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

Bertrand Ameline de Cadeville [2], Olivier Loréal [1], Fouzia Moussouni-Marzolf [1,2,*]

[1]INSERM UMR 991, CHU Pontchaillou, 35033 Rennes, France

[2]Université de Rennes1, Faculté de Médecine, 35043 Rennes, France

Summary

The rapid expansion of biomedical literature has provoked an increased development of advanced text mining tools to rapidly extract relevant events from the continuously increasing amount of knowledge published periodically in PubMed. However, bioinvestigators are still reluctant to use these tools for two reasons: i) a large volume of events is often extracted upon a query, and this volume is hard to manage, and ii) background events dominate search results and overshadow more pertinent published information, especially for domain experts. In this paper, we propose an approach that incorporates the temporal dimension of published events to the process of information extraction to improve data selection and prioritize more pertinent periodically published knowledge for scientists. Indeed, instead of providing the total knowledge associated with a PubMed query, which is usually a mix of trivial background information and non-background information, we propose a method that incorporates time and selects non background and highly relevant biological entities and events published over time for bioinvestigators. Before excluding background events from the total knowledge extracted, a quantification of their amount is also provided. This work is illustrated by a case study regarding Hepcidin gene publications over a decade, a duration that is sufficiently long enough to generate alternative views on the overall data extracted.

1 Introduction

Modern biology is one of the most voluminous areas of science today. This is mainly due to the modern digital revolution and increasing enticements to publish and share new knowledge worldwide. Hundreds of publications appear in MEDLINE each day, resulting in an exponential increase in knowledge over the past few centuries. The volume of publications has increased so rapidly over the past few years that it is no longer possible for a researcher to stay up-to-date with recent discoveries from the literature.

Alternative solutions are required to search for articles, and researchers must learn how to search rapidly and efficiently in this continuously growing collection of knowledge. Text mining coupled with the process of information extraction is a current solution for quickly scanning the volume of literature 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 area of interest [27]. However, these tools still present difficulties, as they are largely developed by information experts. A gap exists between computational linguists, designers of information extraction tools, and biomedical investigators, resulting in many pertinent articles not being pulled in a Medline search. If we consider PubMed, the most popular biomedical search engine for the MEDLINE database, as an example, a survey conducted by Eva Lee et al. in [28] has proved the following statements:

* To whom correspondence should be addressed. Email: fouzia.moussouni@univ-rennes1.fr

doi:10.2390/biecoll-jib-2015-275

1. Few users review results beyond the first page. Most of the selections are on the most recent citations, located on the first result page (by default, PubMed returns 20 results per page).
2. Users seek simple interfaces and have difficulties using the advanced features of PubMed because of the knowledge required to use MeSH terms and the multitude of steps required to build a query.
3. The majority of users retrieve in-depth information focusing on one category of information, such as on a cell, disease, drug or gene.

In addition, when the extracted information consists of recognized biological entities and complex events that are connected in abstracts, the results are even worse. A mass of data is extracted when searching either at a large scale or for a more restrictive search for a small period of time. As an example, when using the extraction tool of [3] and searching for the Hepcidin gene on a 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 time period, a slightly less dense graph is obtained. Moreover, from the view of a Hepcidin expert, the extracted events often include a tremendous amount of background events that have been well established for years and that consequently remain trivial. Domain experts are more reluctant to use such tools as the extraction of massive background information may hide pertinent events involving unexpected genes.

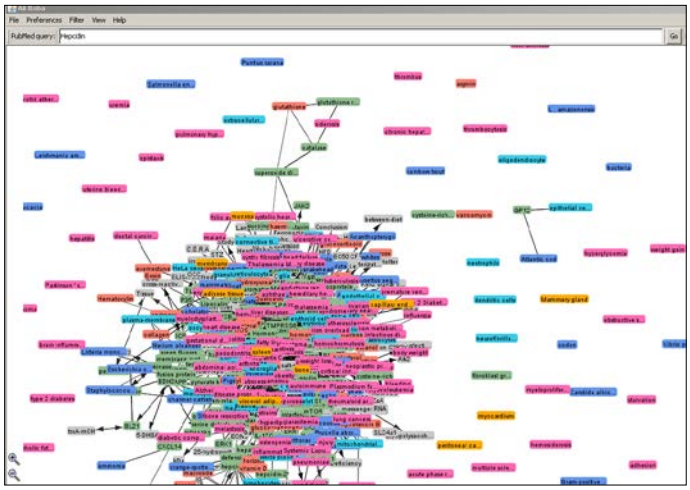

Figure 1. Screen-shot of ali-baba outputs for query Hepcidin.

The tools that are currently in place are mainly devoted to the Bioinformatics community. The complexity of the delivered data and data structure discourage bio-investigators from using them, yet richly annotated events are delivered at the scale of PubMed (mainly protein/gene interactions). From the view of a computational biologist who is able to decipher the output data, such as that of Jari Björne et al. dataset stored in a MySQL database and made available on his web site [25], these algorithms deliver relevant results but require considerable time and a high number of processors. The BioNLP Shared Task series also actively promotes the use of text-mining approaches at a community wide level for information extraction (IE) from biomedical literature [2].

Despite these efforts, few studies have addressed the problem of processing these big-data sets in order for non-computational experts to rapidly convert them into meaningful patterns over time.

To prioritize non background but pertinent events that are published daily over time, we have focused our study on the temporal dimension of the articles. Information Extraction coupled with time allows for a chronological delivery of published events that best mirrors their real course. For example, two genes that have never been associated with one another in the

literature before a certain time t, and that appear unexpectedly after t are easily identified. Chronology in the description of events enhances the comprehensiveness of relative knowledge, especially for non experts.

Biology is a knowledge-centered science, as researchers commonly rely on previously published findings to further produce new experimental hypotheses. These hypotheses, once validated, are in turn published and made available as a set of new events to the community for further investigations. Therefore, there is an implicit chronological dependency between articles, which is imperceptible when time is omitted.

As suggested in a previous survey [28], to avoid missing past published events of possible importance, we propose an approach that pulls a maximum number of events that were extracted over a large period of time. Next, the corresponding events are filtered again before revealing unexpected temporal trajectories of the most pertinent genes, diseases and drugs that best match the user query. The emphasis is given to the relevant information published, as defined by biomedical entities that are highly associated with other entities at time t and that are necessary to reveal at this time, as they may lose their relevance at future time points.

An important proportion of "already known" data corresponding to background knowledge dominates the extracted information and is not relevant for investigation by a domain expert. Indeed, rather than extracting the entire set of articles by mixing background and non background events, as done by major text mining tools, we suggest incorporating time and providing, at regular periods, pertinent biological entities that are associated in non background events. A time unit of one month has been decided upon, as it is the publication period of most life science journals.

In the following text, we will precisely define what we consider as a pertinent biological entity at time t, namely, t-relevant, and what we consider as a trivial “background” event at time t, by customizing information extracted by a selected text mining web service [3]. To illustrate our study, we will consider the hepcidin gene case and the relative literature pertaining to this topic over a decade. New events on this gene have increased rapidly since its discovery in Dec 2000 [4, 5, 6]. The process includes mining relative abstracts, computing and selecting highly relevant biological entities, and allowing for a retrospective investigation 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 and filter the proportion of background events to ultimately provide the major non-background and pertinent events published concerning Hepcidin over the last decade.

