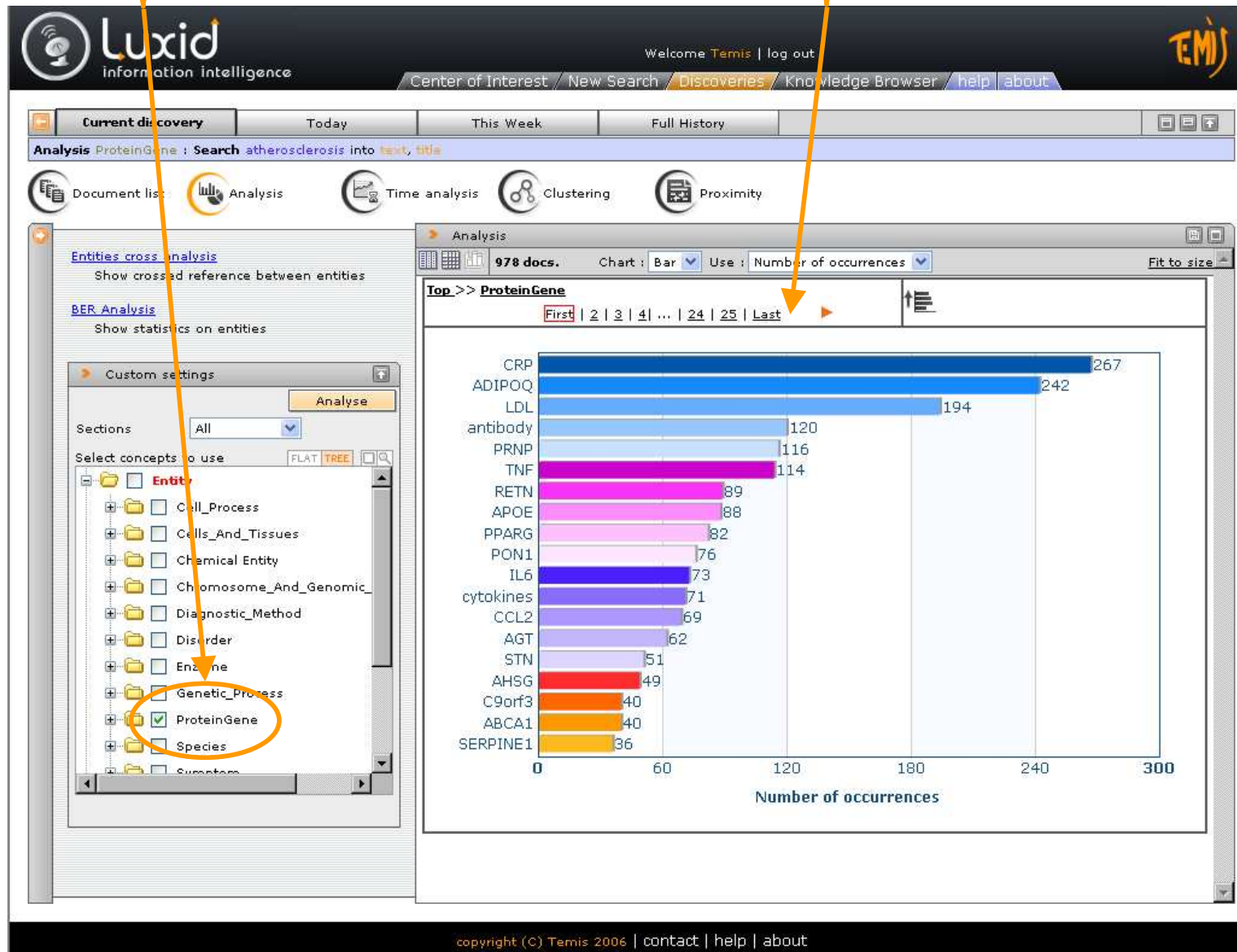
At the bottom of the interface, there is a footer with the text "copyright (C) Temis 2006 | contact | help | about".

I receive a first set of 978 documents with what to start my discovery. I can't read them all so look for Proteins and Genes frequency distribution to identify potential targets and biomarkers

