RetroMine,
or how to provide in-depth retrospective studies from Medline in a glance:
the hepcidin use-case.

Bertrand Ameline de Cadeville
Olivier Loréal
Fouzia Moussouni-Marzolf

Introduction

Life Science is one of the most VOLUMINOUS science

Rapid Expansion of the biomedical literature
Available papers exploding

Increased demand and development of effective text mining tools to find quickly relevant information.

Introduction

Biologists are reluctant to use text mining tools.

why ?

(1) **PubMed Search Engine**
(survey of [Eva Lee et al. 2011])

→ Few users review results beyond the first page.

→ Users seek simple interfaces : difficulties using advanced features of PubMed

→ Users retrieve in-depth information focusing on one category : Gene Cell Disease ...

Many pertinent articles not being pulled in a Medline search

Introduction

Non computationnal biologists are reluctant to use text mining tools.

(2) For fine-grained tools

In abstracts

Identification of **Entities**
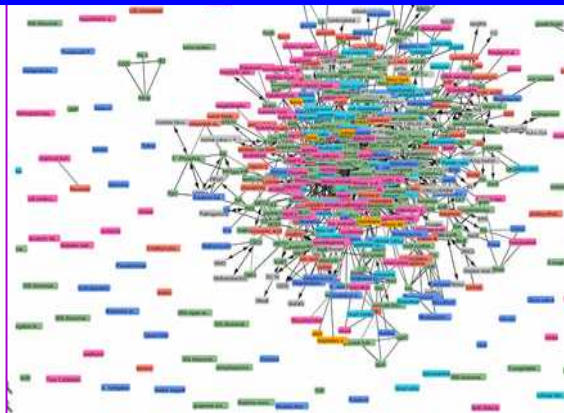
Identification of **Events** connecting Entities
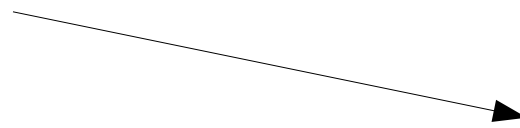
In sentences

Curcumin reduces *hepcidin1*

Introduction

These tools extract a deluge of information
Very dense data

Query « Hepcidin » using Ali-Baba [1]

For non expert
Information is dense
and unreadable

**Pertinent information is
hidden.**
**Non computationnal
biologists are discouraged
from using this kind of tool.**

For an expert
A considerable amount of
background knowledge

[1] Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J. & Leser, U. AliBaba: PubMed as a graph. *Bioinformatics*. **22**, 2444-2445 (2006).

Other Event Extraction Tools

- TEES for Turku Event Extraction System of [Jari Björne *et. al*]

- EVEX web site and database

- BioNLP initiatives . . .

Mostly devoted to the computational biology and bioinformatics community.

Despite these efforts ...

few studies addressed the problem of processing these **big datasets** in order for **non-computational experts** to rapidly convert them into **meaningful patterns** over time.

?
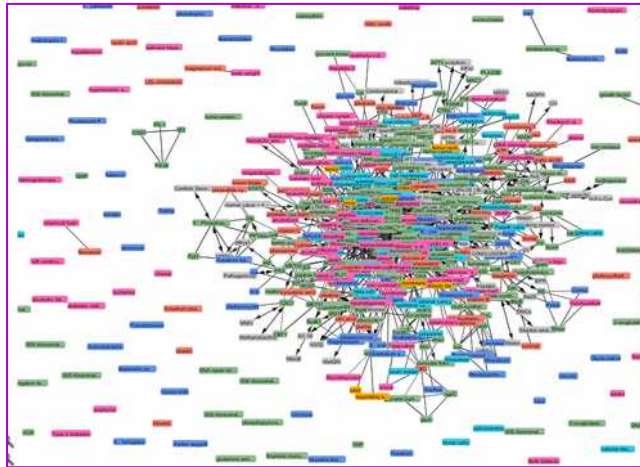
PROPOSAL

Improve Data Selection

Identify **background** knowledge and filter it to reduce the density of information

Make time explicit : perception of events chronology inhance comprehension

time

Select **relevant** non background events over time and provide them to bio-investigators.