2 Related works

2.1 Event Extraction from PubMed

Event Extraction Systems (EES) aim to extract 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” [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 and hepcidin1 and Peptide (PpYLKTK) are recognized as being bio-molecular entities, including genes and proteins. Returning to the source sentences enables the checking or completing of erroneous extracted facts.

There are 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 the same sentence, may have no (or a weak) relationship. The recognized entities are also normalized (i.e., linked to standard data source identifiers) to enable the interoperability of the extracted information with other data [8, 9].
Despite increased interest, only a few tools are publicly available for extracting more precise bio-molecular 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].

2.2 Time-based pertinence of the biological entities in PubMed

The amount of knowledge published periodically on various biomedical items varies over time. Globally, the number of publications evolves exponentially for some topics or keywords (Figure 2a), whereas for others, it fluctuates over time with a pattern of rapid peaks at successive time points (Figure 2b).

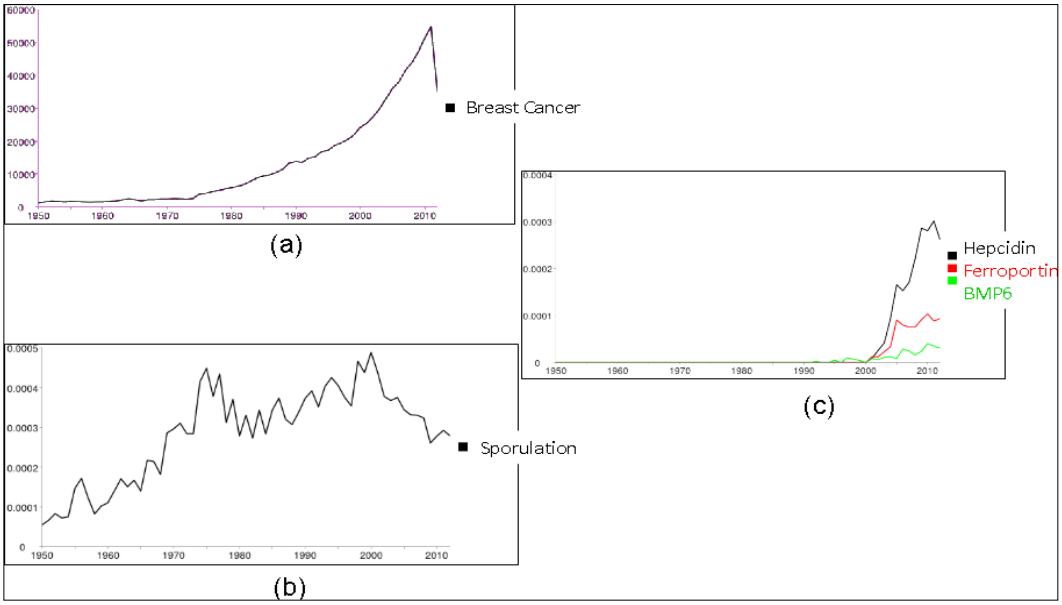

Figure 2. Keyword Trends over time in Medline using MLTrends

Various subjects are also not published with the same intensity. There are tens of thousands of papers for the keyword "breast cancer", but only a few dozen per year for the keyword "Sporulation".
In the context of iron metabolism, the discovery on Dec 2000 of the hepcidin gene having a central regulation function at the organism level has created a strong trend of relative publications. This has increased the complexity of the understanding of the bio-molecular mechanisms of associated diseases, even for experts. As shown in Figure 2c, the rapid increase of publications since 2000 generated a cascading increase of publications on other genes, such as Ferroportin and BMP6, revealing a strong functional link between them. Using MLTrends of a previous study [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, when applied not only to ordinary keywords but to recognized entities and events connecting them, leads to more informative functions to supplement traditional text

mining tools. At each period a focus may be made on biomedical entities that are highly 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 partners, the delivery of the bio-entities linked to it in PubMed, including proteins, diseases, drugs, cells and species, allows for a better temporal visibility of the gene's global research.

3 Methods

3.1 Event Extraction applied to Hepcidin use-case

Table 1. shows hepcidin events extracted over two consecutive years by customizing Ali-Baba web service results. Ali-Baba uses a pattern-based approach offering higher precision at the cost of increased complexity compared to regular expression matching or simple co-occurrence [30]. A user-centric presentation can be found in [3] and the pattern learning approach is described in [31].
An Ali-Baba extract of corresponding events was returned by recognizing several entities in the query abstracts. These included different sorts of bio-entities, such as genes, proteins, diseases, cells, drugs, tissues, and species, and several articles associating these entities in a sentence, mostly by co-occurrence. The search then draws, optionally, the extracted relationships within a graph (Figure 1.), delivered in a GraphML.

Number of Events / Month	2005 Jan	2005 Mars	2005 Mai	2005 Jul	2005 Sep	2005 Nov	2005 Dec	2006 Jan	2006 Mars	2006 Mai	2006 Jul	2006 Sep	2006 Nov
Total Events	2357	814	1711	1536	2108	1869	700	675	1253	2204	1163	1337	47
Hepcidin	884	576	824	962	878	606	476	334	350	1046	376	438	17
Interleukin6	390	0	225	150	236	240	116	9	122	24	86	75	0
Tranferrin	168	0	48	132	36	83	0	0	50	80	64	134	6
Ferroportin1	216	21	144	134	80	5	0	18	190	55	0	216	0
Hemojuvelin	130	0	369	303	272	64	0	14	198	278	112	92	6
BMPs	0	0	0	0	0	0	14	0	84	322	356	0	0
MIP-1Beta	40	0	0	0	40	0	0	0	0	0	0	0	0
SMAD4	0	0	0	0	0	0	44	0	0	0	16	0	0
SLC4DA1	0	24	0	0	20	0	9	0	0	0	0	0	0
Ceruloplasmin	0	0	10	0	0	19	0	0	0	0	0	0	0
IL_1alpha	22	0	0	0	0	0	0	0	0	0	0	0	0
HIF-1alpha	0	0	0	0	0	0	0	0	0	0	0	19	0

Table 1. Periodic number of events extracted for a sample of genes

The quantification of events has been focused on a sample of genes that are sufficiently balanced with well-known hepcidin partners and others that are uncommon to domain experts. For each gene of our sample, we periodically calculated the number of events in which hepcidin has been associated.
The results displayed in Table 1, show that hepcidin, in dark gray, is associated in the highest proportion of events, which probably includes a considerable amount of background information. The remaining elements are categorized into 2 distinct gene sets. The first set, in medium gray, includes genes that are frequently associated with a large amount of events in almost all of the periods, such as Transferrin, Ferroportin and Hemojuvelin. While these genes are well-known and easy to recognize by domain experts, they remain less familiar to novices and may be worth investigation.
The second set, highlighted in lighter gray, includes genes that are punctually associated in hepcidin articles over time. These genes are unexpected by domain experts and may be

interesting to extract at corresponding time points. In the case of large scale extraction from PubMed, the articles from this set are completely dissolved in the total mass of articles extracted, from which a considerable amount is background data. These events and relative genes are thus imperceptible when time is neglected. Integrating time into the process of information extraction and filtering background information from the extracted events yields a higher chance that these unexpected entities and associated articles will emerge.

