

# Integrating text-mining efforts for pathway curation requires output standards.

Andra Waagmeester

Maastricht University

March 22, 2011



**Maastricht University**

Introduction

Pathway Curation

Pathway Loom demo

Text mining

Text mining standard



## How to expose text-mining results?

- Nifty website?



## How to expose text-mining results?

- Nifty website?
- Sentences with “gold nuggets”?



## How to expose text-mining results?

- Nifty website?
- Sentences with “gold nuggets”?
- Precision/Recall



## How to expose text-mining results?

- Nifty website?
- Sentences with “gold nuggets”?
- Precision/Recall
- SOAP
- REST
- csv/tsv
- Unix pipes (ie: `cat textfile — makelist.sh — sort — uniq -c`)



# Finding novel pathway parts on a pathway (1)

OMICS A Journal of Integrative Biology  
Volume 13, Number 5, 2009  
© Mary Ann Liebert, Inc.  
DOI: 10.1089/omi.2009.0029

**Original Article**

## Pathway Enrichment Based on Text Mining and Its Validation on Carotenoid and Vitamin A Metabolism

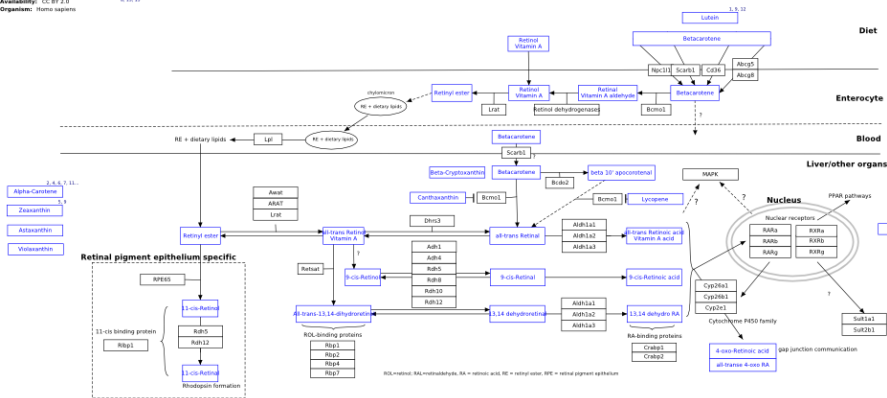
Andra Waagmeester,<sup>1,2,3</sup> Piotr Pezik,<sup>1</sup> Susan Coort,<sup>3</sup> Franck Tourniaire,<sup>4</sup>  
Chris Evelo,<sup>3</sup> and Dietrich Rebholz-Schuhmann<sup>1</sup>



Maastricht University

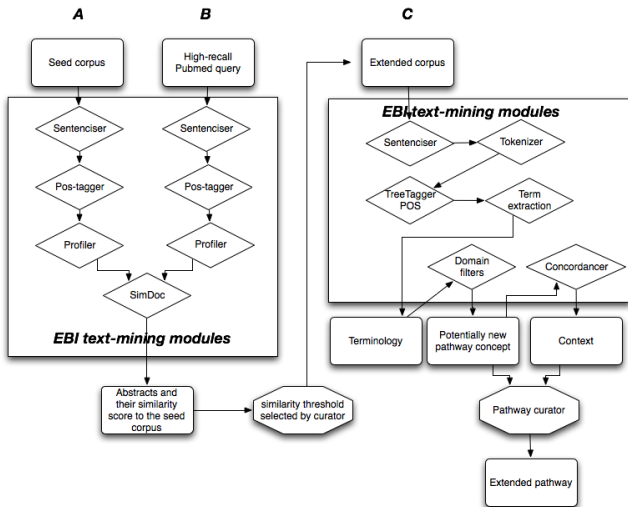
# Finding novel pathway parts on a pathway (2)

Title: Vitamin A and carotenoid metabolism  
 Availability: CC BY 2.0  
 Organism: Homo sapiens





# Identifying novel pathway parts on a carotenoid pathways (3)



## Identifying novel pathway parts on a carotenoid pathways (4)

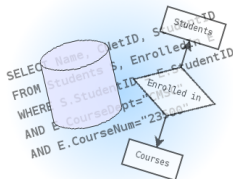
We were able to retrieve 13 novel pathway parts from Pubmed abstracts:

