Full Title
Temporal Trajectories of bio-entities and bio-events extracted from PubMed

Authors and Affiliations
B. Ameline de Cadeville ¹
O. Loréal ¹
F. Moussouni-Marzolf ¹,²,*

¹ INSERM UMR 991, CHU Pontchaillou, 35033 Rennes, France
² Université de Rennes1, Faculté de Médecine, 35043 Rennes, France
Abstract

Motivation: The rapid expansion of biomedical literature has provoked an increased demand of advanced text mining tools to rapidly extract relevant events from the continuously flooding knowledge published periodically. However, biologists are still reluctant to use text mining tools because of the considerable amount of events they extract for their queries. Moreover, the extracted information include significant amount of trivial events corresponding to background knowledge published repeatedly in PubMed abstracts. This dominant knowledge is well established by domain experts and may hide unexpected information extracted.

Results: In this paper, we propose a tool that integrate time to the process of information extraction from PubMed in order to improve the selection of more pertinent knowledge published over time and its comprehension by bio-scientists. Indeed, instead of providing the whole extracted knowledge associated to a PubMed query, usually by mixing trivial and non-trivial information, as done by major text mining tools, RetroMine unfold time and extract non trivial but highly targeted biological entities associated to the user’s query, during time. After filtering trivial information an evaluation of their total amount is provided. This work is illustrated with a use-case study: a decade of publications on Hepcidin, an important gene involved in the regulation of iron metabolism at systemic level.

Availability: relative data and results are available on: http://retromine.univ-rennes1.fr

Introduction

Modern biology is one of the most voluminous sciences today. This is mainly due to the modern digital revolution and increasing incitements to publish and share new knowledge world wide. There are daily in MEDLINE hundreds time more publications than hundred years ago and millions time more than at the beginning of enlightenment. The amount of publications is accumulating so rapidly that it is no longer possible for a researcher to keep up-to-date with recent discoveries from literature.

Alternative solutions are required by bio-investigators for searching, and learning how to search rapidly and efficiently this continuously growing knowledge.

Text mining coupled to the process of information extraction has proved to be a manifest solution for quickly reading this impressive flood of texts by supplementing human readers with automatic tools. Numerous tools exist today for facilitating the application of text mining and information extraction (IE) techniques by researchers to their problem of interest [Entrez-Pubmed, GoPubmed, Ali-baba, Medie, ..., etc]. However, these tools still present difficulties as they are largely developed by information experts. A gap still exists between computational linguists and designers of information extraction tools in one hand, and biomedical investigators in the other hand, making knowledge published in Medline far from being exploited. If we consider PubMed for example, the most popular biomedical search engine for MEDLINE database, a survey conducted by Eva Lee et al. in [] has proved the following statements:
1. Few users review results beyond the first page: most of the clicks happened on the most recent citations, located in the first result page (by default, PubMed returns 20 results per page).

2. Users seek for simple interfaces and have difficulties using advanced features of PubMed because of the knowledge required to use MeSH terms and the several steps they have to go through before building a query.

3. The majority of users retrieve in-depth information focusing on one category of information, as of cell, disease, drug or gene.

Moreover, when the extracted information consists of recognized biological entities and complex events connecting them in abstracts, the situation is even worse. A mass of data is always extracted either for searching at large scale or for more restrictive search on a small period of time. As an example, when using extraction tool of [3] for searching *Hepcidin* gene at large scale on MEDLINE, an unreadable graph representing the extracted events is obtained (cf. Figure 1.). When restricting the search to papers published in a smaller period, a slighter but still very dense graph is obtained. Moreover, from the Hepcidin expert side, the extracted events often include a tremendous amount of background events that are well established since a while and that remain consequently trivial.

**Figure ali-baba**

Domain experts are therefore reluctant to use such tools as massive background events may hide pertinent ones, those that correspond for example, to new discoveries involving unexpected genes.