3.2 How many background events?

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

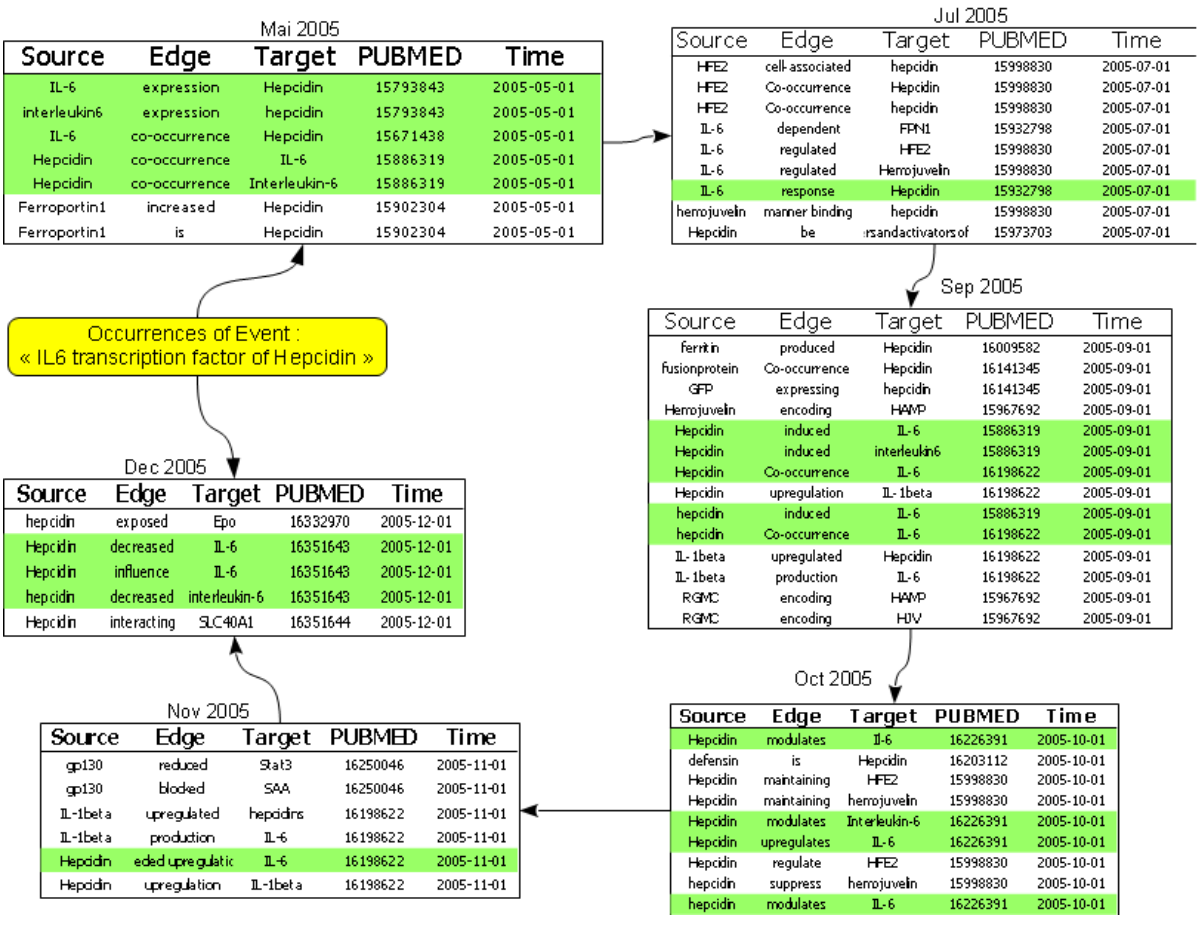

Figure 3. Background events relating hepcidin and interleukin-6 extracted at different periods

Background events are identified by their repetitive occurrence in different abstracts over several periods of time. They are returned continuously in the literature but remain trivial to domain experts. In the case of large volumes, which usually occur, these events may hide unexpected entities and events from bioinvestigators. Figure 3. shows different occurrences of the event representing the induction of hepcidin transcription by interleukin-6, highlighted in green. While this event was published for the first time in April 2003, it still has, until today, several occurrences in the literature as background information. This event is extremely important but is definitely trivial for hepcidin experts. Consequently, a background event may have multiple occurrences in the literature. This is, to our knowledge, the only way to spot the occurrences. We are aware that when using the PubMed search engine or other extraction tool, a background event is always delivered to the user within a few occurrences, probably the ones that are spotted in more recent abstracts. The problem resides in their diversity and

their global amount, which is so tremendous that they pollute more relevant and non-trivial information that is published. Therefore, we propose to identify, quantify (even naively), and filter the background events published in the specified search period to allow for non-trivial information to emerge in the resulting articles.
The problem of background events extracted from PubMed has been briefly reported [1]. To track these articles, authors propose a return to data sources during text mining. Biological "facts" previously occurring in standard databases are considered to be trivial and can consequently be removed, without mentioning whether these facts concern bio-entities, such as proteins or diseases, or biological events associating the entities, such as interaction, regulation, co-occurrence, and others. Published articles are also not recorded in data sources before a necessary delay, which sometimes may require a manual review of relative abstracts.

3.3 Time relevance of a biomedical entity

Definition

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. In a system where different types of biological entities and events are extracted, a *t-relevant* entity is 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 increases with the number of abstracts in which associated events have been spotted.

3.4 A background event

Definition

A recognized event *e* is defined as being a background (or trivial) if it has been spotted repeatedly in the literature at different points in time. In other words, when an event *e* is published for the first time at time *t*, it becomes trivial at subsequent time points.
More formally, let us consider a longer retrospective study made on a duration of time *D* consisting of successive discrete time-points $t_1, t_2 \dots, t_d$. Associated events published at each time-point are represented, respectively, in graphs $G_1, G_2 \dots, G_d$. The identification of the total background events effectively published during *D* is processed gradually by aggregating the trivial events published at each time t_i of *D*. Indeed, given time-point t_i and corresponding graph G_i , the background events published at t_i are represented by subgraph g_i of edge *e*, such as:

$$g_i = \{ e \in G_i \mid (e \in G_{i-1}) \text{ or } (e \in G_{i-2}) \dots \text{ or } (e \in G_1) \} \quad \textbf{(1)}$$

Consequently, the amount of background events published until t_i (i.e. from t_1 to t_i) is equal to:

$$\Sigma |g_j|, \text{ for } j = 1 \rightarrow i \quad \textbf{(2)}$$

3.5 Data pre-processing pipeline

For practical considerations, we have set period *p* to one month because this is the publication period of most life science journals. Given a PubMed Query *Q* and a retrospective duration *D*, in our study, *Q = "hepcidin" and D = [Jan-2001, Dec-2011]*, time-relevance is evaluated for each month of *D* by customizing information extracted by web service of [3]. This system recognizes in resulting abstracts diverse biological entities (including genes, proteins, diseases, species, cells, and drugs) and various relationships associating these entities. 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, such as UniProt for proteins, Drugbank for drugs, and NCBI Taxonomy for species. Relationships between entities are spotted using two different methods. These are the pattern matching technique and predominantly co-occurrence filtering. This method has been chosen for its reuse facilities, particularly its ability to export graphs of events in standard graphML format, its ease of manipulation, and its ability to be executable remotely as a web-service. We have used a workflow centered on web-service to generate massive amounts of data on events for each month of *D*. A pre-processing pipeline prepares data by generating bio-events of successive months using successive calls to the web-service.
Each extracted event is represented by: i) source entity, ii) target entity, and iii) the purpose of their relationship, mainly in the form of co-occurrence in a sentence, or possibly more precise relationships, such as protein-protein interactions, induction, and so on.