- Vitamin D3
- ADH3
- alpha-carotene
- astaxantin
- beta-cryptoxanthin
- Canthaxanthin
- lutein
- lycopene
- zeaxanthin
- PEPCK
- cryptoxanthin
- triglyceride
- lipase



# Pathway curation

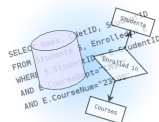
Knowledge is hidden in:



# Computers



× Can't listen



✓ Can Query



× Can't read



Maastricht University

# Pathway Loom demo (1)

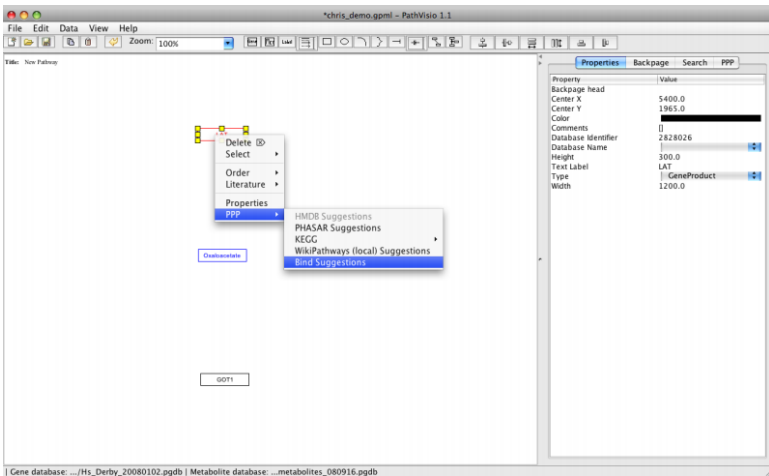
The screenshot shows the PathVisio 1.1 interface. The main window displays a pathway diagram with a central node labeled 'LAT' connected to two other nodes. Below the diagram are two buttons: 'Disable node' and 'GOT1'. The right-hand side features a 'Properties' panel with a table of node attributes.

Property	Value
Backpage head	
Center X	5400.0
Center Y	1980.0
Color	
Comments	
Database Identifier	2828026
Database Name	
Height	300.0
Text Label	LAT
Type	GeneProduct
Width	1200.0

At the bottom of the window, the status bar indicates: Gene database: .../Hs\_Derby\_20080102.pgdb | Metabolite database: ...metabolites\_080916.pgdb



# Pathway Loom demo (2)



The screenshot displays the PathVisio 1.1 interface. The main window shows a pathway diagram with a context menu open over a node. The menu options are: Delete, Select, Order, Literature, Properties, and PPP. The 'PPP' option is selected, opening a sub-menu with the following suggestions: HMDB Suggestions, PHASAR Suggestions, KEGG, WikiPathways (local) Suggestions, and Bind Suggestions. The 'Bind Suggestions' option is highlighted. The properties panel on the right shows the following details for the 'PPP' node:

Property	Value
Backpage head	
Center X	5400.0
Center Y	1965.0
Color	
Comments	
Database Identifier	2828026
Database Name	
Height	300.0
Text Label	LAT
Type	GeneProduct
Width	1200.0

At the bottom of the window, the status bar indicates: Gene database: .../Hs\_Derby\_20080102.pgdb | Metabolite database: ...metabolites\_080916.pgdb

# Pathway Loom demo (3)

The screenshot shows the PathVisio 1.1 interface. The main window displays a pathway diagram with nodes: "Grap", "LAT", "PLCgamma", "G0T1", and "Oxidoreductase". The "LAT" node is connected to "Grap" and "PLCgamma". The "G0T1" node is isolated. The "Oxidoreductase" node is also isolated. The right-hand panel shows the "Properties" tab for the selected "LAT" node, titled "Putative pathway part: Bind". This panel contains a smaller diagram showing "LAT" as a central node connected to "Grap", "PLCgamma", "G0T1", and "Oxidoreductase". The "Grap" and "PLCgamma" nodes in this smaller diagram are highlighted in red. The status bar at the bottom indicates the gene database is "/Hs\_Derby\_20080102.pgdb" and the metabolite database is "...metabolites\_080916.pgdb".