Methods

SOME BASICS

What is a Background Event ?

This concept is relative to a certain time point *t*

A background event is spotted and extracted repeatedly from literature at different time points

Example : IL6 transcription factor of Hepcidin published **first time** in April 2003

May 2005

IL-6	expres
interleukin6	expres
IL-6	co-occur
Hepcidin	co-occur
Hepcidin	co-occur
Ferroportin1	increa
Ferroportin1	is

July 2005

fusionprotein	Co-occurrence	Hepcidin	16141345	2005-09-01
GFP	expressing	hepcidin	1614134!	
Hemojuvelin	encoding	HAMP	1596769;	
Hepcidin	induced	IL-6	1588631!	
Hepcidin	induced	interleukin6	1588631!	
Hepcidin	Co-occurrence	IL-6	1619862;	
Hepcidin	upregulation	IL-1beta	1619862;	
hepcidin	induced	IL-6	1588631!	
hepcidin	Co-occurrence	IL-6	1619862;	
IL-1beta	upregulated	Hepcidin	16198622	2005-09-01
IL-1beta	production	IL-6	16198622	2005-09-01
RGMC	encoding	HAMP	15967692	2005-09-01

Dec 2005

| hepcidin	exposed	Epo	16332970	2005-12-01
Hepcidin	decreased	IL-6	16351643	2005-12-01
Hepcidin	influence	IL-6	16351643	2005-12-01
hepcidin	decreased	interleukin-6	16351643	2005-12-01
Hepcidin	interacting	SLC40A1	16351644	2005-12-01

Background events are continuously returned but remain trivial to Hepcidin experts
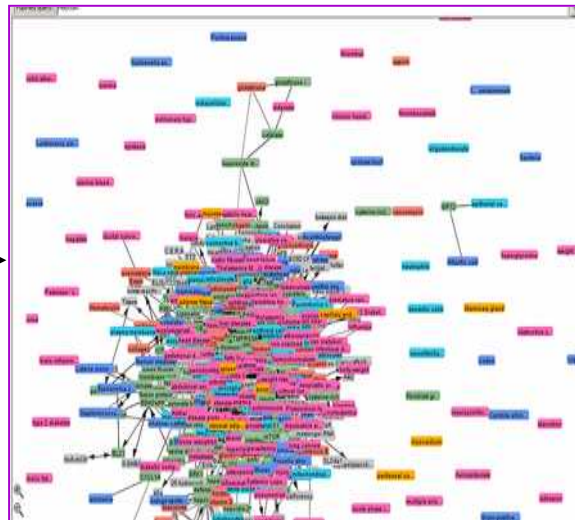
Methods

Definition 1

A recognized event e is defined as being a background if it has been spotted repeatedly in different abstracts and at different time points.
In other words, when an event e is published for the first time at t, it becomes background at time $t+\Delta$.

