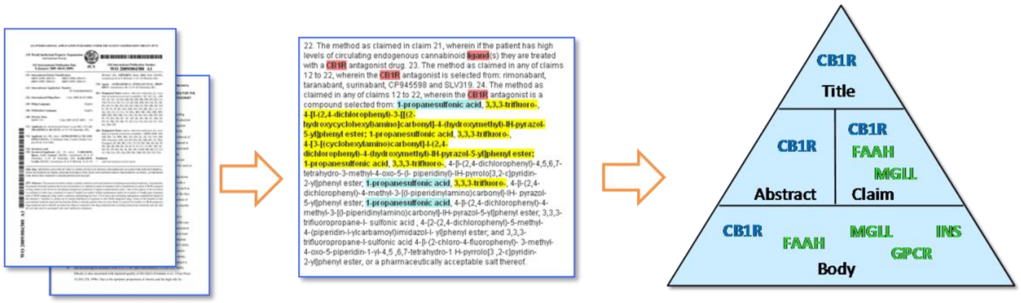


22. The method as claimed in claim 21, wherein if the patient has high levels of circulating endogenous cannabinoid **ligand**(s) they are treated with a **CB1R** antagonist drug. 23. The method as claimed in any of claims 12 to 22, wherein the **CB1R** antagonist is selected from: rimonabant, taranabant, surinabant, CP945598 and SLV319. 24. The method as claimed in any of claims 12 to 22, wherein the **CB1R** antagonist is a compound selected from: **1-propanesulfonic acid**, **3,3,3-trifluoro-, 4-[l-(2,4-dichlorophenyl)-3-[[(2-hydroxycyclohexyl)amino]carbonyl]-4-(hydroxymethyl)-lH-pyrazol-5-yl]phenyl ester; 1-propanesulfonic acid**, **3,3,3-trifluoro-, 4-[3-[(cyclohexylamino)carbonyl]-l-(2,4-dichlorophenyl)-4-(hydroxymethyl)-lH-pyrazol-5-yl]phenyl ester; 1-propanesulfonic acid**, **3,3,3-trifluoro-**, 4-[l-(2,4-dichlorophenyl)-4,5,6,7-tetrahydro-3-methyl-4-oxo-5-(l- piperidinyl)-lH-pyrrolo[3,2-c]pyridin-2-yl]phenyl ester; **1-propanesulfonic acid**, **3,3,3-trifluoro-**, 4-[l-(2,4-dichlorophenyl)-4-methyl-3-[(l-piperidinylamino)carbonyl]-lH- pyrazol-5-yl]phenyl ester; **1-propanesulfonic acid**, 4-[l-(2,4-dichlorophenyl)-4-methyl-3-[(l-piperidinylamino)carbonyl]-lH-pyrazol-5-yl]phenyl ester; 3,3,3-trifluoropropane-l- sulfonic acid , 4-[2-(2,4-dichlorophenyl)-5-methyl-4-(piperidin-l-ylcarbamoyl)imidazol-l- yl]phenyl ester; and 3,3,3-trifluoropropane-l-sulfonic acid 4-[l-(2-chloro-4-fluorophenyl)- 3-methyl-4-oxo-5-piperidin-1-yl-4,5 ,6,7-tetrahydro-1 H-pyrrolo[3 ,2-c]pyridin-2-yl]phenyl ester, or a pharmaceutically acceptable salt thereof.

The Identification of Target Proteins from Patents

Mining of biological entities from a full-text patent database

Master of Science Thesis in Bioinformatics and Systems Biology

ITHIPOL SURIYAWONGKUL

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
Göteborg, Sweden, June 2010

Identification of Target Proteins from Patents
Mining of biological entities from full-text patent database

ITHIPOL SURIYAWONGKUL

Examiner: GRAHAM J. L. KEMP

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

[Cover: Schematic representation of the extraction of target protein names from a patent document.]

Department of Computer Science and Engineering
Göteborg, Sweden June 2010

Abstract

Modern drug discovery involves identifying therapeutic targets (i.e. enzymes, receptors, and other proteins) that are relevant to a disease of interest, and searching for small-molecule compounds that are able to regulate these targets in order to cure or relieve the disease. Therefore, target-assay-compound relationships are key technical information driving modern drug discovery processes. As a consequence, every pharmaceutical R&D organization must seek for this information both from internal and external sources. One of the important information sources is patent documents which were disclosed for public access largely by pharmaceutical enterprises and academic institutions. While the target-assay-compound relationships have long been extracted from patents manually by expert curators, progress in text-mining shows possibilities to automate the extraction. Although the accuracy of human experts cannot yet be matched by automated text mining, access to a substantial corpus of full-text patents is clearly of complementary value for its scale, coverage, and speed.

This work presents results from investigating a full-text patent resource, that AstraZeneca has access to via a collaboration consortium arrangement with IBM. The work focused on exploring challenges and possible solutions for retrieval of target-containing patents and extraction of target protein names from individual patents. The work also serves as a pilot project in exploring and pioneering the mining of structure activity relationship (SAR) data from patents. Several information retrieval and extraction approaches were benchmarked against a source of SAR information manually extracted from patents which is a licensed product from GVKBIO. It was shown that target-containing patents could be retrieved by using keywords in titles with acceptable recall and precision. It was also shown that proximal co-occurrence between protein names and chemical modulation keywords could be used to identify target protein names in full-text patents with different recall and precision particular to patent sections. Assessments of false positives and false negatives from these approaches suggested that the extraction of target protein names required advance text mining techniques in order to interpret semantic context. Furthermore, it was also suggested that the writing practice of pharmaceutical patents is less scientific compared to biomedical journals. This leads to several issues in information retrieval and extraction specific to pharmaceutical patents. In summary, this work has successfully explored possible solutions for the retrieval of target-containing patents and the extraction of target protein names, and paints the way forward for SAR data mining from pharmaceutical patents.

Keywords. Biomedical text mining, Patent information

Acknowledgments

This project was carried out within the Computational Chemistry section (DECS-GCS Comp. Chem) at AstraZeneca Mölndal. The project was part of the early exploitation of a full-text patent data source that the section has access to via a collaboration agreement with the IBM Almaden Services Research (ASR).

I would like to express my appreciation to those who supported me to complete this master thesis. I am indebted to my supervisors, Sorel Muresan and Christopher Southan for overseeing my work and providing insights on various data sources, cheminformatics, and also practices in the pharmaceutical R&D. Many thanks, to Niklas Blomberg and Christian Tyrchan for reviewing my thesis paper and their kind support. Also many thanks to Graham Kemp from the Chalmers side for his supportive supervision and being the examiner for this project.

Further thanks to Plamen Petrov, Péter Várkonyi, Xie Paul for providing accesses to valuable data sources and helping me overcome technical difficulties throughout the project. Thanks to the Patent Analytics Platform team from the IBM Almaden Services Research for their collaboration. Also thanks to all staffs in the GCS Computational Chemistry section for warm welcome and knowledgeable discussions both formally and informally.

Lastly, I would like to dedicate this work to my family who always support my decision and encourage me to contribute to society while enjoying healthy work-life balance.

Contents

1 Introduction

This project has been carried out with the Computational Chemistry section within the Discovery Enabling Capabilities - Global Compound Sciences (DECS-GCS Comp. Chem.) in AstraZeneca Mölndal. This section provides tool development, databases, molecular modelling and cheminformatics support to the lead generation and computational chemistry communities across AstraZeneca Research and Development (AZ R&D).

Within a major global pharmaceutical company the crucial importance of patents is self-evident. The scope of their relevance is broad and encompasses not only small molecules but also genes, antibodies, assays, therapeutic uses, chemical synthesis methods, manufacturing processes, formulations, enantomeric separations, diagnostics and other areas. However, the crown-jewels of the AZ patent estate (as originator or licensee) are those filings that ensure commercial protection for both the product portfolio [1] and the development pipeline [2].

The therapeutic small-molecule (as opposed to the biological) component of the AZ patent portfolio falls within the scope of filings related to medicinal chemistry that claim novel chemical structures for the treatment of disease. The cumulative extent of this is very difficult to quantitate for some of the reasons that will be touched on in this thesis. However, very rough estimates suggest that in the order of 0.5 million distinct (i.e. not patent family related) medicinal chemistry patents, exemplifying around 5 million unique compound structures have been filed globally by approximately 2000 institutions including, but not restricted to, pharmaceutical companies over the last 40 years. The latter half of this time span has shown a significant acceleration in both primary data generation and patent filing rates. These have mainly been driven by developments in high-throughput screening, combinatorial chemistry, new drug targets from the human genome, the expanding contribution from the biotechnology sector and academic institutions as well as a big increase in filings from Asian countries. Annual statistics produced by the World Intellectual Property Organization (WIPO) shows that there are approximately 12,000 World-wide (WO) pharmaceutical patent applications per year.

The expertise of GCS-Comp. Chem. covers a broad spectrum of chemical bioactivity and drug target-related data exploitation with a focus on structure-activity-relationships (SAR). These support not only the strategic expansion and optimisation of the AZ compound collection for High Throughput Screening (HTS) but also for individual drug design project teams across R&D. While the internal work of the section is obviously proprietary, a small selection of recent journal publications provides an overview of their operating methodology (PubMed IDs 20298516, 19879759, 19434898, 19283339, 16711740)

The expertise has recently been broadened by the Knowledge Engineering Program for the comprehensive integration of internal and external bioactivity data. The external components come predominantly from three sources, medicinal chemistry journals, public assay data repositories (e.g. PubChem) and patent documents, with the latter being by far the largest.

The exploitation of patent data by GCS-Comp. Chem. is therefore a strategic priority because it not only scientifically delineates the most extensive SAR space but also explores the prior art competitive landscape that AZ has to circumvent for their own crucial inventive patents. Like other large pharmaceutical companies, AZ utilizes a spectrum of patent information sources.

The most direct are the major patent office portals such as EPO [3], USPTO [4] and WIPO [5] but these are largely limited to the basic indexing information. There are also increasing numbers of independent public indexing efforts that include some full-text extraction such as Google Patents [6], Free Patents Online [7] and PatentLens [8]. These are joined by the biotechnology patent indexing resources at the European Bioinformatics Institute such as CiteXplore [9] and Patent Abstracts [10]. Significantly, there is also a public resource from SureChem [11] that allows chemical searches of Free Patents Online for matches to chemical names that have been automatically converted to structures.

However, the prerequisite for exploitation at any significant scale is to be able to operate cheminformatically on sources where chemical structures from examples and/or claims have been extracted into a database format that, as a minimum, is searchable via chemical structure and provides direct links between the compound and the patent number. While AZ licences a number of on-line patent sources that facilitate limited searches of this type, there is a necessity to host these sources internally within GCS Comp.Chem. for integration and providing large-scale exploitation. Major patent sources are available from Thomson [12,13] and GVKBIO[14]. The value of these is well established (and detailed for GVKBIO later in this report) and substantially derived from the manual extraction of chemical structures identified by inspecting the documents and annotations such as targets and mechanism-of-action (MOA) that are not available from the basic indexing.

Despite this utility it was becoming clear that automated patent extraction was transitioning from proof of concept into a viable complement to manual curated sources. The technical drivers included a) increasing availability of full-text feeds from major patent offices, b) improvements in the recognition of chemical entities in text c) advances in conversion of chemical images to structures and d) technical progress in all aspects of biomedical text-mining and named entity recognition (NER). Thus, the strategic decision was made for GCS Comp.Chem. to enter a collaboration with IBM to access a full-text corpus of patents and automatically extracted chemical structures .The analysis and initial exploitation of this pioneering and unprecedented scale of data source is the subject of this thesis.

1.1 Patents as a data source

A patent is a legal document describing details of a new invention. As an incentive for disclosing details of the invention, the inventor is granted a patent right to exclude others from commercializing the same invention for a period of time – typically 20 years from the application date. To get the patent rights granted, the inventor must clearly describe how the invention can be reproduced by a skilled person. This makes them valuable documents for information mining. The format of patents is set by requirements from each patent office. They typically contain a number of sections as described in Table 1 [15]. Details of patent systems in the context of pharmaceutical and biotechnological industries are described by Webber [15].

Table 1. Anatomy of a patent [15].

Subsection of the patent	Function
Title and Abstract	Facilitates searching by the public and by patent offices.
Background description	Assists the understanding of the invention.
Statements of invention	Basis for claims/restrictions to claims.
Examples and Figures	How to put the invention into practice; support for the claims.
Claims	Explicitly define the scope of the invention.

Within a major pharmaceutical company such as AstraZeneca (AZ) the patent landscape is monitored across many areas relevant to R&D so it is necessary to define the sub-set of patents that this study was focused on. This set focuses on patents that exemplify novel chemical structures supported by activity data towards defined molecular targets *in vitro* together with supporting *in vivo* efficacy data for known diseases (or, alternatively, known compounds for new diseases). A lot of effort is invested to identify this type of patent as soon as possible after publication, including licensing multiple commercial databases. Patents become the subject of intense scrutiny where their content intersects at some level with internal drug discovery projects, disease area interests or marketed products. They are thus a key information source in the pharmaceutical industry not only for competitive intelligence *per se* but also for target and compound related data integration [16].

In principle, a patent document should provide sufficient detail of information of how to reproduce a particular invention as part of the patent deal in exchange for patent rights to the inventor. While there are relative advantages and disadvantages among different technical information carriers (i.e. patents, publications, peoples, products/processes), patents provide public access, codification of knowledge, and depth of coverage [17]. In addition, compared to peer-reviewed journals, patents have the relative advantage of timeliness of information, particularly in certain fields in which new advances are disclosed in patents long before they are published in peer-reviewed journals [18].

Nevertheless, compared to peer-reviewed journals, patents also have relative disadvantages on information adequacy due to reasons including:

1) the competitive incentive of patentees to minimize or obfuscate the information they disclosed [17]
2) lack of theoretical discussion of why the invention works [18]
3) uses of non-standard nomenclature and newly-coined technical terms [18]
4) using of ‘paper examples’ which were never actually carried out [18]
5) embedding the description and results of the performed experiments within 100’s of pages of irrelevant text
6) reporting only ranged, binned or qualitative results rather than quantitative data
7) use of the “shotgun” approach to claiming many potential diseases for treatment
8) using Markush-type specifications for large classes of compounds

These challenges are exemplified by the fact that it is not uncommon for drug discovery project teams to spend up to several man-days combing out the details from a large patent, particularly for enumerating chemical structures, much longer that required for a medicinal chemistry journal paper.

Although there are inadequacies and delay in patent information due to patent filing practices, patents remain key information source for competitive intelligence activities in the pharmaceutical industry [16]. As a consequence, patents are an important public information source for extracting the compound-to-protein relationships.

1.2 The IBM patent database

Towards the end of 2009, AstraZeneca entered into a collaboration consortium agreement with the IBM Almaden Research Center to access and exploit the IBM full-text patent documents and chemical structure extraction resource. This included access to their Strategic Information Mining PLatform for IP Excellence (SIMPLE) [19]. SIMPLE consists of a set of analytic tools and two underlying interconnected data warehouses - a patent data warehouse and a Medline data warehouse. The patent data warehouse contains full-text documents extracted and transformed from three patent offices including USPTO, EPO and WIPO. These data warehouses are structured into a snowflake schema [20] facilitating efficient business-intelligence-type aggregation and reporting [19]. Within the collaboration, AZ has access to these data warehouses for further in-house processing. The system architecture and upstream processing is illustrated in Figure 1. A full-text patent document is structured into four main sections which are 1) title, 2) abstract, 3) claim section and 4) description. They also includes associate information such as patent number, application filing date, publication date, inventor, assignee and International Patent Classification classes (see Appendix A).

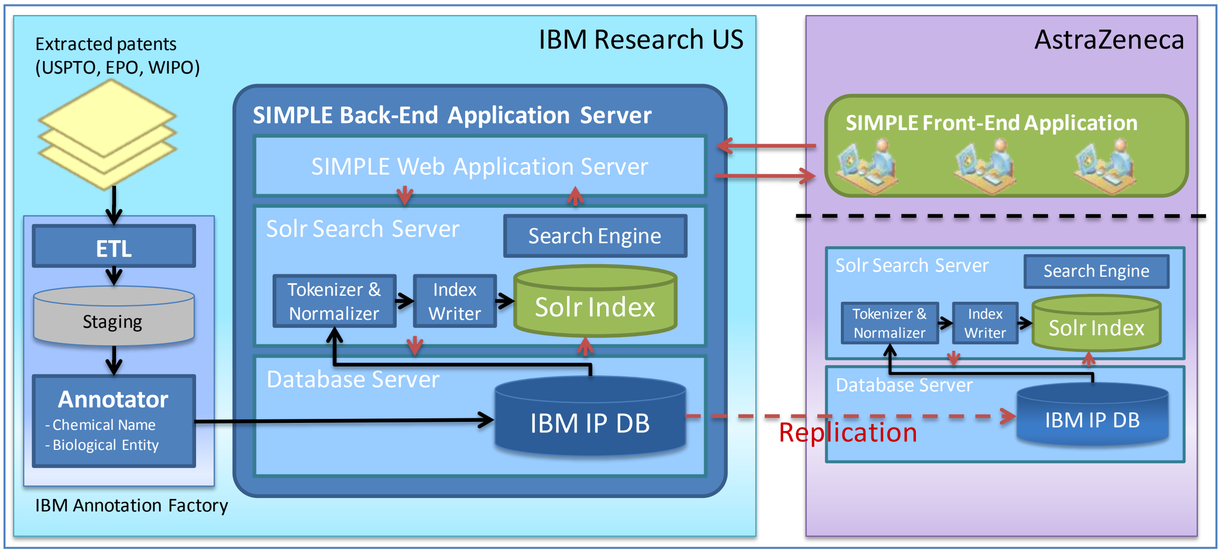


Figure 1. Simplified system architecture of the IBM patent database.

In addition to full-text search and analytic functionalities, SIMPLE also posses chemical and biological entity annotators that extract semantic entities such as chemical names, gene names and other biological terms. These extracted entities are then normalized into standardized terms using domain-specific approaches. All trivial and IUPAC systematic chemical names are converted into structures represented by SMILES codes[1][21] using the name=struct program® from CambridgeSoft Corporation [22], resulting in mapping of different synonyms of a chemical to the same structure. Gene synonyms are converted into their standardized symbols using a dictionary-lookup approach[23]. Semantic entities obtained from these annotators are stored back into the data warehouses for further search and analysis [24]. However, during the period of this study, only chemical annotation results were available within AstraZeneca, while biological annotation results (e.g. gene names) were not available as such.

As of February, 2010, the AZ instance of the IBM patent database contained ~11.1 million patent documents, with ~2.2 million including claimed chemical structures. Of these patents, ~1.5 million unique SMILES codes were included in their claim sections. The data source will henceforth be referred to as “IBM”.

1.3 The GVKBIO patent database

AstraZeneca licenses a number of drug discovery related databases from GVKBIO [14,25]. The database relevant to this study is the GVKBIO target class inhibitor database which comprises compound-to-assay-to-target relationships manually extracted from both journals and patents (Figure 2). A subset of the GVKBIO target inhibitor database extracted from patents was used in this study; henceforth referred to simply as “GVKBIO”. It contained 43,085 patent documents as of February, 2010.

[1] SMILES (Simplified Molecular Input Line Entry System) is a line notation for entering and representing molecules and reactions. e.g. O=C=O (carbon dioxide), CCN(CC)CC (triethylamine). [21]

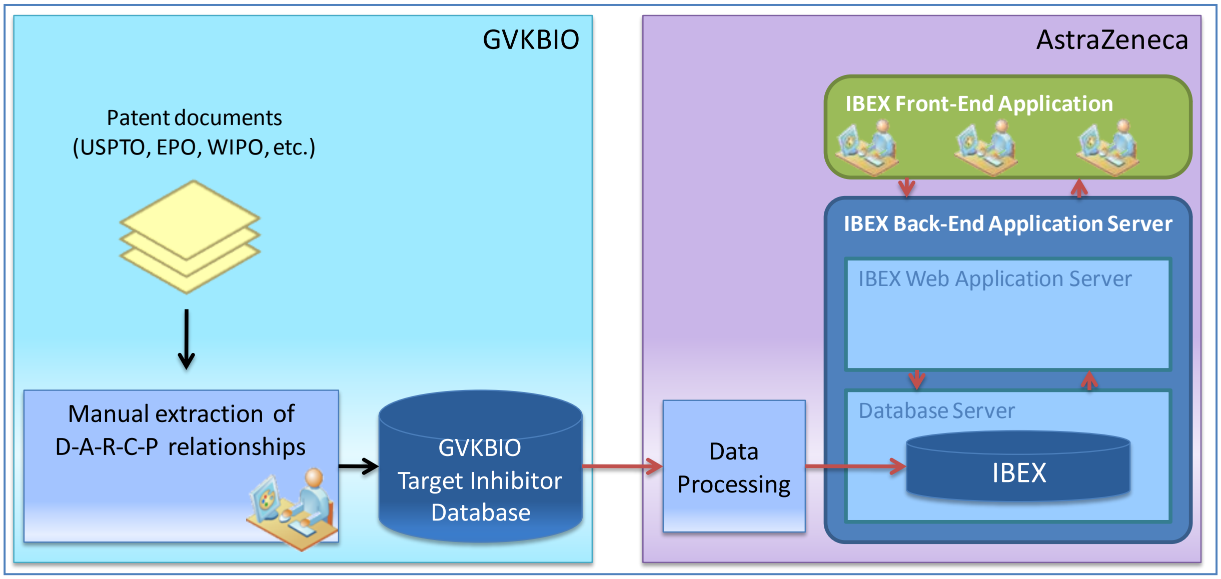


Figure 2. Simplified system architecture of the GVKBIO target inhibitor database.

In the process of compiling GVKBIO database, pharmaceutical patents were manually extracted by expert curators and organized into an entity-relationship data structure. This has five main entities that are document (D), assay description (A), assay result (R), compound structure (C) and target protein (P) (Figure 3). An example of relationship between these five entities is a document "D" describes a compound "C" as an inhibitor of a target protein "P", within a biochemical assay "A" with quantitative result "R". The resultant database from the collection of these D-A-R-C-P relationships is typically referred to as a large-scale structure activity relationship (SAR) database [26]. This database can be search by target protein, compound, and document, for example. More analytical searches could be "what compounds bind to a target protein P with an IC50 greater than R?".

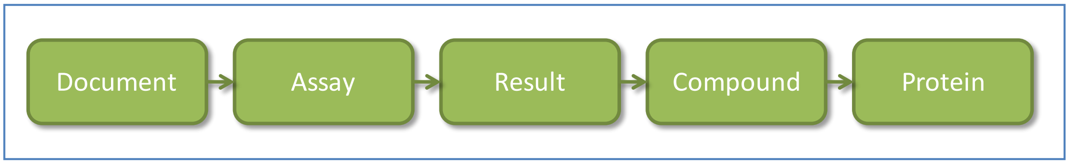


Figure 3. Illustration of the D-A-R-C-P relationship from [26].

1.4 The AZ-interest set of patents

In addition to IBM and GVKBIO, another small corpus of patents has also been utilized. The data source has been compiled internally over a number of years from many different sources including, but not restricted to, products licensed from SciFinder [27], Thomson Reuters Prous Integrity [12], and Investigational Drug Database (IDdb) [13]. This corpus will be subsequently referred to as "AZ-interest set". While this source includes at least some extracted chemical structures linked to patent numbers its main value in this work is the historically broad focus on the molecular target landscape. It thus provides a second corpus of target-enriched patents selected independently of the triage used by GVKBIO.

1.5 Purpose of the study

The ongoing initiative at AstraZeneca R&D is to automate extraction of compound-to-assay-to-protein relationships (D-A-R-C-P relationships) from pharmaceutical patents stored in the IBM patent database. Although the automate extraction is expected to provide lower accuracy than the manual extraction done by GVKBIO expert curators, the high-volumes of patent documents in the IBM database could be beneficial in terms of better coverage of pharmaceutical patents and a faster extraction process.

In order to get the compound-to-assay-to-protein relationship in a similar manner to GVKBIO, five entities and their relationships need to be extracted (section 1.3). Nevertheless, relevant to these five entities, SIMPLE only provides extraction of compound structures and protein names (assuming that extracted gene names infer their protein products). For example, compounds and protein names annotated by IBM could be non-drug-like compounds, non-target proteins, or semantically have no relationship to each other. Therefore, there are several operations to be done on the IBM database to achieve the goal of 1) retrieval of pharmaceutical patents (containing target chemical modulation data), 2) extraction of the five entities, and 3) extraction of the relationships between these entities. For instance, chemical compounds annotated within a pharmaceutical patent could be classified into drug-like compounds and non-drug-like compounds by using Lipinski's rule of five [28]. Similarly, protein names annotated could also be classified into target proteins and non-target proteins; dependent on the context of a particular pharmaceutical patent. Several techniques in text mining have been applied for mining chemical and biological information from biomedical literatures and reviewed by Banville, Krallinger et al., and Cohen et al. [29,30,31].

The aim of this study was to explore the possibilities to automate extraction of target protein names from full-text patents available in IBM. For this purpose, GVKBIO was used as a reference corpus to benchmark the results from the automatic extraction. Automatic target protein name extraction could be beneficial for setting alerts for new target proteins as soon as they are published in patents. In the longer term, the target protein names extracted can be used for establishing complete linkages of D-A-R-C-P relationships in subsequent projects at AstraZeneca.

In addition to extraction of target protein names, this work also served as a pilot project for assessing full-text processing challenges. During the study, several common problems were realized and overcome. These common problems included 1) errors resulting from conversion of PDF formatted documents to plain text, 2) special character handling, 3) chemical name and sentence boundary detection errors, and 4) incompleteness of available protein synonym resources, for example [32].

1.6 Problem statements of the study

As part of the early exploitation phase, this study aimed at exploring challenges and possible solutions for automatic target protein name extraction. Relevant resources available to AZ such as GVKBIO were utilized for this purpose with the main research questions stated as "How can target protein names and corresponding patents be identified and extracted from the IBM patent database?".

Aligned to the main research question, the following secondary research questions were addressed.

1) Was it possible to retrieve patents with chemical modulation data for target proteins from the IBM patent database? *(Information retrieval)*
2) Within each retrieved patent, was it possible to recognize protein names? *(Information extraction)*
3) From a list of protein names recognized in each patent, was it possible to identify target protein names? *(Information extraction)*

The secondary research questions were classified according to a general framework in text mining as described by Banville [29]. The information retrieval (IR) involves finding the most relevant set of documents. The information extraction (IE) involves extracting relevant information within the document identified by IR.

2 Methods

This work was divided into two phases. The first phase focused on exploration of data sources and possible solutions for the retrieval of target-containing patents and the extraction of target protein names. Patents were inspected both manually and computationally in this phase to identify and evaluate possible solutions. Several retrieval techniques were developed. The second phase focused on applying potential solutions identified in the first phase. The problem was broken down into 1) retrieval of target-containing patents and 2) extraction of target protein names from a target-containing patent.

2.1 Target terminology

It is necessary to define some terms to keep the subsequent descriptions concise. While the term "drug target" is well understood it can be ambiguous when used in the context of information retrieval and extraction from pharmaceutical patents. The term "*bona fide* target" will be used to refer to a primary target protein whose chemical modulation is proposed to be the therapeutic mechanism. For example, a patent may report IC50 data on novel chemical inhibitors for renin and claim these for the treatment of hypertension (i.e. the patent includes the D-A-R-C-P relationship where P is renin). Nevertheless, there may be a number of A-R-C-P relationships in the same document for other proteins. For example it is not uncommon for patents claiming renin inhibitors to report their cross-reactivity against other biochemically significant human aspartyl proteases such as Cathepsin D. The situation can be reversed where patents claiming the use of Cathepsin D inhibitors in Alzheimer's disease or cancer may include cross-screening data against renin. For clarity, the term "target" will be used to refer to any proteins whose chemical modulation data were reported in the patent (i.e. the A-R-C-P relationships for that protein). This is a pragmatic definition that encompasses protein complexes, multiple targets (polypharmacology), undefined targets and other subtle distinctions that cannot be detailed here. The term "non-target" will be used to refer to a protein specified in the patent without chemical modulation data.

2.2 Database queries

Both the IBM, GVKBIO and the AZ-interest set data sources are available on Oracle 10g databases and accessible by SQL (structured query language). IBM contains full-text description (i.e. patent number, title, abstract, claim section and description) of all patents published in USPTO, EPO and WIPO. GVKBIO contains target protein names and chemical structures manually tagged for a selected subset of pharmaceutical patents resulting high quality in patent-to-chemical-to-target relationships. AZ-interest set contains chemical structures extracted for selected patents resulting in patent-to-chemical relationships. Therefore, data for each patent document in these three databases can be linked logically by the patent number. For results shown in this article, all full-text patents were accessed from IBM. Some results were calibrated via target protein names curated by GVKBIO for the same patents, as indicated. Experiments shown in this report are categorized in Table 2. Detailed explanations of data selection and additional data sources are given in each result sections.

Table 2. Experimental objectives and data sources.

Result section			Description	Assessment areas	Data sources
3			Phase I - Results and discussion		
	3.1		Data Source Exploration	full-text documents, chemical name annotation, protein name annotation	IBM, GVKBIO, AZ-interest set
	3.2		Selection and compilation of protein name dictionary		UniProt, BioThesaurus, BioLexicon, HGNC
	3.3		Manual inspection of patents	protein name annotation by IBM, possible solution for target protein name identification	IBM, GVKBIO
	3.4		Protein names in titles		
		3.4.1	Manual identification of target protein names in titles	recall *(retrieval of patents with targets)*	Search corpus: GVKBIO
		3.4.2	Automatic identification of target protein names in titles	recall & precision *(retrieval of patents with targets)*	Search corpus: IBM Reference corpus: GVKBIO
	3.5		Protein names in titles, abstracts and claim sections		
		3.5.1	Assessment of improved recall by extending search to abstracts and claim sections	recall *(retrieval of patents with targets)*	Search corpus: GVKBIO
		3.5.2	Assessment of potential false positives from extending search to abstracts and claim sections	recall & precision *(retrieval of patents with targets)*	Search corpus: IBM Reference corpus: GVKBIO
	3.6		Selection and evaluation of filters	false positive reduction (predictive), remaining recall *(retrieval of patents with targets)*	IBM, GVKBIO
	3.7		Testing the use of selected filters	recall & precision *(retrieval of patents with targets)*	Search corpus: IBM Reference corpus: GVKBIO
4			Phase II - Results and discussion		
	4.1		Retrieval of target-containing patents		
		4.2.1	Selection and evaluation of keywords in titles specific to target-containing patents	false positive reduction (predictive), remaining recall *(retrieval of patents with targets)*	IBM, GVKBIO, a small corpus of patents without target proteins
		4.2.2	Testing the use of the data filtration pipeline	recall & precision *(retrieval of patents with targets)*	IBM, GVKBIO, a small corpus of patents without target proteins

Result section			Description	Assessment areas	Data sources
	4.2		Extraction of target protein names from full-text document	recall & precision *(retrieval of targets from a patent with targets)*	Search corpus: GVKBIO

3 Phase I - Results and discussions

3.1 Data source exploration

The aim of this section was to explore patent data sources available to the project, which were IBM, GVKBIO and AZ-interest set. Basic statistics and descriptive analysis which could be useful in subsequent analysis were obtained. These included overlapping between data sources; data quality in each data source; and general characteristics of chemicals and target proteins mentioned in each patent document.

3.1.1 Database content and intersections

As mentioned previously, this study is about information retrieval and extraction from the IBM with GVKBIO, a pharmaceutical-rich corpus, as a reference. Therefore, it is necessary to understand data volume and overlapping between these data sources. In addition to these, the AZ-interest set was also included in this exploration. Although, the AZ-interest set was not used much in subsequent analysis within the project, it was worth to explore at least to comparing chemical annotation between itself, GVKBIO and IBM as explained in section 3.1.3.3. Data content in each data source and overlapping between them were captured and assessed (on February 22, 2010).

Figure 4, the growth of the IBM over time, shows a continuous increase in patent publication per year from the three patent offices.

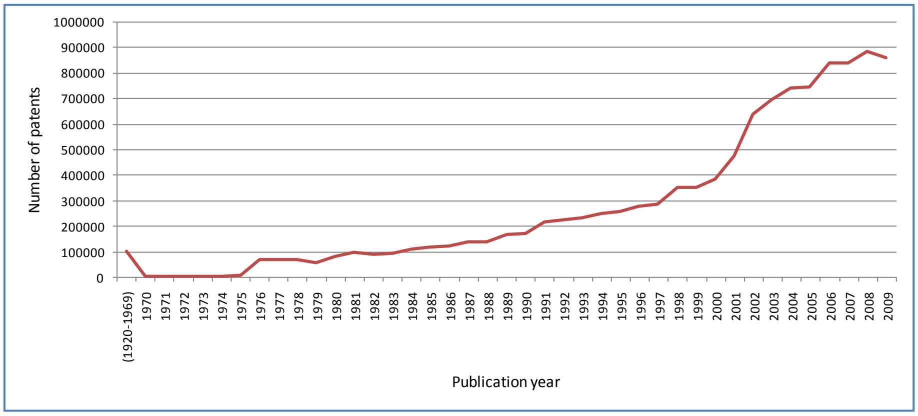


Figure 4. Volume of annual patents published over time for the whole IBM database (11,364,533 patents counted by patent number and KIND[2] code).

Nevertheless, from the whole IBM data source, there were only 1,782,181 patents in which were believed to contain pharmaceutical patents of interest. This set of pharmaceutical patents was selected by 19 pharmaceutical IPC codes which will be described later in Table 19 (i.e. {A61K,C07D,A61P,...,A61F}).

[2] KIND is a group of letter codes (e.g. A1, B1) used for distinguishing patent documents published by industrial property offices, as well as categorizing patent documents derived from the same patent application. For example, the patent number WO2009091542 was published twice with KIND codes A2 (first publication) and A9 (republication with some alterations) resulting WO2009091542A2 and WO2009091542A9.

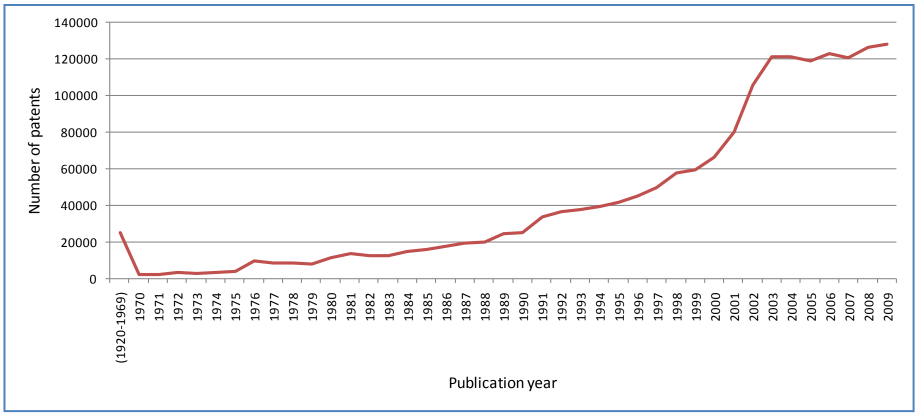


Figure 5. Volume of annual patents published over time for the pharmaceutical patents of interest (selected by IPC codes) in the IBM database (1,782,181 patents counted by patent number and KIND[2] code).

Table 3 shows volume of patents in each data source. Patent documents were counted by patent number without KIND[2] code (e.g. US20020012946, WO2004050024).

Table 3. Patent content in IBM, GVKBIO and AZ-interest set.

Patent country codes	Number of patents in each data source					
	IBM		GVKBIO		AZ-interest set	
US	6542923	(63.34%)	19853	(46.08%)	1640	(17.70%)
EP	2115160	(20.48%)	4511	(10.47%)	281	(3.03%)
WO	1670980	(16.18%)	18699	(43.40%)	7062	(76.21%)
JP	n/a		10	(0.02%)	119	(1.28%)
DE	n/a		8	(0.02%)	69	(0.74%)
FR	n/a		2	(0.00%)	45	(0.49%)
GB	n/a		2	(0.00%)	18	(0.19%)
KR,CN,LI,NL,AU,BE,CA,JA, CS,DD,HU,IL,LS,PR,SI,ZA	n/a		n/a		32	(0.35%)
Total	10329063	(100.00%)	43085	(100.00%)	9266	(100.00%)

Shown in Figure 6 is overlapping between the three data sources. Patent documents were counted and matched between data sources by patent number without KIND[2] code (e.g. US20020012946, WO2004050024). It shows that these three data sources collectively encompass 10,330,126 patents mainly from three major patent offices (i.e. USPTO, EPO and WIPO).

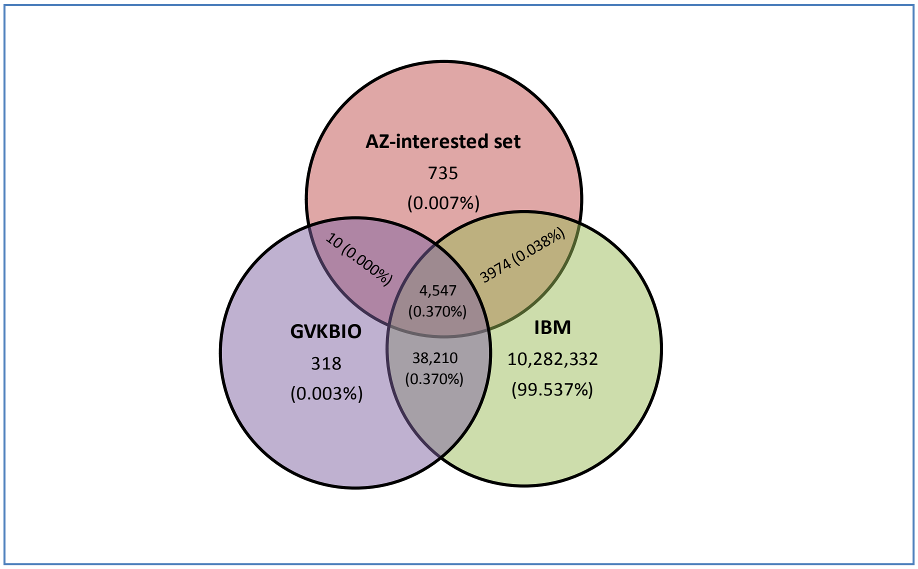


Figure 6. Venn diagram for coverage of patent documents in the three data sources (i.e. IBM, GVKBIO, and AZ-interest set). These three data sources collectively encompass 10,330,126 patent documents.

The overlap between IBM and GVKBIO (Figure 6) supports our assumption that the IBM covers almost all patents in GVKBIO (42,757 patents out of 43,085 patents in total). Investigating the 318 GVKBIO patents which were not in the IBM database shows that the major reason for this was a level of failure in upstream OCR processes (section 3.1.2). These 318 patents were later excluded from subsequent analysis where full-text documents were required.

Within the 42,757 patents-in-common between GVKBIO and IBM, there were 34,575 patents with human target proteins or their homologues in other species (the other 8182 patents are mostly antibiotic patents). The set of 34,575 patents with targets was used for calibration in this study as a corpus of patents with target proteins, and henceforth referred to as "GVKBIO patents with targets". Figure 7 shows the volume of annual patents published over time for this corpus.