Tools currently in place are mainly devoted to the Bioinformatics community, [Nactem tools, BioNLP tasks], etc. The complexity of data extracted by these tools discourage bio-investigators from using them, yet richly annotated events are delivered at the scale of PubMed (mainly protein/gene interactions). From the side of the computational biologist, capable to decipher the output data, as of Jari Björne et. al. dataset stored in a MySQL database and made available on his web site [], these algorithms deliver relevant results and require considerable time machine evaluated in numbers of processors. The BioNLP Shared Task series also promotes actively the use of text-mining approaches at a community wide toward information extraction (IE) from biomedical literature. [2]

Despite these efforts, few studies have addressed the problem of processing this big-data in order to turn it into meaningful information of unexpected patterns over time. In order to prioritize non trivial and pertinent events published daily all over time, we have focused in our study on their temporal dimension. Information Extraction coupled to time allows a chronological delivery of published events best mirroring their real course over time. For example, two genes that have never been associated in literature before a certain time t and that appear unexpectedly after t are easily identified. Chronology in the description of events enhances comprehensiveness of relative knowledge especially for non experts. Biology is a knowledge-centred science. Researchers commonly rely on previously published knowledge represented in a set of factual events in order to produce further experimental new hypotheses. These hypotheses once validated, are in turn published and made available as a set of new events to the community for further investigations. There is therefore an implicit chronological dependency between events, which is imperceptible when time is omitted.
As suggested in the survey of [Lee E.K. AMIA 2011], to avoid missing past published events, possibly important, we propose an approach that uptakes a maximum of events extracted over a large period of time specified by the user, then mine corresponding events once again before revealing unexpected temporal trajectories of most pertinent genes, diseases and drugs that correspond to the user-query. The accent is made on relevant information published, defined by biomedical entities highly associated to other entities at time t and necessary to bring up to the user at this time-point as they may loss their relevance in future time points.

An important proportion of "already known" data corresponding to background knowledge dominates the extracted data and is not relevant to investigate by domain expert. Indeed, rather then extracting the whole events by mixing trivial and non trivial events, as done by major text mining tools, we suggest to unfold time and provide at each time point pertinent published entities associated in non-trivial events. A time unit of one month has been fixed as it is the publication period of most life science journals.

In the following, we will define precisely what we consider as being a pertinent biological entity at time t, namely t-relevant, and what we consider as being a trivial biological event at time t, by customizing information extracted by text mining web service of [3]. To illustrate our study, we will consider the hepcidin gene use-case and relative literature during a decade. New events on this gene are increasing rapidly since its discovery in Dec 2000 [4, 5, 6]. Mining relative abstracts, then computing and selecting highly relevant biological entities, allow turning back in time to recall key events published on this gene since its discovery, including biomolecular partners, associated diseases, location in other species, effects of mutations, and associated clinical treatments. During this process, we will quantify the proportion of trivial events during the decade, filter them and finally provide major non-trivial and pertinent events published during the Hepcidin decade.

Materials

Event Extraction from PubMed

Event Extraction Systems (EES) aim at extracting recognized biological entities and identify relationships between them, namely biological events from PubMed abstracts. For example, in the sentence “STAT3 inhibitors, including curcumin, AG490 and a peptide (PpYLKTK), reduced hepcidin1 mRNA” of [7], a reasonable EES system should deduce the following facts:

- Curcumin reduces hepcidin1
- AG490 reduces hepcidin1
- Peptide (PpYLKTK) reduces hepcidin1

When supported by contextual dictionaries, different types of biological entities are recognized. In this example, curcumin and AG490 are recognized as being biochemical substances or drugs, hepcidin1 and Peptide (PpYLKTK) are recognized as being bio-molecular entities, including genes and proteins. Returning to the source sentences enables checking or completing erroneous extracted facts.

There exist two fundamentally different approaches for extracting such relationships: the “co-occurrence” approach and the Natural Language Processing (NLP) approach. The first approach is straightforward and consists of identifying entities that simply co-occur in a sentence of the
abstract. Co-occurrence methods have better recall but worse precision than NLP ones, as co-occurring entities, even in a same sentence, may have no (or weak) relationship. The recognized entities are also normalized (i.e. linked to standard data source identifiers) in order to enable interoperability of extracted information with other data [8, 9].

Despite increased interests, only few tools are publicly available for extracting more precise biomolecular events [10]. Text Mining coupled to Machine-Learning methods have been applied for extracting rapidly different sorts of relationships, mainly physical protein-protein interactions [11,12,13], relationships involving diseases and GO Terms [14] and NLP-based specific relationships for extracting information on gene regulation and protein phosphorylation [15,16].

**Time-based pertinence of a biological entity**

The amount of knowledge published periodically on various biomedical items behaves differently during time. Globally, the number of publications evolves exponentially for some topic or keywords (Figure 2.a) whereas for others it fluctuates over time by the presence of rapid peaks occurring at successive time points (Figure 2.b). Various subjects are also not published with the same intensity: there are tens of thousands papers for keyword "breast cancer" against few tens per year for "Sporulation" keyword.