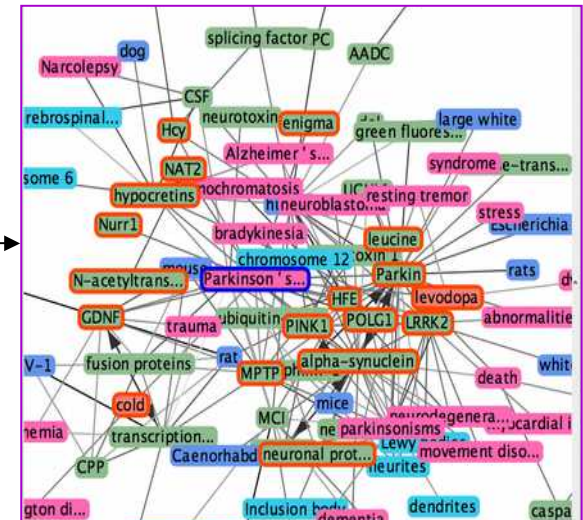G_0 = Graph of events

obtained for time t_0
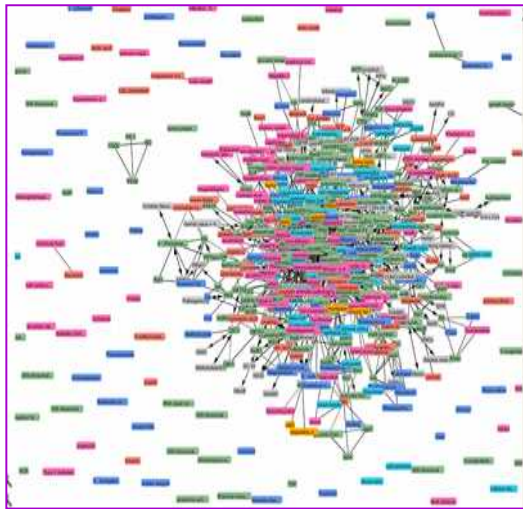
G_1 = Graph of events
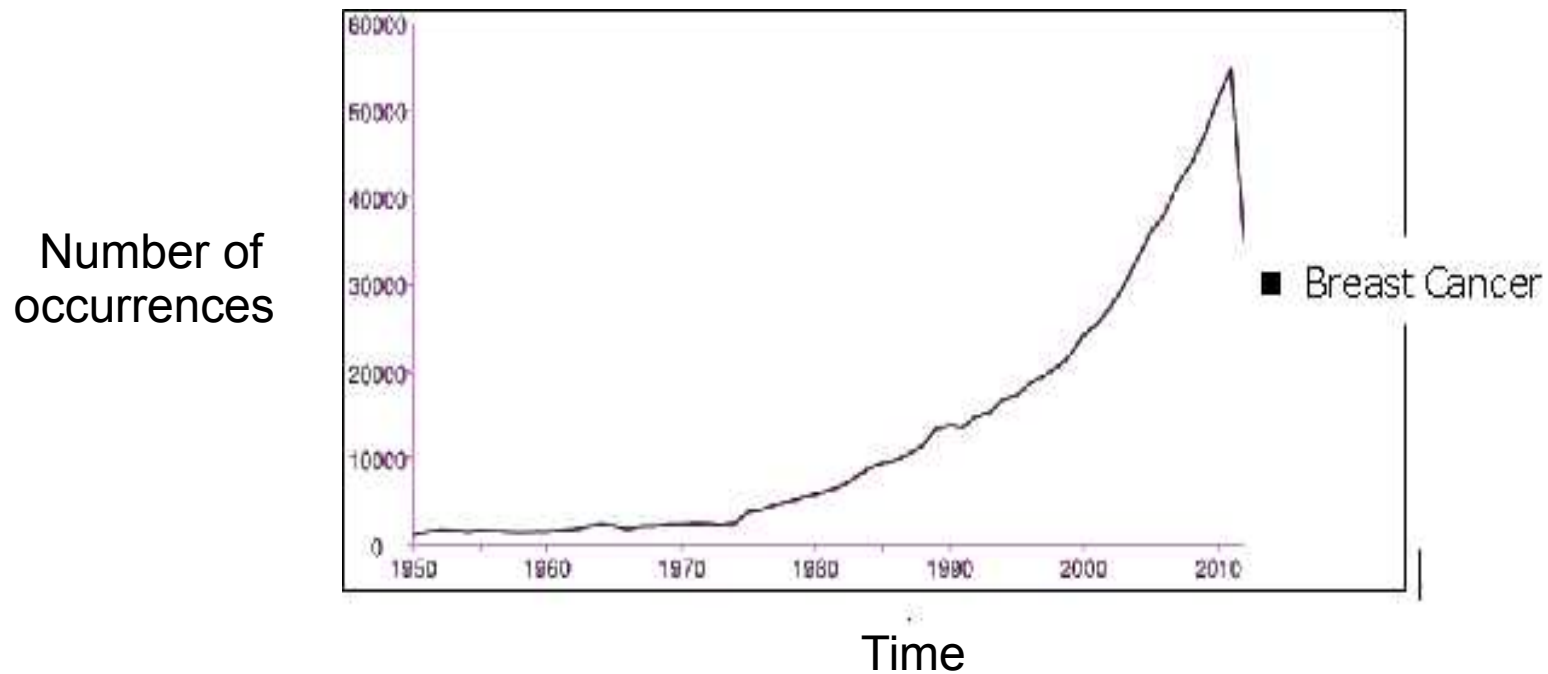
$t_1 = t_0 + p$

G_2 = Graph of events

$t_2 = t_1 + p$

$\cdots$

$(e \in G_1)$ **and** $(e \in G_0)$

$(e \in G_2)$ **and** $(e \in G_1$ **or** $e \in G_0)$

Methods

Time Relevance of a Biological Entity ?