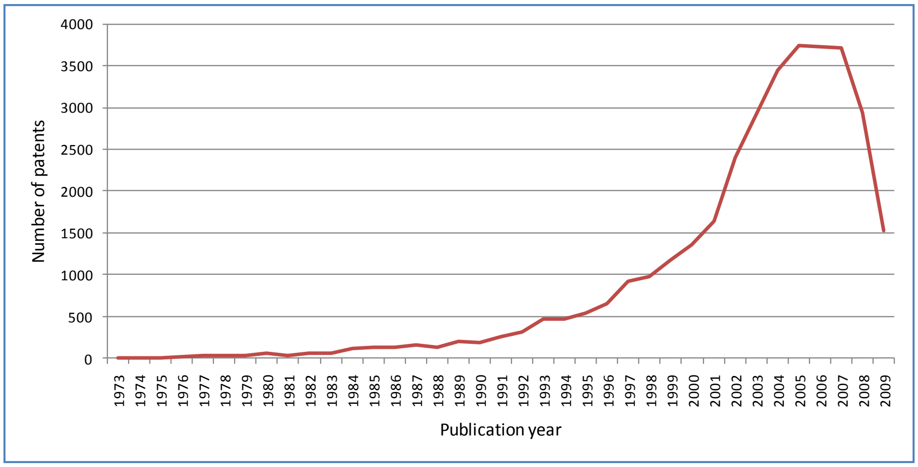


Figure 7. Volume of annual patents published over time for 34,575 GVKBIO patents with targets.

3.1.2 IBM patent document completeness

In order to get a full-text document stored in the IBM database, there are upstream processes for converting patent documents (containing text, tables, figures, etc.) from different formats (e.g. PDF, image archives, etc.) into XML files. Optical character recognition (OCR) technology is the main tool for this process. Errors in the OCR parsing for an entire document or some document sections were expected to some extent. The aim of this section was to investigate the level of this conversion problem. From each of the three patent offices, it was suggested by the IBM Almaden team that conversion issues are different. Each patent document was assessed for their completeness of having a title, abstract, abstract in English, claim section and body description. Results shown here were from a snapshot on April 29, 2010. For each patent office and application filing year in Figure 8 (absolute counts), Figure 9 (percentage per year), and summarized in Table 4.

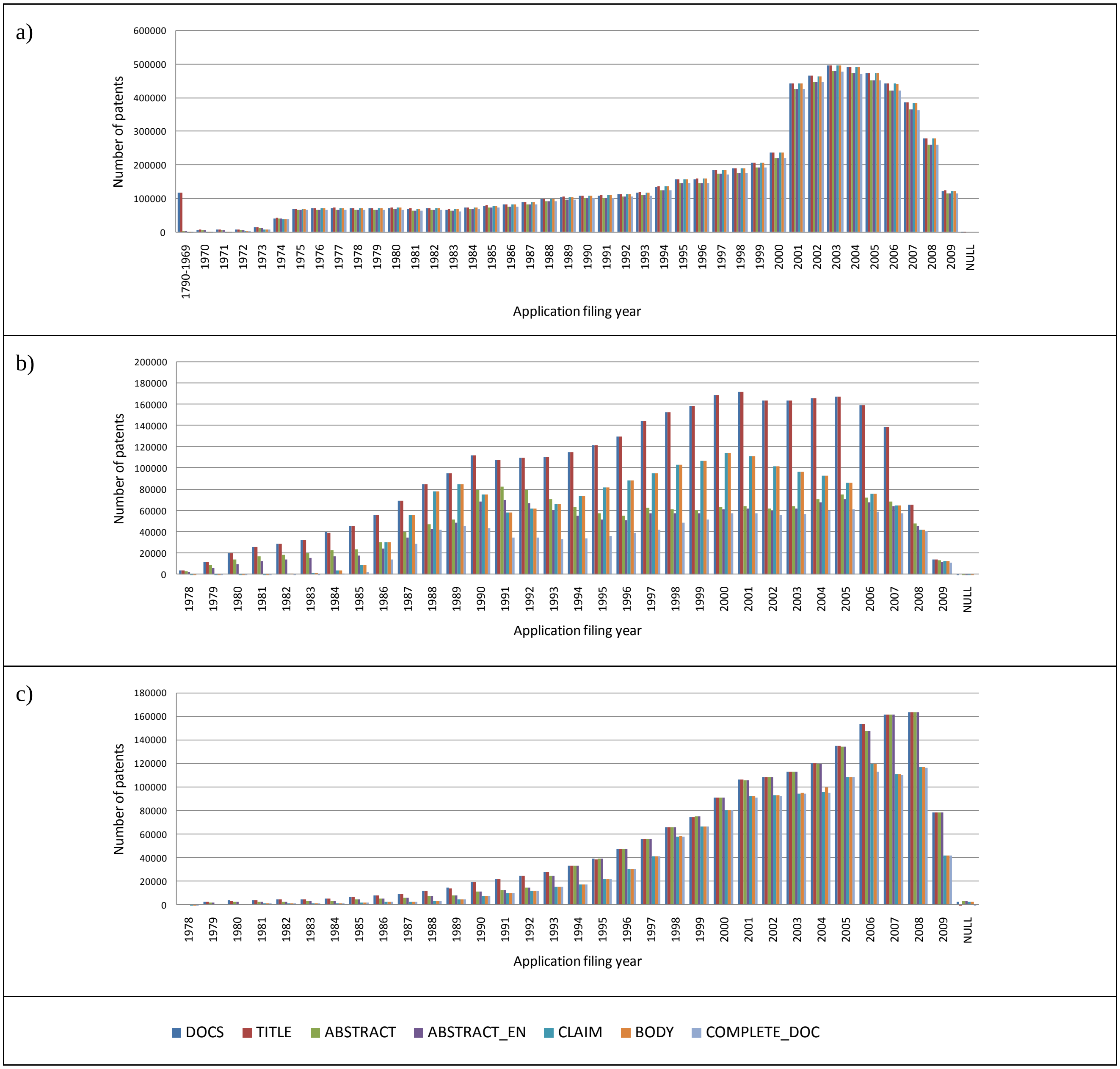


Figure 8. Assessment of patent document section completeness in IBM database (shown in absolute number of patents per year) for a) USPTO patents, b) EPO patents, c) WIPO patents.

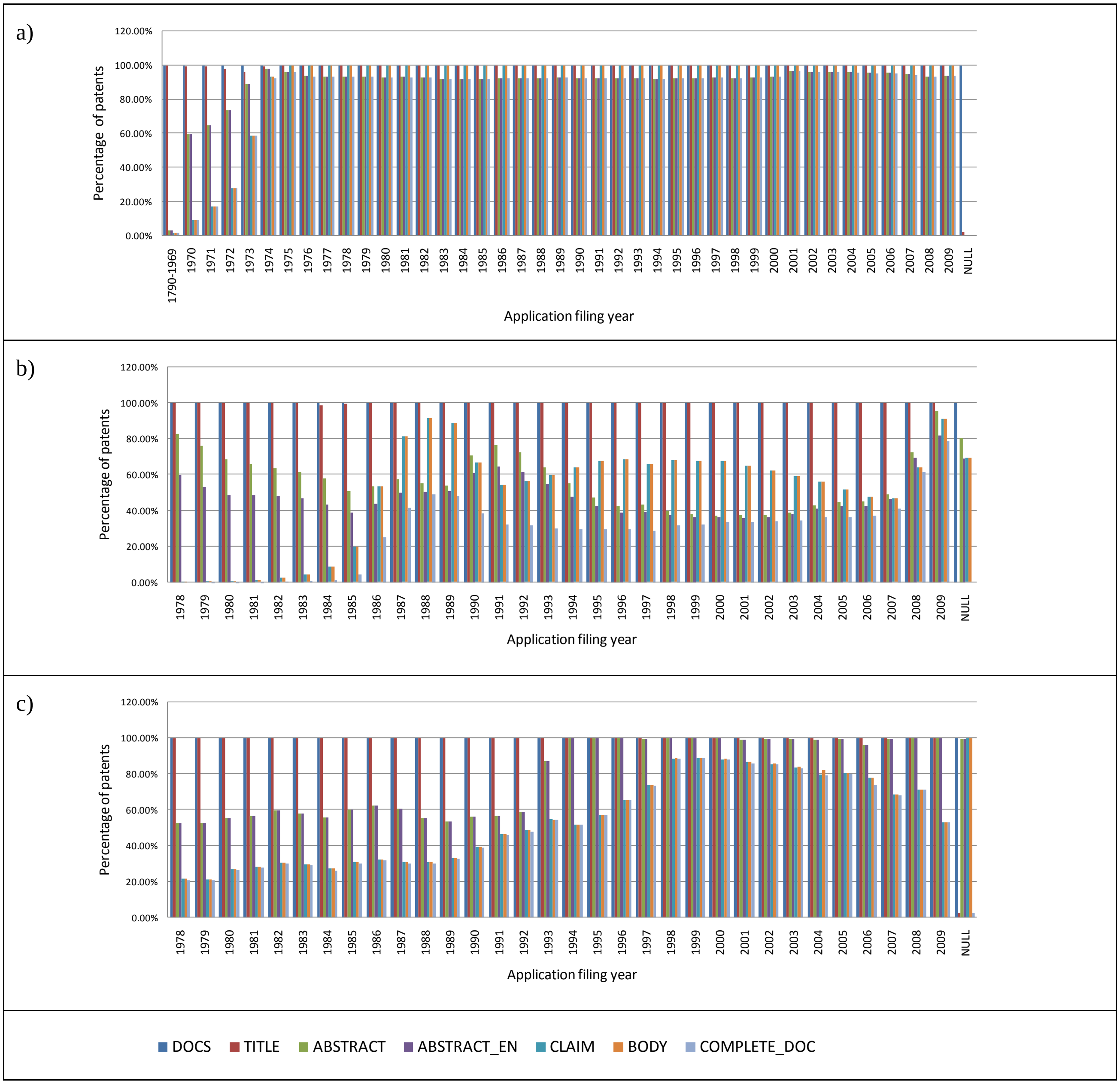


Figure 9. Assessment of patent document section completeness in IBM database (shown in percentage of patents per year) for a) USPTO patents, b) EPO patents, c) WIPO patents.

Table 4. Assessment of patent document completeness in IBM database for each patent office.

Document sections	Patent document completeness from each patent offices					
	USPTO		EPO		WIPO	
Total number of documents	6681105	(100.00%)	3159051	(100.00%)	1718005	(100.00%)
Documents with title	6678746	(99.96%)	3158278	(99.98%)	1715310	(99.84%)
Documents with abstract	6184500	(92.57%)	1572009	(49.76%)	1643704	(95.68%)
Documents with abstract in English	6184500	(92.57%)	1417095	(44.86%)	1643667	(95.67%)
Document with claim section	6530462	(97.75%)	1875546	(59.37%)	1260042	(73.34%)
Document with body section	6529380	(97.73%)	1875568	(59.37%)	1264355	(73.59%)
Document with all sections	6158601	(92.18%)	1047492	(33.16%)	1246315	(72.54%)

The results show that numbers of patents that could be transformed and loaded into the IBM database successfully with all sections (complete documents) were 92.18% for USPTO, 33.16% for EPO and 72.54% for WIPO. This shows that there were problems for upstream processing of EPO patents. As a consequence, not all patent documents were available as full-text. Subsets of complete patent documents were used in some experiments to ensure that results were not influenced by this form of data incompleteness.

3.1.3 Chemical annotation

Chemical structure extraction is one of the main reasons for AZ to enter into the collaboration with IBM. Hence, chemical annotation quality and their aggregate statistics were assessed as part of the project. In addition, extracted chemical structures could also be used to facilitate the identification of target protein names, for example, via co-occurrence of drug-like chemical names with target protein names within the same sentence.

3.1.3.1 IBM patent-chemical statistics

This section shows basic statistics relevant to annotated chemical structures in the IBM database as part of the exploration phase of the project, collected on February 5, 2010.

Document-centric statistics

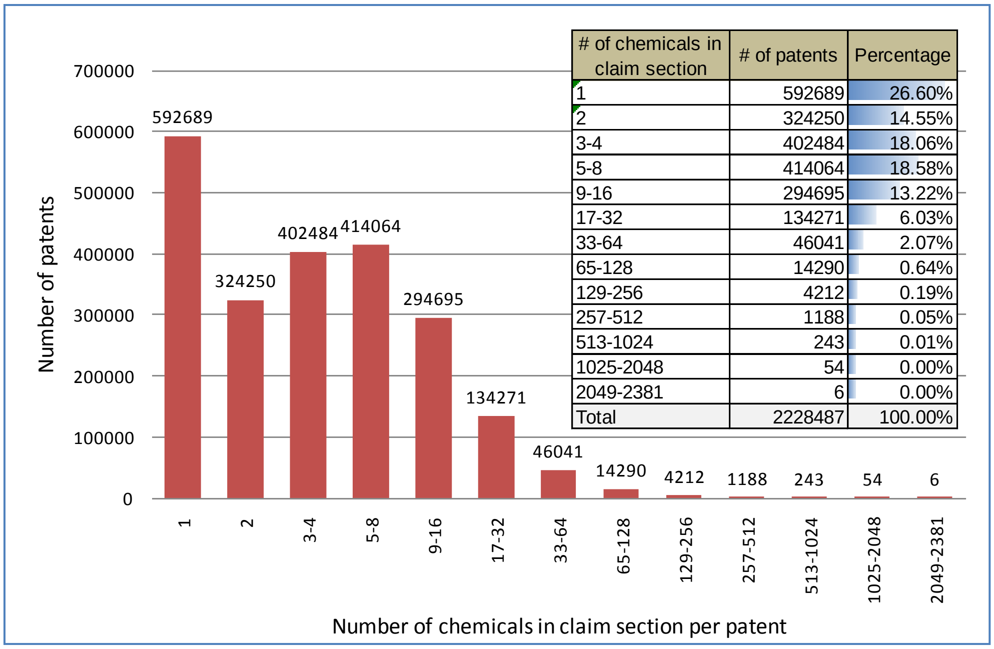
- 11,141,811 patents in total
- 4,352,347 patents with chemicals
- 2,228,487 patents with chemicals in claim sections

Chemical-centric statistics

- 144,886,117 chemical entries from all patents (avg. 33 chemical names per patent)
    - 16,278,062 chemical entries from claim sections (avg. 7 chemical names per claim section)
- 6,539,993 unique SMILES codes in all patents found in the following divisions:
    - 6,204,956 in body sections
    - 1,548,059 in claim sections
    - 96,377 in abstracts
    - 28,331 in titles

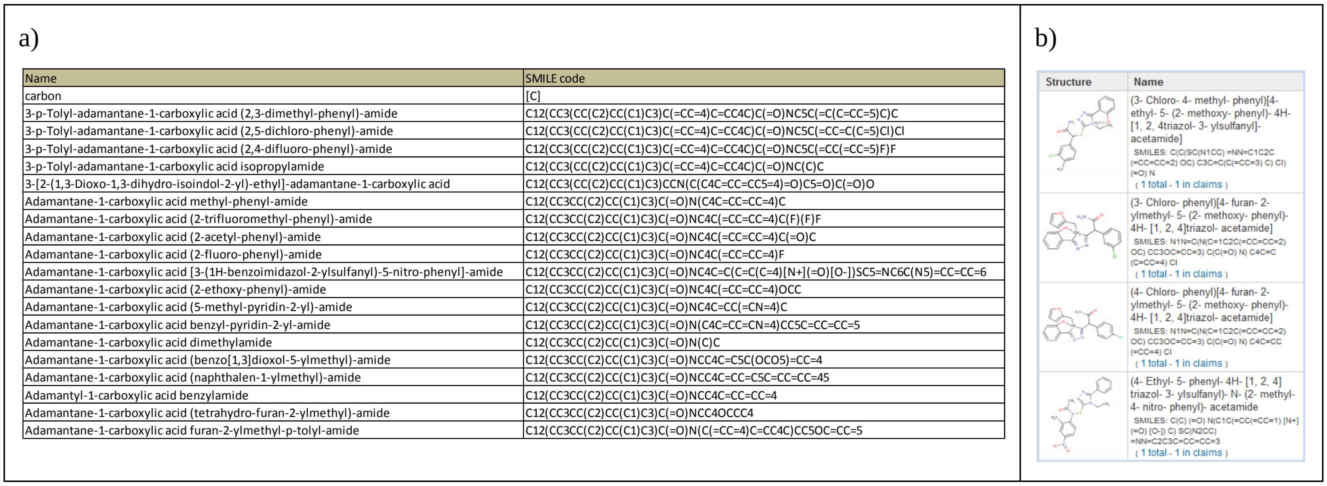
From the set of 2,228,487 patents with chemicals in claim sections, the number of chemicals in claim section for each patent was counted. Distribution of these numbers is shown in Figure 10. Of these patents, ~91% of them have 1-16 chemicals in claim section; and ~99% of them have 1-64 chemicals in claim section. Approximately only 1% of them have more than 64 chemicals in claim section. An extreme case is US20060111348[3], a pharmaceutical patent containing 2,221 chemicals in its claim section with only minor difference between them (Figure 11).

[3] US20060111348 “Combination therapy using an 11beta-hydroxysteroid dehydrogenase type 1 inhibitor and an antihypertensive agent for the treatment of metabolic syndrome and related diseases and disorders” (Assignee. Novo Nordisk A/S)



# of chemicals in claim section	# of patents	Percentage
1	592689	26.60%
2	324250	14.55%
3-4	402484	18.06%
5-8	414064	18.58%
9-16	294695	13.22%
17-32	134271	6.03%
33-64	46041	2.07%
65-128	14290	0.64%
129-256	4212	0.19%
257-512	1188	0.05%
513-1024	243	0.01%
1025-2048	54	0.00%
2049-2381	6	0.00%
Total	2228487	100.00%

Figure 10. Distribution of the number of chemicals in claim sections for each patent document.

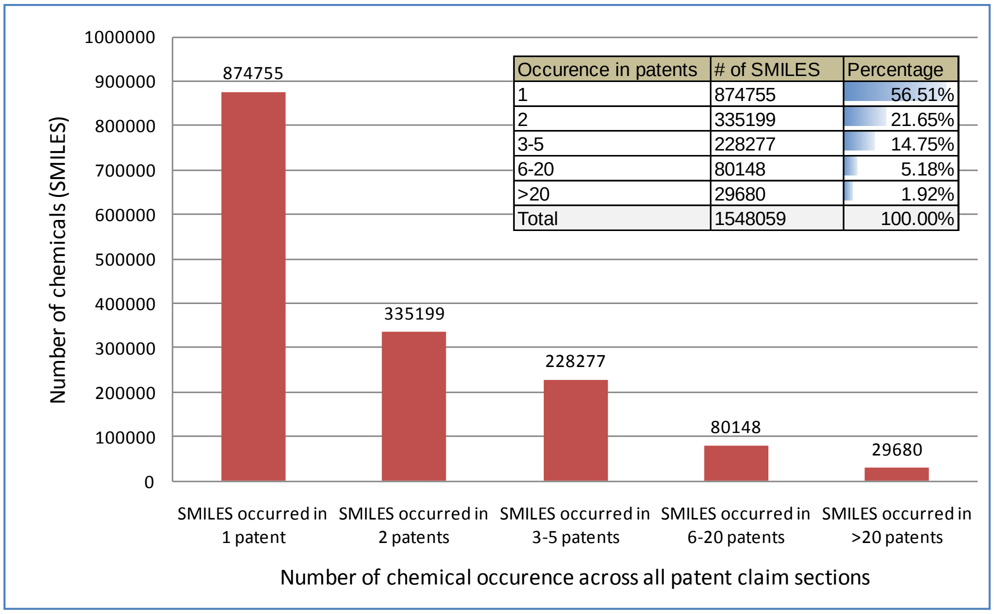


Name	SMILE code
carbon	[C]
3-p-Tolyl-adamantane-1-carboxylic acid (2,3-dimethyl-phenyl)-amide	C12(CC3(CC(C2)CC(C1)C3)C(=CC=4)C=CC4C)C(=O)NC5C(=C(C=CC=5)C)C
3-p-Tolyl-adamantane-1-carboxylic acid (2,5-dichloro-phenyl)-amide	C12(CC3(CC(C2)CC(C1)C3)C(=CC=4)C=CC4C)C(=O)NC5C(=CC=C(C=5)Cl)Cl
3-p-Tolyl-adamantane-1-carboxylic acid (2,4-difluoro-phenyl)-amide	C12(CC3(CC(C2)CC(C1)C3)C(=CC=4)C=CC4C)C(=O)NC5C(=CC(=CC=5)F)F
3-p-Tolyl-adamantane-1-carboxylic acid isopropylamide	C12(CC3(CC(C2)CC(C1)C3)C(=CC=4)C=CC4C)C(=O)NC(C)C
3-[2-(1,3-Dioxo-1,3-dihydro-isoindol-2-yl)-ethyl]-adamantane-1-carboxylic acid	C12(CC3(CC(C2)CC(C1)C3)CCN(C(C4C=CC=CC5=4)=O)C5=O)C(=O)O
Adamantane-1-carboxylic acid methyl-phenyl-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)N(C4C=CC=CC=4)C
Adamantane-1-carboxylic acid (2-trifluoromethyl-phenyl)-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NC4C(=CC=CC=4)C(F)(F)F
Adamantane-1-carboxylic acid (2-acetyl-phenyl)-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NC4C(=CC=CC=4)C(=O)C
Adamantane-1-carboxylic acid (2-fluoro-phenyl)-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NC4C(=CC=CC=4)F
Adamantane-1-carboxylic acid [3-(1H-benzoimidazol-2-ylsulfanyl)-5-nitro-phenyl]-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NC4C=C(C=C(C=4)[N+](=O)[O-])SC5=NC6C(N5)=CC=CC=6
Adamantane-1-carboxylic acid (2-ethoxy-phenyl)-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NC4C(=CC=CC=4)OCC
Adamantane-1-carboxylic acid (5-methyl-pyridin-2-yl)-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NC4C=CC(=CN=4)C
Adamantane-1-carboxylic acid benzyl-pyridin-2-yl-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)N(C4C=CC=CN=4)CC5C=CC=CC=5
Adamantane-1-carboxylic acid dimethylamide	C12(CC3CC(C2)CC(C1)C3)C(=O)N(C)C
Adamantane-1-carboxylic acid (benzo[1,3]dioxol-5-ylmethyl)-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NCC4C=C5C(OCO5)=CC=4
Adamantane-1-carboxylic acid (naphthalen-1-ylmethyl)-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NCC4C=CC=C5C=CC=CC=45
Adamantyl-1-carboxylic acid benzylamide	C12(CC3CC(C2)CC(C1)C3)C(=O)NCC4C=CC=CC=4
Adamantane-1-carboxylic acid (tetrahydro-furan-2-ylmethyl)-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)NCC4OCCC4
Adamantane-1-carboxylic acid furan-2-ylmethyl-p-tolyl-amide	C12(CC3CC(C2)CC(C1)C3)C(=O)N(C(=CC=4)C=CC4C)CC5OC=CC=5

Structure	Name
	(3- Chloro- 4- methyl- phenyl)[4- ethyl- 5- (2- methoxy- phenyl)- 4H- [1, 2, 4triazol- 3- ylsulfanyl]- acetamide] SMILES: C(C(SC(N1CC) =NN=C1C2C (=CC=CC=2) OC) C3C=C(C(=CC=3) C) Cl) (=O) N (1 total - 1 in claims)
	(3- Chloro- phenyl)[4- furan- 2- ylmethyl- 5- (2- methoxy- phenyl)- 4H- [1, 2, 4]triazol- acetamide] SMILES: N1N=C(N(C=1C2C(=CC=CC=2) OC) CC3OC=CC=3) C(C(=O) N) C4C=C (C=CC=4) Cl (1 total - 1 in claims)
	(4- Chloro- phenyl)[4- furan- 2- ylmethyl- 5- (2- methoxy- phenyl)- 4H- [1, 2, 4]triazol- acetamide] SMILES: N1N=C(N(C=1C2C(=CC=CC=2) OC) CC3OC=CC=3) C(C(=O) N) C4C=CC (=CC=4) Cl (1 total - 1 in claims)
	(4- Ethyl- 5- phenyl- 4H- [1, 2, 4] triazol- 3- ylsulfanyl)- N- (2- methyl- 4- nitro- phenyl)- acetamide SMILES: C(C) (=O) N(C1C(=CC(=CC=1) [N+] (=O) [O-]) C) SC(N2CC) =NN=C2C3C=CC=CC=3 (1 total - 1 in claims)

Figure 11. Example of chemicals annotated in the claim section of US20060111348 encompassing 2,221 chemical names in its claim section. a) Chemical names and their SMILES codes, b) Chemical structures.

From the set of 1,548,059 unique SMILES codes collected from all patent claim sections, the number of patents for a particular SMILES code was counted. Distribution of these numbers is shown in Figure 12. Of these SMILES codes, ~57% of them appeared in only one patent; and ~93% in 1-5 patents (approximately a patent family size). This supports the fact that a novel chemical structure should be claimed by only one patent family. On the other hand, ~7% of these unique SMILES codes were found across more than 6 patents. Extended queries show that these SMILES codes were claimed for their application, rather than its structure. Example of these SMILES codes were [Si], [Cu], [C], and [Al].



Occurence in patents	# of SMILES	Percentage
1	874755	56.51%
2	335199	21.65%
3-5	228277	14.75%
6-20	80148	5.18%
>20	29680	1.92%
Total	1548059	100.00%

Figure 12. Distribution of the number of patents in which a particular chemical mentioned in claim section.

3.1.3.2 IBM chemical annotation quality

During this study, there were some issues in the IBM automated chemical annotation that have to be considered when performing an analysis relevant to the annotation result. Although most of these issues found are common in text mining, it is worth to document the most important ones as listed below. (Examples shown here were captured from IBM on February, 2010).

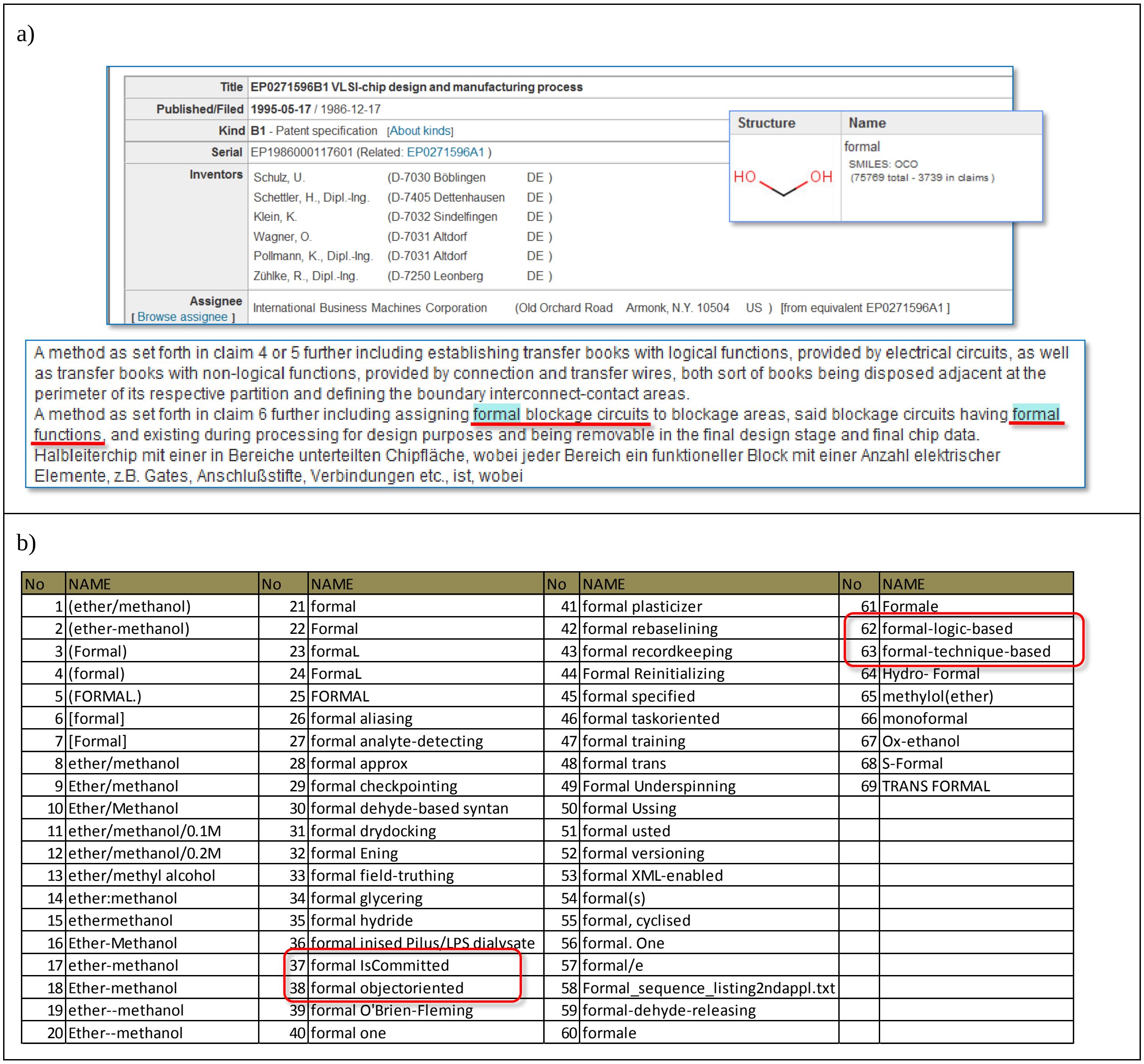
1. Non-chemical terms annotated as chemicals Shown in Figure 13 (a) is an electronics patent which was inaccurately annotated as having a chemical structure “OCO” (SMILES code) in claim section. Further investigation showed that there was a systematic error where every occurrences of the term “formal” was annotated as this chemical. Shown in Figure 13 (b) are terms that were annotated as this chemical. Some of these terms are clearly non-chemical terms. Collectively, the “OCO” was annotated in 74,511 patents, and 3,788 claim sections.

2. Chemical terms not annotated Shown in Figure 14 is a pharmaceutical patent containing generic chemical names for several polymer types that were not annotated. One of the reasons could be that these chemical names could not be converted into SMILES codes (name=struct®[22]). In addition, it was shown in in additional cases that the coverage of the chemical dictionary used by IBM to detect trivial names was incomplete.

3. Annotation of substructures but missed full structures Shown in Figure 15 is a pharmaceutical patent for Lipitor with two chemical structures in its abstract. However, the first was split into two substructures, instead of its full structure. This leads to distortion of the chemical structures extracted, and could result in inconsistent drug-to-structure links.

4. Duplicate chemical entries Shown in Figure 16 (a) is a pharmaceutical patent with 2,221 chemical names annotated in the claim section. However, there are only 2,208 unique structures as SMILES codes. The reason is that, for each patent document, the IBM database stores multiple chemical synonyms and text variants for the same SMILES codes (see Chemical table in Appendix A). This can lead to small difference

between the number of chemical names per patent and the number of chemical structures per patent. Figure 16 (b) shows the numbers of duplicate chemical entries per patent for each patent section.



a)

Title	**EP0271596B1 VLSI-chip design and manufacturing process**
Published/Filed	**1995-05-17** / 1986-12-17
Kind	**B1** - Patent specification [About kinds]
Serial	EP1986000117601 (Related: EP0271596A1)
Inventors	Schulz, U. (D-7030 Böblingen DE) Schettler, H., Dipl.-Ing. (D-7405 Dettenhausen DE) Klein, K. (D-7032 Sindelfingen DE) Wagner, O. (D-7031 Altdorf DE) Pollmann, K., Dipl.-Ing. (D-7031 Altdorf DE) Zühlke, R., Dipl.-Ing. (D-7250 Leonberg DE)
Assignee [Browse assignee]	International Business Machines Corporation (Old Orchard Road Armonk, N.Y. 10504 US) [from equivalent EP0271596A1]

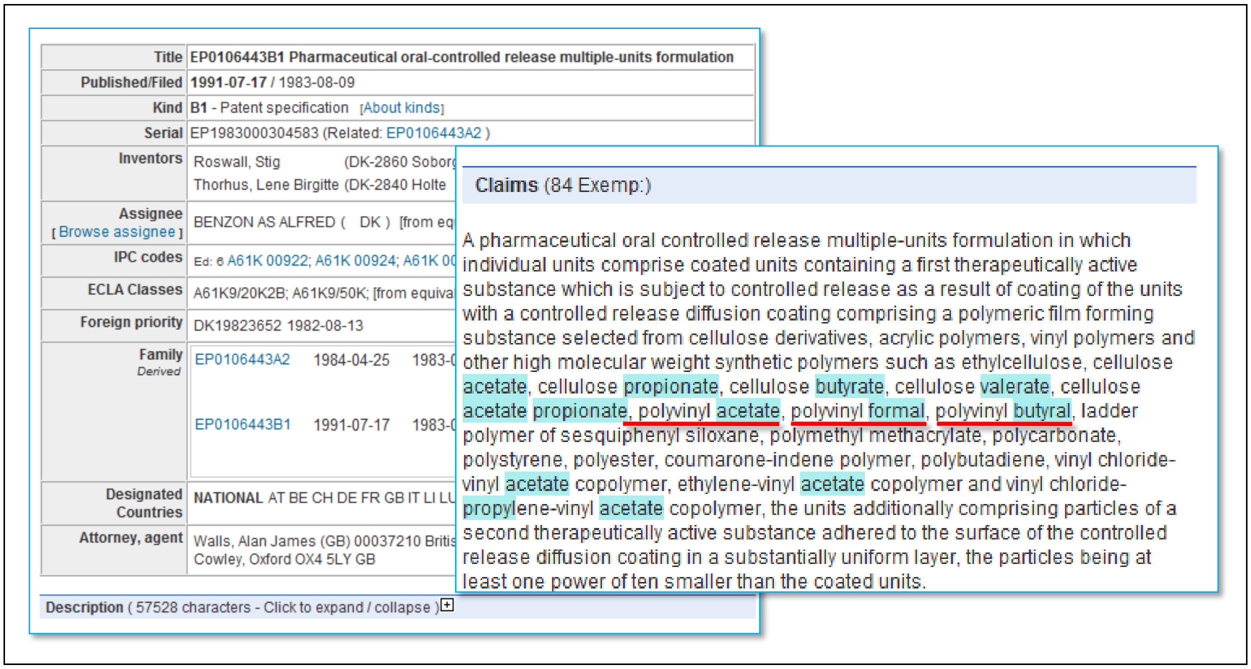
Structure	Name
	formal SMILES: OCO (75769 total - 3739 in claims)

A method as set forth in claim 4 or 5 further including establishing transfer books with logical functions, provided by electrical circuits, as well as transfer books with non-logical functions, provided by connection and transfer wires, both sort of books being disposed adjacent at the perimeter of its respective partition and defining the boundary interconnect-contact areas.
A method as set forth in claim 6 further including assigning formal blockage circuits to blockage areas, said blockage circuits having formal functions, and existing during processing for design purposes and being removable in the final design stage and final chip data.
Halbleiterchip mit einer in Bereiche unterteilten Chipfläche, wobei jeder Bereich ein funktioneller Block mit einer Anzahl elektrischer Elemente, z.B. Gates, Anschlußstifte, Verbindungen etc., ist, wobei

b)

No	NAME	No	NAME	No	NAME	No	NAME
1	(ether/methanol)	21	formal	41	formal plasticizer	61	Formale
2	(ether-methanol)	22	Formal	42	formal rebaselining	62	formal-logic-based
3	(Formal)	23	formaL	43	formal recordkeeping	63	formal-technique-based
4	(formal)	24	FormaL	44	Formal Reinitializing	64	Hydro- Formal
5	(FORMAL.)	25	FORMAL	45	formal specified	65	methylol(ether)
6	[formal]	26	formal aliasing	46	formal taskoriented	66	monoformal
7	[Formal]	27	formal analyte-detecting	47	formal training	67	Ox-ethanol
8	ether/methanol	28	formal approx	48	formal trans	68	S-Formal
9	Ether/methanol	29	formal checkpointing	49	Formal Underspinning	69	TRANS FORMAL
10	Ether/Methanol	30	formal dehyde-based syntan	50	formal Ussing		
11	ether/methanol/0.1M	31	formal drydocking	51	formal usted		
12	ether/methanol/0.2M	32	formal Ening	52	formal versioning		
13	ether/methyl alcohol	33	formal field-truthing	53	formal XML-enabled		
14	ether:methanol	34	formal glycering	54	formal(s)		
15	ethermethanol	35	formal hydride	55	formal, cyclised		
16	Ether-Methanol	36	formal inised Pilus/LPS dialysate	56	formal. One		
17	ether-methanol	37	formal IsCommitted	57	formal/e		
18	Ether-methanol	38	formal objectoriented	58	Formal_sequence_listing2ndappl.txt		
19	ether--methanol	39	formal O'Brien-Fleming	59	formal-dehyde-releasing		
20	Ether--methanol	40	formal one	60	formale		

Figure 13. A case of non-chemical terms annotated as chemical entities. a) EP0271596B1 was wrongly annotated to have a chemical structure “OCO” (SMILES code) in claim section, b) 69 terms in IBM database which were annotated to this chemical structure.