**Figure 2 MLTrends**

In the context of iron metabolism, the discovery on Dec 2000 of hepcidin gene as having a central regulation function at systemic level has created a strong trend of relative publications on this gene. This has made more complex the understanding of bio-molecular mechanisms of associated diseases, even by experts. As shown in Figure 2.c, the rapid increase of publications since year 2000, generated a cascading increase of publications on other genes like Ferroportin and BMP6, revealing a strong functional link between them. Using MLTrends of [17] from which the graphics of Figure 2 have been drawn, keyword-based publication trends are calculated by counting word frequency over time.

This process applied not only to ordinary keywords but to recognized entities and events connecting them leads to more informative functions supplementing traditional text mining tools. At each time-point (a month in our study) a focus is made on biomedical entities associated to other recognized entities in PubMed abstracts, namely relevant ones. For bio-investigators seeking to undertake in-depth retrospective studies regarding a gene and its context, the global delivery of the bio-entities linked to it over time, including proteins, diseases and drugs, allows a better visibility of the gene context research.

**Hepcidin use-case**

**Bio-events extraction:** we have periodically measured the number of events over two consecutive years by customizing web service results of [3]. Given a PubMed query, Ali-Baba extract rapidly corresponding events by recognizing in the query abstracts: i) different sorts of bio-entities as genes, proteins, diseases, cells, drugs, tissues, and species, ii) several events associating them in a sentence, mostly by co-occurrence or by more precise associations between entities like protein interactions, gene regulation, etc. It then draws the extracted relationships within a graph (cf. Figure 1), delivered in a GraphML format.
The quantification of events has been focused on a sample of genes, sufficiently balanced with well-known hepcidin partners and others that are uncommon to domain experts. For each gene of our sample, we have calculated periodically the number of events in which it has been associated. The results presented in the table below show that hepcidin – in dark gray – is associated to the highest proportion of events which probably include a considerable part of background events. The remaining elements are categorized into 2 distinct sets of genes. The first set – in medium gray – includes genes frequently associated to a large amount of events in almost all the periods, like Transferrin, Ferroportin and Hemojuvelin. Whereas these genes are well-known and easy to select by domain experts, they remain less familiar to neophytes and may be worth to investigate. The second set, highlighted in lighter gray, includes genes punctually associated to hepcidin events during time. These genes are unexpected by domain experts and may be interesting to bring out to them at this time point. In case of large scale extraction from PubMed, events of this set are completely dissolved in the total mass of events extracted, from which a considerable amount of background data. These events and relative genes are thus imperceptible when time is neglected. Integrating time to the process of information extraction and filtering background from the extracted events, give more chance to these unexpected entities and associated events to emerge.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Events</td>
<td>2357</td>
<td>814</td>
<td>1711</td>
<td>1536</td>
<td>2108</td>
<td>1869</td>
<td>700</td>
<td>675</td>
<td>1253</td>
<td>2204</td>
<td>1163</td>
<td>1337</td>
<td>47</td>
<td>300</td>
</tr>
<tr>
<td>Hepcidin</td>
<td>884</td>
<td>576</td>
<td>824</td>
<td>962</td>
<td>878</td>
<td>606</td>
<td>476</td>
<td>334</td>
<td>350</td>
<td>1046</td>
<td>376</td>
<td>438</td>
<td>17</td>
<td>140</td>
</tr>
<tr>
<td>Interleukin6</td>
<td>390</td>
<td>0</td>
<td>225</td>
<td>150</td>
<td>236</td>
<td>240</td>
<td>116</td>
<td>9</td>
<td>122</td>
<td>24</td>
<td>86</td>
<td>75</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tranferrin</td>
<td>168</td>
<td>0</td>
<td>48</td>
<td>132</td>
<td>36</td>
<td>83</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>80</td>
<td>64</td>
<td>134</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Ferroportin1</td>
<td>216</td>
<td>21</td>
<td>144</td>
<td>134</td>
<td>80</td>
<td>5</td>
<td>0</td>
<td>18</td>
<td>190</td>
<td>55</td>
<td>0</td>
<td>216</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hemojuvelin</td>
<td>130</td>
<td>0</td>
<td>369</td>
<td>303</td>
<td>272</td>
<td>64</td>
<td>0</td>
<td>14</td>
<td>198</td>
<td>278</td>
<td>112</td>
<td>92</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>BMPs</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>84</td>
<td>322</td>
<td>356</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MIP-1Beta</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SMAD4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SLC4DA1</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ceruloplasmin</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IL_1alpha</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HIF-1alpha</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

How much background events?