The concept of **time relevance** has already been reported in [Palidwor GA, et al.] for graphing MEDLINE **keywords** over time using MLTrends.

Number of occurrences

Time

When applied to recognized entities and events connecting them, this approach leads to much more informative functions

Methods

So what is a time relevant (t-relevant) biological entity ?

Set of abstract for query « Hepcidin 2005/05 [dp] »

graph of events extracted (ali-baba)

e is Highly Targeted by *other bio-entities* **at time t**
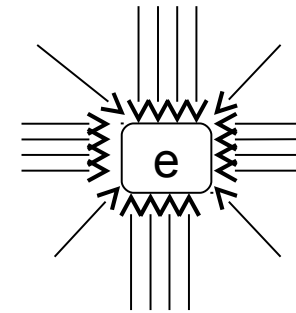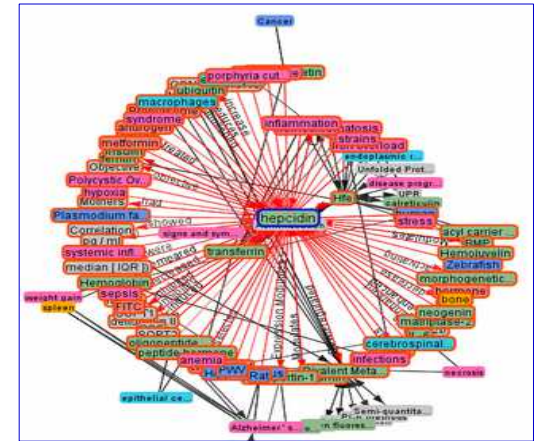
Relevance should be revealed at real time, as entities may lose their relevance in future time points.

Methods

Definition 2

A recognized biological entity e is defined as being relevant at time t (or *t-relevant*) if it achieves a maximum of relationships at time t with other recognized biological entities.

Time relevance may be provided for different sorts of biological entities

EVENT   Source Entity   **Relationship**   Target Entity
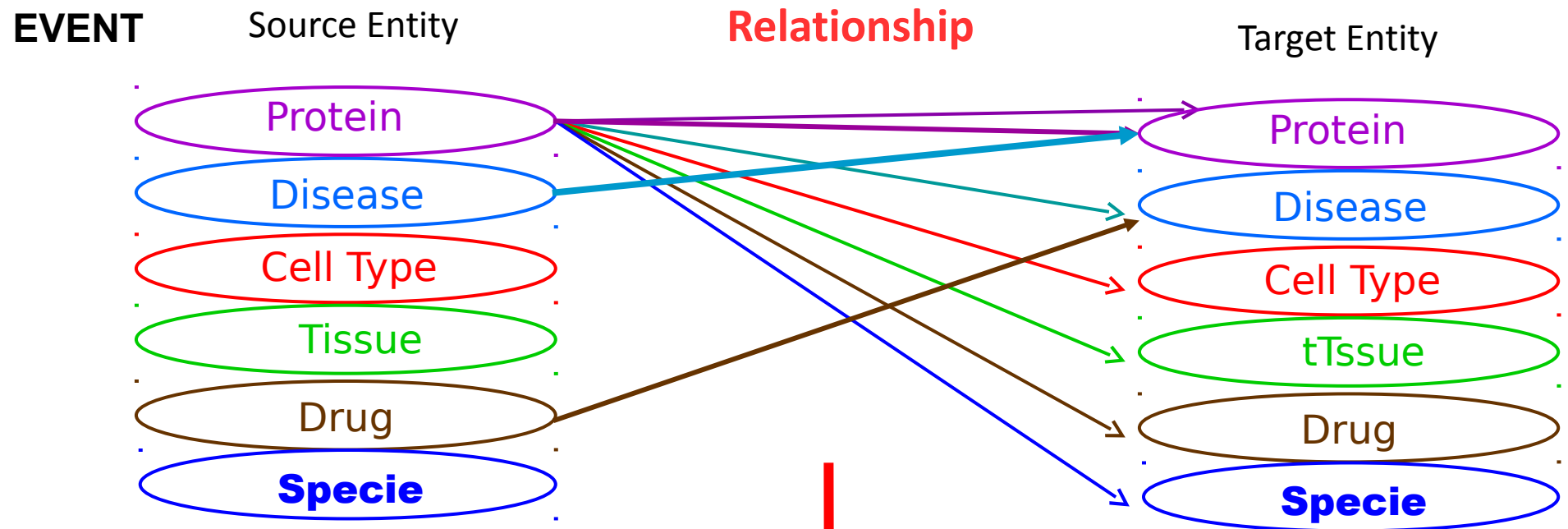
Different valuable information