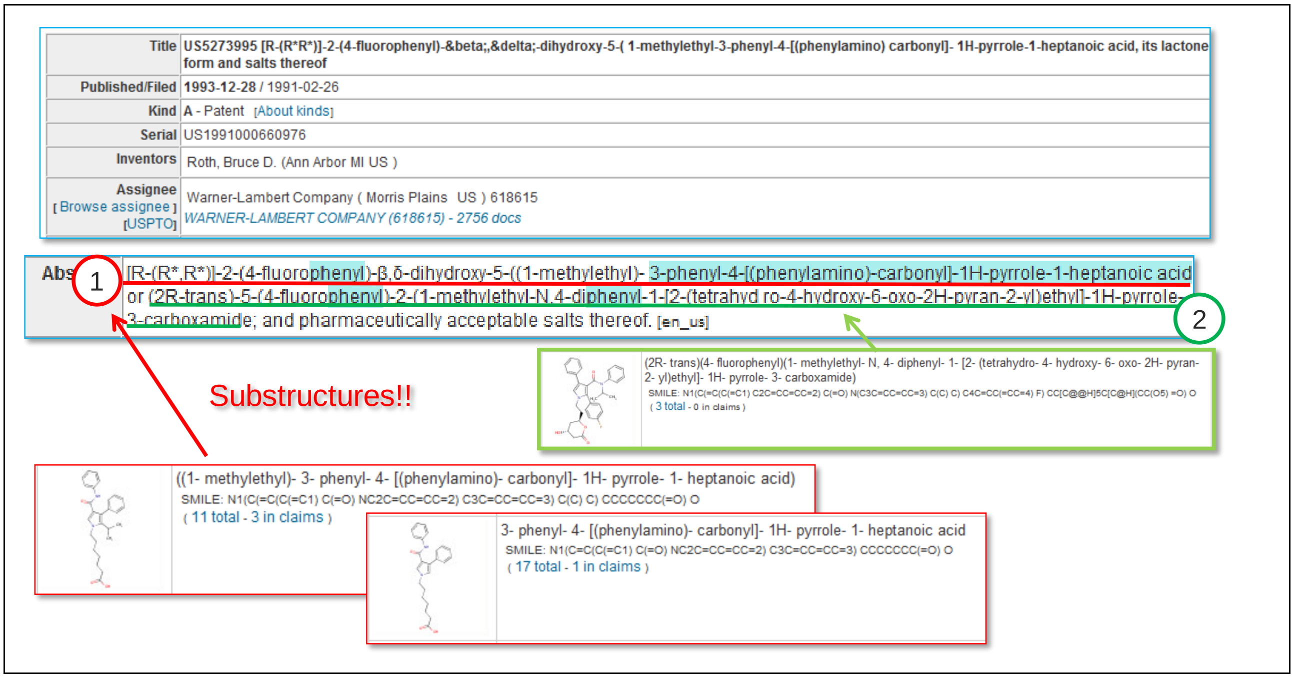
Title	**EP0106443B1 Pharmaceutical oral-controlled release multiple-units formulation**
Published/Filed	**1991-07-17** / 1983-08-09
Kind	**B1** - Patent specification [About kinds]
Serial	EP1983000304583 (Related: EP0106443A2)
Inventors	Roswall, Stig (DK-2860 Sobor Thorhus, Lene Birgitte (DK-2840 Holte
Assignee [Browse assignee]	BENZON AS ALFRED (DK) [from eq
IPC codes	Ed: 6 A61K 00922; A61K 00924; A61K 00
ECLA Classes	A61K9/20K2B; A61K9/50K; [from equiva
Foreign priority	DK19823652 1982-08-13
Family *Derived*	EP0106443A2 1984-04-25 1983-0 EP0106443B1 1991-07-17 1983-0
Designated Countries	**NATIONAL** AT BE CH DE FR GB IT LI LU
Attorney, agent	Walls, Alan James (GB) 00037210 Britis Cowley, Oxford OX4 5LY GB

Description (57528 characters - Click to expand / collapse)⊞

Claims (84 Exemp:)

A pharmaceutical oral controlled release multiple-units formulation in which individual units comprise coated units containing a first therapeutically active substance which is subject to controlled release as a result of coating of the units with a controlled release diffusion coating comprising a polymeric film forming substance selected from cellulose derivatives, acrylic polymers, vinyl polymers and other high molecular weight synthetic polymers such as ethylcellulose, cellulose acetate, cellulose propionate, cellulose butyrate, cellulose valerate, cellulose acetate propionate, polyvinyl acetate, polyvinyl formal, polyvinyl butyral, ladder polymer of sesquiphenyl siloxane, polymethyl methacrylate, polycarbonate, polystyrene, polyester, coumarone-indene polymer, polybutadiene, vinyl chloride-vinyl acetate copolymer, ethylene-vinyl acetate copolymer and vinyl chloride-propylene-vinyl acetate copolymer, the units additionally comprising particles of a second therapeutically active substance adhered to the surface of the controlled release diffusion coating in a substantially uniform layer, the particles being at least one power of ten smaller than the coated units.

Figure 14. A case of missed chemical annotation. Patent number EP0106443B1 contains generic chemical names (e.g. polyvinyl acetate, polyvinyl formal, and polyvinyl butyral) which were not recognized and annotated by the IBM.



Title	**US5273995 [R-(R*R*)]-2-(4-fluorophenyl)-&beta;,&delta;-dihydroxy-5-(1-methylethyl-3-phenyl-4-[(phenylamino) carbonyl]- 1H-pyrrole-1-heptanoic acid, its lactone form and salts thereof**
Published/Filed	**1993-12-28** / 1991-02-26
Kind	**A** - Patent [About kinds]
Serial	US1991000660976
Inventors	Roth, Bruce D. (Ann Arbor MI US)
Assignee [Browse assignee] [USPTO]	Warner-Lambert Company (Morris Plains US) 618615 *WARNER-LAMBERT COMPANY (618615) - 2756 docs*

Abs [R-(R*,R*)]-2-(4-fluorophenyl)-β,δ-dihydroxy-5-((1-methylethyl)- 3-phenyl-4-[(phenylamino)-carbonyl]-1H-pyrrole-1-heptanoic acid or (2R-trans)-5-(4-fluorophenyl)-2-(1-methylethyl-N,4-diphenyl-1-[2-(tetrahyd ro-4-hydroxy-6-oxo-2H-pyran-2-yl)ethyl]-1H-pyrrole-3-carboxamide; and pharmaceutically acceptable salts thereof. [en_us]

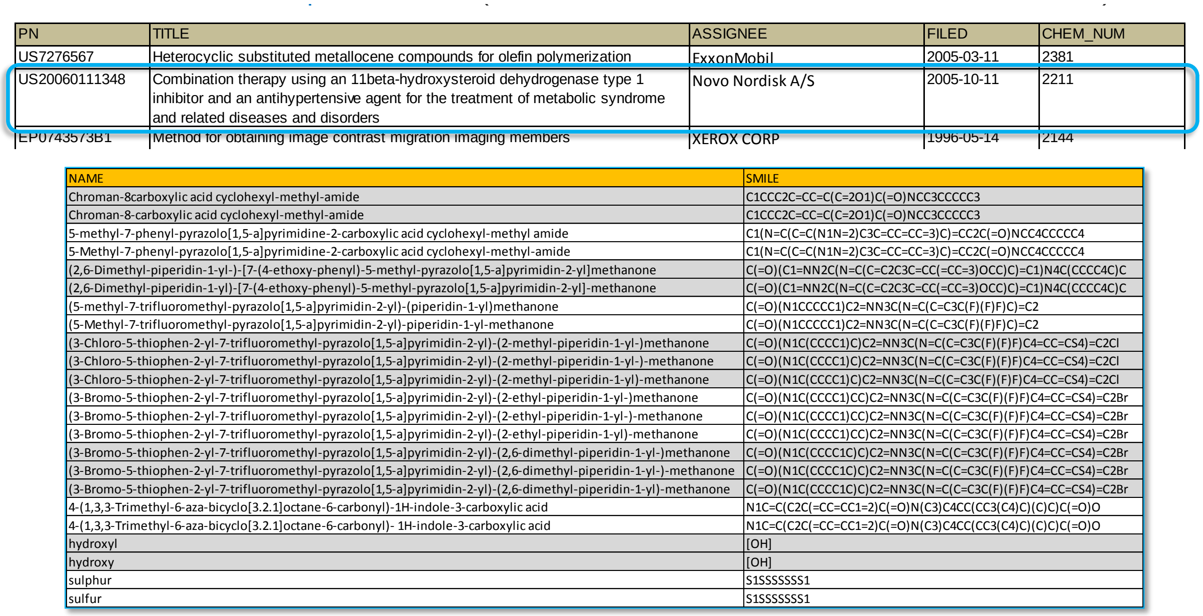
(2R- trans)(4- fluorophenyl)(1- methylethyl- N, 4- diphenyl- 1- [2- (tetrahydro- 4- hydroxy- 6- oxo- 2H- pyran- 2- yl)ethyl]- 1H- pyrrole- 3- carboxamide)
SMILE: N1(C(=C(C(=C1) C2C=CC=CC=2) C(=O) N(C3C=CC=CC=3) C(C) C) C4C=CC(=CC=4) F) CC[C@@H]5C[C@H](CC(O5) =O) O
(3 total - 0 in claims)

((1- methylethyl)- 3- phenyl- 4- [(phenylamino)- carbonyl]- 1H- pyrrole- 1- heptanoic acid)
SMILE: N1(C(=C(C(=C1) C(=O) NC2C=CC=CC=2) C3C=CC=CC=3) C(C) C) CCCCCCC(=O) O
(11 total - 3 in claims)

3- phenyl- 4- [(phenylamino)- carbonyl]- 1H- pyrrole- 1- heptanoic acid
SMILE: N1(C=C(C(=C1) C(=O) NC2C=CC=CC=2) C3C=CC=CC=3) CCCCCCC(=O) O
(17 total - 1 in claims)

Figure 15. A case of annotation of substructures instead of full structures. US5273995 contains 2 chemicals in the original abstract, but the IBM annotate these chemicals into 3 chemical entities.

a)

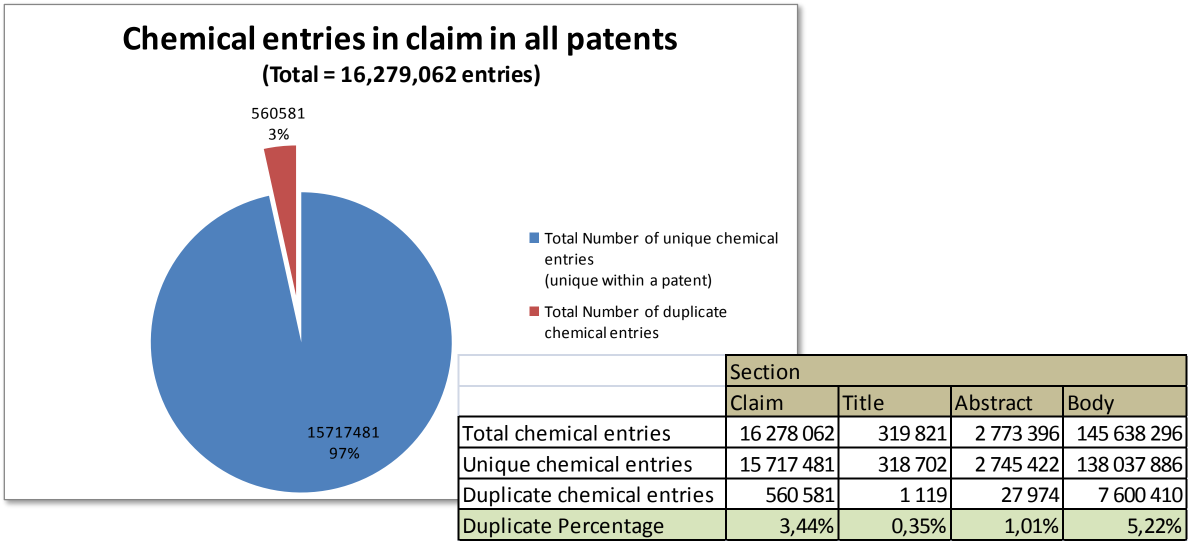


PN	TITLE	ASSIGNEE	FILED	CHEM_NUM
US7276567	Heterocyclic substituted metallocene compounds for olefin polymerization	ExxonMobil	2005-03-11	2381
US20060111348	Combination therapy using an 11beta-hydroxysteroid dehydrogenase type 1 inhibitor and an antihypertensive agent for the treatment of metabolic syndrome and related diseases and disorders	Novo Nordisk A/S	2005-10-11	2211
EP0743573B1	Method for obtaining image contrast migration imaging members	XEROX CORP	1996-05-14	2144

NAME	SMILE
Chroman-8carboxylic acid cyclohexyl-methyl-amide	C1CCC2C=CC=C(C=2O1)C(=O)NCC3CCCCC3
Chroman-8-carboxylic acid cyclohexyl-methyl-amide	C1CCC2C=CC=C(C=2O1)C(=O)NCC3CCCCC3
5-methyl-7-phenyl-pyrazolo[1,5-a]pyrimidine-2-carboxylic acid cyclohexyl-methyl amide	C1(N=C(C=C(N1N=2)C3C=CC=CC=3)C)=CC2C(=O)NCC4CCCCC4
5-Methyl-7-phenyl-pyrazolo[1,5-a]pyrimidine-2-carboxylic acid cyclohexyl-methyl-amide	C1(N=C(C=C(N1N=2)C3C=CC=CC=3)C)=CC2C(=O)NCC4CCCCC4
(2,6-Dimethyl-piperidin-1-yl-)-[7-(4-ethoxy-phenyl)-5-methyl-pyrazolo[1,5-a]pyrimidin-2-yl]methanone	C(=O)(C1=NN2C(N=C(C=C2C3C=CC(=CC=3)OCC)C)=C1)N4C(CCCC4C)C
(2,6-Dimethyl-piperidin-1-yl)-[7-(4-ethoxy-phenyl)-5-methyl-pyrazolo[1,5-a]pyrimidin-2-yl]-methanone	C(=O)(C1=NN2C(N=C(C=C2C3C=CC(=CC=3)OCC)C)=C1)N4C(CCCC4C)C
(5-methyl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(piperidin-1-yl)methanone	C(=O)(N1CCCCC1)C2=NN3C(N=C(C=C3C(F)(F)F)C)=C2
(5-Methyl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-piperidin-1-yl-methanone	C(=O)(N1CCCCC1)C2=NN3C(N=C(C=C3C(F)(F)F)C)=C2
(3-Chloro-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2-methyl-piperidin-1-yl-)methanone	C(=O)(N1C(CCCC1)C)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Cl
(3-Chloro-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2-methyl-piperidin-1-yl-)-methanone	C(=O)(N1C(CCCC1)C)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Cl
(3-Chloro-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2-methyl-piperidin-1-yl)-methanone	C(=O)(N1C(CCCC1)C)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Cl
(3-Bromo-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2-ethyl-piperidin-1-yl-)methanone	C(=O)(N1C(CCCC1)CC)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Br
(3-Bromo-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2-ethyl-piperidin-1-yl-)-methanone	C(=O)(N1C(CCCC1)CC)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Br
(3-Bromo-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2-ethyl-piperidin-1-yl)-methanone	C(=O)(N1C(CCCC1)CC)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Br
(3-Bromo-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2,6-dimethyl-piperidin-1-yl-)methanone	C(=O)(N1C(CCCC1C)C)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Br
(3-Bromo-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2,6-dimethyl-piperidin-1-yl-)-methanone	C(=O)(N1C(CCCC1C)C)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Br
(3-Bromo-5-thiophen-2-yl-7-trifluoromethyl-pyrazolo[1,5-a]pyrimidin-2-yl)-(2,6-dimethyl-piperidin-1-yl)-methanone	C(=O)(N1C(CCCC1C)C)C2=NN3C(N=C(C=C3C(F)(F)F)C4=CC=CS4)=C2Br
4-(1,3,3-Trimethyl-6-aza-bicyclo[3.2.1]octane-6-carbonyl)-1H-indole-3-carboxylic acid	N1C=C(C2C(=CC=CC1=2)C(=O)N(C3)C4CC(CC3(C4)C)(C)C)C(=O)O
4-(1,3,3-Trimethyl-6-aza-bicyclo[3.2.1]octane-6-carbonyl)- 1H-indole-3-carboxylic acid	N1C=C(C2C(=CC=CC1=2)C(=O)N(C3)C4CC(CC3(C4)C)(C)C)C(=O)O
hydroxyl	[OH]
hydroxy	[OH]
sulphur	S1SSSSSSS1
sulfur	S1SSSSSSS1

b)

- ~ 2 million patents with chemicals in claim section
  - containing 16.3 million chemical entries
    - Of these, 0.56 million entries (3%) are duplicate SMILES within the same patent



	Section			
	Claim	Title	Abstract	Body
Total chemical entries	16 278 062	319 821	2 773 396	145 638 296
Unique chemical entries	15 717 481	318 702	2 745 422	138 037 886
Duplicate chemical entries	560 581	1 119	27 974	7 600 410
Duplicate Percentage	3,44%	0,35%	1,01%	5,22%

Figure 16. A case of duplicate chemical entries within a patent. a) US20060111348 contains 2,211 chemical entities annotated in claim section. Of these, 23 chemical entities are sharing 10 SMILES codes (13 duplicate entries), b) division of the error type by patent sections.

3.1.3.3 Chemical annotation comparison between IBM, GVKBIO and the AZ-interest set

It was provided that there would be different in chemical annotation between IBM automated extraction, GVKBIO expert curation, and the AZ-interest set. While GVKBIO extracts target-linked chemicals and AZ collects any chemicals depends on the source, IBM automatically processes all the chemical names in each patent. This resulted in different numbers of chemicals extracted from the same patent between these three data sources. In order to quantify the difference, 4,547 patents which were common to IBM, GVKBIO and AZ-interest set (Figure 6) were assessed (based on database capture on February 22, 2010). Samples of these patents are shown in Table 5. Differences between the numbers of chemicals extracted are shown in Table 6 and Figure 17.

Table 5. Comparison of chemicals curated by IBM, GVKBIO and AZ-interest set (Example of 20 patents).

Patent No.	Titles	Assignees	Target Proteins (GVKBIO curated)	Number of Chemicals Extracted			
				IBM		GVKBIO	AZ-interested set
				All sections	Claim section		
WO2009126861	TRIAZOLOPYRIDINE COMPOUNDS USEFUL AS DGATI INHIBITORS	Bristol-Myers Squibb	DGAT1	293	8	12	30
WO2009126624	TRIAZOLO COMPOUNDS USEFUL AS DGAT1 INHIBITORS	Bristol-Myers Squibb	DGAT1	336	13	66	148
WO2009112445	METHOD OF INCREASING CELLULAR PHOSPHATIDYLCHOLINE BY DGAT1 INHIBITION	Novartis	DGAT1	965	214	66	269
WO2009081195	CARBAMOYL COMPOUNDS AS DGAT1 INHIBITORS 190	AstraZeneca	DGAT1	367	36	79	340
WO2009075857	GLYCINE TRANSPORTER-1 INHIBITORS	Amgen	SLC6A9	493	51	46	89
WO2009047558	BENZOTHIAZOLES AS GHRELIN RECEPTOR MODULATORS	AstraZeneca	GHSR	792	31	270	523
WO2009046802	N-(PYRAZOLE- 3 -YL)-BENZAMIDE DERIVATIVES AS GLUCOKINASE ACTIVATORS	Merck & Co.	GCK	275	9	10	37
WO2009046784	PYRIDINE DERIVATIVES USEFUL AS GLUCOKINASE ACTIVATORS	Merck & Co.	GCK	337	14	99	155
WO2009040410	OXADIAZOLE- AND OXAZOLE-SUBSTITUTED BENZIMIDAZOLE- AND INDOLE-DERIVATIVES AS DGAT1 INHBITORS	Novartis	DGAT1	343	18	72	228
WO2009034062	COMPOUNDS WHICH INHIBIT THE GLYCINE TRANSPORTER AND USES THEREOF IN MEDICINE	GlaxoSmithKline	SLC6A9	139	7	4	24
WO2009034061	SPIRO-CONDENSED IMIDAZOLONE DERIVATIVES INHIBITING THE GLYCINE TRANSPORTER	GlaxoSmithKline	SLC6A9	211	7	23	50
WO2009016462	SUBSTITUTED BICYCLOLACTAM COMPOUNDS	Pfizer (with Wyeth)	DGAT1	188	12	58	118
WO2009010794	2,4-DIAMINO-PYRIMIDINE DERIVATIVES	AstraZeneca	EPHB4	228	10	18	75
WO2009005460	SMALL MOLECULE INHIBITORS OF NAVL.7 SODIUM CHANNELS FOR THE TREATMENT OF PAIN DISORDERS	AstraZeneca	SCN9A	215	19	44	111
WO2009004380	METHOD AND USE OF CIRCULATING LEVELS OF ENDOCANNABINOID LIGANDS FOR THE DETERMINATION OF PATIENT IN NEED AND/OR SUITABLE FOR CB1R ANTAGONIST DRUG TREATMENT AS WELL AS METHOD FOR INDUCING WEIGHT LOSS/MAINTENANCE/GAIN	AstraZeneca	CNR1	36	1	7	16
WO2008152089	NOVEL COMPOUNDS	GlaxoSmithKline	GRM5	118	8	33	83
WO2008141976	DIACYLGLYCEROL ACYLTRANSFERASE INHIBITORS	Hoffmann–La Roche	DGAT1	1199	38	342	548
WO2008138876	PYRAZOLE DERIVATIVES AS P2X7 MODULATORS	GlaxoSmithKline	P2RX7	360	14	183	278
WO2008137436	[6,5]-BICYCLIC GPR119 G PROTEIN-COUPLED RECEPTOR AGONISTS	Bristol-Myers Squibb	GPR119	411	7	19	58
WO2008137435	[6,6] AND [6,7]-BICYCLIC GPR119 G PROTEIN-COUPLED RECEPTOR AGONISTS	Bristol-Myers Squibb	GPR119	562	7	100	259

Table 6. Statistics of chemicals curated by IBM, GVKBIO and AZ-interest set.

Number of chemicals/patent	IBM		GVKBIO	AZ-interest set
	All sections	Claim section		
Average	337	34	89	199
Standard deviation	291	62	113	275
Median	271	17	63	108
Minimum	0	0	0	1
Maximum	3664	1187	1796	3553

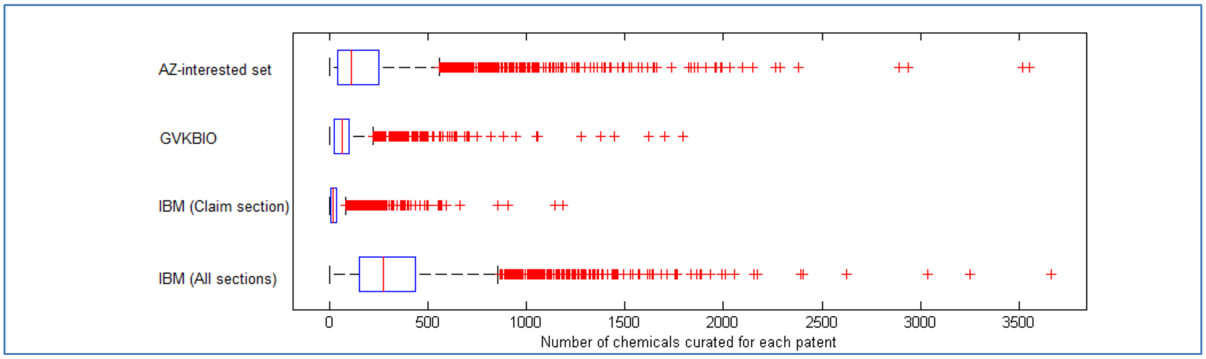


Figure 17. Box plot displaying difference in the number of chemicals curated per patent between IBM (All sections), IBM (Claim section), GVKBIO and AZ-interest set.

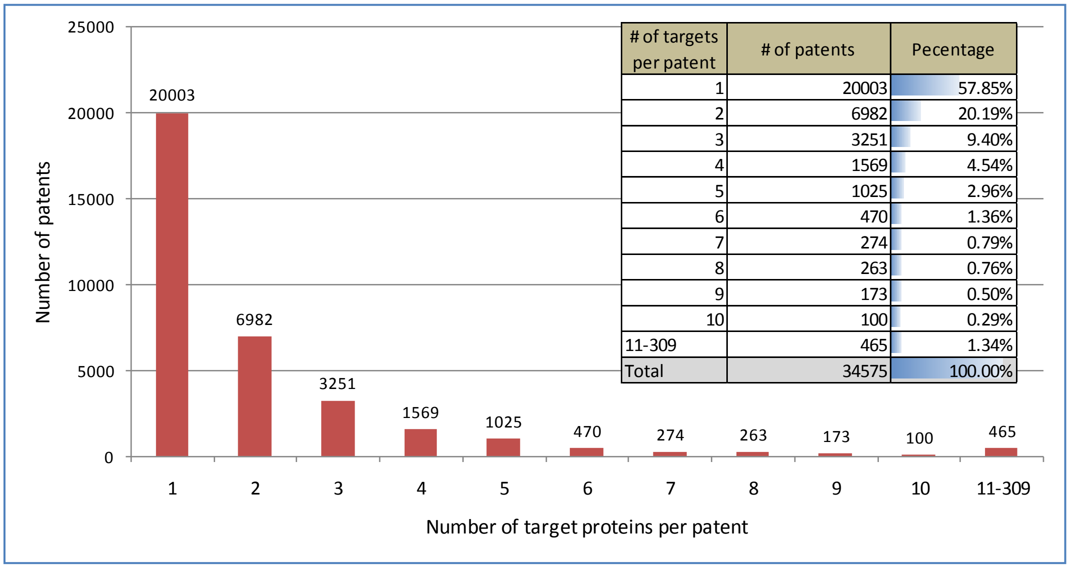
For the retrieval of patents with target proteins, it was assumed that a novel target-modulating chemical structure should be mentioned in claim section. Therefore, a condition of having at least one chemical in claim could be used in the retrieval. However, in this set of patents with target proteins according to GVKBIO curation, there were 319 patents (~7%) without chemicals annotated in claim. This might be resulted from imperfect chemical annotation of IBM. As a consequence, using the existence of chemicals in claim for patent retrieval could lead to false negatives; leaving out ~7% patents with targets.

Interestingly, there was one patent without any GVKBIO-curated chemical structures (WO2000025766A2: “Treatment of Gastric Asthma”). The patent claims uses of chemicals targeting NK2 receptor for treatment of gastric asthma. Manual inspection revealed that there was no structure in claim section, but there were at least three drug-like chemical structures in the patent body. Further investigation on the patent body descriptions showed that this patent obfuscates the NK2 receptor antagonists’ chemical structures by referring to other patents, using “blinding” words such as “Example I” and “Compound X”.

3.1.4 GVKBIO-annotated target proteins

This section aims to explore the number of target proteins mentioned in each patent. The query was performed on a subset of the GVKBIO containing only patents with non-antibiotic targets (mostly human proteins with some inter-species homologues) published between 1973 - 2009, extracted on March 15, 2010. This consisted of 34,575 patents encompassing 1,646 target proteins; resulting in 74,732 document-target links.

Shown in Figure 18 is the distribution of the number of target proteins per patent. It shows that ~58% of patents mention only one target protein in a patent (i.e. primary target protein). On the other hand, ~42% of patents mention multiple target proteins. Examples of these are patents claiming chemical modulation for factor Xa as primary target, while having chemical modulation data for other cross-screening targets such as factor VIIa and thrombin.



# of targets per patent	# of patents	Pecentage
1	20003	57.85%
2	6982	20.19%
3	3251	9.40%
4	1569	4.54%
5	1025	2.96%
6	470	1.36%
7	274	0.79%
8	263	0.76%
9	173	0.50%
10	100	0.29%
11-309	465	1.34%
Total	34575	100.00%

Figure 18. Distribution of the number of target proteins per patent from GVKBIO.

Figure 19 shows 40 popular target proteins ranked by the number of relevant patents. These 40 targets (2.4% of 1,646 targets in total) were contained in 24,558 document-target links (~33% of 74,732 links in total).

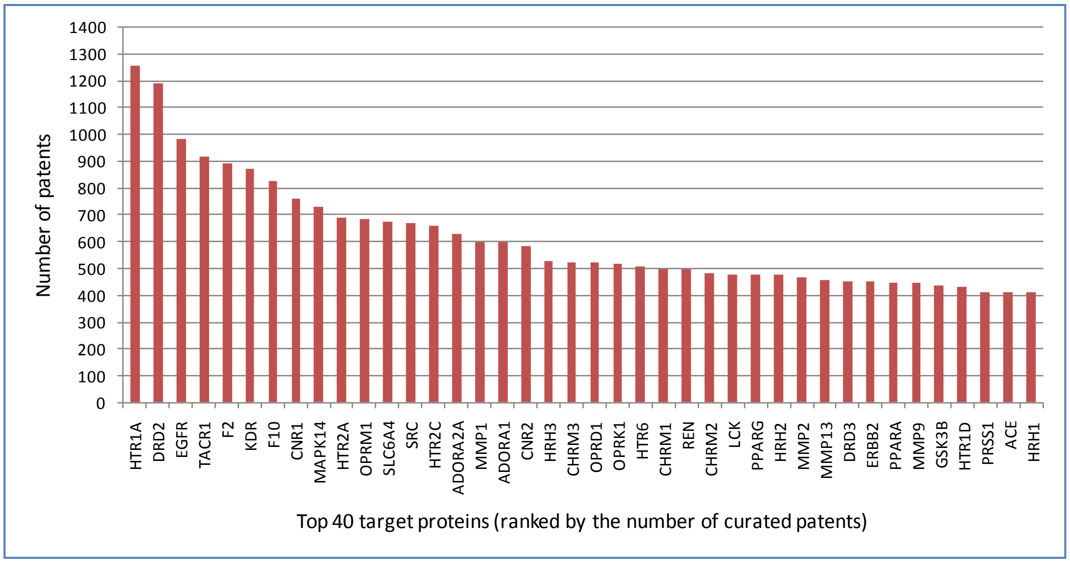


Figure 19. Top 40 target proteins ranked by the number of GVKBIO patents including each protein.

3.2 Selection and compilation of a protein name dictionary

Relevant to the target identification problem addressed in this study, there were requirements for identifying protein names in patents, and standardizing these into unique identifiers. Several approaches for these two requirements were studied and evaluated critically in the first BioCreative challenge (task1A: gene mention task, and task1B: gene normalization task)[33]. In order to address these two requirements, a dictionary-based approach [23] was employed. Three public protein synonym sources including UniProt [34], BioThesaurus [35] and BioLexicon[36] were assessed for their coverage (based on data on March 30, 2010). It is shown later in this section that the BioThesaurus covers the most synonyms for human proteins, compared to the other two sources. Therefore, the BioThesaurus was selected to build a synonym dictionary for subsequent usage.

While UniProt itself was not designed to provide protein/gene synonym information, BioThesaurus maps collection of protein and gene synonyms to protein entries in the UniProt Knowledgebase (UniProtKB). The BioThesaurus was compiled from many online resources, and in 2006, claimed to cover 2.6 million synonyms for 1.8 million UniProtKB entries [37]. However, the BioThesaurus downloaded during this study appeared to contain 29 million synonyms. By mapping these synonyms to HGNC symbols (by UniProtAC), it was found that the BioThesaurus covers 363,331 synonym records for 19,037 HGNC symbols with protein products) (i.e. 19 synonyms per protein in average). Note that there were 19,359 HGNC symbols with protein products as curated by the HGNC.

The third thesaurus evaluated, BioLexicon, was developed to cover several semantic entities in biomedical domain including gene, protein, chemical compounds, organisms and diseases, for example. Protein and gene synonyms were mainly extracted from the BioThesaurus. Nevertheless, it was shown that the BioLexicon contained only 269,401 synonym records for 20,085 human proteins (13 synonyms per protein on average). This reflects the fact that some synonyms in BioThesaurus were removed in the compilation process of the BioLexicon.

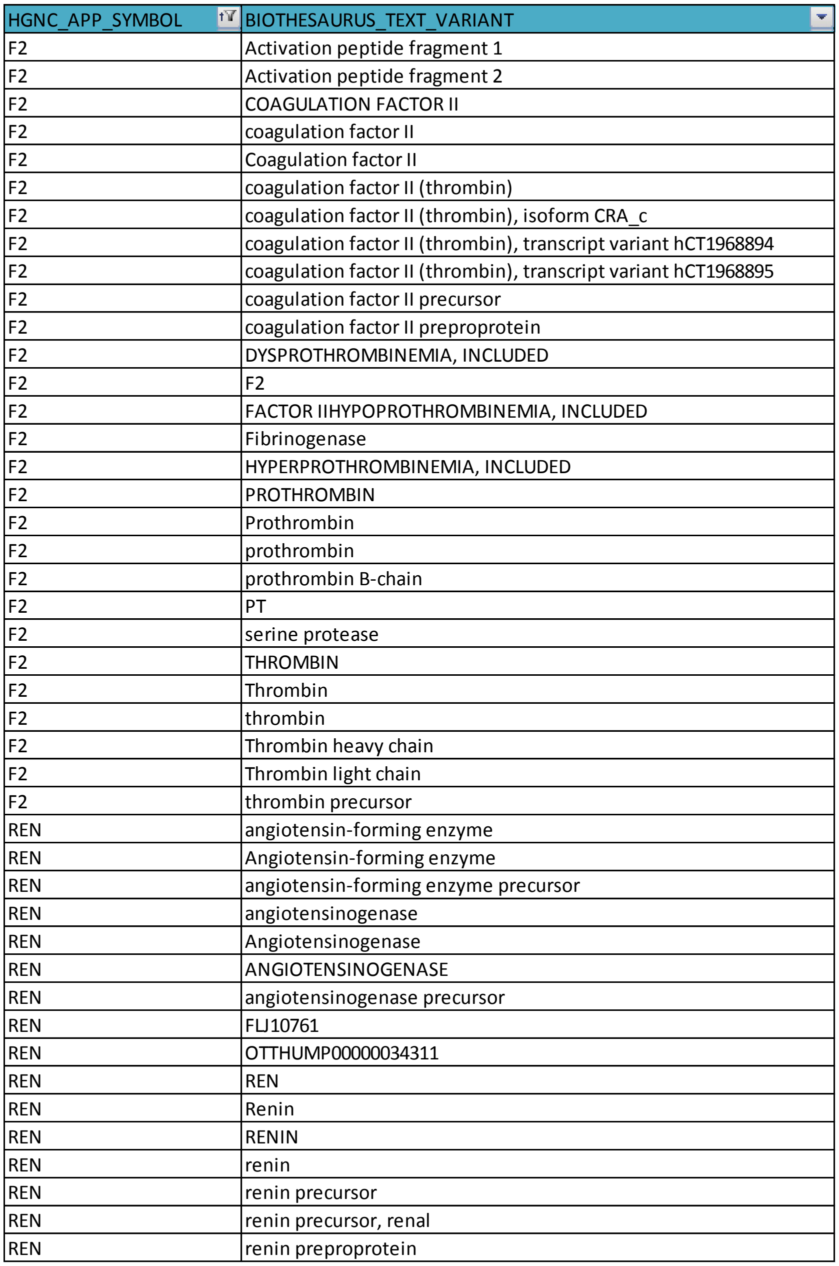
By manual inspection of several examples of protein names and their synonyms collected in each of the sources, it was shown that BioThesaurus covers the most text variants for each protein name, compared to the other two sources (exemplified in Table 7). For instance, the term “CB1R” which is a synonym for CNR1 protein is curated only in the BioThesaurus. Furthermore, it was shown that BioLexicon contained some mismatches between HGNC symbols and their synonyms. For instance, CNR1 was mapped to “Citrate synthase” which refers to another protein (Table 7).

Table 7. Comparison of protein synonyms for CNR1 obtained from three sources. a) UniProt, b) BioThesaurus, c) BioLexicon.

HGNC Symbol (UniProt_AC)	Synonym Dictionaries		
	(a) UniProt	(b) BioThesaurus	(c) BioLexicon
CNR1	CANN6	CANN6	gltA
(P21554)	CB1	cann6	Ta0169
	CB-R	CB1	(R)-citric synthase
	CNR	CB1A	CB1 cannabinoid receptor
	CNR1	CB1K5	Citrate (Si)-synthase
	Cannabinoid receptor 1	CB1R	citrate (si)-synthase
		CB-R	Citrate condensing enzyme
		CGBS08	Citrate oxaloacetate-lyase ((pro-3S)-CH(2)COO(-)->acetyl-CoA)
		CNR	Citrate oxaloacetate-lyase, CoA-acetylating
		CNR1	Citrate synthase
		CNR1 protein	citrate synthase
		CB1 cannabinoid receptor	Citrate synthetase
		CB1 RECEPTOR	Citric synthase
		OTTHUMP00000016838	Citric-condensing enzyme
		OTTHUMP00000016839	Citrogenase
		OTTHUMP00000016840	Condensing enzyme
		brain	methylcitrate synthase
		cannabinoid receptor	Oxalacetic transacetase
		CANNABINOID RECEPTOR 1	
		Cannabinoid receptor 1	
		cannabinoid receptor 1	
		cannabinoid receptor 1 (brain)	
		cannabinoid receptor 1 (brain), isoform CRA_a	
		cannabinoid receptor 1 (brain), transcript variant hCT2332433	
		cannabinoid receptor 1 (brain), transcript variant hCT2332435	
		cannabinoid receptor 1 long isoform	
		cannabinoid receptor 1 splice variant CB1b	
		cannabinoid receptor CB1	
		central cannabinoid receptor	
		central cannabinoid receptor isoform a	
		central cannabinoid receptor isoform b	
		central cannabinoid receptor, isoform a	

A synonym dictionary required in subsequent study included a dictionary for all human proteins with mapping to HGNC symbols (the same format as GVKBIO curation). As stated earlier, the BioThesaurus maps protein synonyms to UniProtAC rather than HGNC symbols. According to HGNC [38], there were 19,359 HGNC symbols with protein products (out of 28,965 symbols in total). A mapping from UniProtAC to HGNC symbols was downloaded from UniProt and HGNC websites [34,38]. The mapping from BioThesaurus into HGNC symbols resulted in a dictionary of 363,331 synonym records encompassing 19,037 human proteins (Table 8).

Table 8. Sample entries from the BioThesaurus dictionary mapped to HGNC symbols (i.e. F2, REN).



HGNC_APP_SYMBOL	BIOTHESAURUS_TEXT_VARIANT
F2	Activation peptide fragment 1
F2	Activation peptide fragment 2
F2	COAGULATION FACTOR II
F2	coagulation factor II
F2	Coagulation factor II
F2	coagulation factor II (thrombin)
F2	coagulation factor II (thrombin), isoform CRA_c
F2	coagulation factor II (thrombin), transcript variant hCT1968894
F2	coagulation factor II (thrombin), transcript variant hCT1968895
F2	coagulation factor II precursor
F2	coagulation factor II preproprotein
F2	DYSPROTHROMBINEMIA, INCLUDED
F2	F2
F2	FACTOR IIHYPOPROTHROMBINEMIA, INCLUDED
F2	Fibrinogenase
F2	HYPERPROTHROMBINEMIA, INCLUDED
F2	PROTHROMBIN
F2	Prothrombin
F2	prothrombin
F2	prothrombin B-chain
F2	PT
F2	serine protease
F2	THROMBIN
F2	Thrombin
F2	thrombin
F2	Thrombin heavy chain
F2	Thrombin light chain
F2	thrombin precursor
REN	angiotensin-forming enzyme
REN	Angiotensin-forming enzyme
REN	angiotensin-forming enzyme precursor
REN	angiotensinogenase
REN	Angiotensinogenase
REN	ANGIOTENSINOGENASE
REN	angiotensinogenase precursor
REN	FLJ10761
REN	OTTHUMP00000034311
REN	REN
REN	Renin
REN	RENIN
REN	renin
REN	renin precursor
REN	renin precursor, renal
REN	renin preproprotein

3.3 Manual inspection of patents

The aim of this section was to gain insight on the problems and possible solutions for target protein name extraction. Five patents common to IBM and GVKBIO were inspected manually (Table 9).