Background events correspond to the background knowledge necessarily reported in multiple abstracts by the authors and usually found in the background section (if any) of the abstract. This part of the abstract is similarly mined and associated events extracted.

Background events are identified by their repetitive occurrence in different abstracts and at several periods of time. They return continuously in literature but remain trivial to domain experts. In case of large amounts, which is usually the case, these events may hide unexpected entities and events to bio-investigators. We show in Figure XXX different occurrences of extracted event representing the induction of hepcidin transcription by interleukin-6, highlighted in green. While this event has been published for the first time in April 2003, it still have, until today, several occurrences in literature as a background information. This event is extremely important but is definitely trivial for
which sometimes may require manual reading of relative abstracts. Published events are also not recorded in data sources before a necessary delay, consequently be removed, without mentioning whether these facts concern bio-entities like proteins or diseases, or biological events associating them like interaction, regulation, co-occurrence, etc. Biological "facts" previously occurring in standard databases are considered as trivial and can consequently be removed, without mentioning whether these facts concern bio-entities like proteins or diseases, or biological events associating them like interaction, regulation, co-occurrence, etc. Published events are also not recorded in data sources before a necessary delay, which sometimes may require manual reading of relative abstracts.

Consequently, a trivial event is an event that occurs in literature more than once in a given interval of time. This is to our knowledge the only way to spot it. Being aware that when using PubMed search engine, or any text mining tool for extracting bio-events as of [], a background event is always delivered to the user within about one occurrence, probably the one that corresponds to the most recent corresponding event. The problem resides in their diversity and their global amount, which is so tremendous that it pollutes more relevant and non trivial information published. We thus propose to the user to calculate, even naively, the total number of trivial events extracted from PubMed has been briefly reported in [1]. For tracking them, authors propose in a perspective proposal a return to data sources during text mining. The problem of trivial events extracted from PubMed has been briefly reported in [1]. For tracking them, authors propose in a perspective proposal a return to data sources during text mining.
Methods

Definition1 - Time relevance of a biomedical entity

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 biological entities recognized with the same IE system. Or similarly, in a system where different types of biological entities and events are extracted, a \( t \)-relevant entity is the one that participates to a maximum of events at time \( t \).

To balance this definition, the number of abstracts in which the potentially relevant entity has been identified, is required, as time relevance increase with the number of abstracts in which associated events have been spotted.

Definition2 – A background event

A recognised event \( e \) is defined as being a background (or trivial) event, if it has been spotted repeatedly in literature at different time points. In other words, when an event \( e \) is published for the first time at time \( t \), it becomes trivial at following time points. More formally, given two distinct time points \( t_n \) and \( t_m \), \( t_m > t_n \), and associated events extracted and represented in graphs \( G_n \) and \( G_m \), background events published at time \( t_m \) are represented by the set of edges present repeatedly in graph \( G_n \) and \( G_m \).

Let us consider a longer retrospective study made on a duration of time \( D \) consisting of successive discrete time-points \( t_1, t_2, \ldots, t_d \) and associated events represented respectively in graphs \( G_1, G_2, \ldots, G_d \).

The identification of total background events published effectively during \( D \) is processed gradually by aggregating trivial events published at each time \( t_i \) of \( D \). Indeed, given time-point \( t_i \) and graph \( G_i \) of corresponding events, trivial events published at \( t_i \) are represented by subgraph \( T_i \) of edge \( e \), such as:

\[
T_i = \{ (e \in G_i) \text{ and } ( (e \in G_{i-1}) \text{ or } (e \in G_{i-2}) \ldots \text{ or } (e \in G_1) ) \} \tag{1}
\]

Consequently, the total amount of background events published all along \( D \) corresponds to:

\[
|T_1| + |T_2| + \ldots + |T_d| \tag{2}
\]

Figure pour démontrer les recouvrements ...

Data Pre-processing pipeline

For practical considerations we have set period \( p \) to one month, as it is the publication period of most life science journals. Given a PubMed Query \( Q \) and a retrospective duration \( D = [t_1, \ldots, t_d] \) for example, \( Q = \text{“hepcidin”} \) and \( I = [\text{jan-2001, dec-2011}] \), time-relevance is evaluated for each month of \( D \) by customizing information extracted by text mining tool of [3]. This latter identifies from resulting abstracts diverse biological entities (including genes, proteins, diseases, species, cells and drugs), and various relationships associating them. It then represents the extracted bio-entities and relative events within a graph. A dictionary-based approach for recognizing various bio-entities is used, with dictionaries collected from different sources, as UniProt for proteins, Drugbank for drugs, and NCBI Taxonomy for species. Relationships between entities are spotted using two
different methods: pattern matching technique and predominantly co-occurrence filtering. It has been chosen for its reuse facilities particularly the ability to export graphs of events in standard graphML format, easy to manipulate, and for being executable remotely as a web-service. Using a workflow centered on Ali-baba web-service (Figure 4.), data on events has been generated for each month of D.