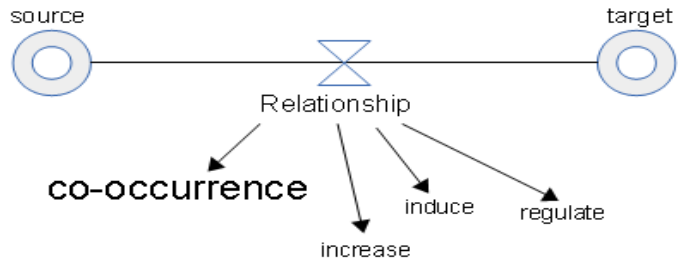

Figure 4. Graphical representation of an extracted event

Source and target entities are identified by their official symbol (official name and synonyms) and accession numbers in standard databases or in standard terminologies (for example, MesH for diseases and Swissprot for proteins). Each event is represented by an Edge-Label corresponding to the spotted relationship between source and target entities in abstract sentences and is additionally stamped with its publication date. The collected events are then transformed and integrated into a mysql data warehouse of events devoted to our large-scale retrospective analyses.

3.6 Data processing of time relevance

When data pre-processing is completed, the collected events are stored in the integrated data warehouse and are available for a large scale extraction of global trends of events over the decade. Time relevance is 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 sort of 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.
However, time relevance may change according to different criteria on source and target entities and on the type of events connecting them. For example, when the source entity type is fixed to a protein, we can either select source proteins that are highly connected to all sorts of target entities or to specific type of target entities.
Indeed, the selection of protein sources highly related to protein targets may return different relevance results. Similarly, the selection of proteins highly connected to disease targets or diseases highly connected to drug targets may be preferred. By applying this process to different combinations of source and target entity types (Figure 7.), a multitude of valuable information can be derived allowing for a variety of knowledge according to the user's focus. Similarly, the time relevance of graph G_i varies according to the type of the represented events. Indeed, when graph G_i includes background events, relevant entities are *a fortiori* associated with background events, as these events are pre-dominant. Time relevance after filtering background events must be re-evaluated to highlight unexpected entities to bioinvestigators.

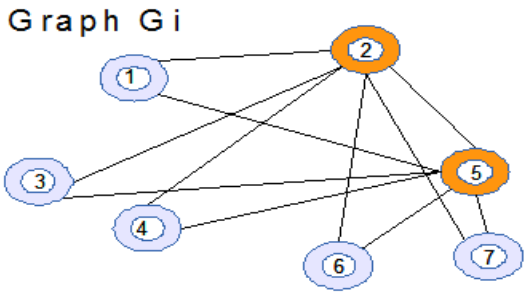

Figure 5. Nodes 2 and 5 are *t_i-relevant* as they have a maximum degree of *n = 6*.

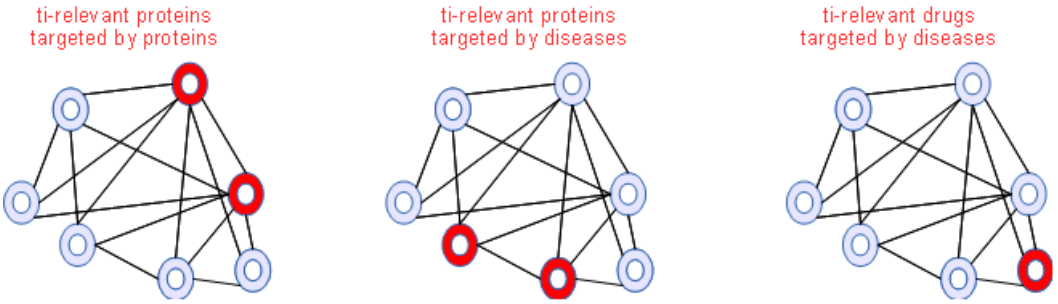

Figure 6. Time relevance variation after combining different type of source and target entity.

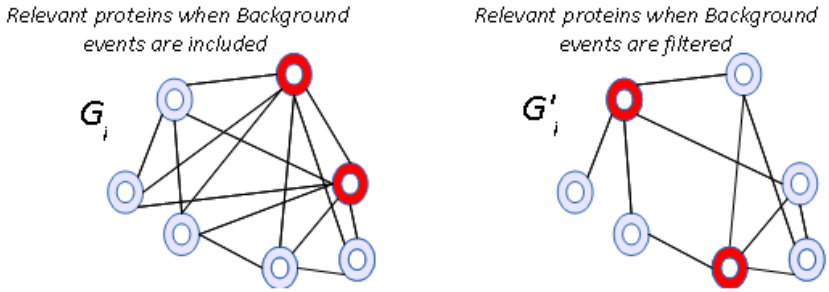

Figure 7. Time relevance with and without background events

Filtering background information 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, $i = 1 \rightarrow d$. Each event e of G_i is compared to events of previous graphs G_{i-1}, G_{i-2}, … . The spotted background events are consequently removed. Using the same principle of maximum degree, time relevance is repeatedly computed on the resulting graph G'_i.
These graph-based calcultaions have been carried-out using mainly SQL and Java on the top of the MySQL data warehouse that stores the extracted and transformed Hepcidin events.

4 Results

Our study may be generalized to any query and using any text mining tool for extracting bio-events from PubMed. However, we present major results obtained using ali-baba text mining tool (which is no more available) applied to Hepcidin use-case over a decade of publications, as this duration is sufficiently long enough to gain an interesting perspective on the overall extracted data.

The main knowledge on hepcidin is presented in the biomedical literature. It describes interactions of a large number of distinct bio-molecular entities and their relationships with different cell compartments and 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.

4.1 Most important discoveries linked to Hepcidin during the last decade

The 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 that is synthesized in the liver as an 84 amino acid protein [4, 5, 6]. It plays a major role in maintaining the iron balance in organisms, as a slight iron excess is toxic to the body. When iron enters liver cells from the blood, ***hepcidin*** is produced 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 previously [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. Hepcidin transcription deficiency causes most of the forms of iron disorders, either due to mutations of the hepcidin gene itself or due to mutations of its regulators. Hepcidin appears as the pathogenic factor in most systemic iron disorders and provides important tools for improving diagnosis and treatment [18, 20].
A retrospective study of the gene has been more recently reported by hepcidin experts [29]. The review turns back on the most important events published during the last decade of research since hepcidin discovery, arguing that the community is in constant demand of such initiatives.

4.2 Background events of the hepcidin decade

A large corpus of hepcidin articles has been searched and integrated to build our data warehouse of events devoted to hepcidin. This stores a considerable amount of biological entities and articles, connecting them throughout the whole decade. In total, approximately 200,000 events were extracted pertaining to our gene. In this data-set, a considerable amount of background events have been identified. We thus attempt to measure their total amount to extract this background from the total hepcidin events and allow more pertinent items to emerge. For this purpose, we have operated a step by step cumulative quantification of trivial information published at each month t_i of the decade.
Background events of month t_i are identified by comparing them to events published at previous months t_{i-1}, t_{i-2}, … t_0. Two events, e_1 and e_2, are similar if and only if:

$$Source(e_1) = Source(e_2)$$
$$Target(e_1) = Target(e_2)$$
$$EdgeLabel(e_1) = EdgeLabel(e_2)$$

We are aware that these assumptions are necessary but not sufficient to affirm a similarity between two published bio-events. A bio-event may be much more complex than a simple connection between a pair of bio-entities. But, using Ali-baba web service, the extracted events are predominantly labeled “co-occurrence“, and a strict equivalence between edge labels is sufficient in this case to affirm event similarities. More precise relationships between entities, such as “protein-protein interaction”, “induce”, ”involved in”, and so on, are possible but remain minor. From the whole hepcidin decade, approximately 93% of the total events extracted for hepcidin are labeled “co-occurrence”, favoring simple event comparisons. The identification of trivial events published at each period t_i , for $i = 1 \rightarrow n$, with n being the

number of periods, enables the calculation of their total amount throughout the decade. A reasonably high amount has been revealed when applied to the whole decade.
As shown in Figure 8. the total proportion for the whole hepcidin decade (blue shape cut to Dec 2011) is of approximately 58% when using “co-occurrence” event labels. More than 58% of the background information may consequently be excluded from the total information after eliminating duplicates. This proportion is sensibly higher when extended to the submission date of this paper.