The screenshot displays the Luxid Information Intelligence web interface. At the top left is the Luxid logo with the tagline 'Information Intelligence'. To the right, it says 'Welcome Temis | log out'. Below this is a navigation bar with links for 'Center of Interest', 'New Search', 'Discoveries', 'Knowledge Browser', 'help', and 'about'. The main interface has a search bar containing 'Search atherosclerosis into 3047 titles'. Below the search bar are several analysis tool icons: 'Document list', 'Analysis' (circled in orange), 'Time analysis', 'Clustering', and 'Proximity'. An orange arrow points from the 'Analysis' icon to a yellow box containing the text 'Activate 'Analysis''. The 'Document List' window shows 978 documents. The first document is titled '[Osteopontin and atherosclerosis]' with top terms: atherosclerosis, Osteopontin, atherosclerosis, atherosclerosis, Osteopontin, SPP1, Osteopontin. The second document is '[Atherosclerosis: butter on the arteries?]' with top terms: Atherosclerosis, atherosclerosis, Atherosclerosis, butter, butter on the artery. The third document is 'Prediabetes & atherosclerosis: what's the connection?' with top terms: atherosclerosis, atherosclerosis, Prediabetes, atherosclerosis, connection. The fourth document is '[Postprandial hyperglycemia as a risk factor of atherosclerosis]' with top terms: hyperglycemia, atherosclerosis, Postprandial hyperglycemia, atherosclerosis, risk factor of atherosclerosis, Postprandial hyperglycemia, risk factor. The fifth document is 'Visual vignette. Unilateral carotid vascular atherosclerosis.' with top terms: atherosclerosis, Unilateral carotid vascular atherosclerosis, atherosclerosis, Unilateral carotid vascular atherosclerosis, vignette, visual vignette, unilateral carotid vascular atherosclerosis. The sixth document is 'Why does atherosclerosis occur where it occurs?' with top terms: atherosclerosis, atherosclerosis, atherosclerosis. The seventh document is 'Racial/Ethnic differences in subclinical atherosclerosis among adults with diabetes: the multiethnic study of atherosclerosis.' with top terms: Racial/Ethnic ... in subclinical atherosclerosis among adults with ... study of atherosclerosis.

copyright (C) Temis 2006 | contact | help | about

The distribution gives you an idea of strongly associated proteins/genes (potential targets) : CRP, ADIPOQ, LDL...



Now we need to look for the “hottest” in the last two years of literature

Activate 'Time Analysis'

The screenshot shows the Luxid Information Intelligence interface. At the top, there is a search bar with the text "Analysis ProteinGene, 2005, 2006 : Search atherosclerosis into text, title". Below the search bar, there are several icons for "Document list", "Analysis", "Time analysis", "Clustering", and "Proximity". The main window displays a table with 978 documents, filtered by "ProteinGene". The table has columns for various protein genes: ADIPOQ, CD36, CRP, cytokine, IL6, LDL, LDLR, STN, AGT, ALB, anticoagulant, antioxidant, APLN, C3, CCL2, and chemokine. The rows represent monthly intervals from 2005-03 to 2006-09. An orange oval highlights the ADIPOQ, CRP, and LDL columns, and a callout box points to these columns with the text "Notice the strong trends about ADIPOQ, CRP and LDL in the last two years".

	ADIPOQ	CD36	CRP	cytokine	IL6	LDL	LDLR	STN	AGT	ALB	anticoagulant	antioxidant	APLN	C3	CCL2	chemokine
2005-03	1	1		1	1	1		1	2			1				
2005-04	2		2		1				1							
2005-05	2		3	1	1	6			2			2			1	
2005-06	2		4		1	3		3	3	1	1					
2005-07	1		5		2	5		5	5	1		1				
2005-08	2	1	6		1	7	2		3			2			1	2
2005-09	1		2			3	1	1				2				
2005-10	2		4	1	1	4		5	1			2				1
2005-11	3		8	3		4			1						1	1
2005-12			6			2					2	2		1		
2006-01			6		2	4		2	1	1	1	1				
2006-02	4		5		2	2		2	2	2	2	2				1
2006-03	6		5		3	2		2	2	2	2	2				1
2006-04	6		5		3	2		2	2	2	2	2				1
2006-05	1	1	9		1	1		1	1	1	1	1				1
2006-06	3		3	1	2	2	1	1	1	1	1	1			1	
2006-07	4	2	6		2	5	2	2	2	2	2	2			1	
2006-08	7		10		2	6	1	1	1	1	1	3			3	1
2006-09	4	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1

I now want to visualize all the diseases in the vicinity of ADIPOQ (Adiponectin)

Activate 'Proximity Analysis'