Each event is represented by : i) the source entity , ii) the target entity, and iii) the purpose of their relationship, mainly in the form of co-occurrence in a sentence, or possibly more precise relationships like protein-protein interaction, induction, etc.

Source and target entities are identified by their official symbol (official name and synonyms), accession numbers in standard databases or in standard terminologies (for example, MesH for diseases and Swissprot for proteins). Each event is represented by the Edge-Label of corresponding relationship between source and target entities, and is additionnally stamped with its publication date.

Collected events are then transformed and integrated into a MySQL database of events devoted to our large-scale retrospective study. As shown in Figure 4. preprocessing pipeline prepares data by generating bio-events of successive months using successive calls to Ali-Baba web service. Extracted events of time-point $t_i$ are represented by the graph $G_i$ of nodes $E_i$ each representing a recognized biological entity, and edges $V_i$ each representing a recognized event connecting two nodes of $E_i$.

**Processing of time relevance**

Time relevance is then calculated for various types of entities thereby giving different sorts of valuable information to the user. Recall that in our context, a protein (or any other biological entity) is $t$-relevant, if the amount of events in which it participates is the highest at time $t$. Therefore, relevant entities at time $t_i$ correspond to nodes of $G_i$ with higher degrees.
Nodes 2 and 5 are set t-relevant as they correspond to maximum degree of $n = 6$.

However, time relevance may change according to different criteria on the source and target entities and on the type of events connecting them. For example, when the source entity type is fixed to protein, we can either select source proteins highly connected to all sorts of target entities, or to target entities of specific type. Indeed, the selection of source proteins highly connected to target proteins may return different relevance results. Similarly, the selection of proteins highly connected to target diseases or diseases highly connected to target drugs may be preferred. By applying this selection process to different combinations of source entity types and target entity types (Figure xxx), a multitude of valuable information can be derived allowing a variety of knowledge in accordance to the user focus.

Similarly, time relevance of graph $G_i$ vary according to the type of the events represented. Indeed, when the events of graph $G_i$ include background events, selected relevant entities are a fortiori associated to trivial events as these events are pre-dominant. Time relevance calculation after filtering background events must be re-calculate in order to highlight unexpected biological entities to the bioinvestigator.

Filtering background events is a time and space consuming procedure, as events of graph $G_i$ that validate assertion (1) are identified, for all time-point $t_i$ in duration $D$. Each event $e$ of $G_i$ is compared to events of previous graphs $G_{i-1}$, $G_{i-2}$, ... . When spotted, $e$ is declared as a background event and is consequently removed. Using the same principle of maximum degree, time relevance is re-calculated on the resulting filtered graph $G'_i$.

**Results and discussion**

This approach may be applied to any PubMed query $Q$, on any temporal duration $D$ and using any text mining tool for extracting bio-events from PubMed. However, we present here major results obtained using ali-baba text mining tool, for Hepcidin use-case over a decade of publications, as
this duration is sufficiently long to get interesting sight on the overall extracted data.

Main knowledge on hepcidin is presented in the biomedical literature. It describes interactions of large number of distinct bio-molecular entities and their relationships with different cell compartments, cell types, different species and diseases, and different drugs. Linked together, these elements represent a systemic view of iron homeostasis for which hepcidin is a major actor.

The most important discoveries on Hepcidin during the last decade: Human HAMP gene (HAMP, HEPC, OMIM 606464) is located on the long arm of chromosome 19 at position 13.1 and codes for a protein called hepcidin. Hepcidin is a circulatory antimicrobial peptide, synthesized in the liver as an 84 amino acid protein [4, 5, 6]. It plays a major role in maintaining iron balance in organism as a slight iron excess is toxic to the body. When iron enters liver cells from the blood, hepcidin is produced by the liver and released in the blood to travel throughout the body. It then interacts with Ferroportin, the unique cellular iron exporter [18] and induces its internalization and degradation to inhibit intestinal iron absorption, iron is controlled by the BMP/Hemojuvelin complex and SMAD signaling pathway. It has been demonstrated in [19] that the bone morphogenetic protein BMP6 has a preponderant role in the activation of the SMAD signaling pathway leading to hepcidin synthesis in vivo. Hepcid recycling by macrophages, and iron mobilization from hepatic stores. Hepcidin transcriptioin deficiency causes most of the forms of iron disorders, either due to mutations of hepcidin gene itself or to mutations of its regulators. Hepcidin appears as being the pathogenic factor in most systemic iron disorders and provide important tools for improving their diagnosis and treatment [18, 20].