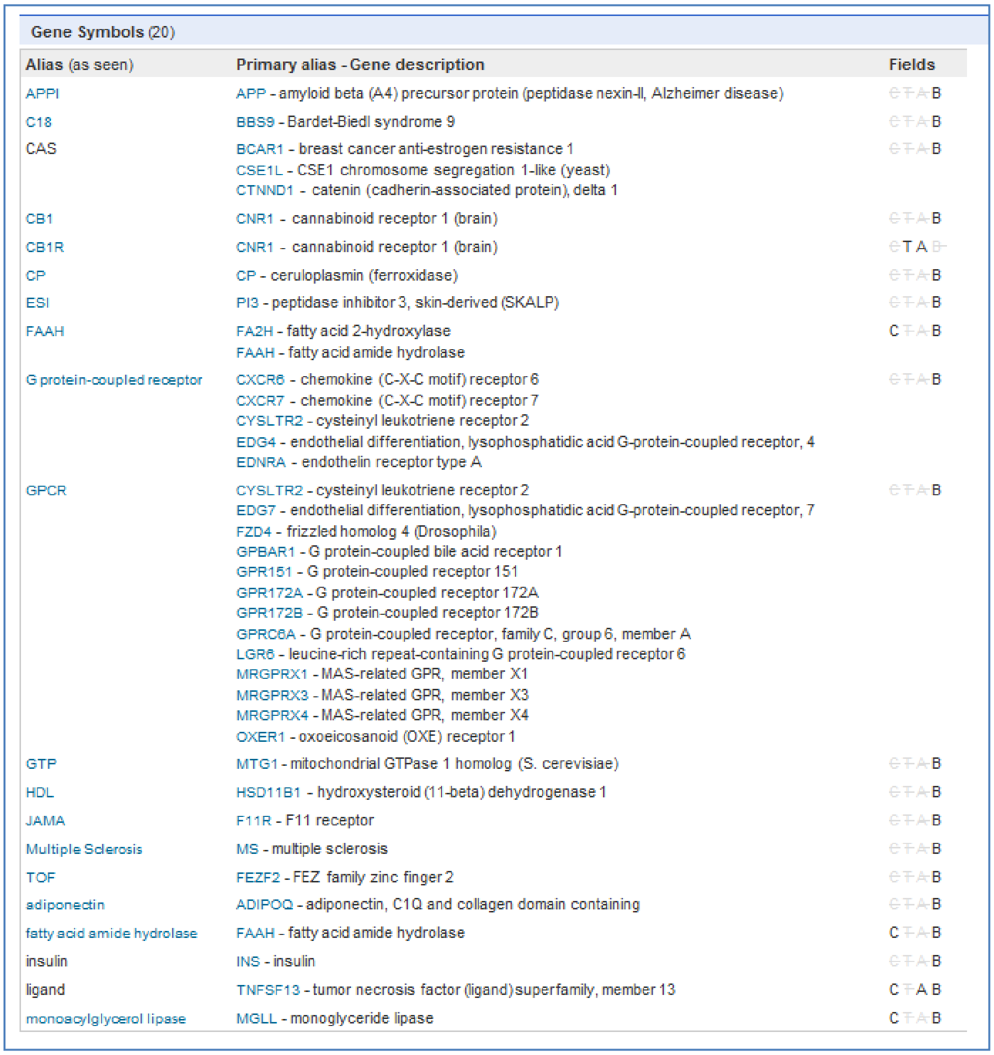
Table 9. A list of five patents which were inspected manually.

Patent No.	Published year	Title	Assignee	GVKBIO -curated target
WO2009004380A1	2009	Method and use of circulating levels of endocannabinoid ligands for the determination of patient in need and/or suitable for cb1r antagonist drug treatment as well as method for inducing weight loss/maintenance/gain.	AstraZeneca	CNR1
WO2008116814A1	2008	Pyrrole and isoindole carboxamide derivatives as p2x7 modulators.	GlaxoSmithKline	P2RX7
US20080090876	2008	Use of thianecarboxamides as dgat inhibitors.	Bristol-Myers Squibb	DGAT1
US20080015213	2008	Macrocyclic aminopyridyl beta-secretase inhibitors for the treatment of alzheimer's disease.	Merck & Co.	BACE1
EP1556034B1	2008	Indole derivatives as beta-2 agonists.	Pfizer	ADRB2

Given target-containing patents were retrieved, questions to be addressed were 1) how to extract a list of the mentioned protein names; and 2) how to extract target protein names from the list of mentioned protein names. In other words, within a target-containing patent, one needs to classify terms into the following hierarchy.

a) Non-protein terms
b) Protein names
    i. Non-target protein names
    ii. Target protein names

Classification of terms into (a) and (b) could be done by dictionary-based named entity recognition [23] which is the same approach used by the IBM to annotate gene names and provide their corresponding HGNC symbols on the IBM SIMPLE web application (shown in Table 20). Nevertheless, it was shown that there were false positive and false negative problems which shall be investigated in section 3.3.1.



Gene Symbols (20)

Alias (as seen)	Primary alias - Gene description	Fields
APPI	APP - amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)	~~C~~ ~~T~~ ~~A~~ B
C18	BBS9 - Bardet-Biedl syndrome 9	~~C~~ ~~T~~ ~~A~~ B
CAS	BCAR1 - breast cancer anti-estrogen resistance 1 CSE1L - CSE1 chromosome segregation 1-like (yeast) CTNND1 - catenin (cadherin-associated protein), delta 1	~~C~~ ~~T~~ ~~A~~ B
CB1	CNR1 - cannabinoid receptor 1 (brain)	~~C~~ ~~T~~ ~~A~~ B
CB1R	CNR1 - cannabinoid receptor 1 (brain)	~~C~~ T A ~~B~~
CP	CP - ceruloplasmin (ferroxidase)	~~C~~ ~~T~~ ~~A~~ B
ESI	PI3 - peptidase inhibitor 3, skin-derived (SKALP)	~~C~~ ~~T~~ ~~A~~ B
FAAH	FA2H - fatty acid 2-hydroxylase FAAH - fatty acid amide hydrolase	C ~~T~~ ~~A~~ B
G protein-coupled receptor	CXCR6 - chemokine (C-X-C motif) receptor 6 CXCR7 - chemokine (C-X-C motif) receptor 7 CYSLTR2 - cysteinyl leukotriene receptor 2 EDG4 - endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 EDNRA - endothelin receptor type A	~~C~~ ~~T~~ ~~A~~ B
GPCR	CYSLTR2 - cysteinyl leukotriene receptor 2 EDG7 - endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 7 FZD4 - frizzled homolog 4 (Drosophila) GPBAR1 - G protein-coupled bile acid receptor 1 GPR151 - G protein-coupled receptor 151 GPR172A - G protein-coupled receptor 172A GPR172B - G protein-coupled receptor 172B GPRC6A - G protein-coupled receptor, family C, group 6, member A LGR6 - leucine-rich repeat-containing G protein-coupled receptor 6 MRGPRX1 - MAS-related GPR, member X1 MRGPRX3 - MAS-related GPR, member X3 MRGPRX4 - MAS-related GPR, member X4 OXER1 - oxoeicosanoid (OXE) receptor 1	~~C~~ ~~T~~ ~~A~~ B
GTP	MTG1 - mitochondrial GTPase 1 homolog (S. cerevisiae)	~~C~~ ~~T~~ ~~A~~ B
HDL	HSD11B1 - hydroxysteroid (11-beta) dehydrogenase 1	~~C~~ ~~T~~ ~~A~~ B
JAMA	F11R - F11 receptor	~~C~~ ~~T~~ ~~A~~ B
Multiple Sclerosis	MS - multiple sclerosis	~~C~~ ~~T~~ ~~A~~ B
TOF	FEZF2 - FEZ family zinc finger 2	~~C~~ ~~T~~ ~~A~~ B
adiponectin	ADIPOQ - adiponectin, C1Q and collagen domain containing	~~C~~ ~~T~~ ~~A~~ B
fatty acid amide hydrolase	FAAH - fatty acid amide hydrolase	C ~~T~~ ~~A~~ B
insulin	INS - insulin	~~C~~ ~~T~~ ~~A~~ B
ligand	TNFSF13 - tumor necrosis factor (ligand) superfamily, member 13	C ~~T~~ A B
monoacylglycerol lipase	MGLL - monoglyceride lipase	C ~~T~~ ~~A~~ B

Figure 20. Example of gene symbols annotated by IBM SIMPLE for patent number WO2009004380A1.

On the other hand, classification of protein names into targets and non-targets is context-dependent and was not provided by IBM. Occurrence characteristics of the target protein names were observed, and possible solutions were proposed in section 3.3.2.

3.3.1 Protein name recognition

In this section, five patents were inspected manually for accuracy of the IBM dictionary-based protein name recognition. By scanning each document manually with facilitation by a text analysis tool called TextSTAT [39], it was possible to classify gene/protein symbols annotated by IBM into three categories (Table 10). Example of protein names annotated by IBM and that were missed for patent number WO2009004380A1 are shown in Table 11. Summary statistics for the IBM annotation performance is shown in Table 12.

Table 10. Classification of ability to recognize protein names by the IBM dictionary-based recognition.

No.	Categories		Description
1.	True positives		Protein names were correctly annotated
	1.1	One-to-one mapping	Annotated protein name can be linked to a particular protein sequence (e.g. GPR119)
	1.2	One-to-many mapping	Annotated protein name cannot be linked to a particular protein sequence (generic or aggregate protein name) (e.g. GPCR)
2.	False positives		Non-protein terms were annotated as protein names
	2.1	Non-biological terms	Non-biological terms such as analytical methods and chemical names were annotated as protein names (e.g. "atmospheric pressure photo ionization (APPI)" was annotated as APP protein)
	2.2	Biological terms	Biological terms such as assay types, drug names and disease names were annotated as protein names (e.g. "multiple sclerosis disease" was annotated as MS protein)
3.	False negatives		Protein names mentioned were missing

Table 11. Classification of protein names annotated and missed by IBM for patent number WO2009004380A1.

Categories	Alias (as seen)	Corresponding HGNC symbols	Text example	# of occurences
(1.1) True positives One-to-one mapping	CB1	CNR1 - cannabinoid receptor 1 (brain)	DETERMINATION OF PATIENT IN NEED AND/OR SUITABLE FOR CB1 RECEPTOR ANTAGONIST ..	1
	CB1R	CNR1 - cannabinoid receptor 1 (brain)	.. treatment with a cannabinoid receptor (CB1R) antagonist drug ..	20
	fatty acid amide hydrolase	FAAH - fatty acid amide hydrolase	For example, the enzyme fatty acid amide hydrolase (FAAH) degrades AEA ..	3
	monoacylglycerol lipase	MGLL - monoglyceride lipase	Similarly, detection of monoacylglycerol lipase, the degrading enzyme of 2-AG ..	3
(1.2) True positives One-to-many mapping	FAAH	FA2H - fatty acid 2-hydroxylase FAAH - fatty acid amide hydrolase	.. the enzyme fatty acid amide hydrolase (FAAH) degrades AEA, and it is believed that ..	5
	G protein-coupled receptor	CXCR6 - chemokine (C-X-C motif) receptor 6 CXCR7 - chemokine (C-X-C motif) receptor 7 CYSLTR2 - cysteinyl leukotriene receptor 2 EDG4 - endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 EDNRA - endothelin receptor type A	.. G protein-coupled receptor (GPCR) action generally requires binding of an agonist, ..	2
	GPCR	CYSLTR2 - cysteinyl leukotriene receptor 2 EDG7 - endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 7 FZD4 - frizzled homolog 4 (Drosophila) GPBAR1 - G protein-coupled bile acid receptor 1 GPR151 - G protein-coupled receptor 151 GPR172A - G protein-coupled receptor 172A GPR172B - G protein-coupled receptor 172B GPRC6A - G protein-coupled receptor, family C, group 6, member A LGR6 - leucine-rich repeat-containing G protein-coupled receptor 6 MRGPRX1 - MAS-related GPR, member X1 MRGPRX3 - MAS-related GPR, member X3 MRGPRX4 - MAS-related GPR, member X4 OXER1 - oxoeicosanoid (OXE) receptor 1	.. G protein-coupled receptor (GPCR) action generally requires binding of an agonist, ..	1
(2.1) False positives Non-biological terms	APPI	APP - amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)	.. atmospheric pressure photo ionization (APPI) ..	1
	C18	BBS9 - Bardet-Biedl syndrome 9	.. HyPurity C18 column ..	1
	CP	CP - ceruloplasmin (ferroxidase)	.. using CP-55940 as agonist ligand ..	1
	ESI	PI3 - peptidase inhibitor 3, skin-derived (SKALP)	.. Electrospray ionization (ESI) ..	1
	JAMA	F11R - F11 receptor	The obesity epidemic affects all demographic groups including children (Hedley et al., JAMA, 291:2847-2850,2004) ..	2
	TOF	FEZF2 - FEZ family zinc finger 2	.. time-of-flight (TOF) mass spectrometers ..	1
	ligand	TNFSF13 - tumor necrosis factor (ligand) superfamily, member 13	.. circulating endogenous cannabinoid ligand in the mammal ..	83
	CAS	BCAR1 - breast cancer anti-estrogen resistance 1 CSE1L - CSE1 chromosome segregation 1-like (yeast) CTNND1 - catenin (cadherin-associated protein), delta 1	.. aka SRl 41716 /Acomplia® ; (5-(4-chlorophenyl)- 1-(2,4-dichlorophenyl)-4-methyl-N-(1-piperidinyl)-lH-pyrazole-3-carboxamide, CAS NO: 158681-13-1 ..	2
(2.2) False positives Biological terms	GTP	MTG1 - mitochondrial GTPase 1 homolog (S. cerevisiae)	.. with a Ki value of <10µM in a GTPγS assay using CP-55940 as agonist ligand ..	1
	HDL	HSD11B1 - hydroxysteroid (11-beta) dehydrogenase 1	.. metabolic disorders (e.g. low HDL- and/or high LDL-cholesterol levels) ..	1
	Multiple Sclerosis	MS - multiple sclerosis	.. neurodegenerative disorders (e.g. Multiple Sclerosis, ..	1
	adiponectin	ADIPOQ - adiponectin, C1Q and collagen domain containing	.. metabolic disorders (e.g. low adiponectin levels) ..	2
	insulin	INS - insulin	.. insulin resistance, insulin resistance syndrome, metabolic syndrome, ..	2
(3) False negatives	CBIR *(misspelling)*	CNR1 - cannabinoid receptor 1 (brain)	.. disorders or conditions, and to whom a CBIR antagonist drug is to be administered ..	89

Table 12. Summary of IBM protein annotation performance for the sample of five patents: a) protein term occurrences, b) unique protein term (HGNC symbol) occurrences.

Patent no.	(a) Number of protein term occurences					(b) Number of unique protein terms				
	(1) True positives	(2) False positives	(3) False negatives	% Recall	% Precision	(1) True positives	(2) False positives	(3) False negatives	% Recall	% Precision
WO2009004380A1	35	98	89	28.2%	26.3%	4	13	1	80.0%	23.5%
WO2008116814A1	58	19	0	100.0%	75.3%	14	8	0	100.0%	63.6%
US20080015213	35	22	18	66.0%	61.4%	9	11	2	81.8%	45.0%
US20080090876	145	20	9	94.2%	87.9%	51	14	3	94.4%	78.5%
EP1556034B1	13	77	72	15.3%	14.4%	12	6	4	75.0%	66.7%
Total	286	236	188	60.3%	54.8%	90	52	10	90.0%	63.4%

Inspection of these five patents revealed protein annotation problems and possible solutions summarized as follow.

False negative problems

1) Incomplete coverage of the dictionary leads to false negatives. For example:
   - "5-HT" (5-hydroxytryptamine; serotonin) (US20080015213)
   - "PPAR delta" (US20080090876)
   - "PDE3" (EP1556034B1)

   Expansion of the dictionary to cover more synonym variants for protein would be required to solve this problem. Nevertheless, the dictionary was able to recall 90% of protein names in sample patents (Table 12 (b)). Moreover, most of the false positive cases were not from the dictionary coverage, but how the dictionary was used such as an ability to match Greek letters (shown below).

2) Inability to recognize Greek letters and Roman numbers leading to false negatives. For example:
   - "CBIR" (WO2009004380A1)
   - "β-secretase" (US20080015213)
   - "PPAR γ", "PPAR δ" and "PPAR α" (US20080090876)

   Clearly similarity matching [40] in addition to exact matches is required in order to expand the coverage of the dictionary for example a dictionary entry "beta secretase" will able to match their text variants such as "beta-secretase", "β secretase", and "β-secretase". Pre-processing of documents to transform Greek letters and Roman numbers into standard format is a possible solution. An alternative is to perform the transformation on-the-fly during the annotation process.

3) Implicit mentions of protein names are impossible to be recognized by the dictionary-based approach. For example, in patent number EP1556034B1, the protein the beta-2 adrenergic receptor (ADRB2) was mentioned in the following text. More sophisticated text mining algorithm such as natural language processing (NLP) is required to detect the protein name in this case.
   - The using of "β2" to refer the ADRB2 starts in the beginning of the patent by the following sentence: *"Adrenoceptors are members of the large G-protein coupled receptor super-family. The adrenoceptor subfamily is itself divided into* α *and* β *subfamilies with the* β *sub-family being composed of at least 3 receptor subtypes:* β*1,* β*2 and* β*3 .."*.
   - Then throughout the patent, ADRB2 is implicitly referred to using the term "β2". For example, *" .. throughout the experiment, with the exception of when a* β*2 agonist according to the present invention is added .. "*.

False positive problems

1) Acronyms for non-biological terms annotated as protein names. Examples are shown in Table 11. For instance, the acronym ESI could be a synonym for the protein PI3 or the electrospray ionization method, depending on context requiring disambiguation by more sophisticated algorithm [41].
2) Biological terms such as disease names and assay types annotated as protein names. This is due to the fact that some gene/protein are named following their disease (functional naming) such as MS (multiple sclerosis), DIANPH (diabetic nephropathy) and (COPD) chronic obstructive pulmonary disease found in patent number WO2008116814A1. This problem can only be solved by parsing the context.
3) HTML/XML tags annotated as protein names. For example, "R^1" which is encoded as "R<SUP>1</SUP>" was annotated as protein HERPUD1 (patent no. US20080090876). Therefore, annotation of protein names should be done on a plain text after removal of HTML/XML formatting tags.

One-to-multiple protein name mappings

It can be seen in category 1.2 in Table 11 that generic protein name such as GPCR could be mapped to many protein sequences. This leads to a problem in identifying an exact target name. Furthermore, the mapping was shown to be inconsistent. For instance, the synonym terms "GPCR" and "G protein-coupled receptor" were mapped to different set of protein sequences (Table 11).

3.3.2 Target protein name recognition

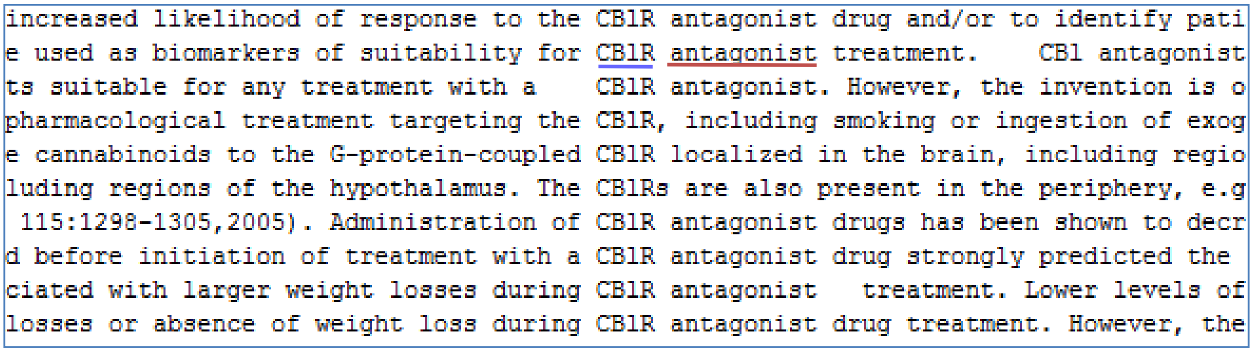
This section investigates the possibility to identify target protein names out of the list of protein names obtained from section 3.3.1. By scanning through the sample of five patents manually, occurrence characteristics of target protein names were identified. Possible extraction solutions were then proposed as shown below.

1) Within a patent, target protein names are likely to distribute uniformly across paragraphs as compared to non-target protein names. For instance, WO2009004380A1, in which the target protein CB1R was claimed for chemical modulation, mentions the term "CB1R" consistently in across the body description as compared to non-target proteins (i.e. GPCR, FAAH, etc.). Therefore, target protein names could possibly be identified by detecting protein names that are mentioned uniformly across paragraphs in the full-text document.
2) Within a patent, target protein names are more frequently mentioned than non-target protein names. Therefore, the most frequently mentioned protein is likely to be the target (Table 13).

Table 13. Frequency of protein names mentioned in each sample patent (manual inspection).

Patent no.	GVKBIO-curated target	protein names mentioned	# of occurences	Patent no.	GVKBIO-curated target	protein names mentioned	# of occurences
WO2009004380A1	CNR1	CNR1	110	US20080090876	DGAT1	DGAT1	37
		FAAH	8			ACACA	1
		GPCR	3			ACAT	21
		MGLL	3			ACE	2
		Total	124			ACP1	1
WO2008116814A1	P2RX7	P2RX7	40			AKR1B1	1
		ALS	1			ALDH7A1	1
		AME	1			ATP8A2	2
		CD20	2			CCK	1
		CD62L	1			CD2	1
		COX-2	1			CD4	2
		IL-1	2			CD40	5
		IL-6	1			DGAT2	2
		JAK3	1			etc. (freq < 5)	77
		M-CSF	1			Total	154
		TACE	1	EP1556034B1	ADRB1	ADRB1	63
		TNF	4			ADRB2	3
		P38	1			ADRB3	6
		tyrosine kinase	1			ALDH7A1	2
		Total	58			COX	1
US20080015213	BACE1	BACE1	33			COX-2	1
		5-HT	1			FLAP	1
		ALB	1			GPCR	1
		APP	12			KNG1	1
		CDK5	1			LTB	1
		COX-2	1			PDE3	1
		p25	1			PDE4A	1
		PDE	1			PDE5A	1
		SAP	1			PECAM1	1
		TAU	1			TNF	1
		Total	53			Total	85

3) Target protein names often co-occur with chemical modulation keywords such as inhibitor, antagonist, etc. (Figure 21). Therefore, target protein names could possibly be identified by searching for co-occurrence of protein names and the keywords within the same sentence or within close proximity.



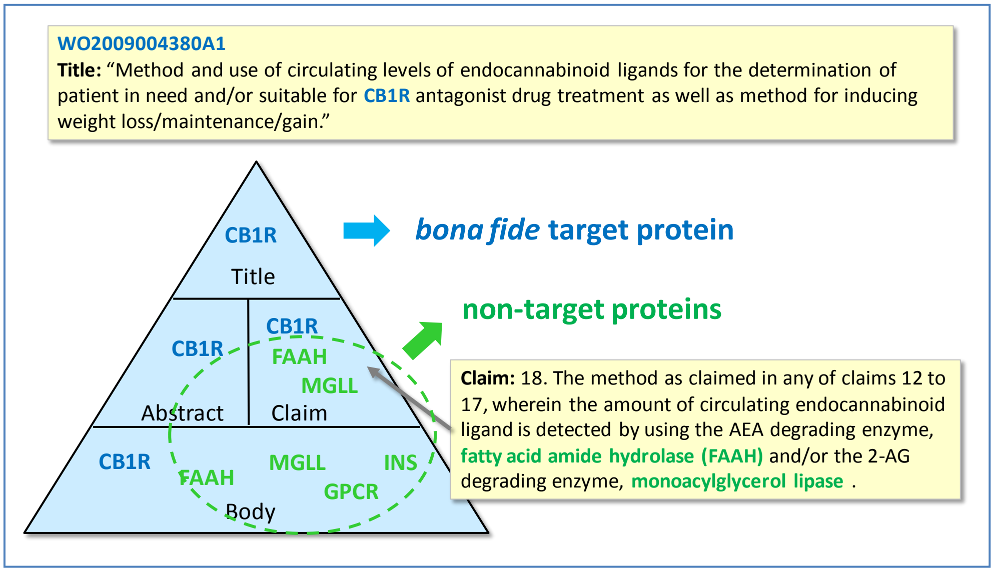
```
increased likelihood of response to the CB1R antagonist drug and/or to identify pati
e used as biomarkers of suitability for CB1R antagonist treatment.   CB1 antagonist
ts suitable for any treatment with a    CB1R antagonist. However, the invention is o
pharmacological treatment targeting the CB1R, including smoking or ingestion of exog
e cannabinoids to the G-protein-coupled CB1R localized in the brain, including regio
luding regions of the hypothalamus. The CB1Rs are also present in the periphery, e.g
 115:1298-1305,2005). Administration of CB1R antagonist drugs has been shown to decr
d before initiation of treatment with a CB1R antagonist drug strongly predicted the
ciated with larger weight losses during CB1R antagonist   treatment. Lower levels of
losses or absence of weight loss during CB1R antagonist drug treatment. However, the
```

Figure 21. Co-occurrence between a target protein name and chemical modulation keywords often found in sample patent (WO2009004380A1).

In order to implement these three solutions, the required database functionalities are 1) searching for terms at the paragraph and sentence level ("section search"), 2) searching for co-occurrence of terms within defined proximity ("proximity search"), 3) counting of occurrence frequency of terms. All these can be done by using Oracle Text features [42]. On the other hand, only the second requirement can be met out-of-the-box by Solr indexes provided along with the IBM databases.

3.4 Protein names in titles

This section evaluates the possibility to use protein names found in titles to retrieve relevant patents containing target-chemical-modulation data (target-containing patents). The approach was based on our initial assumption that a protein name in title is likely to represent the *bona fide* target for a target-containing patent (Figure 22).



WO2009004380A1
Title: "Method and use of circulating levels of endocannabinoid ligands for the determination of patient in need and/or suitable for **CB1R** antagonist drug treatment as well as method for inducing weight loss/maintenance/gain."

Claim: 18. The method as claimed in any of claims 12 to 17, wherein the amount of circulating endocannabinoid ligand is detected by using the AEA degrading enzyme, **fatty acid amide hydrolase (FAAH)** and/or the 2-AG degrading enzyme, **monoacylglycerol lipase** .

Figure 22. Schematic representation of patent sections and occurrences of target and non-target protein names. Exemplified in the figure is patent number WO2009004380A1 having CB1R as *bona fide* target.

3.4.1 Manual identification of target protein names in titles

In an attempt to retrieve patents containing a specific target protein, prior knowledge suggested it could be done by simply searching the protein synonyms in patent titles. This section aimed to estimate the recall performance of this approach by making a systematic estimation of the frequency of *bona fide* target protein names in patent titles. Sample titles from patents with target proteins were inspected manually in this experiment. A set of 211 titles were selected from GVKBIO with the following criteria:

1) they were published between years 2008 to 2010
2) each patent contains only one human protein
3) the patent numbers were common to GVKBIO, IBM, and the AZ-interest set (Figure 6)

Of these, there were 171 unique titles comprising 50 different target proteins (as indexed by GVKBIO). There were many cases of identical titles for different patent numbers with the same applicants but not in the same patent families (e.g. WO2008116107 and WO200807978 are GSK patents with the same title "glucokinase activators"). Each of patent titles in the sample set was read manually by an expert with extensive target and gene name knowledge (Christopher Southan – the project supervisor) in order to recognize probable *bona fide* targets. A categorization of these inspection results is shown in Table 14 and the results of the inspection in Table 15.

Table 14. Classification of ability to recognize target protein names in titles.

Categories	Descriptions
Positive	Probable *bona fide* target protein name (e.g. WO2008000409 “new CXCR2 inhibitors” or WO2009076337 “gamma secretase modulator”)
Ambiguous	Generic or aggregate target designations that cannot be linked to a protein sequence) (e.g. WO2008081208 “piperidine GPCR agonists” or WO2008115369 “derivatives of 5-amino-4,6-disubstituted indole and WO2009023677 “5-amino-4,6-disubstituted indoline as potassium channel modulators”)
Negative	No target protein name (e.g. WO2009059961 “a method of hormone suppression in humans” or WO2009010794 “2,4-diamino-pyrimindine derivatives”)

Table 15. Manual inspection of ability to recognize target protein names in titles.

Categories	Number of unique patent titles	
Positive	87	(51%)
Ambiguous	15	(9%)
Negative	69	(40%)
Total	171	(100.0%)

The analysis was extended by using Wordle [43] to visually display the term frequencies in the three categories of titles from Table 14 (Figure 23).

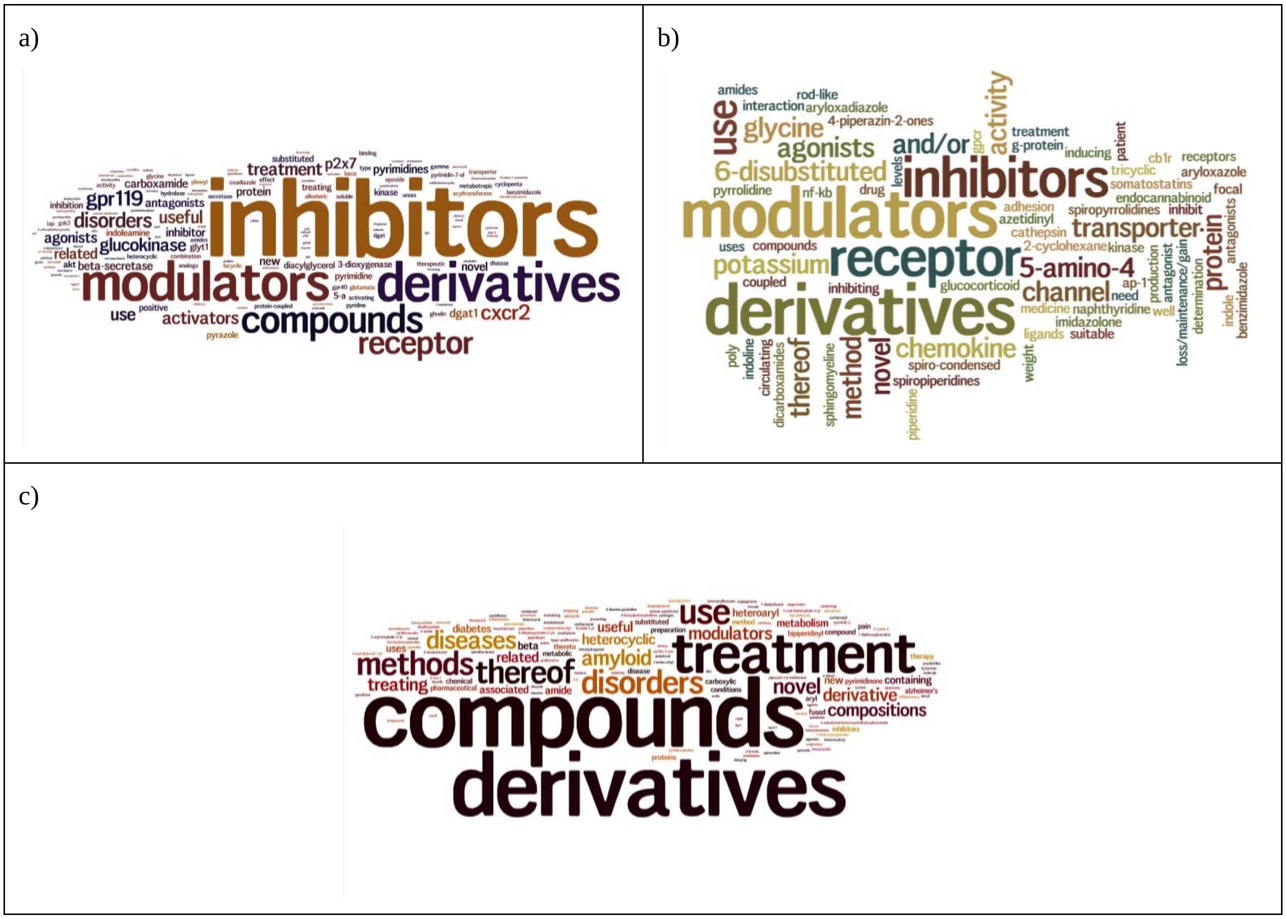


Figure 23. Word clouds of patent titles. a) patent titles in "positive" target recognition category, b) "ambiguous" category and c) "negative" category.

It can be seen from Figure 23 (a) and (b) that patent titles within the "positive" and the "ambiguous" categories contains high frequency of keywords associated with chemical modulation (e.g. "inhibitor" and "modulator"). Therefore, the retrieval of target-containing patents could possibly be done by combining protein names with these keywords for a recall approaching 60% (Table 15, 51% positives, 9% ambiguous). Note that in Figure 23 (a) some probable *bona fide* targets are standing out in the title word clouds (e.g. GPR119, P2X7 and CXCR2). This due to the fact that titles being examined were those common to the AZ-interest set.

It can be seen from Figure 23 (c) that patent titles within the "negative" category usually mention only associated disease and/or chemical series names (e.g. "treatment", "disease", "compounds" and "derivatives"). This category also includes cases where recognizable protein names in titles are not *bona fide* targets. For example, the patent US20080280948 with title "modulator of amyloid beta" actually claims gamma secretase inhibitors rather than direct amyloid chemical modulation (i.e. for disaggregation). The association is thus interpretable but the target name is, strictly speaking, a false positive. It is important to emphasize here that this experiment was performed on a set of sample patents from GVKBIO which were known to have a target protein identified in the body of each patent via manual curation. Therefore, the result cannot predict the precision for retrieving patents using occurrence of protein names in titles. It is also not possible to extrapolate to equivalent precision using a corpus of patents where the presence of a *bona fide* target is unknown (i.e. unknown set of true positives).

3.4.2 Automatic identification of target protein names in titles

Results from the previous section (3.4.2) suggested that the retrieval of target-containing patents could possibly be automated by searching for occurrences of protein names in titles with approximate recall of 60%. Considering the size of IBM, this recall percentage would be acceptable. The objective of this section was to assess the approach for their recall and precision when it is applied to the IBM database.

To avoid the problem of synonym variability, four protein names with relatively clean synonyms were chosen (Table 16) as test cases in this study. Renin and thrombin are *bona fide* target proteins, while albumin and hemoglobin are non-target controls. The data sets used in this experiment were IBM from 1980 - 2009 (10,846,899 patents) and a subset of GVKBIO patents with targets from 1980 - 2009 (34,441 patents). Synonyms of the four names (Table 16) were searched against patent titles in IBM and calibrated against the same (target annotated) GVKBIO patents (Table 17).

Table 16. Protein names, HGNC symbols and their synonyms.

Protein Names	HGNC Symbols	Synonyms
Renin	REN	rennin angiotensinogenase
Thrombin	F2	prothrombin coagulation factor II coagulation factor 2
Albumin	ALB	serum albumin
Hemoglobin	(many subunits)	haemoglobin

Table 17. Retrieved patents using protein synonyms in titles.

Protein Names	Number of Patents Retrieved			Approx. % Recall	Min. % Precision
	GVKBIO	IBM	patents-in-common between GVKBIO & IBM		
Renin	494	813	237	48.0%	29.2%
Thrombin	890	1743	215	24.2%	12.3%
Albumin	5	1200	0	0.0%	0.0%
Hemoglobin	0	1542	0	-	0.0%

In order to estimate percentage recall and precision shown in Table 17, it should be noted that GVKBIO does not cover all patents for a particular target protein as illustrated in Figure 24. Therefore the number of patents-in-common between GVKBIO patents and a search result cannot be used to infer the number of true positives obtained directly. Nevertheless, this can be interpreted as the minimum number of true positives. This can then be used to estimate the minimum percentage precision shown in Formula 1 (below). For example, in the case of renin, there are 813 patents in the IBM search result with 237 patents-in-common with GVKBIO (with renin as a target protein). Therefore, the minimum percentage precision of this retrieval approach can be estimated as 29.2%

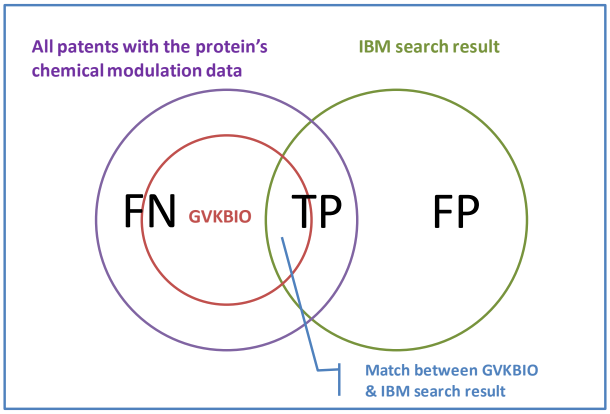

All patents with the protein's chemical modulation data
IBM search result
FN
GVKBIO
TP
FP
Match between GVKBIO & IBM search result

Figure 24. A theoretical Venn diagram showing data coverage of GVKBIO and its use as a calibration for search results from the IBM database. (TP = true positives, FP = false positives, FN = false negatives)

$$\%\,Precision\ = \frac{TP}{\mathrm{TP{+}FP}} \geq \frac{Number\ of\ common\ patents\ between\ GVKBIO\ and\ Search\ Result}{Number\ of\ patents\ in\ the\ Search\ Result} \qquad \text{(Formula 1)}$$

$$\%\,Recall\ = \frac{TP}{TP+FN} \approx \frac{Number\ of\ common\ patents\ between\ GVKBIO\ and\ Search\ Result}{Number\ of\ GVKBIO\ curated\ patents\ for\ a\ particular\ protein} \qquad \text{(Formula 2)}$$

By the same reasoning, the number of false negatives cannot be obtained directly by calibrating the search result against GVKBIO, but the minimum number of false negatives is possible. However, since GVKBIO represents a sample set of all patents with target proteins, an approximate percentage recall can be obtained by using Formula 2 (above). In the case of renin, there are 494 patents in GVKBIO, and 237 of them have renin in titles. Therefore, the approximate percentage recall of this retrieval approach would be 48.0%.

Results show different precision between the two targets. Some differences were due to simple term specificity noise such as kynurenine or antithrombin. This type of false-positive could potentially be removed by using more advanced term recognition rules. For thrombin, some false negatives (w.r.t. GVKBIO) are where some *bona fide* targets for those specific patent documents are, for example, factor Xa or factor VII. (e.g. US20080214495 "heterocyclic sulfonamide derivatives as inhibitors of factor Xa" and US7576098 "heterocyclic compounds as inhibitors of factor VIIa"). The inclusion of cross-screening data against thrombin is not unexpected. This exemplifies cases in which there are data for multiple target proteins in a patent (e.g. factor Xa and thrombin), but the subject and *bona fide* target of the patent may be only one (i.e. factor Xa). There are also some nominal examples of combined targets (e.g. WO2004052851A1, "pyrrolydin-2-one derivatives as inhibitors of thrombin and factor Xa" and EP1294684A2 "thrombin or factor Xa inhibitors") but it is not clear if these represent authentic polypharmacolgy or "claim bet hedging".

Manually inspecting the 1200 patent titles with albumin and 1542 for hemoglobin show they were all, in the *bona fide* target sense, false positives as expected. For albumin, many of them claim methods of conjugation with other proteins (e.g. WO2009121884A1 "insulin albumin conjugates" or EP1745078B1 "method for the purification of albumin conjugates"). Note that Table 17 includes five GVKBIO patents with albumin as a target name. While they could be classified as false-positives in fact they are cases where the chemical modulators of the *bona fide* targets have been tested for albumin binding (e.g. EP1586318 "thiadiazolidinones as GSK-3 inhibitors"). This exemplifies another constitutive challenge for recognition and extraction of target proteins from patents. Most of the hemoglobin applications specify analytical methods (e.g. EP2016390A1 "a method and a system for quantitative hemoglobin determination" or US20090317912 "method of measuring glycated hemoglobin concentration"). Supporting the non-target inferences are the very low frequencies of chemical modulation keywords (e.g. "inhibitor", "modulator", etc.) in titles. Word clouds of the titles retrieved with the four proteins are shown in Figure 25.