The screenshot displays the Luxid Information Intelligence interface. At the top, the search bar contains the query: "Proximity around ADIPOQ and Disorder : Search atherosclerosis into text, title". Below the search bar, it indicates "38/978 docs." and "Found 38 docs and 93 sentences". The main area is a network graph with "ADIPOQ" at the center. Numerous nodes representing diseases and conditions are connected to the center by lines of varying thickness, indicating proximity. Nodes include: Co-Morbid Diabetes, metabolic and circulatory diseases, prediabetes or non-diabetic metabolic syndrome, angina pectoris, Type 2 diabetes, type 2 diabetes mellitus, tumor, glucose-induced hyperinsulinaemia, cardiovascular diseases, myocardial infarction, Linz Peripheral Arterial Disease, adiponectin and metabolic abn, metabolic syndrome, polycystic ovary sync, albuminuria, insulin resistance, type 2 diabetes Disease (CVD), obesity-related metabolic sync, coronary artery disease, coronary atherosclerosis, obesity, treat obesity, metabolic disease, systemic atherosclerotic disease, tinemia, occlusive disease (PAOD), congestive heart failure, dyslipidemia, atherosclerosis, diabetes, congestive heart failure, coronary heart disease, ESRD, hypertension, c dyslipidemia, insulin Resistance, heart failure, assessed atherosclerosis, stage renal disease, cardiovascular disease, Diabetes Mellitus, glycaemia, End-Stage Renal Disease, peripheral arterial occlusive disease, obesity-relate, arterial occlusive disease, symptomatic atherosclerotic peripheral arterial disease (PAU), and lipidemia. On the right side, there are two panels: "Display" and "Selection". The "Display" panel shows a list of diseases with checkboxes for "Display" and "Focus on". The "Selection" panel contains a text snippet: "Therefore, adiponectin replacement in humans may represent a promising approach to prevent and/or treat obesity, insulin resistance and type 2 diabetes; however, clinical studies with adiponectin administration need to be conducted to confirm this hypothesis." An orange arrow points from the text box on the right to the "insulin resistance" node in the network graph.

Type	Display	Focus on
099 Diabetes Mellitus Type I	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
125 Angina Pectoris	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
126 Arteriosclerosis	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
137 Arterial Occlusive Diseases	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
149 Diabetes Mellitus Type II	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

We find that insulin resistance is associated to atherosclerosis and ADIPOQ

Now which compounds are used to treat insulin resistance?
We can use a semantic search to look for treatments that have a positive effect on the disease "insulin resistance"...

Activate 'knowledge browser'

The screenshot displays the Luxid Information Intelligence Knowledge Browser interface. The main window shows a complex semantic network graph with nodes and edges. The node 'thiazolidinediones' is highlighted with a red oval. The interface includes a top navigation bar with 'Center of Interest', 'New Search', 'Discoveries', 'Knowledge Browser', 'help', and 'about'. A right-hand sidebar contains 'Tools' (Remove Node, Remove Sub Graph, Extract sub graph with depth 1, Expand graph on this node), 'Find' (testosterone, Testosterone replacement, thiazolidinediones, Thiazolidinedione agonists, thiazolidinedione treatment), and 'Proof' (...xisome proliferator activated receptor gamma agonists, such as thiazolidinediones which expand the subcutaneous fat depot, may be especially useful in improving insulin resistance in individuals with small fat...).

Copyright (C) Temis 2006 | contact | help | about

Now which compounds are used to treat insulin resistance?
 We can use a semantic search to look for treatments that have a positive effect on the disease "insulin resistance"...

The screenshot shows the Luxid Information Intelligence interface. At the top, there is a navigation bar with "Center of Interest", "New Search", "Discoveries", "Knowledge Browser", "help", and "about". The main area displays a network graph with nodes and edges. A node labeled "thiazolidinediones" is highlighted with a red oval and an orange arrow. To the right, there are toolbars for "Tools" and "Find". The "Tools" section includes options like "Remove Node", "Remove Sub Graph", "Extract sub graph with depth 1", and "Expand graph on this node". The "Find" section lists search results including "testosterone", "Testosterone replacement", "thiazolidinediones", "Thiazolidinedione agonis", and "thiazolidinedione treatme".

Below the graph, a Microsoft Internet Explorer window is open, displaying a document titled "The role of saturation of fat depots in the pathogenesis of insulin resistance." The text in the document is highlighted in orange and reads: "individuals with a large fat depot may become insulin sensitive even when still clinically obese with some amount of weight loss; (2) peroxisome proliferator activated receptor gamma agonists, such as thiazolidinediones which expand the subcutaneous fat depot, may be especially useful in improving insulin resistance; in individuals with small fat".

An orange box with the text "Evidence of the statement" is positioned to the right of the document, with an orange arrow pointing to the highlighted text in the document.

At the bottom of the Internet Explorer window, there is a navigation pane on the left with a tree view showing folders like "Non Drug Related T", "Guessed Treatment", "Drug Related Treat", and "Cell Process (1)". The main content area of the browser shows the text from the document, with a scroll bar on the right. The status bar at the bottom of the browser indicates "Internet" and "copyright (c) Temis 2006 | contact | help | about".