Step1 : Filtering trivial events of the hepcidin decade

A large corpus of hepcidin events has been extracted and integrated to build a specific database devoted to hepcidin. This latter, stores a considerable amount of biological entities and events connecting them throughout the whole decade - about 300,000 events extracted. In this data-set a considerable amount of trivial events has been identified, i.e., events that are repeatedly published at several months. These events may correspond to the background knowledge usually reported in the abstracts and similarly mined. We thus attempt to measure their total amount, in order to extract them from the total hepcidin events and allow more pertinent events to emerge, especially to domain expert bio-investigators. For this purpose, we have operated a step by step cumulative quantification of trivial information published at each month t of the decade.

Trivial events of month t are identified by comparing each event e of t to events published at previous months t_{i-1}, t_{i-2}, ... t_0. When e is found it is marked as trivial.

In our study, two distinct events e_1 and e_2 are similar, if and only if:

1. Source(e_1) = Source(e_2)
2. Target(e_1) = Target(e_2)
3. EdgeLabel(e_1) = EdgeLabel(e_2)

We are aware that these assumptions are necessary but not sufficient to affirm a similarity between two published bio-events as : i) synonymy may be used to compare EdgeLabels, and ii) a bio-event may be much more complex than a simple connection between a pair of bio-entities.

As events extracted using our text mining tool are largely labeled “co-occurrence”, a strict
equivalence between EdgeLabels is sufficient. More precise relationships between entities, like: “protein-protein interaction”, “induce” and “involved in”, are possible but remain minor. From the whole hepcidin decade, about 93% of the total events extracted by ali-baba are labeled “co-occurrence”, favouring simple event comparisons.

The identification of trivial events published at time ti, for i = 1,n, n the number of periods, enables the calculation of the total number of trivial events published throughout the decade. An amount reasonably high has been revealed when applied to the whole decade. As shown in Figure 6, the total proportion of trivial events for the whole hepcidin decade (blue shape cut to dec 2011) is of about 59% when using co-occurrence events. More than 59% of background information may consequently be extracted from the total information after eliminating duplicate information. This proportion is sensibly higher when extended to the writing date of this paper.

Ce nombre peut être plus grands si on tient des évènements réels entre entités. Ceci reste une vue globale basée plus sur les oc-occurrences répétitif d'entités dans des évènements dans le temps, que une comparaison réelle de ces évènements par synonymie par exemple, car les même évènements d'occurrences peuvent être des précisions qui se publient au fur et à mesure du temps. Mais la co-occurences d'entité peut déjà être un révélateur d'un éventuel evenement important entre ces entités, ..etc.

**Step2 : select highly relevant bio-entities of the Hepcidin decade**

From the numerous information one can derive on each entity and each entity type, we choose to illustrate main results obtained on proteins and diseases, major focus of our bioinvestigators. Highly targeted proteins and diseases identified in abstracts citing **hepcidin** over 10 years have been derived. Resulting data for this duration are plotted in Figure 7 and commented in the following.

The identification of trivial events, those who uselessly burden the results of information extracted, helps their filtering and improves the display of unexpected events, previously not provided to the biologist. As shown in Figure 7.a, a high relevance and a permanent visibility is given to hepcidin and perpetual proteins well established by domain experts for their involvement in iron metabolism. For the whole decade, hepcidin, hfe, transferrin and ferritin are expectedly given relevant proteins in almost all periods of time due to the loads of trivial events associating them in the background domain. After filtering, we first noticed a drastic fall of data amount (Figure 7.b), especially in more recent years. Indeed, since hepcidin discovery in 2000, the amount of trivial events achieves 37% until 2003, 49% until 2006, and more than 52% after 2006. Another point is that filtering allows to new entities previously not visible to emerge as being relevant, like SMAD, BMPs proteins and interleukins, more recently described as being important transcription factors of hepcidin gene, et d’autres.