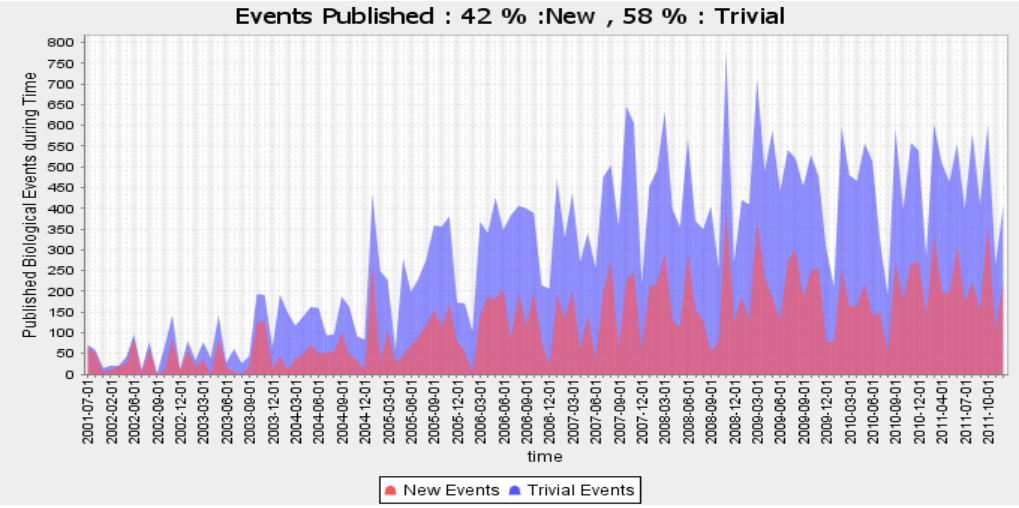

Figure 8. Proportion of background events (blue shape) vs. non background events (red shape) published periodically during the Hepcidin decade.

4.3 Extraction of time relevant entities linked to Hepcidin

From the vast amount of information one can derive on each entity and each entity type, we choose to illustrate the main results obtained on proteins and diseases, the major focus of our bioinvestigators. Highly targeted proteins and diseases identified in abstracts citing *hepcidin* over 10 years have been derived. The resulting data obtained during the decade are plotted below and are commented on in the following.

4.3.1 Highly targeted Proteins

The identification of background events, those who uselessly burden the results of information extracted, aided in the filtering of these articles and improved the display of unexpected events that were previously not visible to the biologist. As shown in Figure 9., a high relevance and a permanent visibility is given to hepcidin and perpetual proteins that are well established and easy to recognize 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. This is due to the high volume of background events associating them.
After filtering, we first noticed a drastic fall in the amount of data (Figure 10.), especially in more recent years. Indeed, since the discovery of hepcidin in 2000, the amount of background events composes 37% until 2003, 49% until 2006, and more than 52% of the results after 2006. Additionally, filtering allows for new bio-entities that were previously not visible to emerge as being relevant, such as SMAD, BMPs proteins and interleukins, which have been more recently described as important transcription factors of the hepcidin gene.

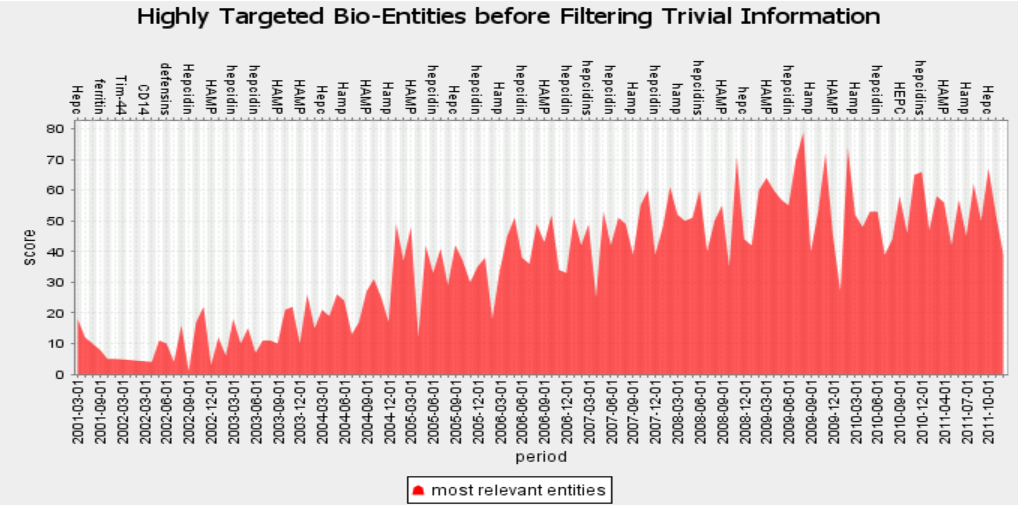

Figure 9. Highly scored proteins over the Hepcidin decade when the background is included

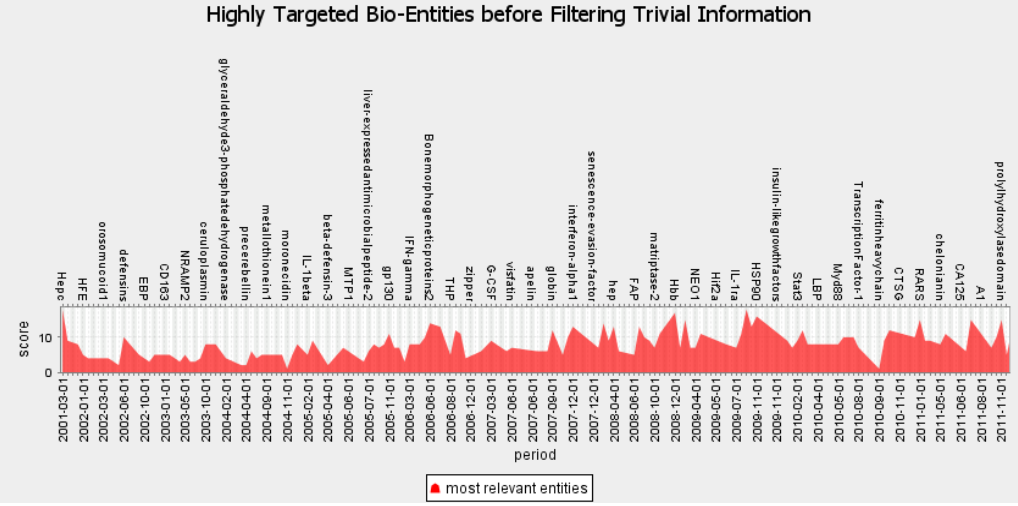

Figure 10. Highly scored proteins over the Hepcidin decade after filtering the background