## PHASAR

### Phrase-based High Accuracy Search and Retrieval



- [home](#)
- [the PHASAR BioMed project](#)
- [the PHASAR TMAP project](#)
- [publications](#)
- [the AGES project](#)
- [the LCS project](#)
- [contact](#)

PHASAR is a new kind of search engine, which does not consider queries and documents as bags-of-words but is based on a deep linguistic analysis and induction of both documents and queries. It is intended for various forms of professional search, which is fully supported by current word-based search engines.

The PHASAR engine provides its users with a wholly new way of searching, using linguistically motivated search terms, giving the user tight control over precision and recall (avoiding long lists of spurious hits) and providing unprecedented support of the search process by information from the index and the thesaur.

The PHASAR search engine is still in the prototype stage. In first application, as an experimental literature search engine for bioinformatics giving access to Medline abstracts, it is a first release of the *Index*, linked by the *stripboxes* and *incompleteness* of the software and *logways*.

PHASAR requires extremely accurate parsers and *fastword*, and is only applicable to languages and domains for which such parsers and thesaur are available. The (partial) development of suitable parsers and thesaur requires a large investment, but it will enable create professional applications of PHASAR.



**Cornelis H.A. Keizer**  
Department of Computing Science  
University of Nijmegen  
6525ED Nijmegen, The Netherlands  
(mailto:ah.keizer@cs.ru.nl)

(g) PHASAR

(h) WHATIZIT (<http://www.ebi.ac.uk/webservices/whatizit/info.jsf>)





 Andra Waagmeester 488  9 | [log out](#) | [faq](#)


# BioStar

[Questions](#)
[Tags](#)
[Users](#)
[Badges](#)
[Unanswered](#)
[Ask Questions](#)

## Which are stable text mining solution for biologists.


  
4
   


  
2

I am working on integrating pathway relations into a pathway diagram. I have two text mining solutions that I use to find interactions. I am looking for new suggestions. The idea is that I provide a pathway entity (ie. gene/protein, metabolite) and all relations based on the literature are returned.

[text-mining](#) [Interaction](#) [pathway](#)
[edit](#) | [rollback](#) | [close](#) | [delete](#) | [flag](#)

 edited Mar 1 at 7:28  
 Lars Juhl Jensen   
 4,441  4  15

 asked Feb 26 at 21:08  
 Andra Waagmeester  
 488  9

Could you please tell us which are the two text mining solutions that you already have? –

[Lars Juhl Jensen](#)  Feb 27 at 7:57

I am using Whatizit: [ebi.ac.uk/webservices/whatizit/info.jsf](http://ebi.ac.uk/webservices/whatizit/info.jsf) and Phasar: [phasar.cs.ru.nl](http://phasar.cs.ru.nl). The first works decently, the second has issues with performance and accessibility. – [Andra Waagmeester](#) Feb 27 at 9:41

[add comment](#)

### 4 Answers

[oldest](#)
[newest](#)
[votes](#)

  
5
   


You might want to take a look at [STITCH](#). It is a database that allows you to query with a gene/protein or a small molecule (e.g. a metabolite) and retrieve the interaction partners. A very large part of the evidence in STITCH comes from text mining, and if you really want to, you can turn off all the other evidence types to *exclusively* get the results obtained through text mining.

tagged

[pathway](#) × 11

[text-mining](#) × 10

[interaction](#) × 8

asked

15 days ago

viewed

205 times

latest activity

11 days ago

Spread the word!

Tell a colleague about BioStar

<http://www.biostars.org>

 Subscribe: [RSS feed](#)

 Admin group: [Biostar Cent](#)

 BioStar content is licensed under  
 Creative Commons Attribution-NonCommercial 3.0 Unported License

**STITCH 2** Search New Settings Saved

**STITCH: Chemical-Protein Interactions**

STITCH is a web-based application for exploring and visualizing interactions between small molecules and proteins. It is based on the STITCH database, which is a comprehensive resource for chemical-protein interactions. STITCH is currently available in two versions: STITCH 2.0 (the current version) and STITCH 1.0 (the previous version).

**How to use STITCH**

STITCH is a web-based application for exploring and visualizing interactions between small molecules and proteins. It is based on the STITCH database, which is a comprehensive resource for chemical-protein interactions. STITCH is currently available in two versions: STITCH 2.0 (the current version) and STITCH 1.0 (the previous version).