We need to find all thiazolidinediones to spot therapeutic opportunities

Activate 'Sub-Structure Search'

The screenshot shows the Luxid information intelligence software interface. At the top, there is a navigation bar with 'Center of Interest', 'New Search', 'Discoveries', 'Knowledge Browser', 'help', and 'about'. The 'New Search' tab is highlighted with an orange circle. Below the navigation bar, there are tabs for 'Document search' and 'Sub structure search'. The 'Sub structure search' tab is active, and a chemical structure of a thiazolidinedione is displayed in the editor. The results pane on the right shows two search results, each with a 3D ball-and-stick model and a text description. An orange callout box on the left points to the first result, and another orange callout box on the right points to the second result.

2-methoxy-5-(2,4-dioxo-5-thiazolidinyl)-n-[[4-(trifluoromethyl)phenyl] methyl] benzamide (C19H15F3N2O4S) - 1 docs.

Absorption, metabolism, and excretion of [14C]MK-0767 (2-methoxy-5-(2,4-dioxo-5-thiazolidinyl)-N-[[4-(trifluoromethyl)phenyl] methyl] benzamide) in humans.

... and excretion of [14C]MK-0767 (2-methoxy-5-(2,4-dioxo-5-thiazolidinyl)-N-[[4-(trifluoromethyl)phenyl] methyl] benzamide) in humans...MK-0767 (KRP-297; 2-methoxy-5-(2,4-dioxo-5-thiazolidinyl)-N-[[4-(trifluoromethyl)phenyl] methyl] benzamide) is a ... type 2 diabetes.

5-(4-hydroxybenzyl)-3-triphenylmethylthiazolidine-2,4-dione (C29H23NO3S) - 1 docs.

Synthesis and biological activity of [[(heterocycloamino)alkoxy] benzyl]-2,4-thiazolidinediones as PPARgamma agonists.

... (2) with 5-(4-hydroxybenzyl)-3-triphenylmethylthiazolidine-2,4-dione and assayed for ... macrophages.

We identify 2 innovative compounds related to thiazolidinediones

copyright (C) Temis 2006 | contact | help | about

It is now time to share all my discoveries with colleagues and managers. The center of interest has kept the most relevant proofs of my analysis.

Activate 'Centers of Interest'

The screenshot displays the Luxid information intelligence interface. At the top, the 'Center of Interest' tab is highlighted with an orange circle. Below the navigation bar, the main content area is divided into several panels:

- Search Results:** Shows 978 documents about atherosclerosis. Three document snippets are visible, each with a list of top terms.
 - Document 1: [Osteopontin and atherosclerosis]. Top terms: atherosclerosis, Osteopontin, atherosclerosis, atherosclerosis, Osteopontin, SPP1, Osteopontin.
 - Document 2: [Atherosclerosis: butter on the arteries?]. Top terms: Atherosclerosis, atherosclerosis, Atherosclerosis, butter, butter on the artery.
 - Document 3: [Postprandial hyperglycemia as a risk factor of atherosclerosis]. Top terms: hyperglycemia, atherosclerosis, Postprandial hyperglycemia, atherosclerosis, risk factor of atherosclerosis, Postprandial hyperglycemia, risk factor.
- Protein Distribution Charts:** A bar chart showing the number of occurrences for various proteins. The x-axis ranges from 0 to 300.

Protein	Number of occurrences
CRP	267
ADIPOQ	242
LDL	194
antibody	120
PRNP	116
TNF	114
RETN	89
APOE	88
PPARG	82
PON1	76
IL6	73
cytokines	71
- Analysis Table:** A table showing the distribution of proteins across different years.

Year	ADIPOQ	CD36	CRP	cytokine
2005-03	1	1		1
2005-04	2		2	
2005-05	2		3	1
2005-06	2		4	
2005-07	1		5	
2005-08	2	1	6	
2005-09	1		2	
2005-10	2	1	4	1
2005-11	3	1	8	3
2005-12			6	
2006-01			6	2
- Proximity Graph:** A network graph showing relationships between terms, with nodes and edges representing connections.



1

Drug Discovery Process

2

Recognition of Chemical Entities

3

Case study - Scenario Highlight

4

TEMIS Presentation

- TEMIS = **TE**xt **MI**ning **S**olutions
 - Software company founded in Sept 2000
- Strong background & experience in Text Mining
 - Seasoned engineers and managers from IBM
 - Acquisition of Xerox Linguistics in 2003 (XeLDA® = 20 years of R&D, 26 worldwide patents)
- Largest European solution provider
- Blue chip customers
 - Life Sciences
 - Publishing
 - Industry



Life Sciences >>



Publishing >>



Industry >>



- A new TEMIS Software

... Your information make you lucid ...

... Enlight your ideas (lux) ...



Luxid®

*... The new generation of TEMIS
Insight Discoverer™ (ID)...*

Beyond Search >>

Recognition of Chemical Entities in Scientific Literature



THANK YOU!