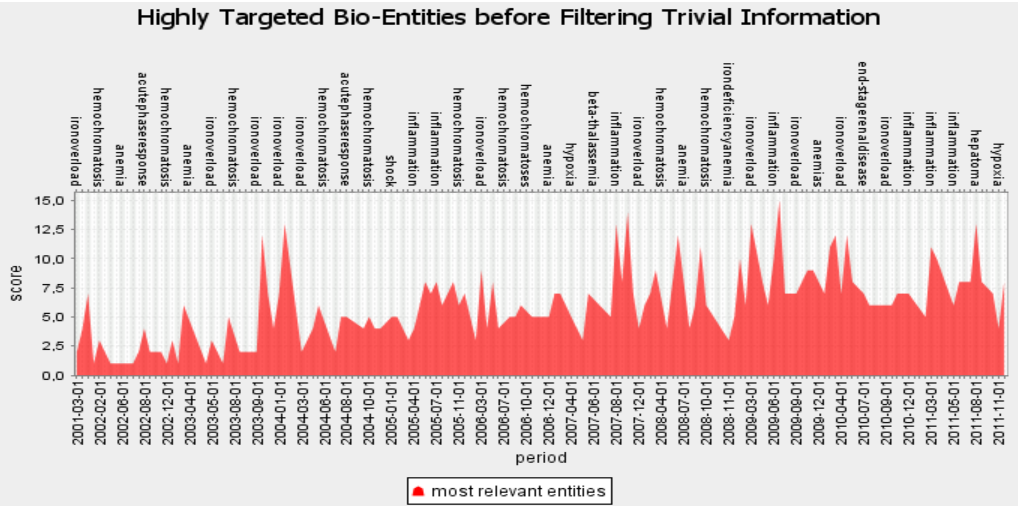

Figure 11. Highly targeted diseases when the background is included

4.3.2 Highly targeted diseases

Expected disease names, such as iron-overload and anemia, are highly cited over time (Figure11.). These results do not provide original information to bioinvestigators, as events associated to these diseases, which are already well known, are often dense and hide unexpected phenotypes.
New phenotypes emerge after filtering background information (Figure12.), such as tuberculosis, colon cancer, fish diseases and non-hemochromatosis rare genetic diseases, identified more recently due to mutations identified on the Ferroportin gene, a main target of hepcidin in the intestine [21].

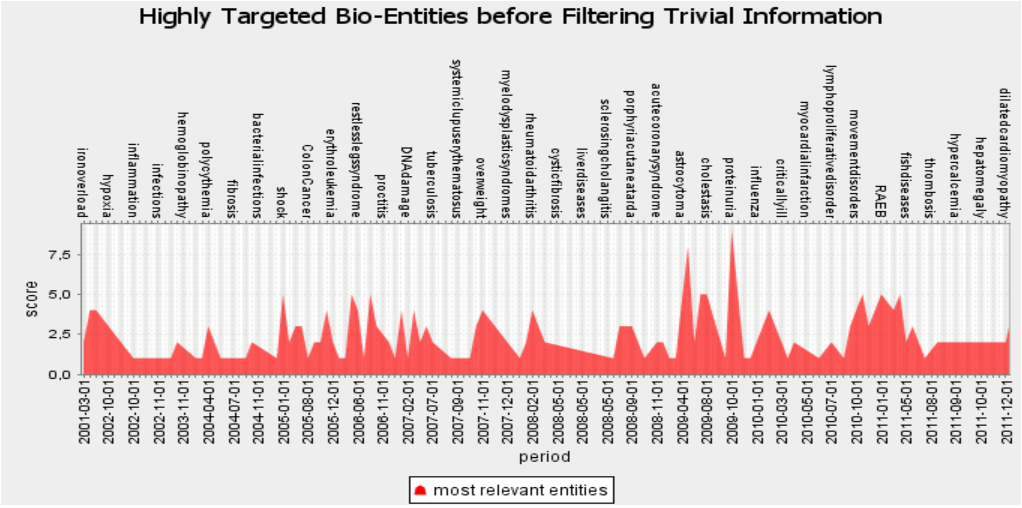

Figure 12. Highly targeted diseases when the background is excluded

In addition to proteins and diseases, other entity types, such 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. The majority of hepcidin studies are devoted to mice, rat and human subjects. Filtering out this dominant knowledge enabled us to view unusual species in which hepcidin has been found to emerge, such as fish species (*Sole* and *Rainbow trout*), bacteria, and diverse mammals, including bovines and Cynomolgus monkeys.
Several combinations of entity types have also been studied for the whole decade and associated information has been filtered and exploited by our collaborators. To allow the bioinvestigator to have more information on the revealed entities, an access to data sources and PubMed_ids is proposed in a supplementary annotation table (Figure 14.). For each period we have access to:

- *Name* of the relevant entity as identified in the abstracts. Spotted entities, such as "*human*", "*patient*" or "*child,*" are equivalent and unified to "*Human*" specie in ali-baba.
- Maximal number of target entities, along with their official names.
- Number of abstracts in which relative events have been found. The user can qualify time relevance using this number as relevance increases with ever-increasing number of distinct abstracts. The user also has direct access to these abstracts in PubMed by a simple click.
- Accession numbers of the relevant entities giving full access to their description in standard databases (e.g., DrugBank for drugs, Swissprot for proteins, etc.)

4.3.3 Comparative study of different backgrounds

This study may be applied for other use-cases given their bio-events extracted with a dedicated tool. We have been able to provide a comparative study of different PubMed queries using ali-baba web service.

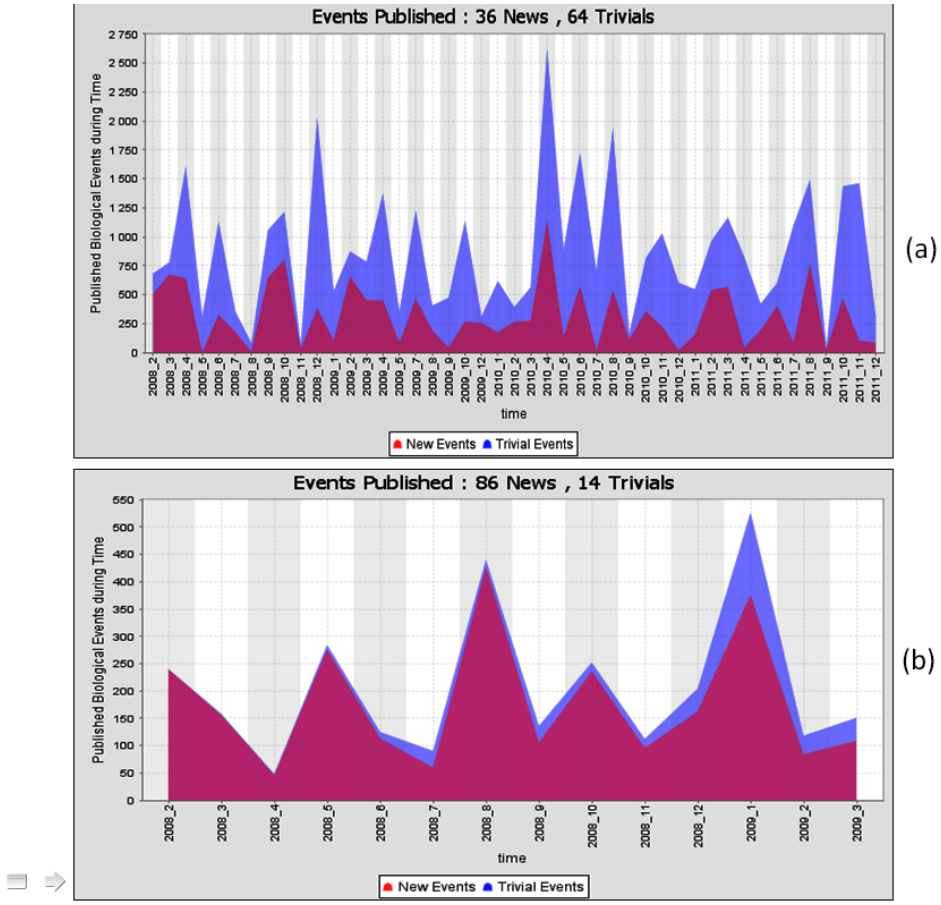

Figure 13. Comparative study of extracted events for queries "BMP6" (as a gene) and "Osteoporosis" (as a disease).