Our Use Case Study

10 years Hepcidin publications

Large Scale Extraction of biological entities and events, using Ali-baba web service

do it for each period p = 1 month, from Dec 2000 to Dec 2011

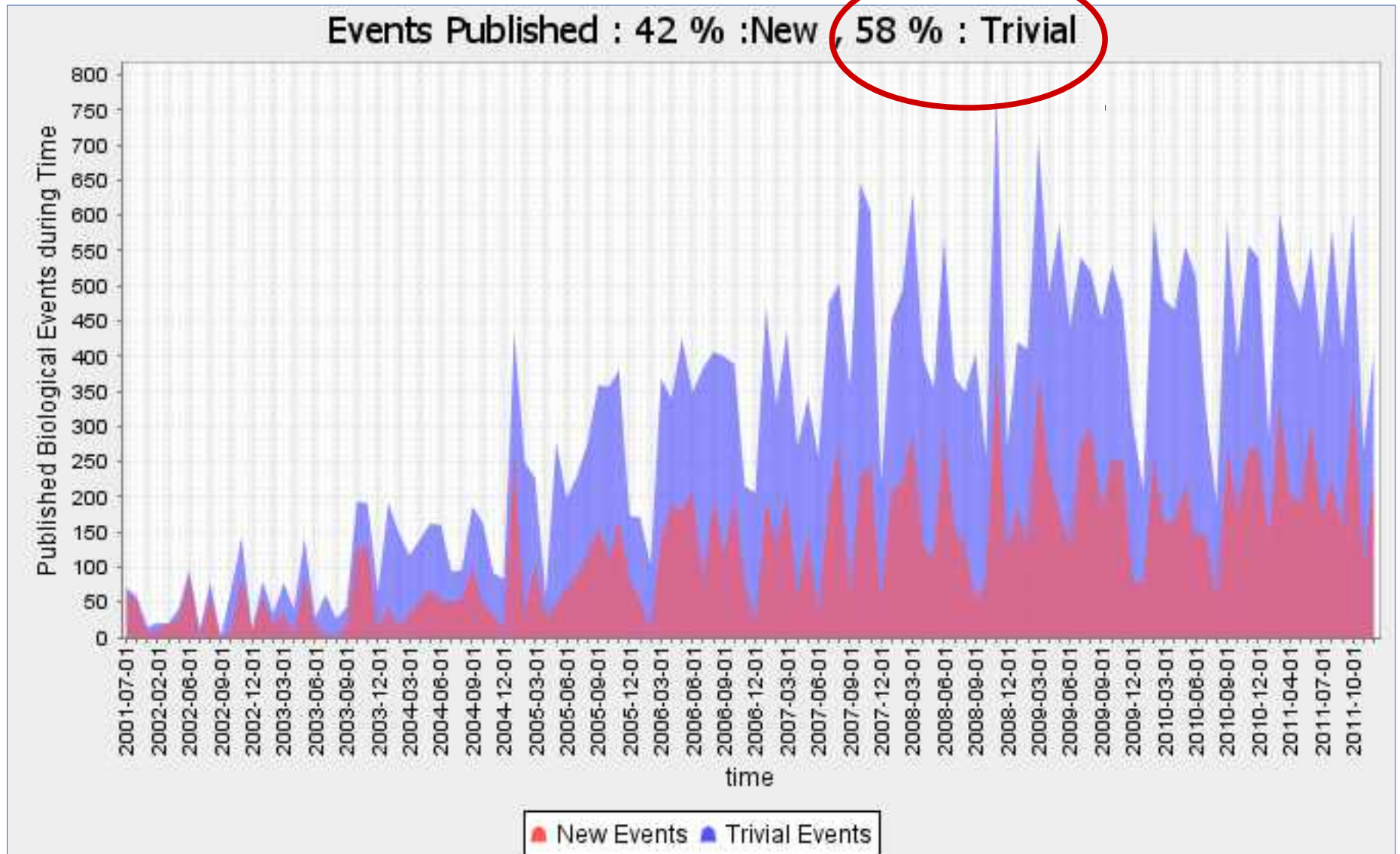Amount of background knowledge published during the Hepcidin decade