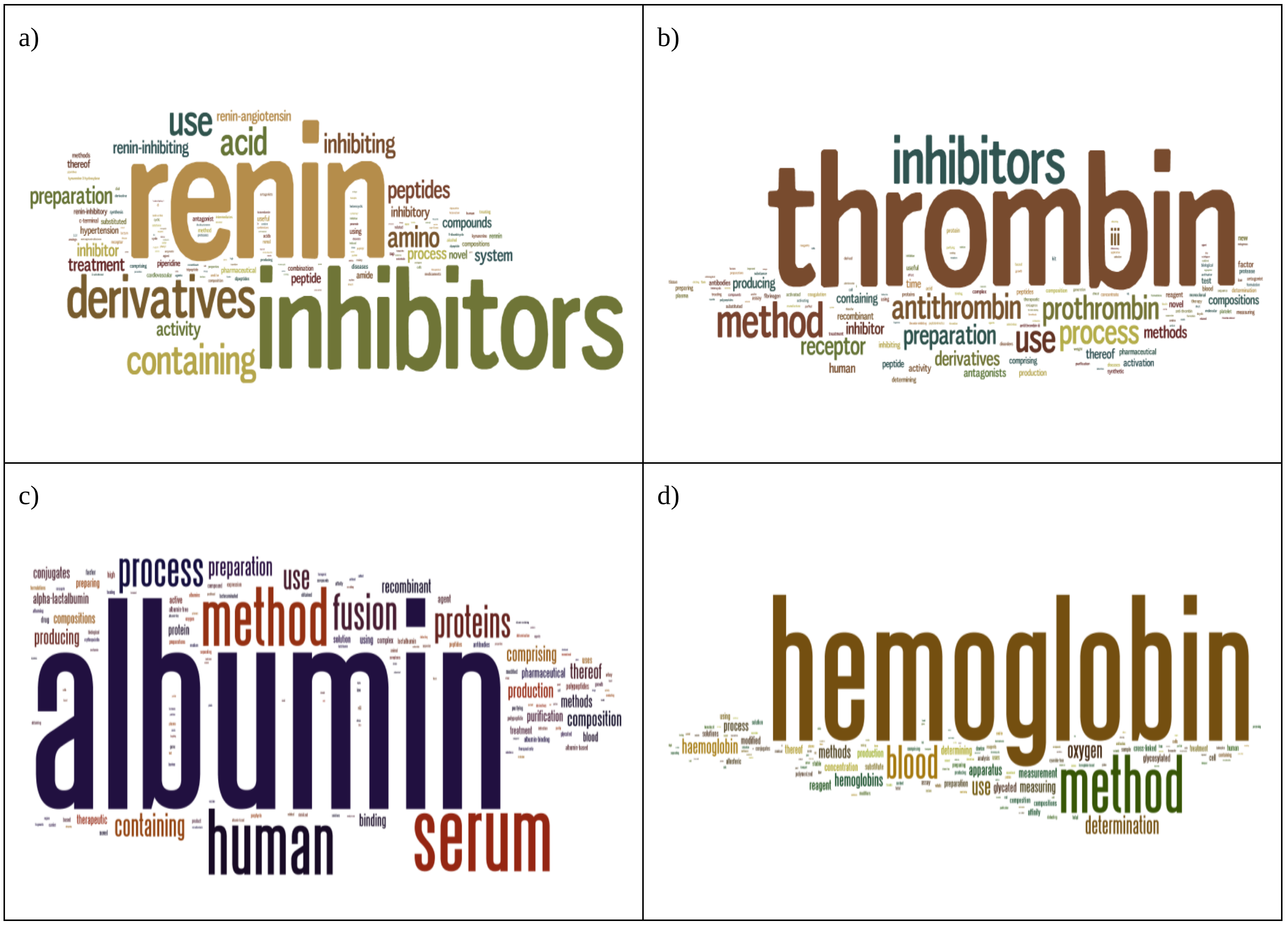


Figure 25. Word clouds of patent titles obtained by searching synonyms of non-target proteins in titles. a) renin, b) thrombin, c) albumin, d) hemoglobin.

One of the conclusions that can be drawn from these results is that using protein names and synonyms alone to search against patent titles is likely to have a high false positive rate for *bona fide* targets. AZ's extensive experience with mining GVKBIO data and other pharmaceutical R&D data sources suggests that *bona fide* targets often show a "time signal" in the sense that the generation rate of published data directly associated with these targets, rather than being constant, will often vary significantly on a year-to-year basis. There are many possible causes of these fluctuations that are difficult to verify formally. However, it is known that declared drug R&D success milestones (e.g. new target validation data, initiation of clinical trials or an NCE submission) invariably trigger some level of follow-on activity that can result in a subsequent "spike" of patent applications. To test this, the frequencies of the four proteins in patent titles from 1980 to 2008 were plotted (Figure 26).

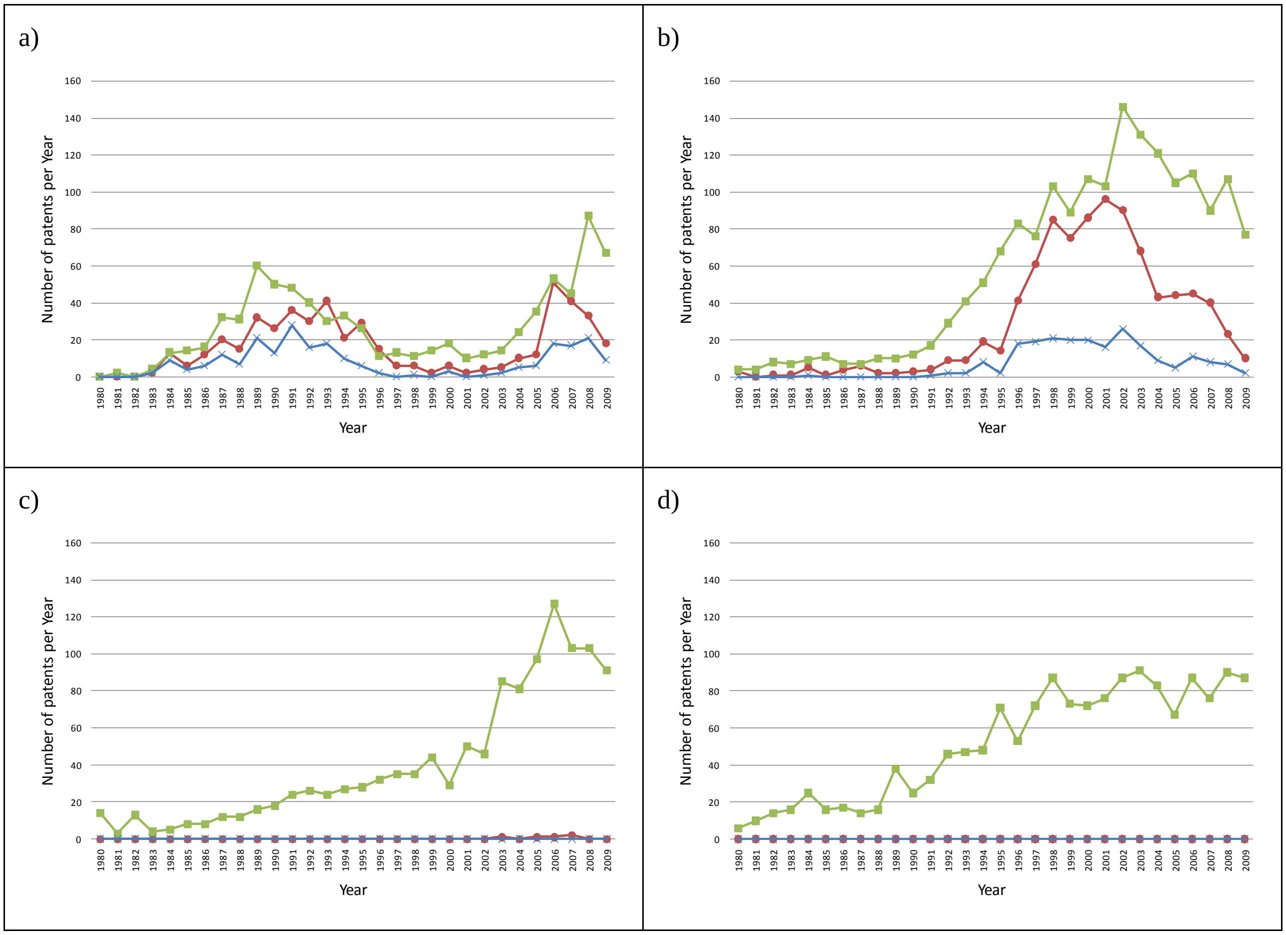


Figure 26. The result of retrieving patents using occurrence of protein synonyms in titles for a) Renin, b) Thrombin, c) Albumin, d) Hemoglobin. (● represents number of GVKBIO patents tagged with a particular protein name, ■ represents number of IBM patents obtained from search, × represents number of patents-in-common between GVKBIO and IBM search results)

It can be seen (Figure 26 (c) and (d)) that the two non-targets show a steady increase. Nonetheless there is suggestion of a peak for albumin at 2006 -2007, although as already explained this may be new-use related but not target related. In contrast, the two targets (Figure 26 (a) and (b)) not only show strong signals but that these appear to be correlated between GVKBIO and IBM. As these are selected by curated target in the former case this suggests the signal in the latter case may be authentic in representing a significant increase in patent publications for these *bone fide* targets.

3.5 Protein names in titles, abstracts and claim sections

3.5.1 Assessment of improved recall by extending search to abstracts and claim sections

Relevant to the problem of target-containing patent retrieval, exploratory queries suggested that extending searching for protein synonyms in abstracts and claim sections might retrieve patents where target names are absent from titles. This promisingly results in improvement of the overall recall performance, with tradeoff of possibly increasing false positives. The objective of this section to explore the improved recall performance by this extended search in titles abstracts and claim sections. The increasing false positives of this extended search will be assessed in the next section (3.4.2).

From the same set of GVKBIO used in section 3.2 (34,411 patents), selected 8,167 patents were published between years 2006 – 2009 where IBM had full-text documents. The reason from filtering by more recent publication dates is that these showed a better quality of text extraction for abstracts and claims that older patents, due to the availability of direct XML feeds from the patent offices. From these, 79 renin and 80 thrombin patents were selected. Within each of these patent sets, occurrences of the protein synonyms in titles, abstracts and claim sections were searched. The results are shown below (Figure 27).

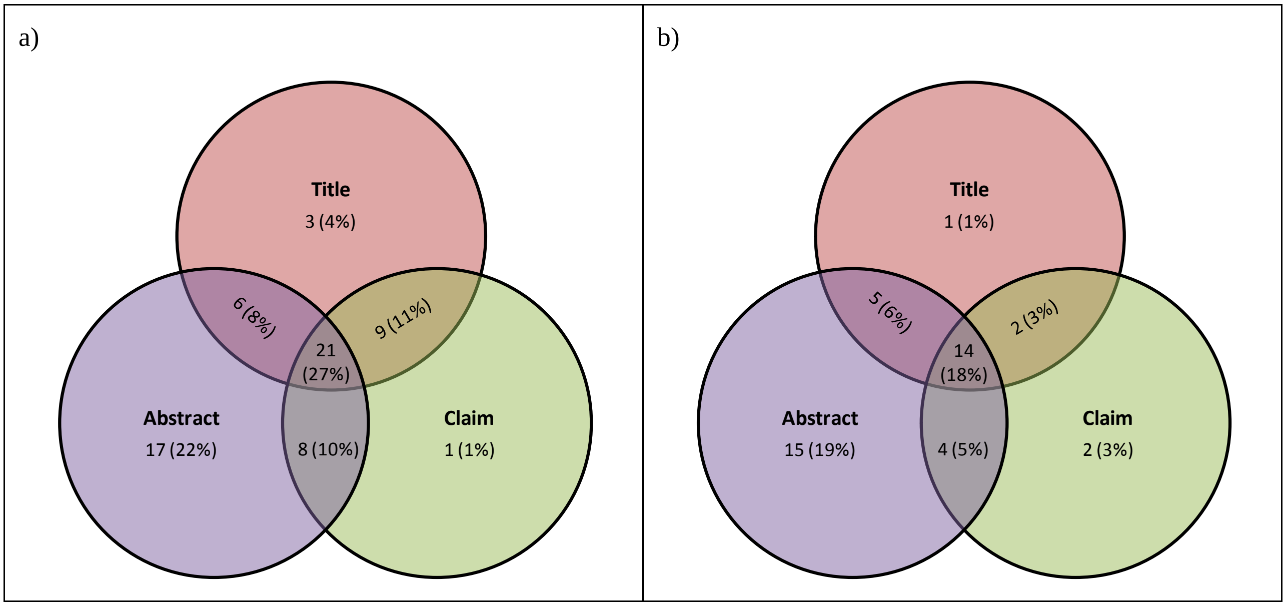


Figure 27. Venn diagram showing containment of target protein synonyms in titles, abstracts and claim sections (counted by patent number) for a) Renin – collectively contains 65 patents (82%) out of 79 sample patents, b) Thrombin – collectively contains 43 patents (54%) out of 80 sample patents.

For 79 renin patents (Figure 27 (a)), searching for the protein synonym could retrieve 39 (49%) in title, 52 (66%) in abstract, 39 (49%) in claim section, and 65 patents (82%) in all three sections. Interestingly, searching in abstracts retrieved 17 unique patents (22%) not found by searching in titles or claim sections. The equivalent figures for the 80 thrombin patents (Figure 27 (b)) were 22 (28%) in title, 38 (48%) in abstract, 22 (28%) in claim section, and 43 patents (54%) in all sections. In this case searching in abstracts was also shown to retrieve 15 unique patents (19%). These show that extending search from titles to abstracts and claim sections could significantly improve the retrieval coverage, mostly due to unique information found in patent abstracts. These two cases suggested that abstracts contain more unique information as compared to claim sections. This supports the fact that pharmaceutical patents usually claim chemical structures for treatment of disease without mentioning target protein names. However, this hypothesis was later rejected in subsequent analysis in this section.

Significantly, the recall percentage by using these three patent sections for renin (82%) and thrombin (54%) are different. This results from the fact that the set of 80 thrombin patents contains both 1) patents with thrombin as a *bona fide* target, and 2) patents with thrombin as a cross-screened target. The latter case usually mention thrombin in their body descriptions, but not in titles, abstracts and claim sections. These are mostly claiming chemical modulation for factor Xa and factor VIIa as primary targets which are usually cross-screened with thrombin. In contrast, the set of 79 renin patents contains mostly patents with renin as a *bona fide* target; since renin is less often cross-screened. Frequent occurrences of the renin synonyms in these three patent sections as compared to thrombin synonyms suggests that 1) *bona fide* target names are likely to show in these sections, 2) other targets are likely to hide in body descriptions.

To confirm the understanding on occurrence of *bona fide* targets and other targets mentioned earlier, these sets of sample patents were extended to include more documents and target protein names. From the same set of 8,167 GVKBIO patents (published between 2006-2009), there were 7,648 selected patents in which synonyms for its curated target protein available in the dictionary complied in section 3.2. From these, a sample set of patents with only *bona fide* targets was created by 4,324 patents with only one target (4,324 document-target links). This set encompasses 466 *bona fide* targets, and will be subsequently referred to as "*bona fide* target patents". On the other hand, from the same set of 7,648 patents, a sample set of patents with targets in general was compiled from all 7,648 patents with one or more targets (16,860 document-target links). This set encompasses 921 targets, and will be subsequently referred to as "mixed-target patents".

Within each of these patent sets, occurrences of their target protein synonyms in titles, abstracts and claim sections were searched. The results are shown below (Figure 28).

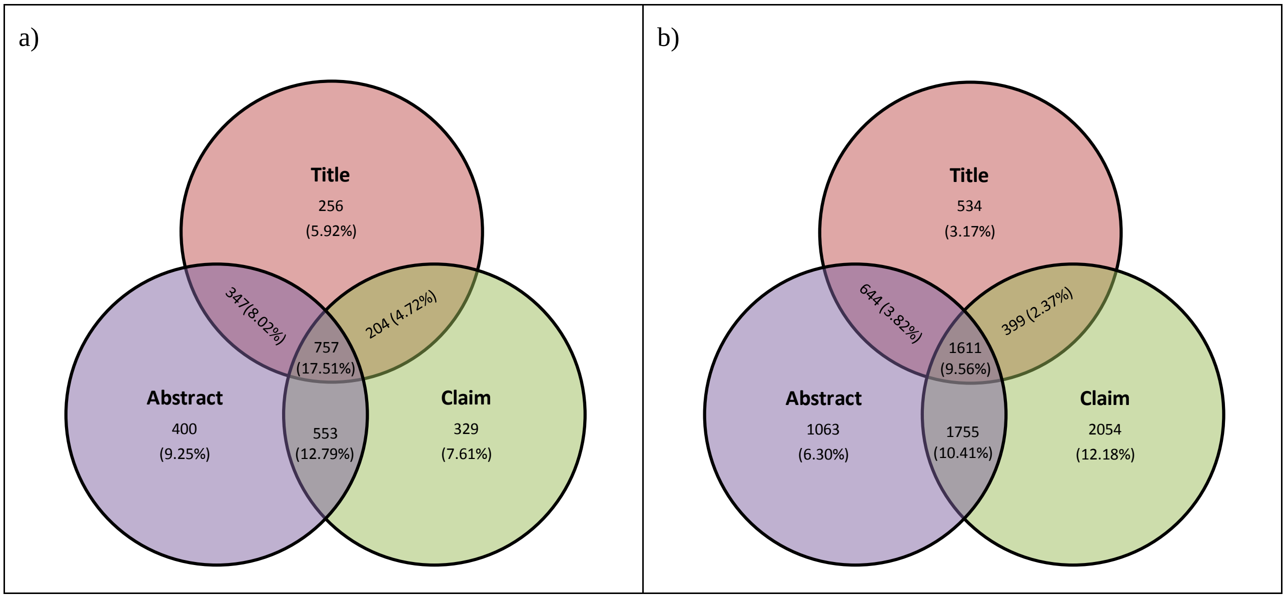


Figure 28. Venn diagram showing containment of target protein synonyms in titles, abstracts and claim sections (in each patent) for a) GVKBIO *bona fide* target patents – collectively contains 2846 document-target links (65.8%) out of 4324 patent-target links, (b) GVKBIO mixed-target patents – collectively contains 8060 document-target links (47.8%) out of 16860 document-target links.

For *bona fide* target patents (Figure 28 (a)), searching for the protein synonym could retrieve 1,564 (36.2%) in title, 2,057 (47.6%) in abstract, 1,843 (42.6%) in claim section, and 2,846 (65.8%) in all three sections. For mixed-target patents (Figure 28 (b)), equivalent figures were 3,188 (18.9%) in title, 5,073 (30.1%) in abstract, 5,819 (34.5%) in claim section, and 8,060 (47.8%) in all three sections. The difference in recall percentage between these two patent sets does support the hypothesis that *bona fide* targets are often shown in these three sections; while other targets are often in body descriptions. Interestingly, for these two patent sets, the level of unique information found in abstracts is not significantly higher than in claim sections. Furthermore, claim section seems to have high content of general target protein mentions as compared to abstracts (Figure 28 (b)).

While comparing results in Figure 27 and Figure 28, it shows that there is a drop in overall recall for the result in Figure 28, this could result from imperfect ability to recognize all target protein names mentioned in each patent texts (incomplete coverage of the protein name dictionary). Nevertheless, in particular to the result shown in Figure 28, the relative information coverage between abstracts and claim sections seems to be reliable; due to 1) the result was performed on a larger sample set as compared to Figure 27, and 2) incomplete recognition of protein names applied to abstracts and claim sections indifferently. Therefore the hypothesis that abstracts contain more unique information as compared to claim sections was rejected.

3.5.2 Assessment of potential false positives from extending search to abstracts and claim sections

To examine the extent of false-positives from protein name searching in abstracts and claim sections, the same target and non-target pairs were used (i.e. renin, thrombin, albumin and hemoglobin)(Table 16). These protein synonyms were searched in titles, abstracts and claim sections in IBM for all USPTO patents published between 2006 – 2009 (1,234,684 patents). Results for each protein name were then calibrated with all USPTO patents found in the same set of GVKBIO explained in section 3.2 (34,441 patents). The results are shown in Figure 29.

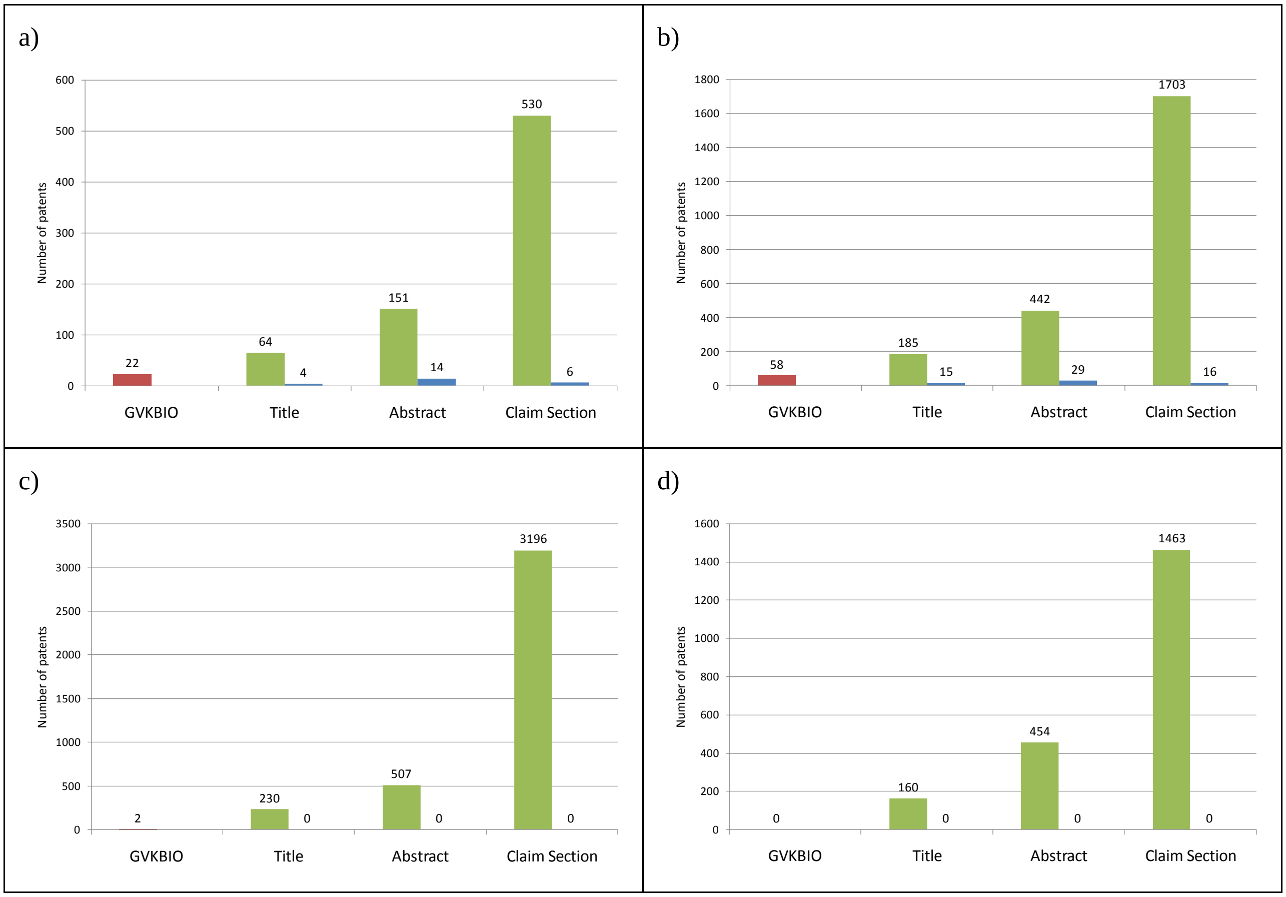


Figure 29. The result of retrieving patents with a particular target protein using occurrence of the protein names in titles, abstracts and claim sections for a) renin, b) thrombin, c) albumin, d) hemoglobin. (● number of GVKBIO curated patents, ■ number of IBM search result, ✕ number of patents-in-common between GVKBIO and IBM)

Results from albumin and hemoglobin (Figure 29 (c) and (d)) suggest that searching protein names in abstracts and claims leads to substantial increases in false positives in claim sections. Similarly, renin and thrombin (Figure 29 (a) and (b)) also show substantial increase of patents unmatched by GVKBIO, which could potentially be false-positives. Moreover, comparing the matches between IBM and GVKBIO suggests abstracts show the highest recall of target-containing patents in line with the results from section 3.4.

3.6 Selection and evaluation of filters

It has been shown in previous sections that retrieving patents with targets by using only protein names results in substantial amount of false positives (i.e. patents without target proteins). The aim of this section was to select and evaluate possible filters that could be used to remove these false positives, while retaining overall recall performance. In addition, the filtration could also reduce search space from ~10 million patents in IBM into a smaller set containing enriched pharmaceutical patents.

Four potential filters were selected for evaluation were 1) IPC codes[4], 2) the number of chemical in claim section, 3) the number of chemical in patent, and 4) chemical modulation keywords.

3.6.1 IPC codes

IPC codes[4] were intentionally tagged and revised by patent offices. Several sets of IPC subclasses relevant to the pharmaceutical industry are suggested by the WIPO as shown in Table 18 [44]. However, it was shown that these IPC subclasses were not specific only to patents with target chemical modulation. They also cover other kinds of patents such as C07B (general method for organic chemistry; apparatus therefore). Most importantly these codes still cover large amount of false positives found in section 3.4.2 for the non-target albumin and hemoglobin patents.

Table 18. IPC subclasses relevant to the pharmaceutical industry [44].

Field of Technology	IPC Codes (subclasses)
Organic fine chemistry	(C07B, C07C, C07D, C07F, C07H, C07J, C40B) not A61K, A61K-008, A61Q
Biotechnology	(C07G, C07K, C12M, C12N, C12P, C12Q, C12R, C12S) not A61K
Pharmaceuticals	A61K not A61K-008

Due to poor specificity of the set of IPC codes provided by WIPO, it was decided to extract a set of IPC codes from the GVKBIO which had high content of patents with targets. From the set of 42,575 GVKBIO patents common to IBM (Figure 6), IPC codes were extracted from 42,192 patents in which IPC code information were available in IBM. The result shows that there were 176 IPC codes in GVKBIO. Top 20 most frequent IPC codes are shown in Table 19.

Figure 30 shows uses of these IPC codes in GVKBIO over time. Note that multiple IPC classes are included in individual patent documents. Therefore, there is an overlap between patents having A61K and patents having C07D, for example. Further analysis shows that top five IPC codes (i.e. A61K,C07D,A61P,C07C,C07K) appeared to cover

[4] IPC codes (International Patent Classification codes) are alphanumerical symbols indicated on each patent document for facilitating retrieval of "prior art". Such retrieval is needed by patent-issuing authorities, potential inventors, research and development units, and others concerned with the application or development of technology. IPC divides technology into eight sections with approximately 120 classes and 640 subclasses. Example of subclasses relevant to pharmaceutical industry is A61K (preparation for medical, dental, or toilet purposes) and C07D (heterocyclic compounds). (URL: www.wipo.int/classifications/ipc/)

97.68% of all GVKBIO sample patents (41,216 out of 42,192 sample patents). Notice that these five IPC codes fall into three WIPO fields of technology in Table 18.

Table 19. Most frequent IPC codes used in GVKBIO patents.

IPC Code (Subclasses)	#Patent (Total = 41296 Patents)		
	in Main&Sub IPC	in Main IPC	in Sub IPC
A61K	81.19%	73.81%	19.93%
C07D	71.44%	58.54%	20.84%
A61P	40.13%	38.58%	3.19%
C07C	10.66%	9.16%	2.08%
C07K	5.53%	4.57%	1.30%
C07F	2.68%	2.34%	0.51%
C12N	2.24%	1.95%	0.43%
G01N	1.99%	1.78%	0.36%
A01N	1.70%	1.39%	0.59%
C07H	1.42%	1.05%	0.42%
C12Q	1.32%	1.19%	0.22%
C12P	0.66%	0.51%	0.19%
C07B	0.63%	0.56%	0.07%
C07J	0.54%	0.45%	0.12%
A01K	0.16%	0.13%	0.03%
C12R	0.15%	0.00%	0.14%
A61Q	0.14%	0.14%	0.00%
A61L	0.13%	0.11%	0.02%
A61F	0.13%	0.11%	0.03%
G06F	0.11%	0.08%	0.04%

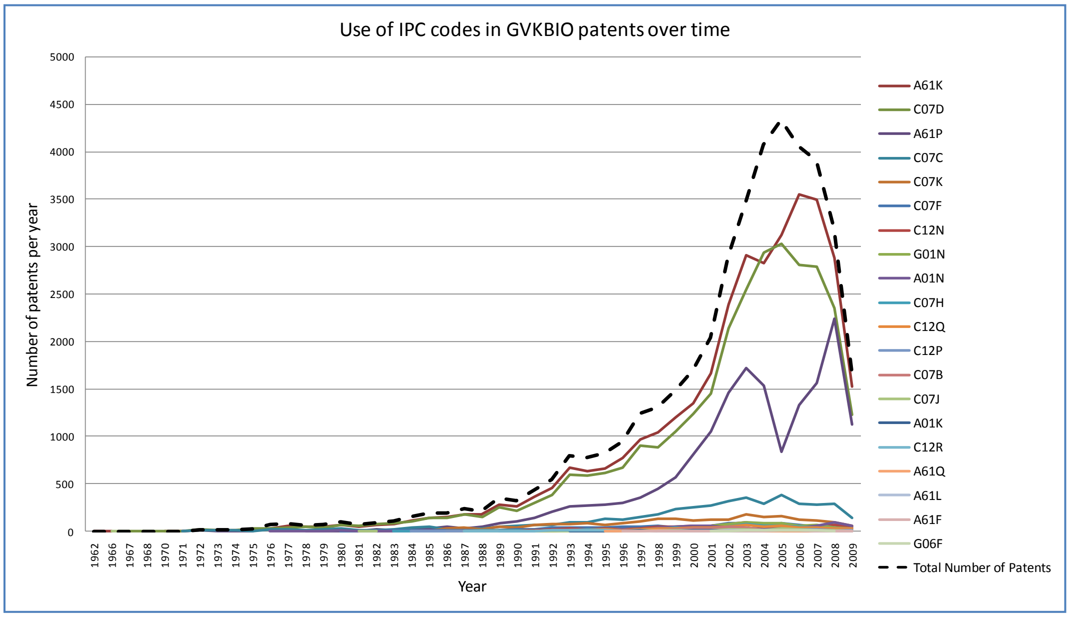
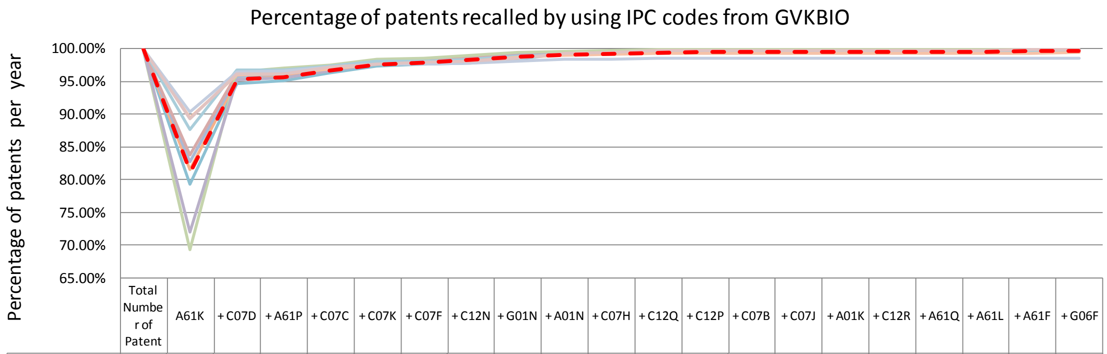


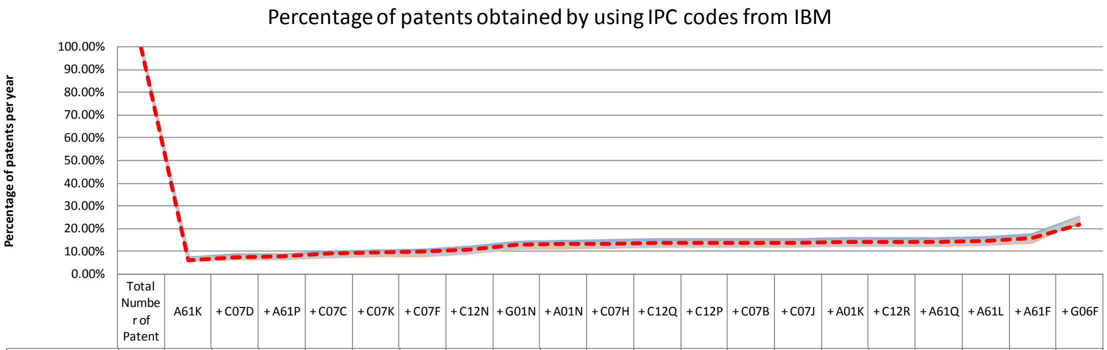
Figure 30. Use of IPC codes in GVKBIO patents over time.

From the list of top 20 IPC codes specific to patents with targets in Table 19, it was necessary to select a set of IPC codes as patent retrieval criteria and address both recall and precision performance. Twenty sets of IPC codes were proposed for further evaluation. These sets are 1) {A61K}, 2) {A61K,C07D}, 3) {A61K,C07D,A61P}, and so on. The recall analysis of these IPC sets was performed on GVKBIO (42,192 patents with IPC code information, published in 1973-2009) (Figure 31). Precision analysis was performed on IBM (11,336,265 patents published in 1920-2009)(Figure 32).



	Total Number of Patent	A61K	+ C07D	+ A61P	+ C07C	+ C07K	+ C07F	+ C12N	+ G01N	+ A01N	+ C07H	+ C12Q	+ C12P	+ C07B	+ C07J	+ A01K	+ C12R	+ A61Q	+ A61L	+ A61F	+ G06F
2000	100.00	79.34%	94.64%	95.06%	96.29%	97.29%	97.65%	97.82%	98.29%	99.23%	99.35%	99.71%	99.76%	99.76%	99.82%	99.82%	99.82%	99.82%	99.82%	99.88%	99.88%
2001	100.00	81.64%	95.25%	95.79%	96.72%	97.70%	97.80%	98.19%	98.68%	99.41%	99.46%	99.61%	99.66%	99.66%	99.71%	99.71%	99.71%	99.71%	99.71%	99.71%	99.76%
2002	100.00	82.72%	95.75%	96.20%	96.89%	97.79%	97.96%	98.48%	98.93%	99.27%	99.34%	99.65%	99.65%	99.65%	99.69%	99.69%	99.69%	99.69%	99.69%	99.72%	99.72%
2003	100.00	83.87%	96.17%	96.40%	96.89%	98.21%	98.33%	98.82%	99.28%	99.45%	99.54%	99.77%	99.77%	99.77%	99.86%	99.86%	99.86%	99.86%	99.86%	99.88%	99.91%
2004	100.00	69.40%	96.41%	97.00%	97.54%	98.38%	98.50%	98.94%	99.39%	99.51%	99.71%	99.80%	99.80%	99.80%	99.83%	99.83%	99.83%	99.83%	99.85%	99.85%	99.90%
2005	100.00	71.98%	95.22%	95.52%	96.61%	97.60%	97.83%	98.32%	98.98%	99.17%	99.33%	99.47%	99.49%	99.52%	99.65%	99.68%	99.68%	99.68%	99.70%	99.84%	99.84%
2006	100.00	87.66%	96.74%	96.79%	97.33%	98.03%	98.10%	98.42%	98.82%	99.11%	99.16%	99.41%	99.48%	99.48%	99.56%	99.56%	99.56%	99.56%	99.58%	99.70%	99.78%
2007	100.00	89.53%	96.03%	96.13%	96.90%	97.59%	97.72%	98.03%	98.64%	99.00%	99.05%	99.23%	99.26%	99.28%	99.28%	99.28%	99.28%	99.28%	99.28%	99.33%	99.36%
2008	100.00	90.32%	96.40%	96.55%	97.12%	97.59%	97.62%	97.84%	98.09%	98.31%	98.40%	98.50%	98.50%	98.50%	98.50%	98.50%	98.50%	98.50%	98.53%	98.56%	98.56%
2009	100.00	89.37%	96.12%	96.18%	97.18%	97.71%	97.88%	98.12%	98.47%	99.24%	99.47%	99.76%	99.76%	99.76%	99.76%	99.76%	99.76%	99.76%	99.76%	99.82%	99.82%
(1962-2009)	100.00	81.19%	95.41%	95.62%	96.77%	97.68%	97.90%	98.34%	98.76%	99.13%	99.29%	99.47%	99.51%	99.51%	99.57%	99.58%	99.58%	99.58%	99.59%	99.64%	99.66%

Figure 31. GVKBIO patents retrieved by using different sets of IPC codes ({A61K}, {A61K,C07D}, …).



	Total Number of Patent	A61K	+ C07D	+ A61P	+ C07C	+ C07K	+ C07F	+ C12N	+ G01N	+ A01N	+ C07H	+ C12Q	+ C12P	+ C07B	+ C07J	+ A01K	+ C12R	+ A61Q	+ A61L	+ A61F	+ G06F
2000	100.00%	6.88%	8.09%	8.14%	9.41%	10.20%	10.49%	11.80%	13.88%	14.33%	14.54%	14.82%	14.92%	14.94%	14.96%	15.31%	15.31%	15.32%	15.77%	17.07%	22.16%
2001	100.00%	7.01%	8.10%	8.18%	9.38%	10.16%	10.40%	11.67%	13.71%	14.12%	14.33%	14.67%	14.79%	14.81%	14.82%	15.12%	15.12%	15.12%	15.55%	16.79%	22.73%
2002	100.00%	6.97%	7.98%	8.05%	9.09%	9.83%	10.05%	11.38%	13.61%	13.96%	14.19%	14.55%	14.66%	14.67%	14.69%	15.01%	15.01%	15.01%	15.42%	16.62%	24.32%
2003	100.00%	7.35%	8.34%	8.41%	9.37%	10.38%	10.61%	11.96%	14.32%	14.66%	14.90%	15.29%	15.39%	15.40%	15.41%	15.72%	15.72%	15.73%	16.16%	17.46%	25.17%
2004	100.00%	6.82%	7.88%	7.97%	8.85%	9.62%	9.82%	11.02%	13.27%	13.56%	13.76%	14.19%	14.29%	14.30%	14.32%	14.61%	14.61%	14.61%	15.03%	16.36%	24.20%
2005	100.00%	6.69%	7.70%	7.76%	8.53%	9.24%	9.43%	10.59%	12.87%	13.15%	13.35%	13.82%	13.91%	13.92%	13.94%	14.26%	14.26%	14.26%	14.67%	16.00%	24.70%
2006	100.00%	6.31%	7.03%	7.06%	7.79%	8.39%	8.58%	9.59%	11.59%	11.85%	12.02%	12.53%	12.63%	12.63%	12.65%	12.95%	12.96%	12.98%	13.32%	14.62%	22.98%
2007	100.00%	6.20%	6.89%	6.92%	7.64%	8.22%	8.40%	9.39%	11.39%	11.67%	11.82%	12.31%	12.42%	12.42%	12.43%	12.70%	12.70%	12.73%	13.10%	14.41%	22.67%
2008	100.00%	6.10%	6.84%	6.87%	7.65%	8.18%	8.37%	9.33%	11.40%	11.69%	11.85%	12.28%	12.41%	12.42%	12.43%	12.68%	12.69%	12.71%	13.08%	14.31%	23.40%
2009	100.00%	6.33%	7.10%	7.14%	7.90%	8.46%	8.65%	9.73%	11.86%	12.17%	12.35%	12.83%	12.99%	12.99%	13.00%	13.25%	13.25%	13.27%	13.67%	14.88%	24.62%
(1920-2009)	100.00%	6.09%	7.55%	7.59%	8.86%	9.45%	9.71%	10.72%	12.86%	13.20%	13.38%	13.71%	13.82%	13.83%	13.86%	14.19%	14.19%	14.20%	14.57%	15.72%	21.90%

Figure 32. IBM patents retrieved by using different sets of IPC codes ({A61K}, {A61K,C07D}, …).