(i) STITCH

**iHop**

Network Hopping

Network Hopping is a web-based application for exploring and visualizing interactions between small molecules and proteins. It is based on the STITCH database, which is a comprehensive resource for chemical-protein interactions. STITCH is currently available in two versions: STITCH 2.0 (the current version) and STITCH 1.0 (the previous version).

(j) iHop

**ChemTagger**

ChemTagger is a web-based application for exploring and visualizing interactions between small molecules and proteins. It is based on the STITCH database, which is a comprehensive resource for chemical-protein interactions. STITCH is currently available in two versions: STITCH 2.0 (the current version) and STITCH 1.0 (the previous version).

(k) CHEMICAL TAGGER

**University of Cambridge** Department of Chemistry University of Cambridge

**ChemicalTagger**

ChemicalTagger is an open-source tool for tagging and parsing experimental sections in the abstract literature. It is written in a Java-based library and is currently only under development.

**Availability:**

Source code: <http://bitbucket.org/chemtagger>

**Sample data sets:**

<http://bitbucket.org/chemtagger/sample-data>

**Chilibot** Mining PubMed for relationships

Chilibot searches PubMed literature database (abstracts) about specific relationships between proteins, genes, or keywords. The results are returned as a graph (see examples). It supports several different search methods.

**Search for relationship between two genes, proteins or keywords**

Keywords:  &

**Search for relationships between many genes, proteins, or keywords**

Keywords:

**Search for relationships between two sets of genes, proteins, or keywords**

Keywords:  &

(l) CHILIBOT

**Agilent Technologies** Agilent Literature Search Software

Agilent Literature Search Software is an enhanced tool for searching Agilent literature. It is designed to help you find the information you need to support your business. It is available in two versions: Agilent Literature Search Software (the current version) and Agilent Literature Search Software (the previous version).

(m) AGILENT

**Reflect**

Reflect is a web-based application for exploring and visualizing interactions between small molecules and proteins. It is based on the STITCH database, which is a comprehensive resource for chemical-protein interactions. STITCH is currently available in two versions: STITCH 2.0 (the current version) and STITCH 1.0 (the previous version).

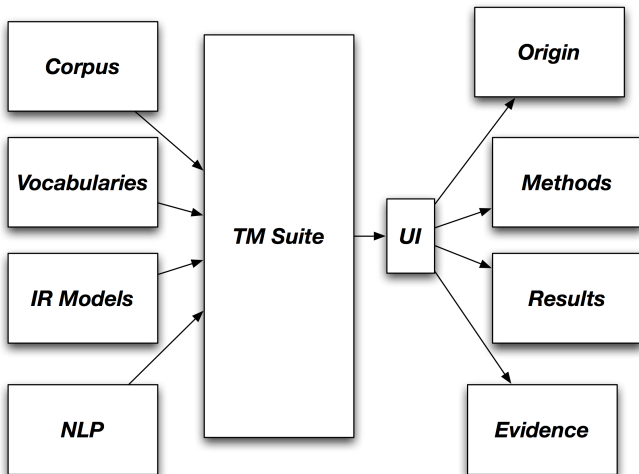
(n) REFLECT

**Knowledge Enhancer**

Knowledge Enhancer is a web-based application for exploring and visualizing interactions between small molecules and proteins. It is based on the STITCH database, which is a comprehensive resource for chemical-protein interactions. STITCH is currently available in two versions: STITCH 2.0 (the current version) and STITCH 1.0 (the previous version).

(o) KNOWLEDGE ENHANCER

# Text-mining standard for pathway curation



## Origin

- Description
- Corpora
- Identifiers

## Methods used

- Description
- Method name
- Vital statistics

## Results

- Source
- Target
- Source Identifiers
- Target Identifiers

## Evidence

- Relevant text parts



# Acknowledgements

## Maastricht University

- Chris Evelo
- Thomas Kelder
- Martina Kutmon
- Jahn Saito
- Martijn van Iersel

## USC - SF

- Alexander Pico

## BIOSTAR <http://biostar.stackexchange.com>

- Lars Juhl Jensen
- Casey Bergman
- Egon Willighagen
- Dominique Noel