Collected Events

MySQL Data Warehouse of Hepcidin Events

Highlight retrospectively

Time relevance : Provide periodically highly targeted proteins diseases, drugs,cell types and tissues.

Results

Cumulative Quantification of Background Information Published
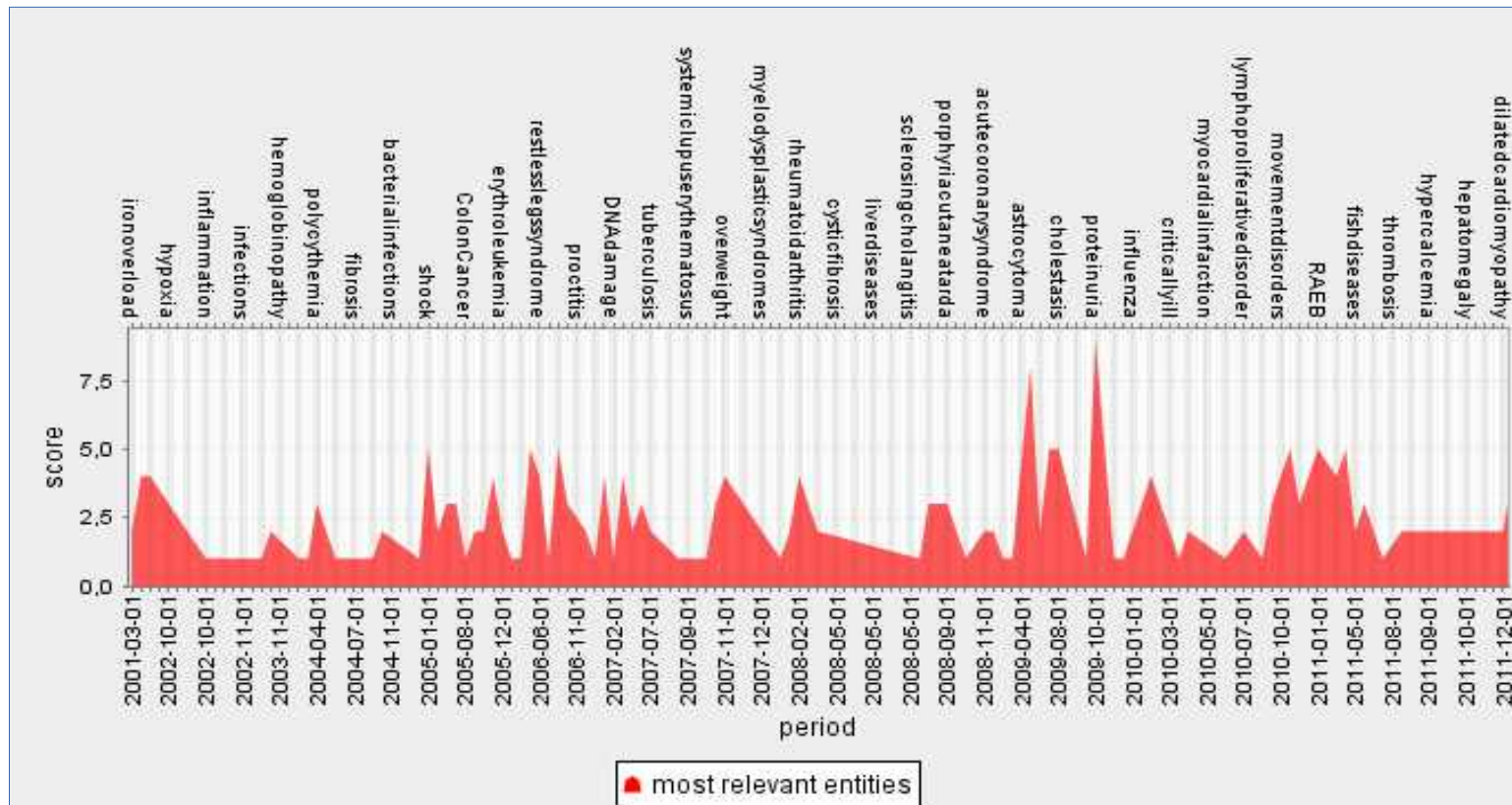
Results