The recall analysis (Figure 31) shows that using the IPC code A61K could retrieve 81.19% of GVKBIO patents, the set of two IPC codes (i.e. {A61K,C07D}) could retrieve 95.41% of GVKBIO patent. This due to the fact that the majority of pharmaceutical patents are falling under these two subclasses (A61K – "preparation for medical, dental, or toilet purposes", C07D – "heterocyclic compounds"). Extending these two IPC code to include more IPC codes appear to improve recall gradually increase until reaching maximum of 99.66%. Furthermore, analysis in different patent publication years shows that these combination of IPC codes were giving comparable recall performance (e.g. For {A61K,C07D}, percentage recall were 94.64% for patents published in 2000, and 96.12% for 2009).

The precision analysis (Figure 32) shows that using just two IPC codes (i.e. {A61K,C07D}) could reduce the IBM search space down to 7.55%. By assuming that factors causing false positives are distributed uniformly across the IBM patent documents, it was possible to infer that 92.45% of false positive patents were removed by using these two IPC codes. For instance, searching for thrombin by using the term "F2" could hit non-pharmaceutical patents (e.g. US20020015431 "F2-laser with line selection"). By using the filtration by IPC codes, these types of false positives patents were removed.

Although the assumption of false positive reduction being proportional to the search space reduction seems to be plausible, there were no other ways to measure precision improvement because there was no false positive patent corpus. Therefore, this assumption will be used for the purpose of filter section and evaluation. However, later in this report when this filter was applied to actual queries, it was shown that false positive factors are more likely to skew towards some patents, rather than being uniformly distributed across the IBM database.

To select the set of IPC codes with high false positive reduction while retaining recall, results in Figure 31 and Figure 32 were transformed and compared in terms of "reduced recall percentage" and "reduced false positive percentage" as shown in Figure 33 (a). In order to facilitate the comparison between sets of IPC codes, a scoring index called "filter utility ratio" shown in Formula 3 (below) was proposed. Results in Figure 33 (a) were then calculated for their filter utility ratios as shown in Figure 33 (b).

$$Filter\ Utility\ Ratio\ = \frac{Reduced\ False\ Positives\ (\%)}{Reduced\ Recall\ (\%)} \approx \frac{Reduced\ Search\ Space\ (\%)}{Reduced\ Recall\ (\%)} \qquad \text{(Formula 3)}$$

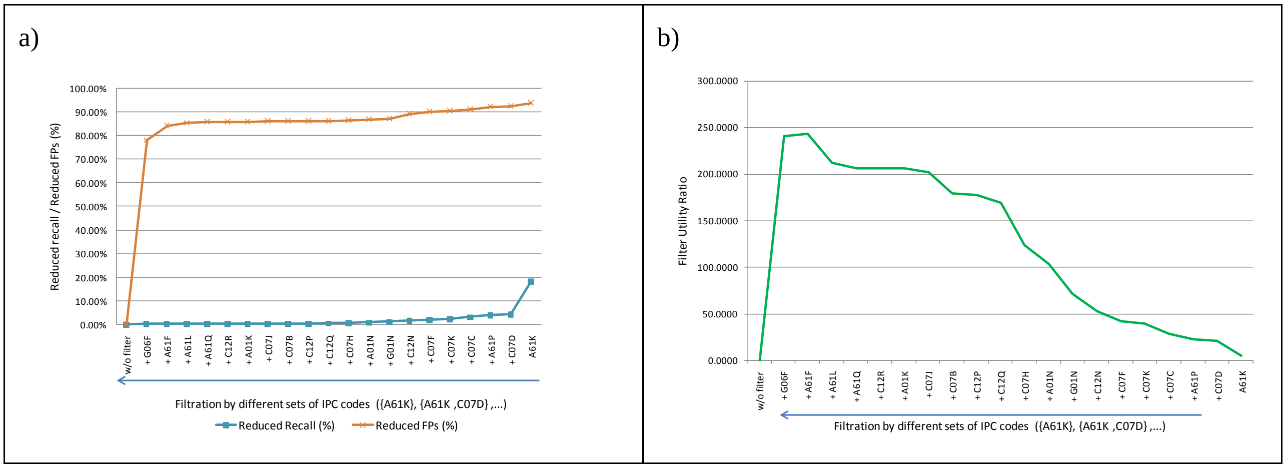


Figure 33. Comparison of false positive reduction and recall reduction that could be obtained by different filtration criteria (sets of IPC codes). a) False positive reduction (%) versus recall reduction (%), b) Filter utility ratios.

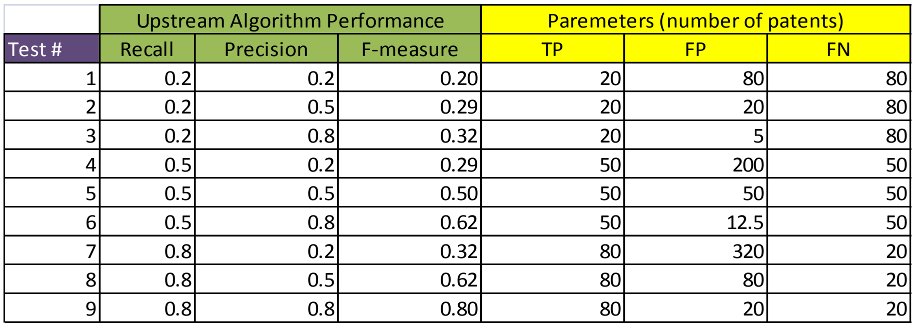
Looking at the filter utility ratios for different sets of IPC codes (Figure 33 (b)), the set of 19 codes (i.e. {A61K,C07D,A61P,...,A61F}) turned to be the best filter. This could remove 84.28% of false positives (or search space reduction, to be precise), while retaining recall at 99.64%.

It might be arguable that the filter utility ratio is not reliable. Nevertheless, using F-measure (Formula 4) to consolidate recall and precision scores led to complication in calculation. Although a constant number for percentage recall specific to each filtration criteria was known (Figure 31), percentage precision is inconstant and depending on recall and precision of the upstream algorithm (i.e. in a sequence of combined filtration criteria). Therefore the percentage precision for each filtration criteria cannot be obtained directly from the data shown in Figure 32. To elaborate, for the precision calculation (Formula 1), absolute numbers of true positives and false positives need to be known. However, Figure 32 only provides a constant ratio for false positive reduction for each filtration criteria, but not the absolute number of false positives. Furthermore, the number of true positives is dependent on content of the search space the filter being applied; resulting in inconstant percentage precision for each filtration criteria.

$$\% \text{ F} - \text{measure } = \frac{2 \cdot Recall \cdot Precision}{Recall \cdot Precision} \qquad \text{(Formula 4)}$$

As a consequence, the calculation of F-measure scores for these filters could only be done by providing starting numbers of 1) true positives (TP), 2) false positives (FP), and 3) false negatives (FN) produced from upstream retrieval algorithm (i.e. searching for protein names in titles). Example sets of TP, FP, and FN parameters assumed for upstream algorithms with various recall and precision are shown in Table 20. For each parameter set, a calculation of F-measure score then can be done by combining data from Figure 31 (true positive reduction) and Figure 32 (false positive reduction). Example of the calculation for the testing parameter no. 1 (Table 20) is shown in Table 21. Calculation results for all parameter sets are shown in Figure 34.

Table 20. Parameter seeding for upstream algorithms with different recall and precision.



	Upstream Algorithm Performance			Paremeters (number of patents)		
Test #	Recall	Precision	F-measure	TP	FP	FN
1	0.2	0.2	0.20	20	80	80
2	0.2	0.5	0.29	20	20	80
3	0.2	0.8	0.32	20	5	80
4	0.5	0.2	0.29	50	200	50
5	0.5	0.5	0.50	50	50	50
6	0.5	0.8	0.62	50	12.5	50
7	0.8	0.2	0.32	80	320	20
8	0.8	0.5	0.62	80	80	20
9	0.8	0.8	0.80	80	20	20

Table 21. Example of F-measure score calculation for an upstream algorithm with 20% recall and 20% precision.

Filters	GVKBIO - Recall %		IBM - False Positives (%)		1							
IPC Codes	#Patents	#Patents (%)	#Patents	#Patents (%)	TP	FP	FN	Positive	Actual	Recall	Precision	F-measure
w/o filter	37164	100.00%	11336265	100.00%	20.0000	80.0000	80.0000	100.0000	100.0000	0.2000	0.2000	0.2000
+ G06F	37043	99.67%	2483200	21.90%	19.9349	17.5239	80.0651	37.4588	100.0000	0.1993	0.5322	0.2900
+ A61F	37035	99.65%	1782181	15.72%	19.9306	12.5768	80.0694	32.5074	100.0000	0.1993	0.6131	0.3008
+ A61L	37014	99.60%	1651596	14.57%	19.9193	11.6553	80.0807	31.5746	100.0000	0.1992	0.6309	0.3028
+ A61Q	37009	99.58%	1609234	14.20%	19.9166	11.3564	80.0834	31.2729	100.0000	0.1992	0.6369	0.3034
+ C12R	37009	99.58%	1608299	14.19%	19.9166	11.3498	80.0834	31.2663	100.0000	0.1992	0.6370	0.3035
+ A01K	37009	99.58%	1608197	14.19%	19.9166	11.3490	80.0834	31.2656	100.0000	0.1992	0.6370	0.3035
+ C07J	37005	99.57%	1570736	13.86%	19.9144	11.0847	80.0856	30.9991	100.0000	0.1991	0.6424	0.3040
+ C07B	36985	99.52%	1568129	13.83%	19.9037	11.0663	80.0963	30.9700	100.0000	0.1990	0.6427	0.3039
+ C12P	36983	99.51%	1566761	13.82%	19.9026	11.0566	80.0974	30.9592	100.0000	0.1990	0.6429	0.3040
+ C12Q	36974	99.49%	1553800	13.71%	19.8978	10.9652	80.1022	30.8629	100.0000	0.1990	0.6447	0.3041
+ C07H	36904	99.30%	1517083	13.38%	19.8601	10.7061	80.1399	30.5661	100.0000	0.1986	0.6497	0.3042
+ A01N	36852	99.16%	1496526	13.20%	19.8321	10.5610	80.1679	30.3931	100.0000	0.1983	0.6525	0.3042
+ G01N	36708	98.77%	1457657	12.86%	19.7546	10.2867	80.2454	30.0413	100.0000	0.1975	0.6576	0.3038
+ C12N	36536	98.31%	1214762	10.72%	19.6620	8.5726	80.3380	28.2346	100.0000	0.1966	0.6964	0.3067
+ C07F	36365	97.85%	1100807	9.71%	19.5700	7.7684	80.4300	27.3384	100.0000	0.1957	0.7158	0.3074
+ C07K	36303	97.68%	1071451	9.45%	19.5366	7.5612	80.4634	27.0979	100.0000	0.1954	0.7210	**0.3074**
+ C07C	35971	96.79%	1004360	8.86%	19.3580	7.0878	80.6420	26.4457	100.0000	0.1936	0.7320	0.3062
+ A61P	35641	95.90%	860042	7.59%	19.1804	6.0693	80.8196	25.2497	100.0000	0.1918	0.7596	0.3063
+ C07D	35553	95.67%	856230	7.55%	19.1330	6.0424	80.8670	25.1754	100.0000	0.1913	0.7600	0.3057
A61K	30355	81.68%	689978	6.09%	16.3357	4.8692	83.6643	21.2049	100.0000	0.1634	0.7704	0.2696

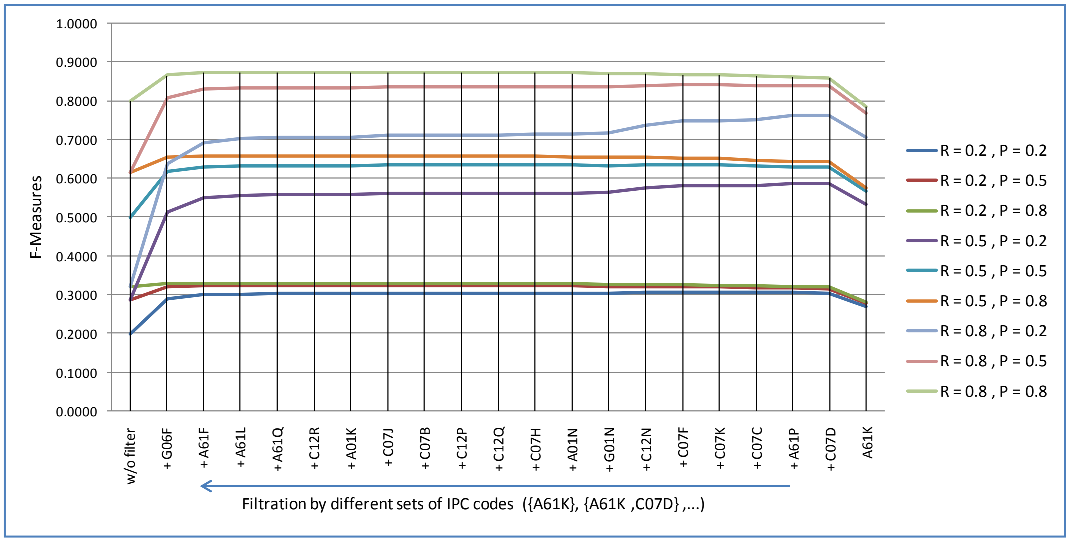


Figure 34. Predictive improvement in F-measures by different filtration criteria (sets of IPC codes) for different starting recall and precision of upstream algorithms.

Results shown in Figure 34 emphasize the fact that different filtration criteria suit different situations. For instance, in cases of upstream algorithms with poor precision (P=0.2), sets of IPC codes with best F-measures contained 3-5 IPC codes (i.e. {A61K,…,A61P}, {A61K,…,C07C}, {A61K,…,C07C}), which are relatively stringent filters. On the other hand, cases of upstream algorithms with high precision (P=0.8), sets of IPC codes with best F-measures were containing 14 IPC codes (i.e. {A61K,…,C07J}), which are relatively lower stringent filters. These implies that if there were high proportion of false positives produced from an upstream algorithm (P=0.2), a stringent filter should be used to remove then. Likewise, if there were a low proportion of false positives left by an upstream algorithm (P=0.8), a lower stringent filter could be used to retain recall percentage. These are concerns that should be pragmatically considered when applying a filter in combination with other algorithms (e.g. searching for protein names in titles).

3.6.2 Chemicals in claim section

Relevant to the retrieval of patents with target chemical modulation data, there was a heuristic that each of these patents should contain at least one chemical structure in the claim section (i.e. a patent with target should claim at least a chemical structure modulating the target protein). Therefore, the existence of a chemical in claim section is a potential filter for selecting target-containing patents. In this section, the objective was to evaluate and select the best number of chemicals in claim section to use as filtration criteria. In order to do this, recall and precision analysis were performed for different numbers of chemicals in claim by using the same methodology as for the analysis of IPC codes (section 3.6.1). This was performed on a subset of GVKBIO patents with targets in which their claim sections and chemical annotations were available in IBM (33,108 patents, published in 1973-2009). Precision analysis was performed on IBM patents in which claim sections were available (9,140,527 patents published in 1790-2009).

Firstly, the subset of GVKBIO was investigated for the number of GVKBIO-curated chemicals - extracted from any patent sections (Figure 35). The chemical structures curated in GVKBIO are those having target modulation data. The result shows that every GVKBIO patents has at least one target-modulating chemical structure. Therefore, a patent without a target-modulating chemical could be filtered without affecting recall. Nevertheless, because IBM automated extraction cannot specifically identify target-modulating chemicals, filtering patents by the number of chemicals in the claim section could be a surrogate.

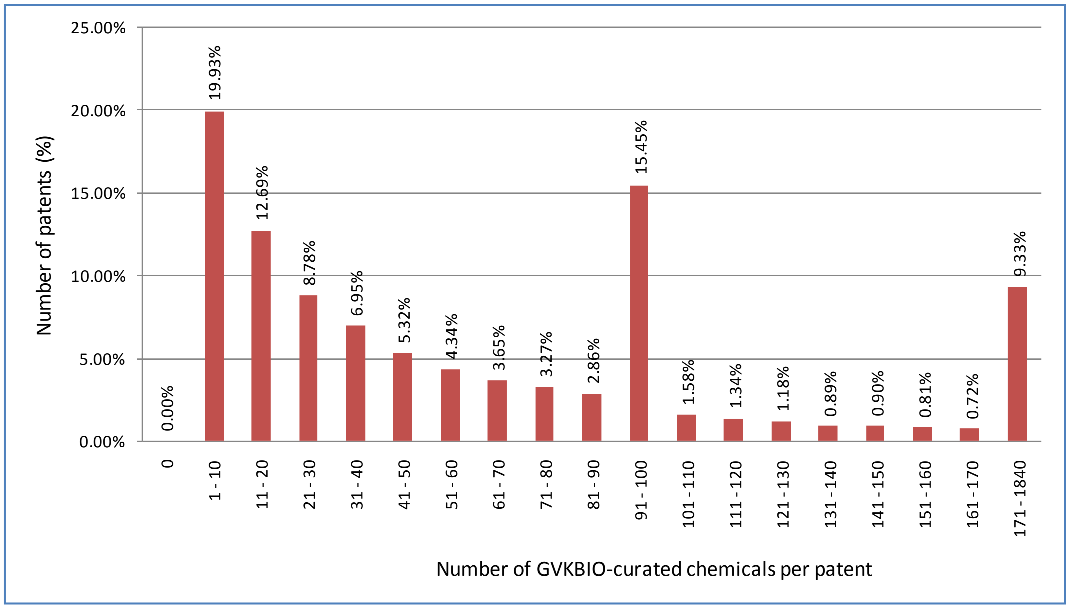


Figure 35. Distribution of the number of GVKBIO-curated chemicals per patent from a set of GVKBIO patents with targets.

Secondly, recall analysis was performed on the subset of GVKBIO for different filtration criteria. These were 1) at least 1 chemical in claim, 2) at least 2 chemicals in claim, and so on. Figure 36 shows percentage recall for these. Note that when the criterion of having at least one chemical in claim was applied, the recall percentage dropped to

94.15%. This implied that there were 5.85% of GVKBIO patents (1,936, patents) without an annotated chemical in claim. Some of these might actually have claimed chemicals in its original documents, but the IBM chemical annotator failed to recognize them (as exemplified in 3.1.3.2).

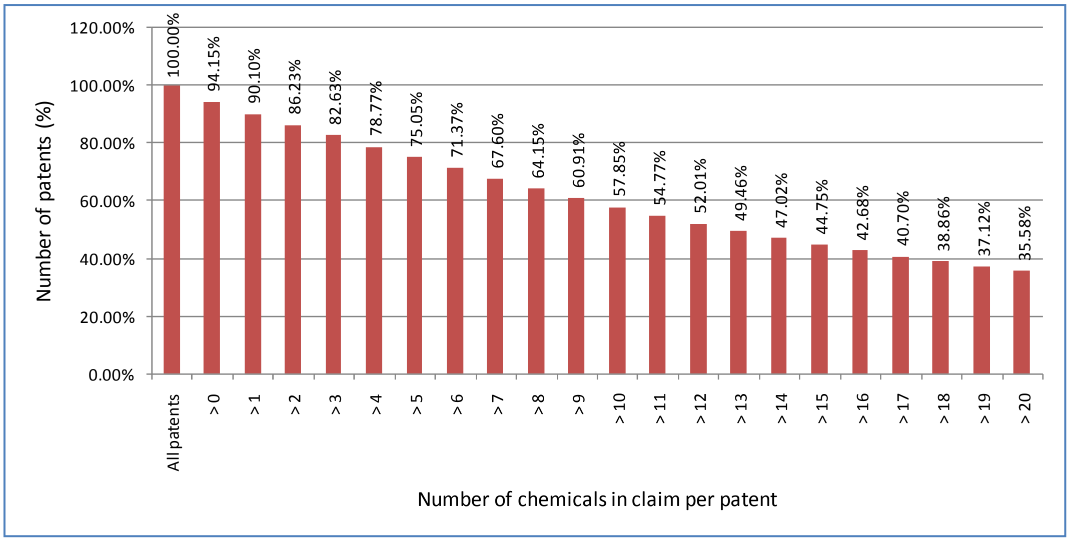


Figure 36. Cumulative distribution of the number of chemicals in claim per patent (annotated by IBM) from a set of GVKBIO patents with targets.

Thirdly, precision analysis was performed on the subset of IBM for different filtration criteria (numbers of chemical in claim). Figure 37 shows reduction in search space for different filtration criteria. Using the same assumption as described in IPC code analysis (section 3.6.1), it was assumed that the reduction in search space reflects a proportional reduction in false positives. Therefore, when the criterion of having at least one chemical in claim was applied, the search space was reduced to 23.28%, and 76.72% of false positive factors were removed.

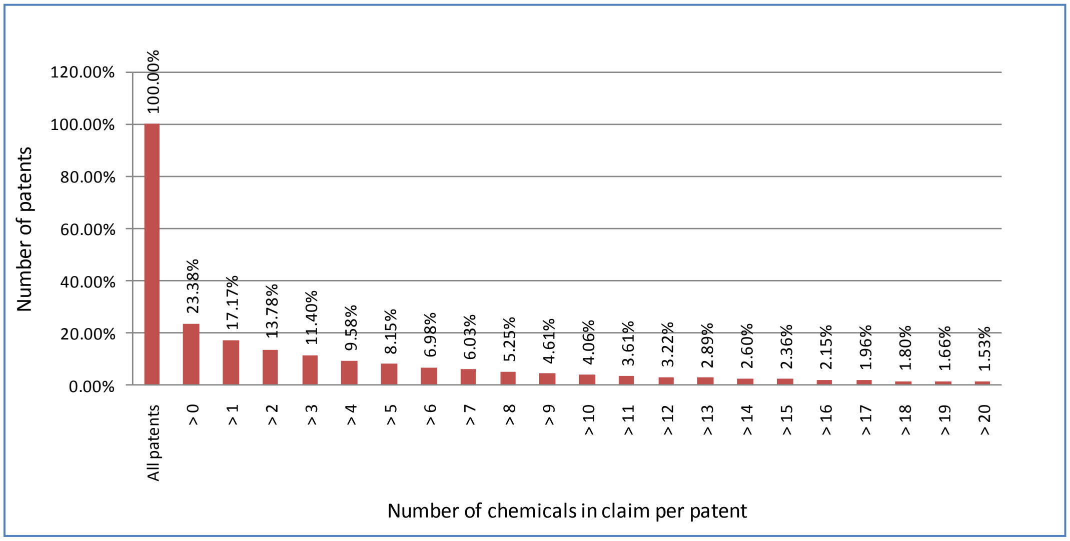


Figure 37. Cumulative distribution of chemicals in claim per patent from a set of IBM patents.

Finally, to select the best filtration criterion for high false positive reduction while retaining recall, results in Figure 36 and Figure 37 were transformed and compared in terms of "reduced recall percentage" and "reduced false positive percentage" (Figure 38 (a)). In order to facilitate the comparison between sets of IPC codes, the "filter utility ratio" shown in Formula 3 (section 3.6.1) was used. Results in Figure 33 (a) were then calculated for their filter utility ratios as shown in Figure 38 (b).

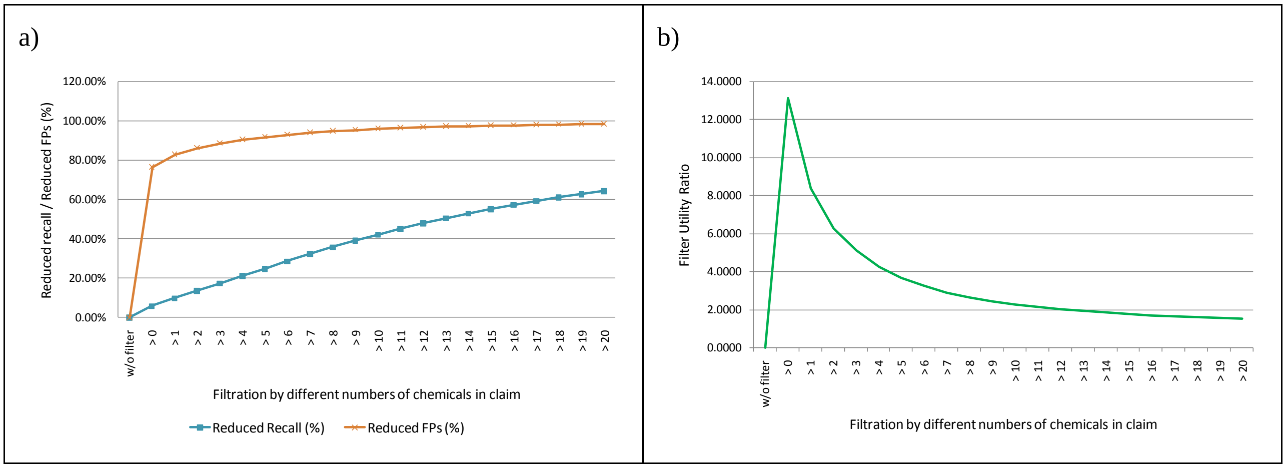


Figure 38. Comparison of false positive reduction and recall reduction that could be obtained by different filtration criteria (the number of chemicals in claim). a) False positive reduction (%) versus recall reduction (%), b) Filter utility ratios.

Looking at the filter utility ratios for different filtration criteria (Figure 38 (b)), the criterion of having at least one chemical in claim turns to be the best filter. This filter could remove 76.72% of false positives (or search space reduction, to be precise), while retaining recall at 94.15%.

In order to measure the performance of filtration criteria based on different upstream algorithms, the same methodology as done in IPC code analysis was conducted, by using data from Figure 37 (true positive reduction) and Figure 38 (false positive reduction). Calculated results are shown in Figure 39 as F-measure scores.

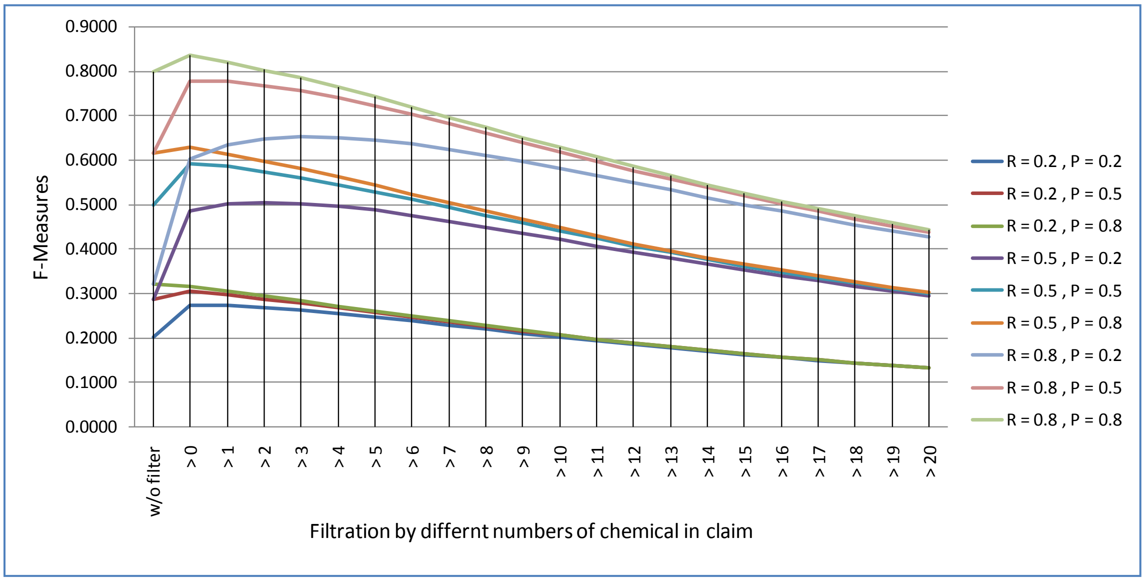


Figure 39. Predictive improvement in F-measures by different filtration criteria (the number of chemicals in claim) for different starting recall and precision of upstream algorithm.

Interestingly, results in Figure 39 shows that low numbers of chemicals in claim (i.e. 0-4 chemicals) are preferable regardless upstream algorithms with various recall and precision performance. Moreover, in the case of an upstream algorithm with poor recall and high precision (R=0.2, P=0.8), all downstream filtration criteria results in worse F-measure scores. This is due to the fact that recall performance drops significantly while increasing the required number of chemicals in claims (Figure 36).

3.6.3 Chemicals in patent

Following the previous section (3.6.2), this section aimed to explore the using of the number of chemicals found in all patent sections (i.e. not just claim section). Recall and precision analyses were done with the same methodology and data sets as the previous section.

The recall analysis on the GVKBIO subset (Figure 40) shows that recall percentage dropped to 98.70% when applying the criterion of having at least one chemical in the patent. This implied that there were only 1.30% of GVKBIO patents (431 patents) without annotated chemicals in the patent. Moreover, the recall percentage did not drop significantly while increasing the required number of chemicals in patent, as compared to the number of chemicals in claims (Figure 36).

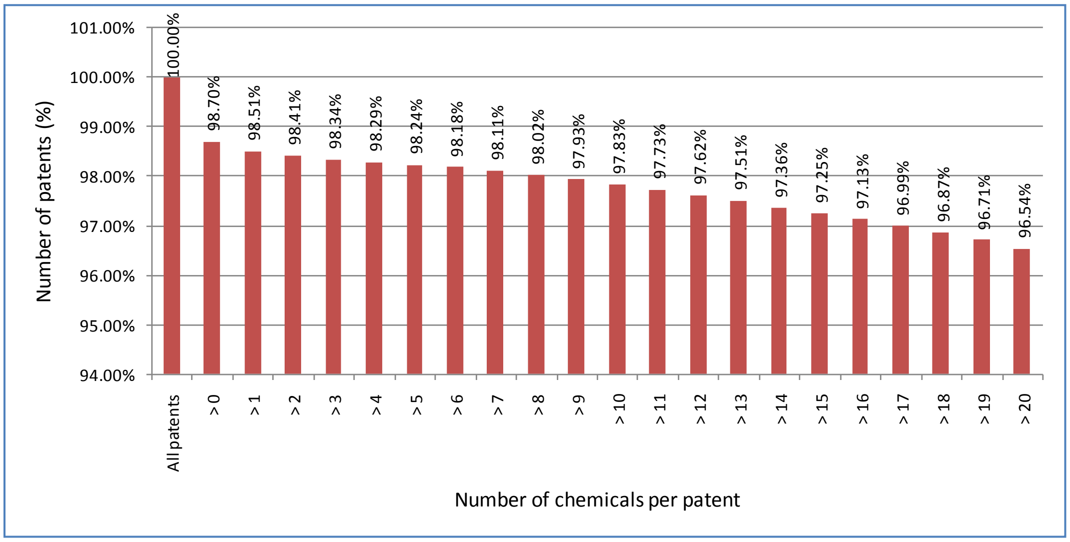


Figure 40. Cumulative distribution of the number of chemicals per patent from a set of GVKBIO patents with targets.

The precision analysis on the IBM subset (Figure 41) shows that the search space was reduced to 43.86% when applying the criterion of having at least one chemical in the patent. With the assumption described in section 3.6.1, this implied that 56.14% of false positives were removed.

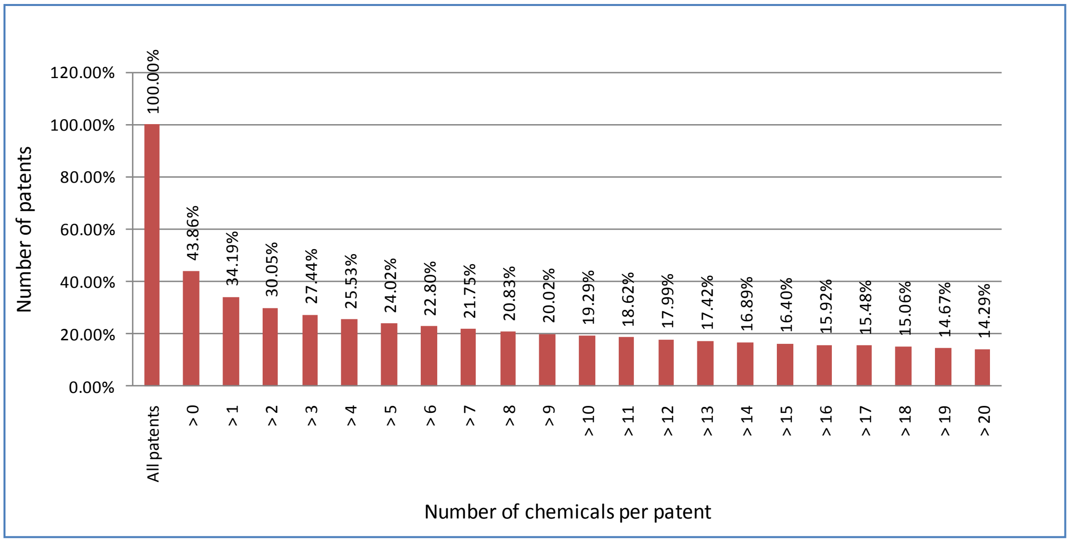


Figure 41. Cumulative distribution of chemicals per patent from a set of IBM patents.

To select the best filtration criterion, results from Figure 40 (true positive reduction) and Figure 41(false positive reduction) were transformed and compared in Figure 42 (a), and calculated for the filter utility ratios (Formula 3) in Figure 42 (b). It was shown in Figure 42 (b) that the criterion of having at least two chemical in patent turns to be the best filter by removing 65.81% of false positives, while retaining recall at 98.51%.

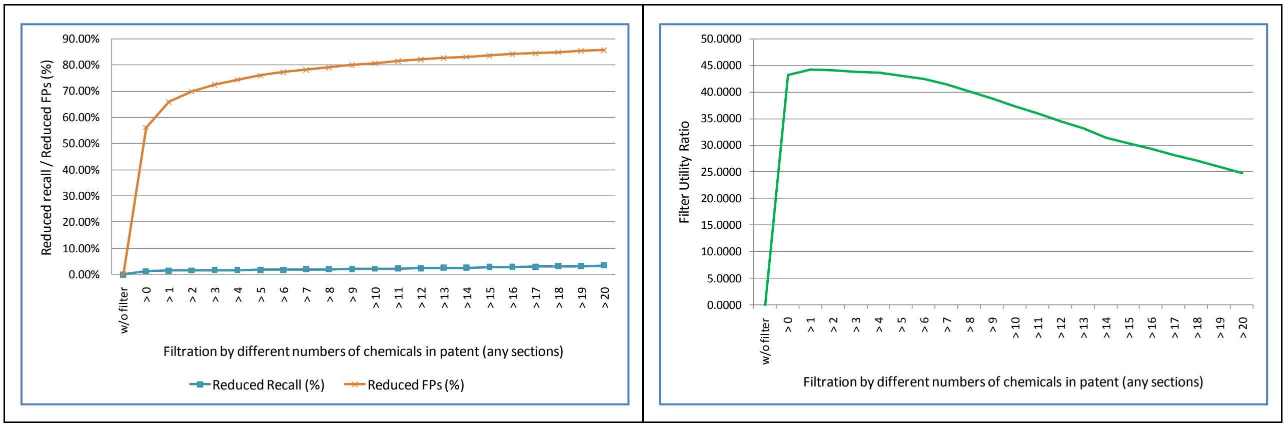


Figure 42. Comparison of false positive reduction and recall reduction that could be obtained by different filtration criteria (the number of chemicals in patent). a) False positive reduction (%) versus recall reduction, b) Filter utility ratios.

Shown in Figure 43 are F-measure scores for different filtration criteria and upstream algorithms (various recall and precision performance). Aligning with previous section, the result suggests that for cases having upstream algorithms with poor precision (P=0.2), the most stringent filtration criteria (having at least 21 chemicals in patent) turns to be the best filter. In constrast, cases having upstream algorithms with high precision (P=0.8), more relaxed filtration criteria (having at least 7-13 chemicals in patent) turn to be the best filter (maximum F-measure score). Furthermore, F-measure scores were generally improved regardless of the beginning precision and recall. This is due to the fact that recall performance did not drop significantly while increasing the required number of chemicals in the patent (Figure 40).

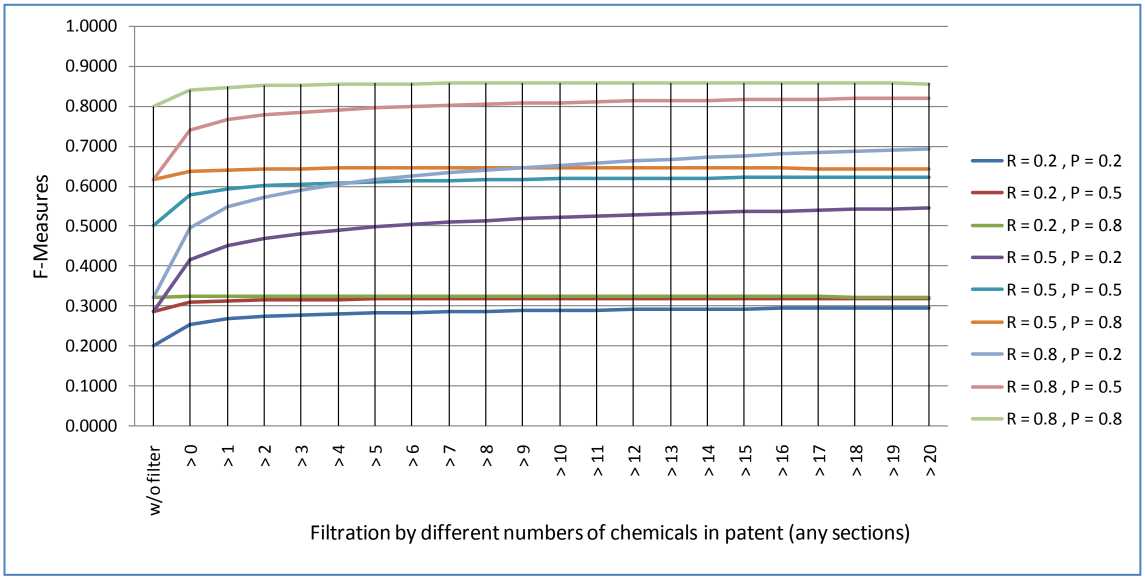


Figure 43. Predictive improvement in F-measures by different filtration criteria (the number of chemicals in patent) for different starting recall and precision of upstream algorithm.