Expected disease names like iron-overload and anemia are given highly relevant over time (Figure 8.a). This result does not provide original information to bioinvestigators as events associated to these diseases, already well known, are often dense and hide unexpected phenotypes. New phenotypes emerge after filtering trivial information (Figure 8.b), like tuberculosis, colon-cancer, fish diseases and non hemochromatosis rare genetic diseases, identified more recently and due to mutations identified on Ferroportin gene, main target of hepcidin in the intestine [21].
In addition to proteins and diseases, other entity types as important as Species, Cell types and Drugs have been tackled. The filtering phase has revealed more peculiar species in which hepcidin has been localized and studied. Massive hepcidin studies are devoted to Mice, Rat and Human. Filtering this dominant knowledge allowed the emergence of unusual species in which hepcidin has been found, like fish species (Sole and Rainbowtrout), Bacteria, and diverse mammals like Bovine and Cynomolgus Monkeys.

Several combinations of entity types have also been studied for the whole decade and associated information filtered and exploited. To allow bioinvestigator having more information on the revealed entities, an access to data sources and PubMed_ids, is proposed in a supplementary annotation table.

As shown in the table of Figure 9, users have also access for each time-point to:

- The name of the relevant entity as identified in the abstracts: spotted entities like “human”, “patient” or “child” are equivalent and unified to “Human” Specie,
- The maximal number of target entities, along with their official names. In Figure 9, we can see that substance VitaminD3 has been highly cited by several diseases and phenotypes like anorexia, diarrhea, body-weight, etc. in abstracts of Sep 2011.
- The number of abstracts where relative events have been extracted. Using this number, the user can qualify the relevance of an entity, as relevance may increase with ever-increasing number of abstracts. User has also access directly to these abstracts in PubMed.
- Standard database identifier of the relevant entity giving full access to its description in corresponding databank (e.g. DrugBank for drugs, Swissprot for proteins, etc.)

**RetroMine results for other use-cases**

**Comparative study from different backgrounds**

As our approach may be used for any query, we have been able to provide a comparative study applied to different PubMed queries. The calculation of the proportion of trivial information published for distinct queries in a relatively similar retrospective duration, has revealed that this rate is conspicuously different from one domain to another (Figure 10.a, Figure 10.b), achieving 64% of the total amount of events for “BMP6” query, and only 14% for query “Osteoporosis”, suggesting that the background knowledge is probably much more important in BMP6 relative abstracts than in Osteoporosis.

**Retromining hot queries**

RetroMine has also been applied to more popular queries, those concentrating massive biomolecular and medical research like “Breast Cancer”. Relative literature on this domain is rapidly increasing and may discourage non experts wishing to be promptly informed. A call to RetroMine for a reasonable duration for data generation (pre-processing), gave priority access and rapidly enough, to highly cited proteins, drugs, cell types and diseases linked to breast cancer over time,
Conclusion

The application at large scale of advanced information retrieval and extraction tools to the biomedical literature has yet confirmed of the extraordinary amount of biological events and entities published daily in Medline. These torrential data, once extracted, look forward to be mined in order to draw unexpected trajectories of behaviours over time. RetroMine, as a prototype, is a straightforward application that provides in a glance, retrospective studies on genes, diseases and drugs. Time has been bound to the process of information extraction to enhance comprehension of the extracted biological events by introducing chronology and giving priority to events highly targeted over time.

RetroMine enabled also to reveal the amount of trivial events published in the biomedical literature. While these events are essential for a non expert wishing to learn common contextual data, they can pollute considerably pertinent knowledge for domain experts. Filtering trivial information published during the hepcidin decade has clearly allowed unexpected information to emerge. Excluding review papers from the process of information extraction possibly prevents the assault of such events from the extracted ones. We propose also to focus text mining, when possible, on the “Results” section.

RetroMine may also be customized to accelerate tasks of database feeders as the amount of publications is increasing exponentially. Database feeders “read” continuously newly published abstracts (or full text articles) and integrate relative events in standard life science databases. As an example, the OMIM database lists established genetic diseases and links them to relevant genes in Human Genome. Time-line reporting of the main relevant events published is visible in OMIM given a query. Chronology of events and their relevance over time are also clearly perceptible using RetroMine. Today, slow feeding is prominent to life science databases (Ex: hepcidin gene is reported in OMIM using publications ending at early 2009!).

In this paper we used the tool of [3] in which events of type “co-occurrence” are predominant. Improving event accuracy and complexity through more specialized methods, like EventMine of [22] and more devoted resources of [23], certainly help reducing the amount of trivial information and diversify the spectrum of queries. Also, extending period usages (from one month to one year for example) certainly lead to more coarse grain relevance and larger retrospective studies.