The calculation of the proportion of background 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 14.a, Figure 14.b), achieving 64% of the total volume of events for query "BMP6" and only 14% for the query "Osteoporosis", suggesting that the background knowledge is probably much more important in abstracts for some queries than for others.

This approach has also been tested to more popular queries, such as those concentrating massive bio-molecular and medical research, including "Breast Cancer". The relative literature on this domain is rapidly increasing and may discourage scientists wishing to be rapidly informed.

Given a reasonable duration for data generation (pre-processing), the results gave rapid priority access to highly cited proteins, drugs, cell types and diseases linked to breast cancer over time, along with their target entities correctly annotated in the supplementary annotation table. In Figure 14., a subset of highly published drugs over time linked to breast cancer, along with their targets, is browsed and richly annotated.

otations of Time Relevant Bio-Entities

Period T...	Source Entity	Max N...	The Targeted Entities	#Abst...	Source Entity in Databanks
2009-01-01	erlotinib	6	BreastCancer,Breastcancer,breast,breast...	1	http://www.drugbank.ca/drug...
2009-01-01	gefitinib	6	BreastCancer,Breastcancer,breast,breast...	1	http://www.drugbank.ca/drug...
2009-01-01	cetuximab	6	BreastCancer,Breastcancer,breast,breast...	1	http://www.drugbank.ca/drug...
2009-01-01	trastuzumab	6	BreastCancer,Breastcancer,breast,breast...	1	http://www.drugbank.ca/drug...
2009-02-01	hormones	3	breast,breastcancer,breasttumors,NHC,EPO	2	http://www.drugbank.ca/drug...
2009-03-01	SERMs	3	BreastCancer,Breastcancer,breastcancer,...	1	http://www.drugbank.ca/drug...
2009-04-01	celecoxib	3	cyclooxygenase-2,histonedeacetylase,VEGF	1	http://www.drugbank.ca/drug...
2009-04-01	rosiglitazone	3	cyclooxygenase-2,histonedeacetylase,VEGF	1	http://www.drugbank.ca/drug...
2009-05-01	cascades	7	BreastCancer,Breastcancer,breast,breast...	2	http://www.drugbank.ca/drug...
2009-06-01	adenosine	4	alkalinephosphatase,nucleosidediphospha...	1	http://www.drugbank.ca/drug...
2009-07-01	prostacyclin	4	lungcancer,cyclooxygenase-2,PGI,PGE	1	http://www.drugbank.ca/drug...
2009-08-01	docetaxel	2	BreastCancer,Breastcancer,breast,breast...	1	http://www.drugbank.ca/drug...
2009-08-01	cholesterol	2	C-reactiveprotein,CRP,high-densitylipopro...	1	http://www.drugbank.ca/drug...
2009-08-01	methylamine	2	IGF,MG115	1	http://www.drugbank.ca/drug...
2009-09-01	cytokeratin	4	myosinheavy-chain,estrogenreceptor,P63...	1	http://www.drugbank.ca/drug...
2009-10-01	vitaminD	5	BC,Breast,breast,breastcancer,breastcan...	1	http://www.drugbank.ca/drug...
2009-10-01	ethanol	5	Neu,estrogenreceptor,adenylatecyclase,E...	1	http://www.drugbank.ca/drug...
2009-10-01	androgen	5	osteopenia,BC,Breast,breast,breastcance...	1	http://www.drugbank.ca/drug...
2009-11-01	Luminal	6	Acutephase-response,BCa,Breastcancer,...	1	http://www.drugbank.ca/drug...
2009-12-01	antiestrogen	2	Breastcancer,Breastcancers,breast,breas...	1	http://www.drugbank.ca/drug...
2010-01-01	ribose	3	PARP,BRCA1,EGFR,epidermalgrowthfacto...	2	http://www.drugbank.ca/drug...
2010-02-01	BCNU	5	BCs,BreastCancer,Breastcancer,breast,br...	1	http://www.drugbank.ca/drug...
2010-03-01	genistein	4	boneresorption,Breast,BreastCancer,Brea...	1	http://www.drugbank.ca/drug...
2010-04-01	collagen	3	actin,desmin,vimentin	1	http://www.drugbank.ca/drug...
2010-06-01	progesterone	5	Breast,BreastCancer,Breastcancer,breast...	1	http://www.drugbank.ca/drug...
2010-06-01	2ME	5	Breast,BreastCancer,Breastcancer,breast...	1	http://www.drugbank.ca/drug...
2010-07-01	proteasome	4	ubiquitin,PARP,epidermalgrowthfactorrec...	1	http://www.drugbank.ca/drug...
2010-08-01	doxorubicin	5	BreastCancer,Breastcancer,breast,breast...	1	http://www.drugbank.ca/drug...
2010-09-01	thalidomide	2	plasmacytoma,diseaseprogression	1	http://www.drugbank.ca/drug...
2010-10-01	Tamoxifen	2	Carcinoma,carcinoma,ER+,estrogenrecep...	1	http://www.drugbank.ca/drug...

Figure 14. Annotation of highly targeted drugs linked to "Breast Cancer" from 2009/01 to 2013/12.

5 Conclusion and Discussion

The application of advanced information retrieval and extraction tools at a large scale in the biomedical literature has confirmed the extraordinary amount of published articles giving birth periodically to new biological entities and events in Medline. This torrential amount of data, once extracted, must be mined in a second round to draw unexpected patterns of biological entity behaviors over time. RetroMine approach is straightforward but is extremely helpful for providing in-depth retrospective studies at a glance to researchers on subjects of their interest.

Time has been bound to the process of information extraction to enhance the comprehension of the extracted events by introducing chronology and giving priority to biological entities highly targeted over time. Based on co-occurrence events, the study revealed the considerable increasing amount of background information published periodically in the biomedical literature that is necessary to filter out. While these events are essential for a non-expert wishing to learn usual contextual data about a domain, they may considerably pollute pertinent knowledge for domain experts. Excluding background hepcidin knowledge from the total events extracted has clearly allowed unexpected information to emerge to bioinvestigators. The user may prevent the assault of such articles by, for example, omitting review papers from the process of information extraction. Another possible solution is to focus on mining texts of section "Results" when available.