3.6.4 Chemical modulation keywords

It has been shown in section 3.4.2 that retrieving of patents with target proteins simply by searching for occurrence of protein synonyms in titles could result in many false positives. However, patent titles within these false positives also have a very low occurrence of keywords signifying chemical modulation (Figure 25 (c) and (d)). In contrast, patent titles within true positives, there is a high co-occurrence of target protein names and these keywords in titles (Figure 23 (a) and (b)). Based on these observation, it was a hypothesized that target-containing patents could be retrieved by searching for the co-occurrence of protein names with these keywords in titles with high precision (i.e. false positives removed) while sustaining percentage recall.

In order to get a set of keywords signifying probable chemical modulation of targets, a word frequency analysis was performed on a corpus of titles of target-containing patents. This used the 34,575 GVKBIO patents with targets used in section 3.4.2 but was expanded to included patents published between years 1973-2009. The result shows these titles include 16,714 word forms, excluding stop words [45]. Those signifying chemical modulation were classified into four groups (Table 22 (a)). Their occurrence frequencies infer the approximate numbers of patent titles matched by these keywords.

Table 22. Chemical modulation keywords extracted from patents with target proteins (GVKBIO patent database). a) word frequency analysis, b) retrieval testing.

Keyword Groups	(a) Word frequency analysis			(b) Retrieval testing using keywords		
	Keywords as found in patent titles	Occurrence frequency (out of 34,575 patent titles)		Derived search terms used in retrieval	Retrieval result (# of patent documents)	
Agonism	agonism, agonist, agonistic, agonist-like, agonists, agonizing	1775	(5.1%)	agonism, agonist, agonistic, agonizing	6270	(18.1%)
Antagonisation	antagonisation, antagonise, antagonising, antagonism, antagonist, antagonistic, antagonistics, antagonist-like, antagonists, antagonize, antagonizing	4702	(13.6%)	antagonisation, antagonise, antagonising, antagonism, antagonist, antagonistic, antagonize, antagonizing	4782	(13.8%)
Inhibition	inhibit, inhibiting, inhibition, inhibitive, inhibitor, inhibitors, inhibitory, inhibits	8756	(25.3%)	inhibit, inhibiting, inhibition, inhibitive, inhibitor, inhibitory	8762	(25.3%)
Modulation	modulate, modulates, modulating, modulation, modulations, modulator, modulators, modulatory	2624	(7.6%)	modulate, modulating, modulation, modulator, modulatory	2640	(7.6%)
	Total (Summation)	17857	(51.6 %)	Union (all keywords)	17503	(50.6%)

To estimate the recall performance for each keyword group, search terms (Table 22 (b)) were derived by stemming their plural forms. Each keyword group was then assessed for percentage recall by searching its derived search terms against the set of 34,575 patent titles (Table 22 (b)). Note that the number of patent titles matched by each keyword group could be different from the corresponding approximate number in shown in Table 22 (a). This is because keywords in each group match multiple keyword variants (e.g. “agonist-specific”, “tumor-inhibiting” and “immunomodulators”). Note that the “agonism” keyword group gave 18.1% recall, more than three times the expected recall (5.1%). This was the result of substring matching between the term “agonist” and occurrence of “antagonist” in titles. This is an inherent problem when trying to differentiate nested terms. The results show these keywords can be used collectively to retrieve 50.6% of patents with protein chemical modulation regardless of protein name occurrence in patent titles. Extending this list to include more terms such as “phosphorylation ” and “activation” might further improve recall, but only these four keyword groups were used in this study.

3.7 Testing the use of selected filters

Following the selection and evaluation of four filters in previous section (3.6), in this section, the use of these filters in actual retrieval of target-containing patents was examimed. Filters evaluated in previous section were applied to search results obtained in sections 3.4.2 and 0 in attempt to remove false positives (patents without target protein). The resulting retrieval performance (i.e. recall and precision) was compared to the performance before filter application.

The four filters selected for this examination were:

1) Co-occurrence of chemical modulation keywords in the same section as protein names (Table 22 (b))
2) Existence of at least one chemicals in claim section
3) Existence of at least two chemicals in patent
4) Existence of 19 pharmaceutical IPC codes i.e. {A61K,C07D,A61P,...,A61F} (Table 19)

3.7.1 Application of filters for identification of target protein names in titles

The four selected filters were applied to each search result obtained in section 3.4.2 (Table 17). The results are shown in Table 23.

Table 23. Results of applying different filters on patents retrieved by occurrence of the protein synonyms in titles.

Protein Names	GVKBIO	IBM Search Results (Protein Synonyms in titles)																				
		Patent Sections	(a) w/o filter				(b) Chemical Modulation Keywords				(c) Chemicals in Claim (>0)				(d) Chemicals in Patent (>1)				(e)IPC Codes			
			Number of Patents		Approx. % Recall	Min. % Precision	Number of Patents		Approx. % Recall	Min. % Precision	Number of Patents		Approx. % Recall	Min. % Precision	Number of Patents		Approx. % Recall	Min. % Precision	Number of Patents		Approx. % Recall	Min. % Precision
			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result		
Renin	494	Titles	813	**237**	48.0%	29.2%	723	**237**	48.0%	32.8%	527	204	41.3%	38.7%	608	219	44.3%	36.0%	795	233	47.2%	29.3%
Thrombin	890	Titles	1743	**215**	24.2%	12.3%	903	**208**	23.4%	23.0%	994	197	22.1%	19.8%	1317	213	23.9%	16.2%	1716	214	24.0%	12.5%
Albumin	5	Titles	**1200**	0	0.0%	0.0%	**21**	0	0.0%	0.0%	575	0	0.0%	0.0%	905	0	0.0%	0.0%	1118	0	0.0%	0.0%
Hemoglobin	0	Titles	**1542**	0	-	0.0%	**15**	0	-	0.0%	882	0	-	0.0%	1137	0	-	0.0%	1408	0	-	0.0%

The result shows that the chemical modulation keywords in titles appear to be the best filtration criterion. In effect, the search result after filtration contains only patents with co-occurrence of the protein name and keywords in titles. Considering cases of non-target protein names, by comparing IBM search results in Table 23 (a) and (b), the keyword filtration appear to remove nearly all false positive patents. For instance, in the case of albumin, it removed nearly 1200 false positive patents.

On the other hand, while considering cases of target protein names (i.e. renin and thrombin), by comparing matched patents between GVKBIO and IBM (Table 23 (a) and (b)), the keyword filtration does not appear to significantly degrade recall. For instance, 237 patents were recalled by using co-occurrence of renin synonyms and chemical modulation keywords, which is the same number of patents as recalled by using only renin synonyms. Indeed, it can be implied that patent titles with target protein names often include these keywords.

The utility of this combination is also demonstrated when applied to the longitudinal analysis already shown in Figure 26. This is shown below in Figure 44. The use of the combination is thus very effective at filtering out the non-target proteins but maintaining the signals of the *bona fide* target names.

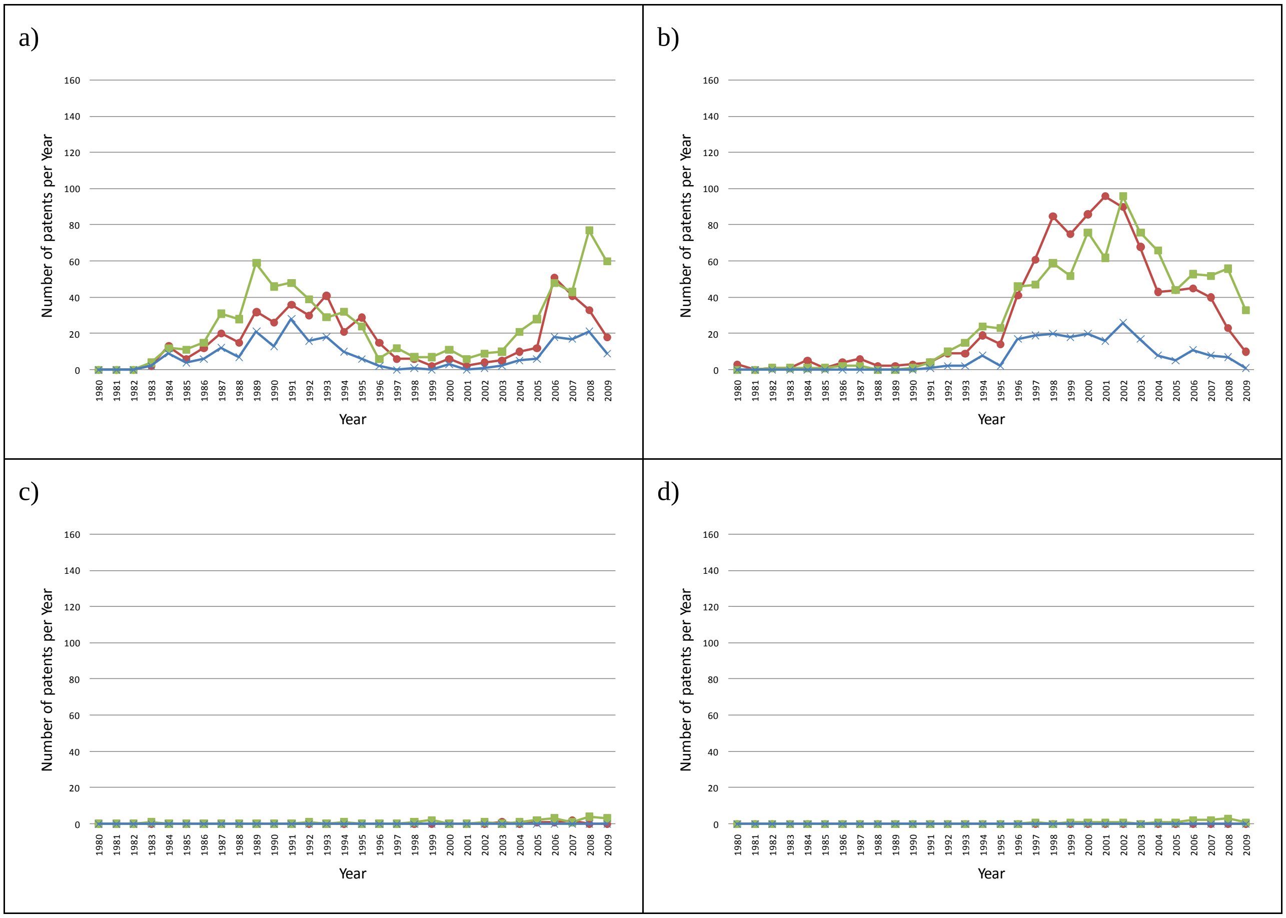
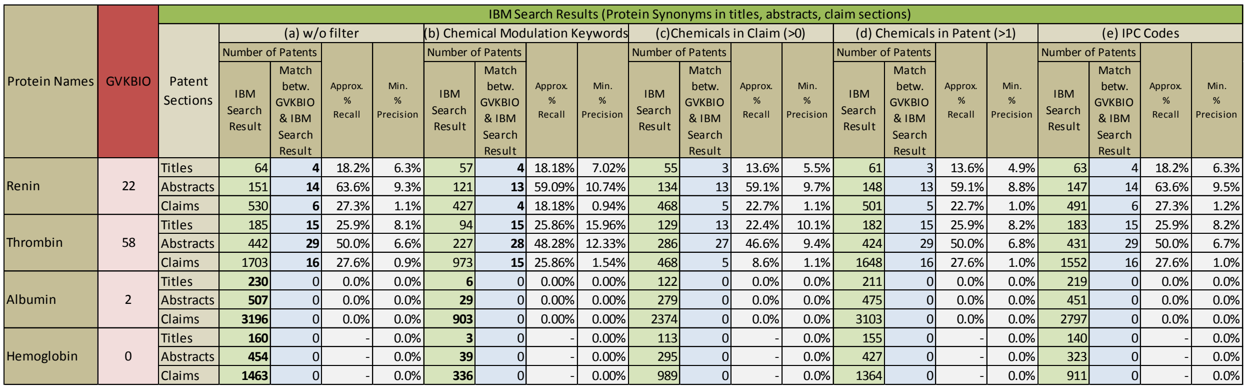


Figure 44. The result of retrieving patents using co-occurrence of protein synonyms and chemical modulation keywords in titles for a) Renin, b) Thrombin, c) Albumin, d) Hemoglobin. (● number of GVKBIO curated patents, ■ number of IBM search result, × number of patents-in-common between GVKBIO and IBM)

3.7.2 Application of filters for identification of target protein names in titles, abstracts and claim sections

The four selected filters were applied to each search result obtained in section 3.4.2 (Figure 29). The results were shown are Table 24.

Table 24. Results of applying different filters on patents retrieved by occurrence of the protein synonyms in titles, abstracts and claim sections.



Protein Names	GVKBIO		IBM Search Results (Protein Synonyms in titles, abstracts, claim sections)																			
		Patent Sections	(a) w/o filter				(b) Chemical Modulation Keywords				(c)Chemicals in Claim (>0)				(d) Chemicals in Patent (>1)				(e) IPC Codes			
			Number of Patents				Number of Patents				Number of Patents				Number of Patents				Number of Patents			
			IBM Search Result	Match betw. GVKBIO & IBM Search Result	Approx. % Recall	Min. % Precision	IBM Search Result	Match betw. GVKBIO & IBM Search Result	Approx. % Recall	Min. % Precision	IBM Search Result	Match betw. GVKBIO & IBM Search Result	Approx. % Recall	Min. % Precision	IBM Search Result	Match betw. GVKBIO & IBM Search Result	Approx. % Recall	Min. % Precision	IBM Search Result	Match betw. GVKBIO & IBM Search Result	Approx. % Recall	Min. % Precision
Renin	22	Titles	64	**4**	18.2%	6.3%	57	**4**	18.18%	7.02%	55	3	13.6%	5.5%	61	3	13.6%	4.9%	63	4	18.2%	6.3%
		Abstracts	151	**14**	63.6%	9.3%	121	**13**	59.09%	10.74%	134	13	59.1%	9.7%	148	13	59.1%	8.8%	147	14	63.6%	9.5%
		Claims	530	**6**	27.3%	1.1%	427	**4**	18.18%	0.94%	468	5	22.7%	1.1%	501	5	22.7%	1.0%	491	6	27.3%	1.2%
Thrombin	58	Titles	185	**15**	25.9%	8.1%	94	**15**	25.86%	15.96%	129	13	22.4%	10.1%	182	15	25.9%	8.2%	183	15	25.9%	8.2%
		Abstracts	442	**29**	50.0%	6.6%	227	**28**	48.28%	12.33%	286	27	46.6%	9.4%	424	29	50.0%	6.8%	431	29	50.0%	6.7%
		Claims	1703	**16**	27.6%	0.9%	973	**15**	25.86%	1.54%	468	5	8.6%	1.1%	1648	16	27.6%	1.0%	1552	16	27.6%	1.0%
Albumin	2	Titles	**230**	0	0.0%	0.0%	**6**	0	0.00%	0.00%	122	0	0.0%	0.0%	211	0	0.0%	0.0%	219	0	0.0%	0.0%
		Abstracts	**507**	0	0.0%	0.0%	**29**	0	0.00%	0.00%	279	0	0.0%	0.0%	475	0	0.0%	0.0%	451	0	0.0%	0.0%
		Claims	**3196**	0	0.0%	0.0%	**903**	0	0.00%	0.00%	2374	0	0.0%	0.0%	3103	0	0.0%	0.0%	2797	0	0.0%	0.0%
Hemoglobin	0	Titles	**160**	0	-	0.0%	**3**	0	-	0.00%	113	0	-	0.0%	155	0	-	0.0%	140	0	-	0.0%
		Abstracts	**454**	0	-	0.0%	**39**	0	-	0.00%	295	0	-	0.0%	427	0	-	0.0%	323	0	-	0.0%
		Claims	**1463**	0	-	0.0%	**336**	0	-	0.00%	989	0	-	0.0%	1364	0	-	0.0%	911	0	-	0.0%

Similar to the case of identification of target protein names in titles, the result shows that co-occurrence of the protein name and chemical modulation keywords in the same section appear to be the best filtration criterion. Considering cases of non-target protein names (i.e. albumin and hemoglobin), by comparing IBM search results between different filters in Table 24, the keyword filtration performed the best in removing false positive patents. For instance, in the case of searching albumin synonyms in abstracts, it reduced IBM search result containing mostly false positives from 507 patents to only 39 patents.

On the other hand, while considering cases of target protein names (i.e. renin and thrombin), by comparing matched patents between GVKBIO and IBM search results in Table 24 (a) and (b), the keyword filtration does not appear to significantly degrade recall. For instance, in the case of searching for renin synonyms in abstracts, 13 patents were recalled by using co-occurrence of renin synonyms and chemical modulation keywords, which is comparable to the number of patents recalled by using only renin synonyms (14 patents).

Shown in Figure 45 is an illustration of keyword filtration results (Table 24 (a) and (b)) which can be used to compare with search results before the filtration in Figure 29. It can be seen from cases of non-target protein names that even the keyword filtration was applied, search in claim sections still led to considerable false positives as compared to titles and abstracts.

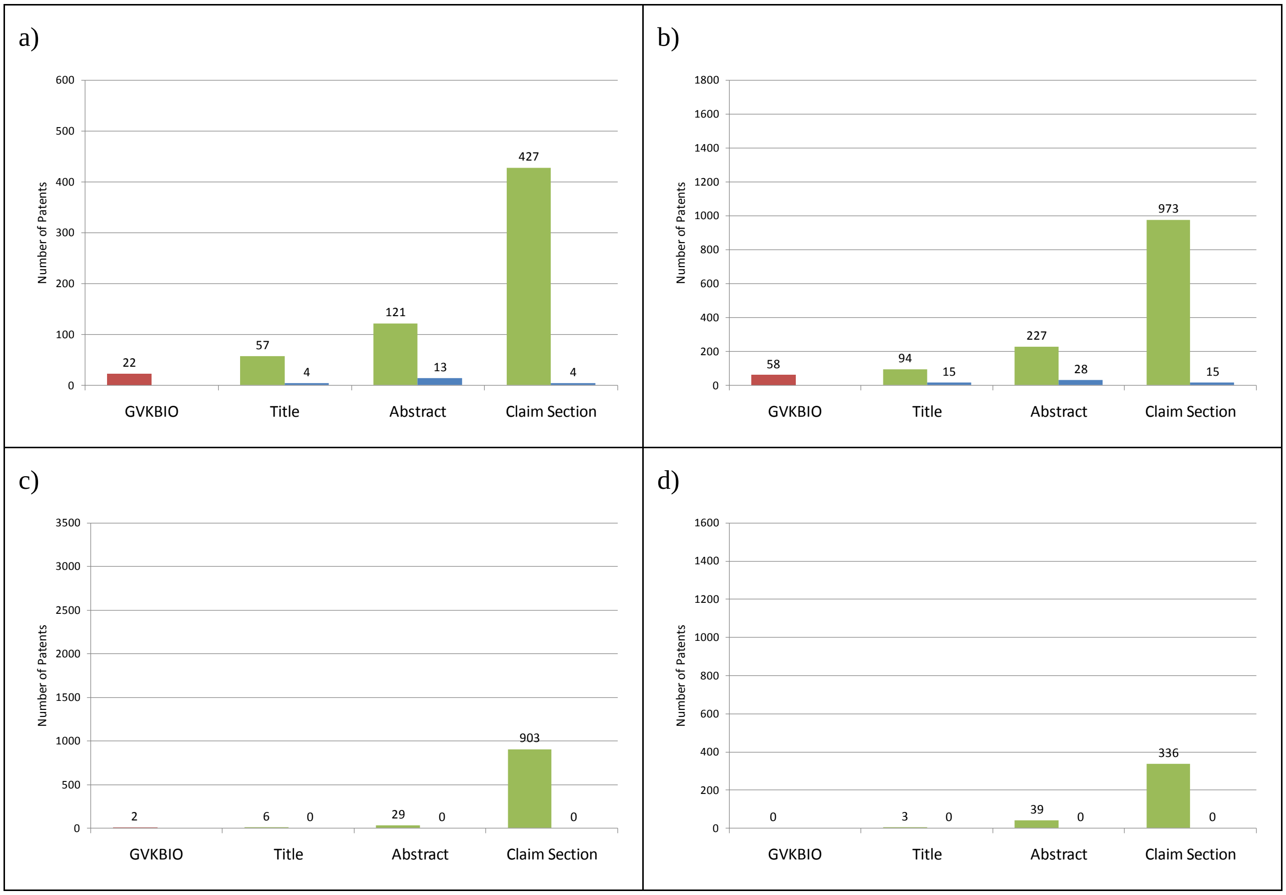


Figure 45. The result of retrieving patents using co-occurrence of protein synonyms and chemical modulation keywords in titles, abstracts and claim sections for a) renin, b) thrombin, c) albumin, d) hemoglobin. (● number of GVKBIO curated patents, ■ number of IBM search result, ✕ number of patents-in-common between GVKBIO and IBM)

4 Phase II - Results and discussions

It was ascertained from the first phase of this study that the tasks of identifying target protein names from a patent database could be classified into two major steps. These are 1) information retrieval (IR) and 2) information extraction (IE). The information retrieval (IR) is relevant to the retrieval of all patents with target proteins from a pool of patent documents. Following the information retrieval, within each patent document, target protein names shall be extracted (information extraction). This classification is aligned with approaches in text mining and also emphasized in the biological text mining domain [29,46]. Therefore, in the second phase, the study was divided into two workstreams that were 1) retrieval of target-containing patents, and 2) extraction of target protein names from a target-containing patent.

4.1 Retrieval of target-containing patents

In this workstream, a data filtration pipeline for retrieval of patents with target proteins (information retrieval) was developed and evaluated. From evaluation of four filters in the first phase of this study, three of them were selected and combined in the order shown in Figure 46. Nevertheless, the set of keywords is not only chemical modulation keywords shown in section 3.6.4 which only retrieve half of patents with targets from GVKBIO. The set of keywords in titles used in the data filtration pipeline was extended to be able to retrieve most of the patent with target proteins while keeping false positives minimal (described in section 4.1.1).

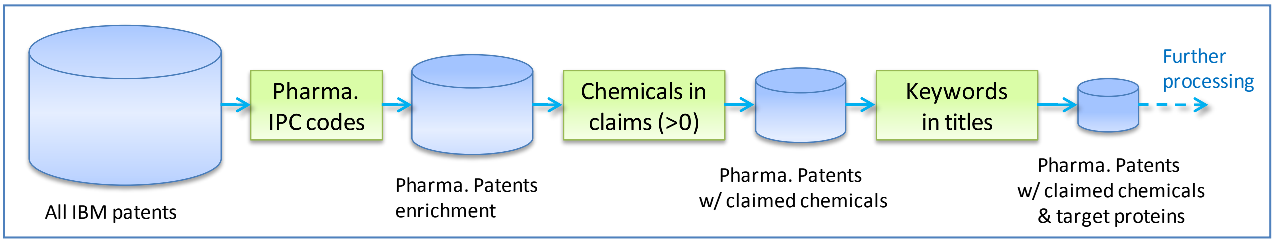

All IBM patents
Pharma. IPC codes
Pharma. Patents enrichment
Chemicals in claims (>0)
Pharma. Patents w/ claimed chemicals
Keywords in titles
Pharma. Patents w/ claimed chemicals & target proteins
Further processing

Figure 46. Data filtration pipeline proposed for retrieval of patents with target proteins.

4.1.1 Selection and evaluation of keywords in titles specific to target-containing patents

Considering the task of retrieving target-containing patents by the proposed data filtration pipeline in Figure 46, the analyses in section 3.7 had shown that even first two steps of filtering by IPC codes and existence of chemicals in claim left a considerable amount of false positives (patents without targets) in the retrieval result. Nevertheless, it was also suggested from the analyses that chemical modulation keywords in titles can be effectively used to classify target-containing patents with high precision. However, these keywords could retrieve only ~51% of GVKBIO target-containing patents (Table 22). It was suggested in Figure 23 that other keywords that could potentially be used for the retrieval such as “compound”, “derivative”, and “treatment”. In this section, the objective was to compile a new set of keywords that able to retrieve target-containing patents with high recall and precision.

In order to extract title keywords that are specific to target-containing patents, word frequency analysis was performed by using TextSTAT[39] on patent titles from 4 corpora as follow:

1) Patents with targets : GVKBIO patents with targets (34,575 patents)
2) Generic patents : All IBM patents published in 2006-2009 (1,234,684 patents)
3) Pharmaceutical patents : IBM patents filtered by 17 pharmaceutical IPC codes and existence of a chemical in claim (821,941 patents)

4) Patents without targets : The corpus was collected from Albumin patents (1,222 patents) and hemoglobin patents (1,560 patents) shown in Table 17 but extended to cover patents published during years 1970-2009.

The aim of keyword frequency analysis on these corpora was to extract title keywords that occur at high frequency in the corpus of patents with targets (GVKBIO), but a low frequency in other corpora. Shown in Table 25 (a) are the top-30 keywords found in the corpus of patents with targets. Corresponding word frequencies in other corpora are shown in Table 25 (b), (c) and (d). It is apparent that the top-five keywords - derivative, inhibitor, receptor, compound, and antagonist - have potential retrieval performance; having a high frequency in patents with targets (Table 25 (a)) and a low frequency in generic patents (Table 25 (b)) and patents without target (Table 25 (d)). In contrast, "method", which has high frequency in patents with targets, tends to be ineffective for retrieval; because it also has high frequency in all other corpora.

Table 25. Word frequency analysis in different sets of corpus to extract keywords in titles specific to patents with targets.

	Database/ Testing corpus	(a) Patents with target (GVKBIO)			(b) IBM 2006-2009			(c) IBM filtered by IPC codes & Number of Chemicals in Claim			(d) Patents without target (Albumin & Hemoglobin patents)		
		34575 patents			1234684 patents			821941 patents			2784 patents		
No.	Words	Word Freq.	Z-Score	Estimate % occurence in patents	Word Freq.	Z-Score	Estimate % occurence in patents	Word Freq.	Z-Score	Estimate % occurence in patents	Word Freq.	Z-Score	Estimate % occurence in patents
1	derivatives	9169	69.16	26.52%	4121	2.90	0.33%	75562	120.46	9.19%	32	0.53	1.15%
2	inhibitors	7109	53.60	20.56%	3403	2.39	0.28%	29051	46.29	3.53%	5	-0.09	0.18%
3	receptor	6096	45.96	17.63%	2327	1.62	0.19%	16969	27.02	2.06%	19	0.23	0.68%
4	compounds	5770	43.49	16.69%	6079	4.30	0.49%	61010	97.25	7.42%	32	0.53	1.15%
5	antagonists	3990	30.05	11.54%	935	0.63	0.08%	10078	16.03	1.23%	1	-0.18	0.04%
6	substituted	2819	21.21	8.15%	1492	1.03	0.12%	22064	35.15	2.68%	6	-0.07	0.22%
7	treatment	2677	20.14	7.74%	10769	7.64	0.87%	37908	60.41	4.61%	93	1.94	3.34%
8	thereof	2527	19.00	7.31%	33147	23.59	2.68%	50832	81.02	6.18%	143	3.09	5.14%
9	methods	2240	16.84	6.48%	70094	49.93	5.68%	62943	100.34	7.66%	168	3.66	6.03%
10	kinase	2220	16.68	6.42%	987	0.67	0.08%	5619	8.92	0.68%	0	n/a	0.00%
11	modulators	2114	15.88	6.11%	1219	0.83	0.10%	5326	8.45	0.65%	0	n/a	0.00%
12	compositions	1763	13.23	5.10%	14005	9.95	1.13%	67690	107.91	8.24%	96	2.00	3.45%
13	activity	1498	11.23	4.33%	2231	1.55	0.18%	12749	20.29	1.55%	19	0.23	0.68%
14	acid	1412	10.58	4.08%	4421	3.11	0.36%	46068	73.43	5.60%	24	0.35	0.86%
15	agonists	1385	10.38	4.01%	424	0.27	0.03%	3604	5.71	0.44%	2	-0.16	0.07%
16	pharmaceutical	1384	10.37	4.00%	2172	1.51	0.18%	27150	43.26	3.30%	52	0.99	1.87%
17	ligands	1374	10.30	3.97%	515	0.33	0.04%	4920	7.81	0.60%	3	-0.13	0.11%
18	preparation	1324	9.92	3.83%	5357	3.78	0.43%	65843	104.96	8.01%	141	3.04	5.06%
19	agents	1186	8.88	3.43%	2578	1.80	0.21%	23067	36.75	2.81%	21	0.28	0.75%
20	treating	1180	8.83	3.41%	5625	3.97	0.46%	20427	32.54	2.49%	35	0.60	1.26%
21	method	1069	7.99	3.09%	354788	252.87	28.74%	113230	180.53	13.78%	754	17.14	27.08%
22	protein	999	7.46	2.89%	2181	1.52	0.18%	12586	20.03	1.53%	71	1.43	2.55%
23	disorders	987	7.37	2.85%	1600	1.10	0.13%	8561	13.61	1.04%	4	-0.11	0.14%
24	protease	899	6.71	2.60%	337	0.20	0.03%	3404	5.39	0.41%	0	n/a	0.00%
25	heterocyclic	872	6.51	2.52%	419	0.26	0.03%	5116	8.12	0.62%	0	n/a	0.00%
26	receptors	829	6.18	2.40%	324	0.19	0.03%	3046	4.82	0.37%	3	-0.13	0.11%
27	process	770	5.73	2.23%	25213	17.94	2.04%	107996	172.18	13.14%	249	5.52	8.94%
28	therapeutic	720	5.36	2.08%	2246	1.56	0.18%	9959	15.84	1.21%	67	1.34	2.41%
29	selective	709	5.27	2.05%	2359	1.64	0.19%	5196	8.25	0.63%	10	0.03	0.36%
30	diseases	677	5.03	1.96%	1270	0.87	0.10%	8557	13.61	1.04%	3	-0.13	0.11%

In order to compare word frequencies between corpora, Z-scores were calculated. (Note that stop words [45], which are frequent words in English such as “about”, “above”, “also” and “although” were removed before the Z-score calculation.) Z-scores from the corpus of patents with targets were then plotted against Z-scores of corresponding words in another corpus as shown in Figure 47 (a), (b) and Figure 48 (a).

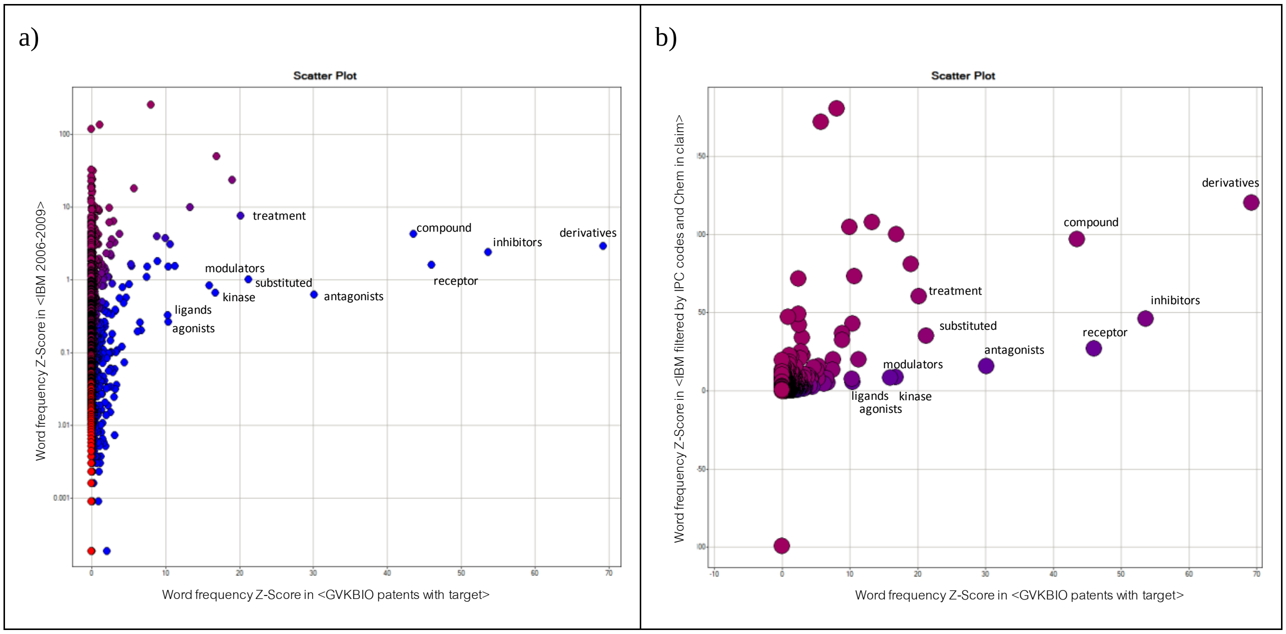


Figure 47. Scatter plots of word frequencies (standard scores) in the corpus of GVKBIO patents with targets against a corpus of generic patents. a) GVKBIO patents against IBM patents published between years 2006-2009, b) GVKBIO patents against all IBM patents filtered by pharmaceutical IPC codes and existence of chemicals in claim section.

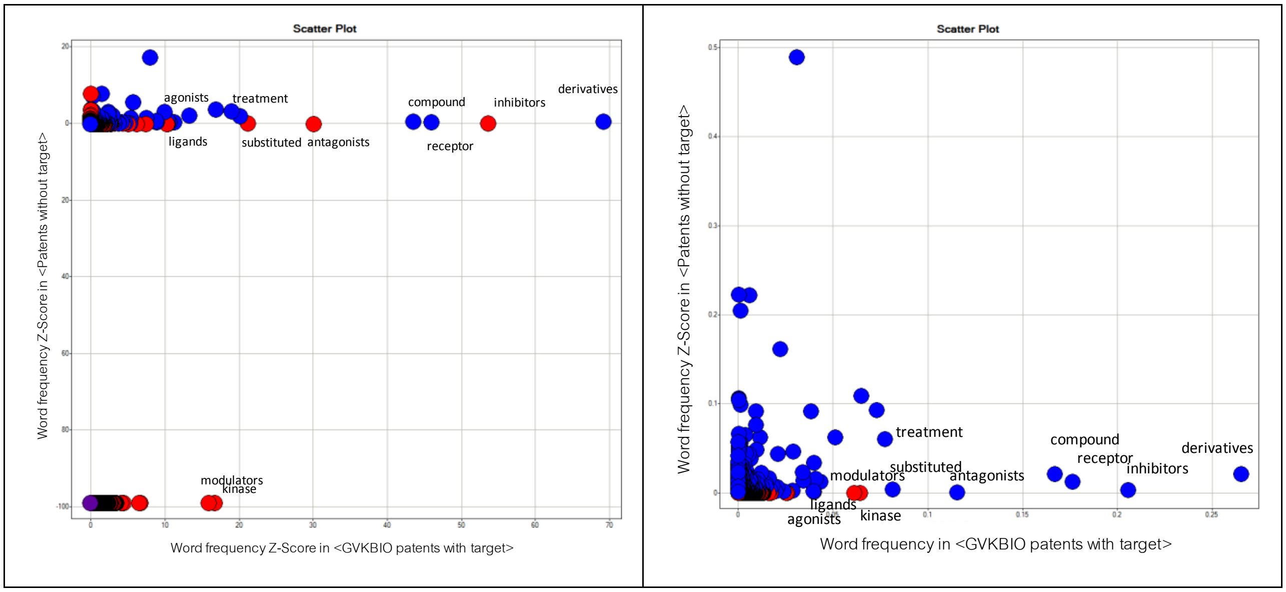


Figure 48. Scatter plots of word frequencies in the corpus of GVKBIO patents with targets against the corpus of patents without target. a) shown in standard scores, b) shown in absolute word frequencies.

Figure 47 (a) shows clearly that keywords specific to target-containing patents (e.g. derivative, inhibitor, receptor, etc.) have high frequency in GVKBIO target-containing patents while having low frequency in IBM generic patents.

Considering the data pipeline proposed (Figure 46), these keywords would be used after selection of pharmaceutical patents with a chemical in claim. Therefore, potential keywords should be extracted by comparing the corpus of patents with targets and the corpus of pharmaceutical patents. The comparison is shown in Figure 47 (b). Unfortunately, from the figure, keywords specific to target-containing patents are not standing out apparently. Furthermore, many chemical modulation keywords appear to have high Z-scores in this corpus of pharmaceutical patents. For example, "inhibitor" scores 46.29 in this corpus (Table 25 (c)). It could be that this corpus has high content of target-containing patents, thereby "submerging" the keywords of interest.

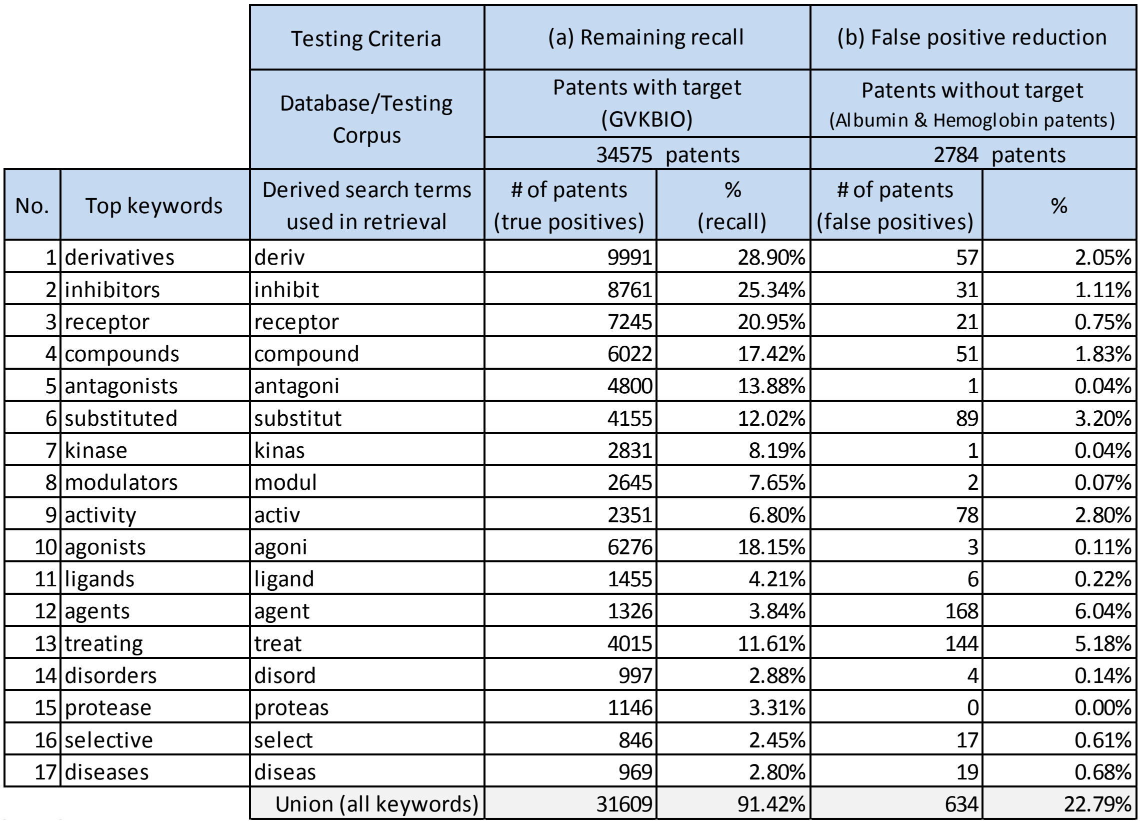
To avoid the problem described above, the corpus of patents with targets was then compared against the corpus of patents without targets (Figure 48). In this case, keywords specific to target-containing patents show up. From these, 17 keywords were selected by three criteria:

1) They are in top-30 used in patents with targets (Table 25 (a)).
2) They have low frequency in patents without targets (less than 1.5%, Table 25 (d)).
3) They are not trivial words (e.g. "acid", "heterocyclic", and "receptor" were considered as trivial words and removed.)

