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
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?

- 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 computational biologists are reluctant to use text mining tools.

For fine-grained tools:

- 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.

For an expert
A considerable amount of background knowledge

Non computational biologists are discouraged from using this kind of tool.

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.
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 enhances comprehension.

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

<table>
<thead>
<tr>
<th>May 2005</th>
<th>July 2005</th>
<th>Dec 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-6 expression</td>
<td>IL-6 expression</td>
<td>IL-6 expression</td>
</tr>
<tr>
<td>interleukin6 expression</td>
<td>interleukin6 induced</td>
<td>interleukin6 induced</td>
</tr>
<tr>
<td>Hepcidin</td>
<td>Hepcidin induced</td>
<td>Hepcidin induced</td>
</tr>
<tr>
<td>Ferroportin1 increased</td>
<td>Ferroportin1 increased</td>
<td>Ferroportin1 increased</td>
</tr>
</tbody>
</table>

**Background events are continuously returned but remain trivial to Hepcidin experts**
**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 $t+\Delta$.

---

**Methods**

- $G_0 = \text{Graph of events obtained for time } t_0$
- $G_1 = \text{Graph of events } t_1 = t_0+p$
- $G_2 = \text{Graph of events } t_2 = t_1+p$

$\quad (e \in G_1) \text{ and } (e \in G_0)$

$\quad (e \in G_2) \text{ and } (e \in G_1 \text{ or } e \in G_0)$
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.

When applied to recognized entities and events connecting them, this approach leads to much more informative functions.
So what is a time relevant (t-relevant) biological entity?

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

Relevance should be revealed at real time, as entities may lose their relevance in future time points.

Methods

Graph of events extracted (ali-baba)

*e is Highly Targeted* by other bio-entities at time t
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.
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

MySQL Data Warehouse of Hepcidin Events

Amount of background knowledge published during the Hepcidin decade

Highlight retrospectively

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

Cumulative Quantification of Background Information Published

Events Published: 42% New, 58% Trivial
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.

<table>
<thead>
<tr>
<th>Number</th>
<th>Date</th>
<th>HighTargetedEntity</th>
<th>Number of targets</th>
<th>#Abstracts</th>
<th>Swissprotid</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>mars 2003</td>
<td>reverse transcriptase</td>
<td>4</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>29</td>
<td>mars 2003</td>
<td>serum transferrin</td>
<td>3</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>30</td>
<td>avril 2003</td>
<td>interleukin-6</td>
<td>3</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>31</td>
<td>avril 2003</td>
<td>tumor necrosis factor alpha</td>
<td>3</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>32</td>
<td>mai 2003</td>
<td>hepcidin 1</td>
<td>3</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>33</td>
<td>mai 2003</td>
<td>transport proteins</td>
<td>3</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>34</td>
<td>juin 2003</td>
<td>beta-2-microglobulin</td>
<td>3</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>35</td>
<td>juin 2003</td>
<td>TR1</td>
<td>1</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>36</td>
<td>jul 2003</td>
<td>Namp</td>
<td>12</td>
<td>2</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>37</td>
<td>sep 2003</td>
<td>ACD</td>
<td>4</td>
<td>2</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>38</td>
<td>oct 2003</td>
<td>ferroportin-1</td>
<td>13</td>
<td>2</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>39</td>
<td>nov 2003</td>
<td>Namp2</td>
<td>12</td>
<td>2</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>40</td>
<td>jan 2004</td>
<td>reductase</td>
<td>8</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>41</td>
<td>fev 2004</td>
<td>ceruloplasmin</td>
<td>8</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>42</td>
<td>mars 2004</td>
<td>ferroportin-1</td>
<td>11</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>43</td>
<td>avril 2004</td>
<td>hemosiderin</td>
<td>5</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>44</td>
<td>mai 2004</td>
<td>IL-6</td>
<td>5</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>45</td>
<td>jun 2004</td>
<td>HIV</td>
<td>7</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>46</td>
<td>jul 2004</td>
<td>hepatocyte growth factor-Ik...</td>
<td>3</td>
<td>1</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>47</td>
<td>28 août 2004</td>
<td>Lmo2</td>
<td>5</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>48</td>
<td>29 août 2004</td>
<td>T3G2</td>
<td>5</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>49</td>
<td>30 août 2004</td>
<td>Gm69</td>
<td>5</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>50</td>
<td>31 août 2004</td>
<td>TTD1</td>
<td>5</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
<tr>
<td>51</td>
<td>sep 2004</td>
<td>Sox6</td>
<td>5</td>
<td>6</td>
<td><a href="http://www.uniprot.org/uniprot/">http://www.uniprot.org/uniprot/</a>...</td>
</tr>
</tbody>
</table>

Serum hepcidin in clinical specimens.

Cataldo G, Fleury T, Meana RT.

Hematology Oncology Division, Department of Medicine, Ralph H. Johnson VA Medical Center, University of South Carolina, Charleston, SC 29425, USA. meana@musc.edu

Abstract

The hepatic antimicrobial protein, hepcidin, is implicated in duodenal iron absorption and mobilization. Overexpression of the hepcidin gene is associated with a hypoferritinemia.

Recommended name: Acyl-CoA dehydrogenase

Gene names

- Name: acdA
- Synonyms: acd
- Ordered Locus Names: ESU37170
**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 Literature
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