RetroMine approach may be helpful to accelerate tasks of database feeders, as the amount of publications is increasing exponentially. Database feeders continuously "read" newly published abstracts (or full text papers) 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 the Human Genome. Given a query, time-line reporting of the main relevant events published is visible in OMIM. Similarly chronology of events and their relevance over time are clearly perceptible using RetroMine. Today, slow feeding is

prominent to life science databases (Ex: the hepcidin gene is reported in OMIM using publications ending in early 2009!).
In this paper, we have used Web service of a previous study [3] in which events of "co-occurrence" are predominant. This Web service is not available any more. Nevertheless, we make available on retromine.univ-rennes1.fr the java application that allowed : i) data mining on the corpus of bio-events extracted from hepcidin publications during a decade and ii) the display of the different graphics resulting from this study.
Accuracy may decrease using RetroMine approach, but this is juxtaposed by the considerable gain in pre-processing time. Our objective in this paper was to demonstrate how to provide rapidly global patterns of relevant recognized bio-entities over time given a query. Improving event accuracy and complexity through more specialized methods, such as the EventMine of [22] and more devoted resources of [23] and [25], will certainly help to reduce the amount of background information published and diversify the spectrum of queries.
Current developments consist of the exploitation of retromine approach to provide in-depth retrospective studies on microbiology and at mid-term its generalisation to any use-case.

Acknowledgements

Ulf Leser and Astrid Rheinländer provided important helps for maintaining ali-baba continuously working at HU-Berlin during our study.
Members of "Iron and Liver" Group, at INSERM U991 provided valuable ideas to improve ongoing developments and outputs.

References

[1] L. J. Jensen, J. Saric, P. Bork. Literature mining for the biologist: from information retrieval to biological discovery, *Nature Reviews Genetics*, 7(2):119-129, 2006.
[2] C. Nédellec, R. Bossy, J. D. Kim, J. J. Kim, T. Ohta, S. Pyysalo, P. Zweigenbaum, "Overview of BioNLP shared task 2013." *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1-7, 2013.
[3] C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, and U. Leser. AliBaba: PubMed as a graph. *Bioinformatics*, 22(19):2444–2445, 2006.
[4] A. Krause, S. Neitz, H.-J. Mägert, A. Schulz, W.-G. Forssmann, P. Schulz-Knappe, and K. Adermann. LEAP-1, a novel highly disulfide-bonded human peptide, exhibits antimicrobial activity. *FEBS Letters*, 480(2):147–150, 2000.
[5] C. H. Park, E. V. Valore, A. J. Waring, and T. Ganz. Hepcidin, a urinary antimicrobial peptide synthesized in the liver. *Journal of Biological Chemistry*, 276(11):7806–7810, 2001.
[6] C. Pigeon, G. Ilyin, B. Courselaud, P. Leroyer, B. Turlin, P. Brissot, and O. Loréal. A new mouse liver-specific gene, encoding a protein homologous to human antimicrobial peptide hepcidin, is overexpressed during iron overload. *Journal of Biological Chemistry*, 276(11):7811–7819, 2001.
[7] N. Fatih, E. Camberlein, M. L. Island, A. Corlu, E. Abgueguen, L. Détivaud, P. Leroyer, P. Brissot, and O. Loréal. Natural and synthetic STAT3 inhibitors reduce hepcidin expression in differentiated mouse hepatocytes expressing the active phosphorylated STAT3 form. *Journal of Molecular Medicine*, 88(5):477–486, 2010.
[8] J. Hakenberg, M. Gerner, M. Haeussler, I. Solt, C. Plake, M. Schroeder, G. Gonzalez, G. Nenadic, and C. M. Bergman. The GNAT library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771, 2011.
[9] M. Huang, J. Liu, and X. Zhu. GeneTUKit: a software for document-level gene normalization. *Bioinformatics*, 27(7):1032–1033, 2011.

[10] M. Gerner, F. Sarafraz, C. M. Bergman, and G. Nenadic. BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, 28(16):2154–2161, 2012.
[11] J. W. Cooper and A. Kershenbaum. Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. *BMC Bioinformatics*, 6(1):143, 2005.
[12] I. Donaldson, J. Martin, B. De Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G. D. Bader, and K. Michalickova. PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4(1):11, 2003.
[13] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4, 2008.
[14] M. Krallinger, A. Valencia, and L. Hirschman. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biology*, 9(Suppl 2):S8, 2008.
[15] M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker. Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics*, 21(suppl 1):i319–i327, 2005.
[16] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. EDGAR: extraction of drugs, genes and relations from the biomedical literature. in *Pac Symp Biocomput*, 2000, 5:514–25.
[17] G. A. Palidwor and M. A. Andrade-Navarro. MLTrends: Graphing MEDLINE term usage over time. *Journal of Biomedical Discovery and Collaboration*, 5:1, 2010.
[18] E. Nemeth, M. S. Tuttle, J. Powelson, M. B. Vaughn, A. Donovan, D. M. Ward, T. Ganz, and J. Kaplan. Hepcidin regulates cellular iron efflux by binding to ferroportin and inducing its internalization. *Science*, 306(5704):2090–2093, 2004.
[19] L. Kautz, D. Meynard, A. Monnier, V. Darnaud, R. Bouvet, R.-H. Wang, C. Deng, S. Vaulont, J. Mosser, and H. Coppin. Iron regulates phosphorylation of Smad1/5/8 and gene expression of Bmp6, Smad7, Id1, and Atoh8 in the mouse liver. *Blood*, 112(4):1503–1509, 2008.
[20] H. N. Hunter, D. B. Fulton, T. Ganz, and H. J. Vogel. The solution structure of human hepcidin, a peptide hormone with antimicrobial activity that is involved in iron uptake and hereditary hemochromatosis. *Journal of Biological Chemistry*, 277(40):37597–37603, 2002.
[21] L. Détivaud, M.-L. Island, A.-M. Jouanolle, M. Ropert, E. Bardou-Jacquet, C. Le Lan, A. Mosser, P. Leroyer, Y. Deugnier, and V. David. Ferroportin diseases: functional studies, a link between genetic and clinical phenotype. *Human Mutation*, 34(11):1529–1536, 2013.
[22] M. Miwa, S. Pyysalo, T. Hara, and J. 'ichi Tsujii. Evaluating dependency representation for event extraction. in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, :779–787.
[23] S. Pyysalo, T. Ohta, M. Miwa, H.-C. Cho, J. 'ichi Tsujii, and S. Ananiadou. Event extraction across multiple levels of biological organization. *Bioinformatics*, 28(18):i575–i581, 2012.
[24] B. Kemper, T. Matsuzaki, Y. Matsuoka, Y. Tsuruoka, H. Kitano, S. Ananiadou, and J. 'ichi Tsujii. PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics*, 26(12):i374–i381, 2010.
[25] J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. Extracting Contextualized Complex Biological Events with rich graph-based Feature Sets. *Computational Intelligence*, 27(4):541–557, 2011.

[26] T. Abeel, S. Van Landeghem, R. Morante, V. Van Asch, Y. Van de Peer, W. Daelemans, and Y. Saeys. Highlights of the BioTM 2010 workshop on advances in bio text mining. *BMC Bioinformatics*, 11(Suppl 5):I1, 2010.
[27] Z. Lu. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011.
[28] E. K. Lee, H.-R. Lee, and A. Quarshie. SEACOIN–An investigative tool for biomedical informatics researchers. in *AMIA Annual Symposium Proceedings*, 2011, 2011:750.
[29] T. Ganz. Hepcidin and iron regulation, 10 years later. *Blood*, 117(17):4425–4433, 2011.
[30] P. Palaga, L. Nguyen, U. Leser, and J. Hakenberg. High-performance information extraction with alibaba. in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, 2009, :1140–1143.
[31] J. Hakenberg, C. Plake, L. Royer, H. Strobelt, U. Leser, and M. Schroeder. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology*, 9(Suppl 2):S14, 2008.