Highly Targeted Proteins Over Time

After filtering Background :
1. Drastic Fall of data (events)
2. From the expert side : non expected proteins emerge as being relevant :SMAD, BMPs … ,
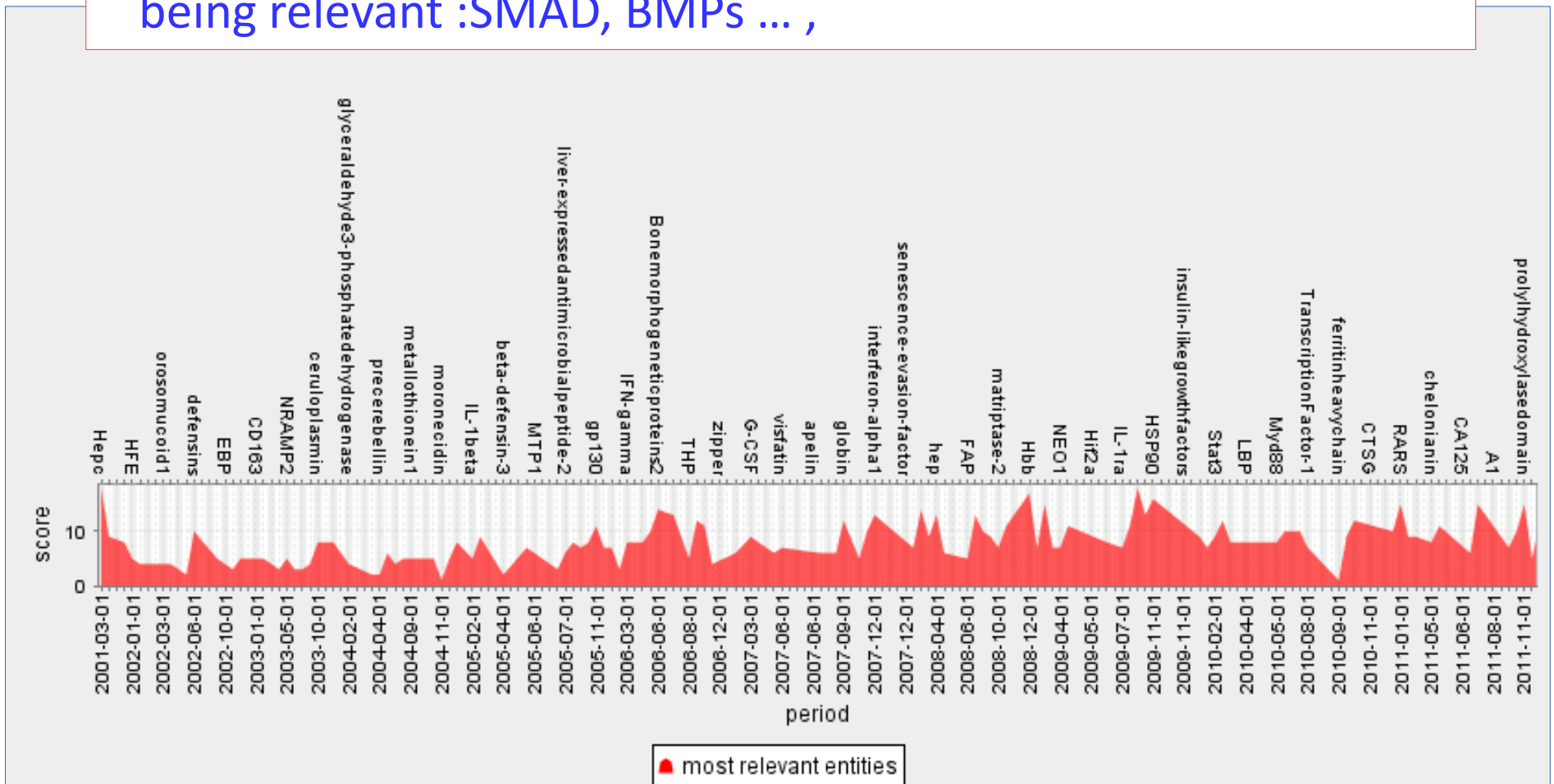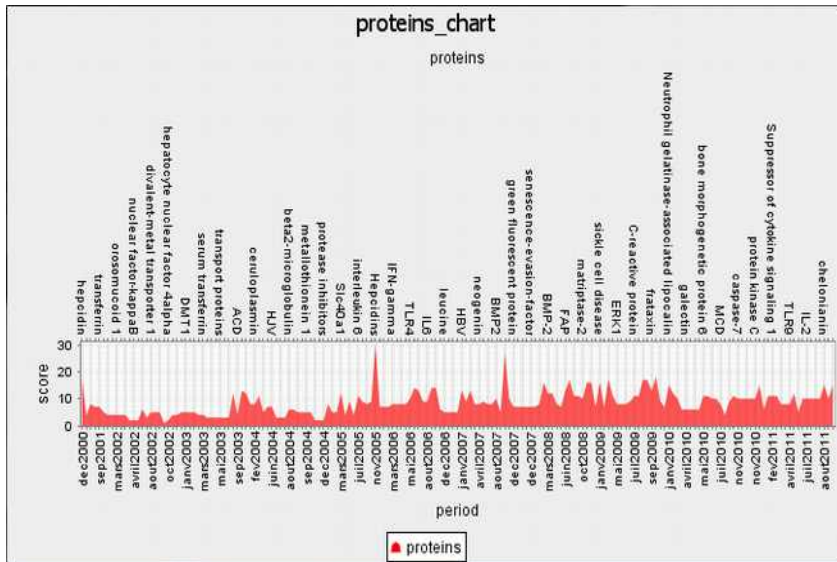
Results

Time Relevant Diseases

After Clearing Background
New diseases linked to Hepcidin and iron
emerge as relevant like fish diseases and
neurological diseases

Results

We may provide for biologists more annotations of the time relevant entities

Conclusion - Discussion

✳ RetroMine approach is straightforward but is extremely helpful for providing **in-depth retrospective studies** to researchers on subjects of their interest.

✳ Torrential data are extracted using advanced text mining tools and **mined in a second round** to draw unexpected patterns of biological entities and events behaviors over time.

✳ RetroMine enhance comprehension of the extracted events :
- by introducing chronology using TIME
- by giving priority to non background biological entities, highly targeted over time

✳ RetroMine revealed the considerable amount of **background information** published periodically in the biomedical literature

✳ This work is still ongoing. Current developments :
- Toward a generalization to any query on biological entities – substitute ali-baba
- Mining Microbiota Litterature

Thanks to ...

Bertrand Ameline de Cadeville
Master **MSB** Université de Rennes1

Olivier Loréal
INSERM U991 **Iron and Liver** group
Supports from ANR IronReg

Ulf Leser
Astrid Rheinländer

Knowledge Management in Bioinformatics group
HU-Berlin

ESF
European Science Foundation