Finally, RetroMine focus primarily on recognized biomedical entities identified in PubMed abstracts to build several retrospective studies, while more complex objects, such as metabolic pathways may be targeted. Putting the scale in RetroMine to pathway data, using text mining tools for pathway curation as of [24], may improve comprehension of their dynamism and complexity over time.

Comme c'est le cas de ce type de méthodes on perd quand même en précision et on gagne en temps … comme c'est toujours le cas quand c'est des co-occurrences (revoir Materials). Très fortement conseillée lorsque le but n'est pas dans la précision mais de capter une vue globale sur une grande quantité de données … à revoir ou bien dans discussion.
Acknowledgements

Astrid Reinländer provided important help for maintaining ali-baba continuously working at HU-Berlin. Thanks to Pierre Le Marre who designed RetroMine web site.

Members of “Iron and Liver” Group, at INSERM U991 provided valuable ideas to improve ongoing developments and outputs of RetroMine.

References


Figure 1. Outputs of Ali-Baba text mining tool of [3] for Query Hepcidin

This figure provides an overview of the high density of information extracted using Ali-Baba text mining tool, for Query 'Hepcidin' on MEDLINE. It includes nodes of recognised extracted proteins in green, diseases in pink, drugs in red, species in blue, cell-types in turquoise, and tissues in orange. It also includes different relationships between these nodes, reflecting their co-occurrences or more precise relationships, identified within abstract sentences.

Figure 2. Keyword Trends over time in Medline using MLTrends

Strong publication trends of keyword “Breast Cancer” are shown in Figure 2.a compared to much weaker occurrences of “Sporulation” keyword in Figure 2.b Comparative graphing of multiple keywords: Hepcidin+BMP6+Ferroportin is provided in Figure 2.c.

Figure 3. $S_n$ is the set of events occurring at time $t_n$, $S_m$ the set of events of time $t_m$, trivial events, $S_n \cap S_m$, is the subset of events occurring repeatedly at time $t_n$ and $t_m$.

Figure 4. Ali-baba based workflow for collecting, transforming and analyzing the extracted events, given a PubMed query Q and a temporal duration D. During pre-processing, Q relative events are generated for each period of D, then transformed and stored in RetroMine database of events. The processing phase is in charge to prioritize t-relevant biological entities and relative events, and proposes to filter them from trivial information published.

Figure 5. Hepcidin across different organs is able to regulate iron amount in the body.

Hepcidin is produced in the liver and released in the blood to go across the body. In case of iron excess, Hepcidin interacts with Ferroportin (Fp), the unique iron exporter and induces Fp degradation to decrease intestinal iron absorption, iron recycling by macrophages, and iron mobilization from hepatic stores.

Figure 6. Proportion of trivial events (blue shape) vs. new events (red shape) published during the Hepcidin decade. An exact quantification of the number of trivial events occurring over 10 years of publications has revealed a proportion of 59% until dec 2011, and more than 60% until the writing date of this paper.

Figure 7. Highly scored proteins over the Hepcidin decade.

In Figure 7.a extracted events include trivial information on Hepcidin, Transferrin, and HFE genes, dominating all along the decade. After filtering, other biological entities and events are highlighted in Figure 7.b. Using filtering, unexpected information may emerge.

Figure 8. Highly scored diseases over the Hepcidin decade.
Similarly, extracted events include trivial information on iron diseases like anemia and hematochromatosis (Figure 8.a) providing non pertinent information to domain experts. After filtering, other unexpected phenotypes like those linked to tuberculosis and fish diseases are highlighted in Figure 8.b.

**Figure 9. Annotation table of highly scored entities along with their targets.**

RetroMine provide in a supplementary annotation table for each monthly revealed entity: i) its official symbol and hyperlink to corresponding standard database including SwissProt for proteins, DrugBank for drugs, NCBI Taxonomy for species and MeSH annotations for diseases, tissues and cell types, ii) the abstracts where the high number of extracted events have been spotted, and iii) official name of the target entities of these events.

**Figure 10. Comparative study of the extracted events for queries “BMP6” and “Osteoporosis”.**

Focus has been made on the calculation of trivial information published in a relatively similar time. The study revealed that the ratio of trivial information published is comparatively more significant in “BMP6” results (Figure 10.a) than in “Osteoporosis” ones (10.b).

**Figure 11. Highly relevant drugs linked to “Breast Cancer” from Jan 2009 to Dec 2013 annotated with RetroMine supplementary table.**
ajouter un titre pour les computationnal requirements à la façon de l'article « complex event extraction at pubmed scale » de jari björne