These 17 selected keywords were then derived into search terms by stemming using Porter's Stemmer [47] and some refinement as shown in Table 27. Resulting search terms were then assessed for their percentage recall by searching them against the corpus of patents with targets (Table 27 (a)). Likewise, their capability to reduce false positives was assessed against the corpus of patents without target (Table 27 (b)). Results show that this set of 17 keywords was able to retrieve patents with 91.42% recall, and able to remove 87.21% of false positive patents.

Further literature study showed that the retrieval of biomedical literature can also be done by machine learning methods such as the support vector machine (SVM) to classify the literature as described by Ghanem et al. [48]. In particular to this study, the SVM can be trained on the words in titles and abstracts from the corpora of patents with targets and patents without targets.

Table 26. Testing of target-containing-patent-specific keywords. a) recall performance, b) false positive reduction.



		Testing Criteria	(a) Remaining recall		(b) False positive reduction	
		Database/Testing Corpus	Patents with target (GVKBIO)		Patents without target (Albumin & Hemoglobin patents)	
			34575 patents		2784 patents	
No.	Top keywords	Derived search terms used in retrieval	# of patents (true positives)	% (recall)	# of patents (false positives)	%
1	derivatives	deriv	9991	28.90%	57	2.05%
2	inhibitors	inhibit	8761	25.34%	31	1.11%
3	receptor	receptor	7245	20.95%	21	0.75%
4	compounds	compound	6022	17.42%	51	1.83%
5	antagonists	antagoni	4800	13.88%	1	0.04%
6	substituted	substitut	4155	12.02%	89	3.20%
7	kinase	kinas	2831	8.19%	1	0.04%
8	modulators	modul	2645	7.65%	2	0.07%
9	activity	activ	2351	6.80%	78	2.80%
10	agonists	agoni	6276	18.15%	3	0.11%
11	ligands	ligand	1455	4.21%	6	0.22%
12	agents	agent	1326	3.84%	168	6.04%
13	treating	treat	4015	11.61%	144	5.18%
14	disorders	disord	997	2.88%	4	0.14%
15	protease	proteas	1146	3.31%	0	0.00%
16	selective	select	846	2.45%	17	0.61%
17	diseases	diseas	969	2.80%	19	0.68%
		Union (all keywords)	31609	91.42%	634	22.79%

4.1.2 Testing a data filtration pipeline

To evaluate the data filtration pipeline performance, three criteria were measure. These are 1) search space reduction, 2) false positive reduction, and 3) remaining recall. For evaluation of the search space reduction, the pipeline was executed on the whole IBM database (11,336,265 patents). To test the false positive reduction, the pipeline was executed on the corpus of patents without target described in section 4.1.1. The recall performance of this pipeline was then confirmed by executing it on the corpus of patents with targets described in section 4.1.1. The results of these three criteria testing are shown in Table 27.

Table 27. Performance evaluation of the data filtration pipeline for different criteria a) search space reduction, b) false positive reduction, c) remaining recall.

Testing Criteria	(a) Search space reduction		(b) False positive reduction		(c) Remaining recall	
Database /Testing corpus	IBM database (1969-2009)		Patents without target (Albumin & Hemoglobin patents)		Patents with targets (GVKBIO)	
	# of patents	%	# of patents	%	# of patents	%
Total patents	11336265	(100.00%)	2782	(100.00%)	34575	(100.00%)
IPC codes	1178281	(10.39%)	2543	(91.41%)	33770	(97.67%)
IPC codes + Chem in claim	821941	(7.25%)	1352	(48.60%)	30722	(88.86%)
IPC codes + Chem in claim + Keywords	340824	(3.01%)	364	(13.08%)	28234	(81.66%)

The results in Table 27 shows that the data filtration pipeline is relatively effective in 1) reducing search space down to only ~3%, 2) removing most of false positives (~87%), and 3) retaining recall at ~82%. The reduction of search space is particularly useful, since the text mining on full-text documents is a computational intensive task. For example, an attempt during this study to recognize all protein names in ~10,000 full-text patents by using a dictionary containing ~300,000 synonyms for ~19,000 human proteins took a day of execution time on the Oracle database. Focusing on a corpus highly enriched for target-containing patents will speed up execution, and enable more advanced mining algorithms. Nevertheless, the recall performance might be considered too low in some application, for example searching for prior art to support patent application. Fine tuning or dropping some filtration criteria could be done to improve the recall percentage.

4.2 Extraction of target protein names from full-text document

Following the retrieval of target-containing patents in previous section (4.1), this section aimed to extract target protein names within individual patents. It was shown earlier by manual inspection that target protein names usually co-occur with chemical modulation keywords in the same sentence (section 3.3.2). A computational search of their co-occurrence in the same patent section also suggested that the keywords rarely co-occur with non-target protein names (done in section 3.7.2, Figure 45). Based on this co-occurrence characteristic, an approach to extract target protein names from a patent was proposed. The hypothesis here was that protein names that proximally co-occur with chemical modulation keywords are likely to be target proteins (illustrated in Figure 49).

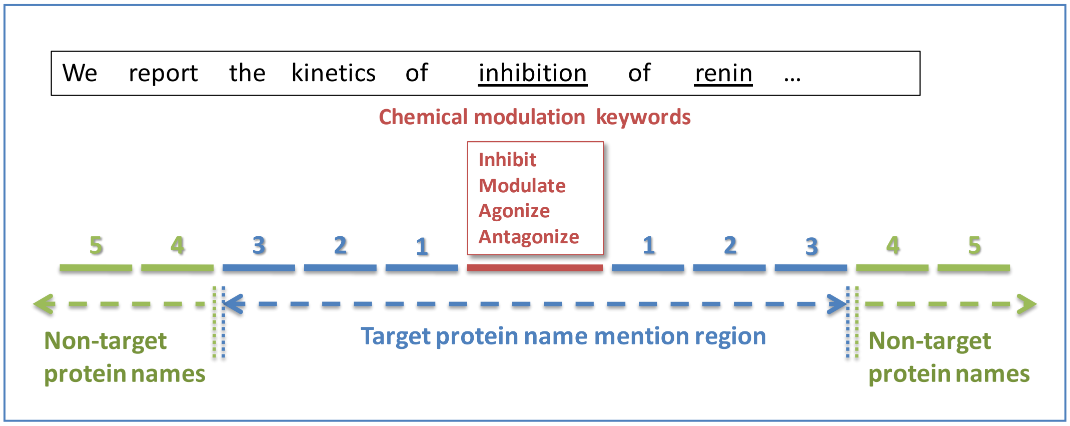

We report the kinetics of inhibition of renin ...
Chemical modulation keywords
Inhibit
Modulate
Agonize
Antagonize
5 4 3 2 1 1 2 3 4 5
Non-target protein names
Target protein name mention region
Non-target protein names

Figure 49. Schematic representation of the target protein name extraction by detecting proximal co-occurrence between protein names and chemical modulation keywords (word co-occurrence within proximity of three words).

The literature study showed that this type of proximal co-occurrence is also used by AMENDA (Automatic Mining of ENzyme DAta) to extract inhibitor-enzyme-organism relationship [49]. It is also an effective approach for retrieving journal articles containing experimental evidence for gene products [31,50].

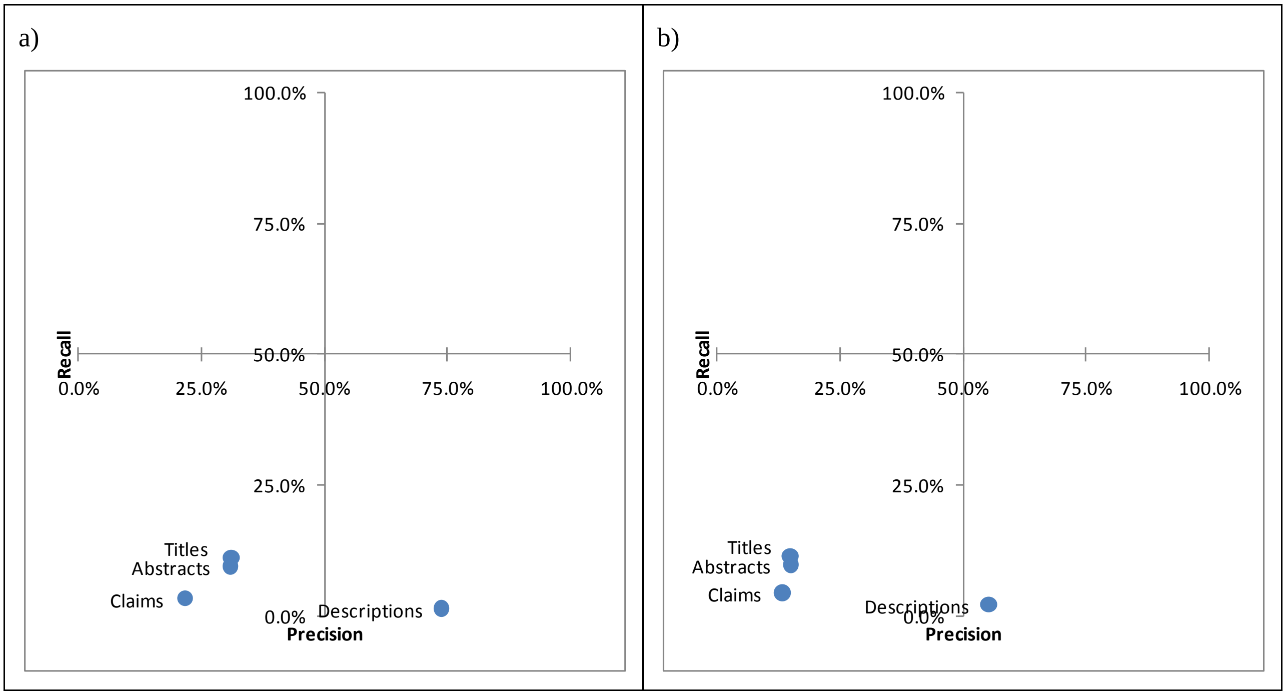
In particular to this study, the performance of this approach was assessed on two sets of target-containing patents which were curated by GVKBIO and used earlier in section 3.5.1. These were 1) GVKBIO *bona fide* target patents (4324 patents, 4324 document-target links), and 2) GVKBIO mixed-target patents (7640 patents, 16860 document-target links).

Protein names in these two patent sets were tagged and normalized to HGNC symbols by using the BioThesaurus dictionary (compiled section 3.2) with both full-word and case-insensitive matching. The resulting protein names which collocate with chemical modulation keywords (Table 22) within proximity of 3 words were then identified as target proteins (Figure 49). This computational target identification resulted in several document-target links (possibly multiple target proteins per patent). The recall and precision performance for each patent set were assessed by benchmarking document-target links obtained computationally and the links curated by GVKBIO (Table 28).

Table 28. Target protein name extraction results by detecting proximal co-occurrence between all human protein synonyms and chemical modulation keywords on corpora of target-containing patents (GVKBIO, published in 2006-2009).

Testing Corpus		Total Number of Doc-Taget Links in the testing corpus	Search Results															
			(a) Titles				(b) Abstracts				(c) Claim sections				(d) Body descriptions			
			# of Doc-Taget Links		% Recall	% Precision	# of Doc-Taget Links		% Recall	% Precision	# of Doc-Taget Links		% Recall	% Precision	# of Doc-Taget Links		% Recall	% Precision
No.	Corpus classes & description		IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result		
1	*bona-fide* -target patents (4,324 GVKBIO patents containing only one target)	4324	12011	1340	31.0%	11.2%	14162	1336	30.9%	9.4%	28239	932	21.6%	3.3%	231842	3191	73.8%	1.4%
2	mixed-target patents (7,648 GVKBIO patents containing one or more targets)	16860	21776	2501	14.8%	11.5%	25902	2530	15.0%	9.8%	50728	2233	13.2%	4.4%	439323	9298	55.1%	2.1%

Figure 50. Scatter plot for recall and precision performance of target protein name extraction by detecting proximal co-occurrence between all human protein synonyms and chemical modulation keyword on corpora of target-containing patents (GVKBIO, published in 2006-2009). a) GVKBIO *bona fide* target patents, b) GVKBIO mixed-target patents.



Results in Table 28 show the best recall performance when using the approach in body descriptions - 1) ~74% for GVKBIO *bona fide* target patents, and 2) ~55% for GVKBIO mixed-target patents. One of the reasons causing a drop in recall for the latter corpus could be that only *bona fide* target names often co-occur with chemical modulation keywords.

Nevertheless, the results also show poor precision performance for all patent sections in both corpora (<15%). The major reason was false positive protein names identified by the BioThesaurus dictionary. Table 29 (a) shows top 10 terms which often co-occur with chemical modulation keywords (in the corpus of GVKBIO mixed-target patents), ranked by the number of patent containing this co-occurrence. It is apparent that these terms are not referring to particular proteins. This resulted in false positive target protein names identified for each patent (Table 29 (b) and (c)). For instance, the protein SLC20A2 was identified as a target in ~69% of GVKBIO mixed-target patents, due to its nonsensical synonym “receptor” included in the BioThesaurus.

Table 29. Top-10 protein synonyms curated in BioThesaurus and frequently detected to co-occur proximally with chemical modulation keywords. This results in considerable amount of false positive target protein identified by the proximal co-occurrence approach.

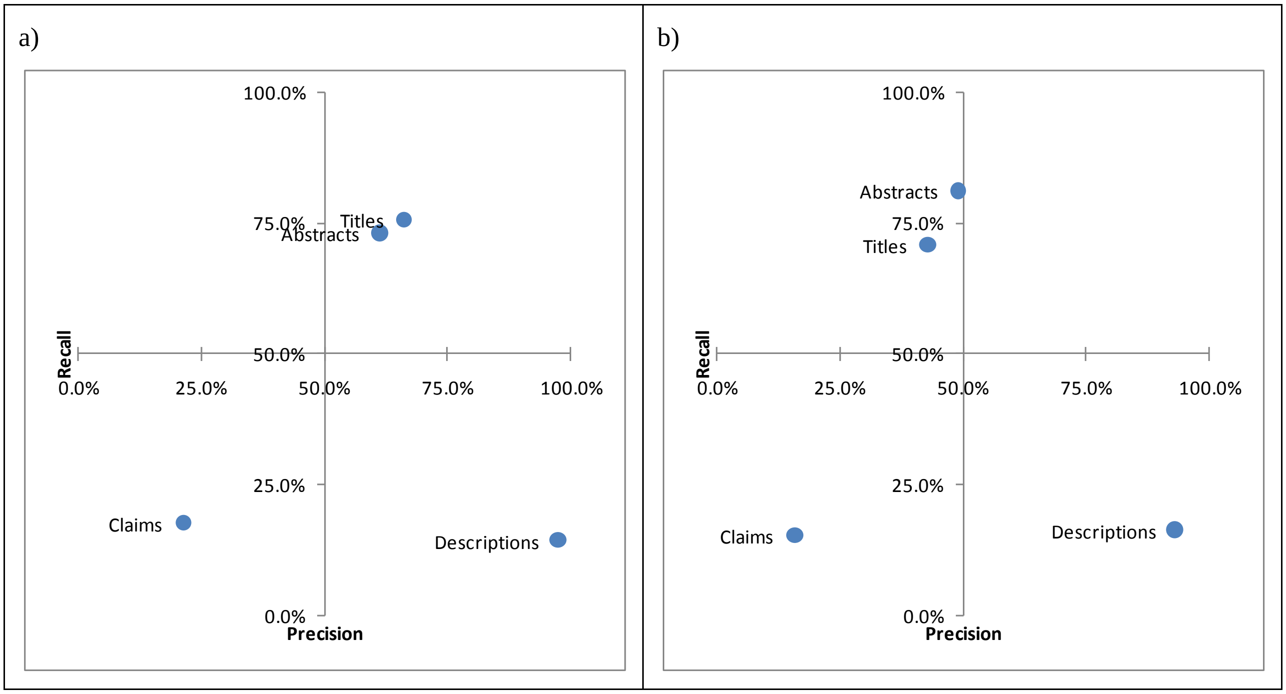
(a) Protein synonyms found in the text	(b) Corresponding HGNC symbols	(c) Patents containing co-occurrence of the protein synonyms and chemical modulation keywords	
		# of Patents	%
receptor	SLC20A2	5310	69.43%
binding	GNB2L1	2213	28.94%
kinase	BRD2, TTK, MAK	2034	26.60%
protein	PPAP2B	2023	26.45%
cell	CEL	1856	24.27%
ii	CD74, GCNT2, VIPR1, MGAT2, TAF8	1492	19.51%
reductase	SDR9C7	1382	18.07%
acid	GBA, GAA, MED25	1275	16.67%
human	RBM38, RP9	1252	16.37%
partial	ZNF71, ABCC1, FGG, ATP4A	1246	16.29%

To avoid this issue, the same target extraction approach was performed again, but restricted to renin and thrombin proteins which have relatively clean synonyms in the BioThesaurus (Table 8). Results are shown in Table 30. On each extraction, either renin or thrombin was searched for co-occurrence with chemical modulation keywords. For instance, searching for co-occurrence between renin synonyms and the keywords in titles in the corpus of *bona fide* target patents resulted in 38 patents predicted to contain renin as a target (Table 30 (a)).

Table 30. Target protein name extraction results by detecting proximal co-occurrence between renin/thrombin synonyms and chemical modulation keywords on corpora of target-containing patents (GVKBIO, published between 2006-2009)

Testing Corpus		Target proteins of interest	Total Number of Doc-Taget Links in the testing corpus Doc-REN links -or- Doc-F2 links	Search Results															
				(a) Titles				(b) Abstracts				(c) Claim sections				(d) Body descriptions			
				# of Doc-Taget Links		% Recall	% Precision	# of Doc-Taget Links		% Recall	% Precision	# of Doc-Taget Links		% Recall	% Precision	# of Doc-Taget Links		% Recall	% Precision
No.	Corpus classes			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result			IBM Search Result	Match betw. GVKBIO & IBM Search Result		
1	*bona-fide-target patents* *(4,324 GVKBIO patents containing only one target)*	Renin (REN)	59	38	36	61.0%	94.7%	37	35	59.3%	94.6%	44	10	16.9%	22.7%	280	59	100.0%	21.1%
		Thrombin (F2)	21	32	17	81.0%	53.1%	30	14	66.7%	46.7%	52	7	33.3%	13.5%	263	19	90.5%	7.2%
2	mixed-target patents (7,648 GVKBIO patents containing one or more targets)	Renin (REN)	79	41	39	49.4%	95.1%	41	39	49.4%	95.1%	78	12	15.2%	15.4%	438	79	100.0%	18.0%
		Thrombin (F2)	80	55	29	36.3%	52.7%	55	39	48.8%	70.9%	85	13	16.3%	15.3%	468	69	86.3%	14.7%

Figure 51. Scatter plot for average recall and precision performance of target protein name extraction by detecting proximal co-occurrence between renin/thrombin synonyms and chemical modulation keyword on corpora of target-containing patents (GVKBIO, published in 2006-2009). a) GVKBIO *bona fide* target patents, b) GVKBIO mixed-target patents.



By comparing results in Figure 50 with Figure 51, it can be seen that overall recall and precision performance improve significantly when restricting the search from all human proteins to only renin and thrombin. This implies that the recall and performance deficiency when searching for all human proteins (Figure 50) could partly come from imperfect synonym coverage and selectivity of the BioThesaurus. However, it could be that target protein names other than renin and thrombin might co-occur with the chemical modulation keywords less than expected. The significant improvement in precision when delimiting search to only renin and thrombin was mainly the result of removing a lot of false positive protein names (non-targets) found in each patent document. Nevertheless, false positives found by searching for these two protein names have highlighted the issue of this approach as will be described subsequently.

Considering recall performance for renin and thrombin extraction, body descriptions provide the best, while claim sections provide the worst (Figure 51). This resulted from the fact that the patent body usually contains experimental results and description of the proposed drug and *in vitro* target proteins, while the claim sections usually contain only the drug chemical structures and disease names as these are the key inventive steps.

Nevertheless, the result also shows that body descriptions and claim sections provide precision below 25%, thus making the claim section of low interest for target identification due to both low precision and recall. On the other hand, the body descriptions provide high recall but low precision. Manual inspection was performed on false positive patents US20090082423 and WO2007146124A3 which were inaccurately identified to have renin as a target protein. It was found that both patents contain the terms “renin inhibitor” in their body descriptions as shown in (Table 31, Table 32). But in both cases, the renin inhibitor was exemplified in a subordinate manner (i.e. not in the target context). Nevertheless, the term “renin inhibitor” occurs only once in each patent description. Therefore, this issue could possibly be addressed by taking the frequency of the co-occurrence between protein names and the keywords into account. This is aligned with the result of manual inspection in section 3.3.2 showing that target protein names are often mentioned in patent descriptions at higher frequency than non-target protein names.

Table 31. Body description of the patent number US20090082423 wrongly identified to has renin as target protein.

Patent No.	US20090082423
Title	Soluble epoxide hydrolase inhibitors
GVKBIO curated target	EPHX2
Body description (excerpt)	Combination therapy includes administration of a single pharmaceutical dosage formulation which contains a compound of Formula (I)-(IV) or (VIa)-(VIc) or of Table 1 and one or more additional active agents, as well as administration of the compound and each active agent in its own separate pharmaceutical dosage formulation. For example, a compound of Formula (I)-(IV) or (VIa)-(VIc) or of Table 1 and one or more angiotensin receptor blockers, angiotensin converting enzyme inhibitors, calcium channel blockers, diuretics, alpha blockers, beta blockers, centrally acting agents, vasopeptidase inhibitors, renin inhibitors, endothelin receptor agonists, AGE (advanced glycation end-products) crosslink breakers, sodium/potassium ATPase inhibitors, endothelin receptor agonists, endothelin receptor antagonists, angiotensin vaccine, and the like; can be administered to the human subject together in a single oral dosage composition, such as a tablet or capsule, or each agent can be administered in separate oral dosage formulations.

Table 32. Body description of the patent number US20090082423 wrongly identified to has renin as target protein.

Patent No.	WO2007146124A3
Title	Substituted PDE5 inhibitors
GVKBIO curated target	PDE5A
Body description (excerpt)	The compounds provided herein can also be administered in combination with other classes of compounds, including, but not limited to, endothelin converting enzyme (ECE) inhibitors, such as phosphoramidon; thromboxane receptor antagonists, such as ifetroban; potassium channel openers; thrombin inhibitors, such as hirudin; growth factor inhibitors, such as modulators of PDGF activity; platelet activating factor (PAF) antagonists; anti-platelet agents, such as GPIIb/ITIa blockers (e.g., abdximab, eptifibatide, and tirofiban), P2 Y(AC) antagonists (e.g., clopidogrel, ticlopidine and CS-747), and aspirin; anticoagulants, such as warfarin; low molecular weight heparins, such as enoxaparin; Factor Vila Inhibitors and Factor Xa Inhibitors; renin inhibitors; neutral endopeptidase (NEP) inhibitors; vasopepsidase inhibitors (dual NEP-ACE inhibitors), such as omapatrilat and gemopatrilat …

5 Conclusions

This work illustrates some of the challenges and possible solutions for retrieval of target-containing patents and extraction of target protein names from full-text patent databases. The work also serves as a pilot project in exploring and pioneering the mining of structure activity relationship (SAR) data from patents. The major question to be addressed was the identification of *bona fide* targets and other targets out of all protein names mentioned in a patent document. By using several text mining approaches such as dictionary-based named entity recognition and term frequency analysis on IBM and GVKBIO, it was possible to develop and evaluate approaches to identify target protein names.

In the first phase, it was shown that protein names in titles can be used to retrieve target-containing patents. However, the recall coverage of this approach is not complete due to the practice of putting chemical or disease names in titles. It was also demonstrated that non-target proteins are also mentioned in titles. To remove these false positives, it was shown that, among many filtration criteria, combining protein names and chemical modulation keywords for searching patent titles improves precision without significant loss of recall. The results also show that extending this type of search from patent titles to abstracts and claim sections significantly improves target retrieval via the extra information found in these sections. A caveat with this approach was a substantial increase in false positives by including claim sections. Therefore, combining title with abstract searches represents a useful compromise because of improved recall performance combined with low false positives.

In the second phase, to address the problem of target-containing patent retrieval, a set of title keywords that are highly specific to target-containing patents was developed. This was proven to retrieve target-containing patents with high recall and precision. This set of keywords was then combined with other filtration criteria into the data filtration pipeline to retrieve target-containing patents. The data filtration pipeline restricted the IBM database, containing ~11m patents, down to a set of ~0.3m patents highly enriched for target-containing patents. This search space reduction pragmatically enabled computational intensive processing on full-text patent documents.

In the second phase, relevant to the problem of target protein name extraction, it was shown that the proximal co-occurrence between protein names and chemical modulation keywords can be used to identify *bona fide* target protein names. It was shown that extending the search from titles to abstracts, and body descriptions improves recall significantly due to unique information found in these sections. Nevertheless, this comes with a precision tradeoff as expected. Interestingly, claim sections appeared to provide the worst recall and precision performance which could be relevant to pharmaceutical patent claiming practices. Furthermore, it was clearly shown that the major underlying source of false positives and false negatives was the poor sensitivity and specificity of the dictionary-based named entity recognition.

In brief, this work has shown that the retrieval of target-containing patent and extraction of target protein names can be done with some level of recall and precision by using basic text mining techniques such as dictionary-based named entity recognition, term co-occurrence and term frequency analysis. Inspection of false positives and false negatives also show that occurrences of target protein names depend on their semantic context in sentences and paragraphs, rather than the term level. Therefore, using more sophisticated text mining techniques especially natural language processing (NLP) [51] could possibly improve recall and precision performance.

The literature study showed that several text mining techniques have been developed in the domain of biomedical journals. Although they mostly process only on titles and abstracts (available on PubMed MEDLINE), several of them have high recall and precision [29,30,31,33]. Unlike biomedical journals, it was shown in this study that considerable target protein names are hidden in patent body descriptions. This mainly due to differences in writing styles of patents and biomedical journals. Therefore, there is a need to extract information from body descriptions. However, this also comes with several false positive issues, calling for more sophisticated information extraction techniques. Nevertheless, it is strongly suggested that proven techniques developed for biomedical journals be applied to patents as a starting point.

The results sections exemplify a range of problems identified in the IBM encompassing a broad spectrum from frank errors by applicants, OCR problems, name-to-structure conversion failures, gene name recognition false-positives and patent office-specific extraction losses. None of these were unexpected and some (e.g. those introduced during the many stages of patent office conversions) are extrinsic to IBM. Given the magnitude, challenges and technical scope of the undertaking, they do not detract from the value of this resource that is supported by the results presented. In addition these findings were regularly fed back to, and gratefully acknowledged by, the IBM team. Thus, improvements have been made, or are planned, as a consequence of this work.

To conclude, this initial exploitation of the IBM full-text patent data source has successfully pointed out challenges for the exploitation of this valuable medicinal chemistry data source. In addition, the work also illustrates well-defined methodologies for addressing some of the challenges through extensive benchmarking with the other reference corpora.

References

1. AstraZeneca product portfolios. Available at: http://www.astrazeneca.com/medicines/

2. AstraZeneca development pipeline. Available at: http://www.astrazeneca.com/research/our-pipeline-summary/

3. EPO (European Patent Office). Available at: http://ep.espacenet.com/

4. USPTO (United States Patent and Trademark Office). Available at: http://patft.uspto.gov/

5. WIPO (World Intellectual Property Organization). Available at: http://www.wipo.int/pctdb/en/index.jsp

6. Google Patents. Available at: http://www.google.com/patents

7. Free Patents Online. Available at: http://www.freepatentsonline.com

8. Patent Lens. Available at: http://www.patentlens.net

9. CiteXplore. Available at: http://www.ebi.ac.uk/citexplore

10. Patent Abstracts. Available at: http://srs.ebi.ac.uk

11. SureChem. Available at: http://www.surechem.org

12. Thomson Reuters Integrity. Available at: http://thomsonreuters.com/products_services/science/science_products/a-z/integrity

13. Thomson Pharma. Available at: http://www.thomson-pharma.com/

14. GVK BIO. Available at: http://www.gvkbio.com/

15. Webber, P.: A guide to drug discovery. Protecting your inventions: the patent system. Nature reviews. Drug discovery 2, 823-830 (2003)

16. Grandjean, N., Charpiot, B., Pena, C., Peitsch, M.: Competitive intelligence and patent analysis in drug discoveryMining the competitive knowledge bases and patents. Drug Discovery Today: Technologies 2, 211-215 (2005)

17. Granstrand, O.: The economics and managment of intellectual property: towards intellectual capitalism. Edward Elgar Publishing Limited (2000)

18. Grubb, P.: Patents for chemicals, phamaceuticals, and biotechnology. Oxford Univ Press, New York (2004)

19. Chen, Y., Spangler, S., Kreulen, J., Boyer, S., D., T., Alba, A., Behal, A., He, B., Kato, L., Lelescu, A., Zhang, L., Kieliszewski, C.: SIMPLE: A Strategic Information Mining Platform for IP Excellence., San Jose, CA, USA

(2009)

20. Levene, M., Loizou, G.: Why is the snowflake schema a good data warehouse design? Information Systems 28(3), 225-240 (2003)

21. SMILES (Simplified Molecular Input Line Entry System). Available at: http://www.daylight.com/smiles/index.html

22. Brecher, J.: Name=struct: A practical approach to the sorry state of real-life chemical nomenclature. Journal of Chemical Information and Computer Science 39, 943–950 (1999)

23. Leser, U., Hakenberg, J.: What makes a gene name? Name entity recognition in the biomedical literature. Briefings in Bioinformatics 6(4), 357-369 (2005)

24. Rhodes, J., Boyer, S., Kreulen, J., Chen, Y., Ordonez, P.: Mining patents using molecular similarity search. In : Pacific Symposium on Biocomputing 2007, Maui, Hawaii, p.304 (2007)

25. Sarma, J., Radha, K.: Database systems for knowledge-based discovery. In : Chemogenomics: Methods and Applications 575. (2009) 159-172

26. Southan, C., Várkonyi, P., Muresan, S.: Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. Journal of cheminformatics 1, 10 (2009)

27. SciFinder. Available at: https://scifinder.cas.org/

28. Lipinski, C., Lombardo, F., Dominy, B., Feeney, P.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews 46, 3-26 (2001)

29. Banville, D.: Mining chemical and biological information from the drug literature. Current Opinion in Drug Discovery & Development 12(3), 376-387 (2009)

30. Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., Valencia, A.: Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biology 9, S1 (2008)

31. Cohen, A., Hersh, W.: A survey of current work in biomedical text mining. Briefings in Bioinformatics 6(2), 57-71 (2004)

32. Krallinger, M., Leitner, F., Rodriguez-Penagos, C., Valencia, A.: Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biology 9, S4 (2008)

33. Krallinger, M., Valencia, A., Hirschman, L.: Linking genes to literature: text mining, information extraction,and retrieval applications for biology. Genome Biology 9, S8 (2008)

34. UniProt (Universal Protein Resource). Available at: http://www.uniprot.org

35. BioThesaurus [PIR - Protein Information Resource]. Available at: http://pir.georgetown.edu/pirwww/iprolink/biothesaurus.shtml

36. BioLexicon [BOOTStrep project, NaCTeM]. Available at: http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html

37. Liu, H., Hu, Z.-Z., Torii, M., Wu, C.: Quantitative assessment of dictionary-based protein named entity tagging., 497-507 (2006)

38. HGNC (HUGO Gene Nomenclature Committee). Available at: http://www.genenames.org/

39. TextSTAT - Simple Text Analysis Tool. Available at: http://neon.niederlandistik.fu-berlin.de/en/textstat/

40. Yang, Z., Lin, H., Li, Y.: Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature., 287–291 (2008)

41. Yu, H., Kim, W., Hatzivassiloglou, V., Wilbur, W.: Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. Journal of Biomedical Informatics 40(2), 150-159 (2007)

42. Oracle® Text Application Developer's Guide 10g Release 2. Available at: http://download.oracle.com/docs/cd/B19306_01/text.102/b14217/toc.htm

43. Wordle. Available at: http://www.wordle.net

44. WIPO IPC-Technology Concordance Table. Available at: http://www.wipo.int/export/sites/www/ipstats/en/statistics/patents/xls/ipc_technology.xls

45. Stop words. Available at: ftp://ftp.cs.cornell.edu/pub/smart/english.stop

46. Muller, H., Kenny, E., Sternberg, P.: Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biololgy 2(11), e309 (2004)

47. Porter's Stemmer Online. Available at: http://maya.cs.depaul.edu/classes/ds575/porter.html

48. Ghanem, M., Guo, Y., Lodhi, H., Zhang, Y.: Automatic scientific text classification using local pattterns-KDD Cup 2002 (task1). ACM SIGKDD Explorations Newsletter 4(2), 95-96 (2002)

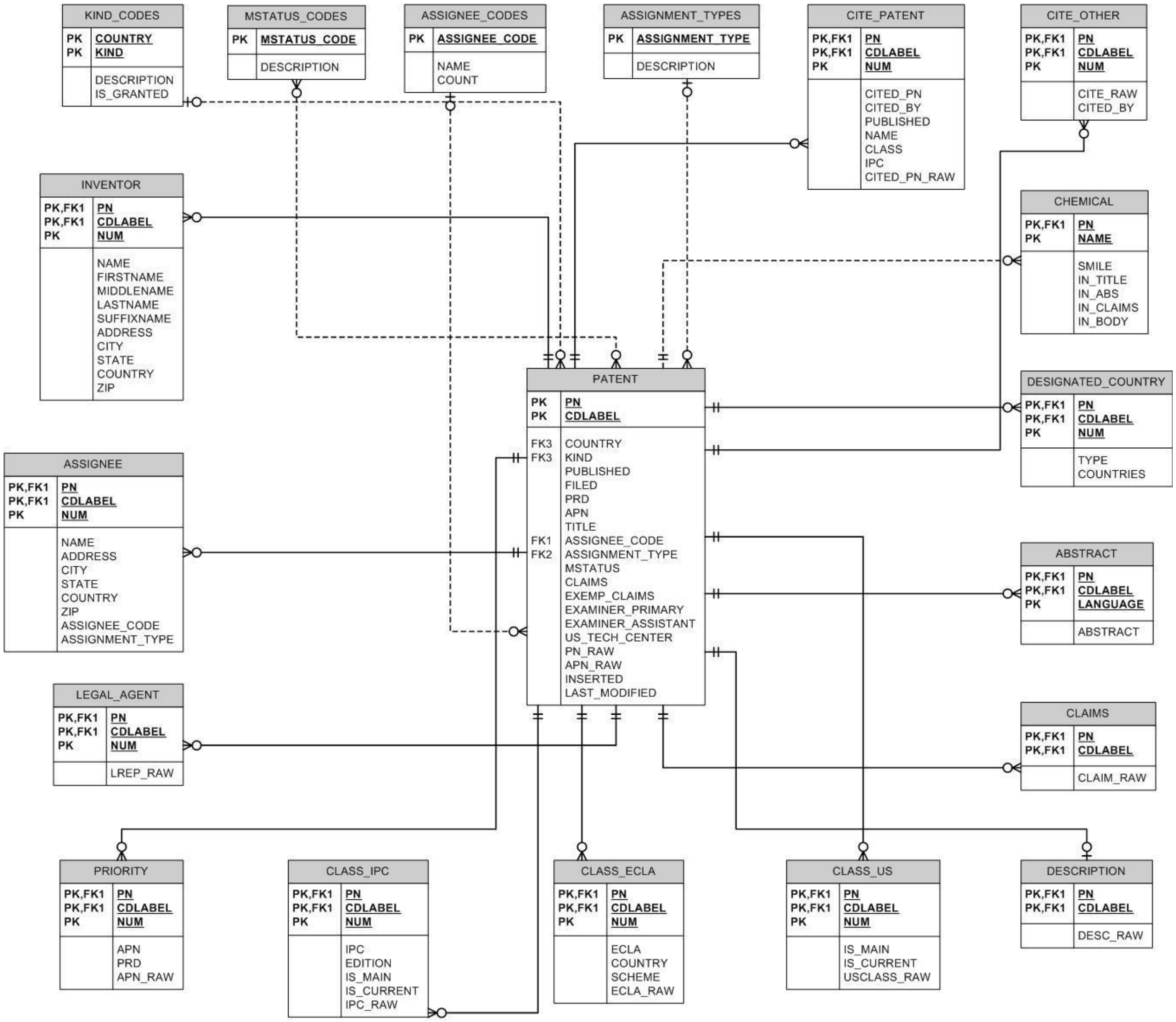
49. Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I., Schomburg, D.: BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. Nucleic Acids Research 35, D511-D514 (2007)

50. Shi, M., Edwin, D. S., Menon, R., Shen, L., Lim, J. Y. K., Loh, H. T.: A machine learning approach for the curation of biomedical literature–KDD Cup 2002 (task 1). ACM SIGKDD Explorations Newsletter 4(2), 93–94 (2003)

51. Kolarik, C., Hofmann-Apitius, M.: Linking chemical and biological information with natural language processing. In : Chemical information mining. CRC Press, New York, NY, USA (2009) 123-150

52. Agarwal, P., Searls, D.: Can literature analysis identify innovation drivers in drug discovery? Nature reviews. Drug discovery 8, 865-878 (2009)

Appendix A – IBM Patent Database Schema


KIND_CODES
PK COUNTRY
PK KIND
DESCRIPTION
IS_GRANTED
MSTATUS_CODES
PK MSTATUS_CODE
DESCRIPTION
ASSIGNEE_CODES
PK ASSIGNEE_CODE
NAME
COUNT
ASSIGNMENT_TYPES
PK ASSIGNMENT_TYPE
DESCRIPTION
CITE_PATENT
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
CITED_PN
CITED_BY
PUBLISHED
NAME
CLASS
IPC
CITED_PN_RAW
CITE_OTHER
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
CITE_RAW
CITED_BY
INVENTOR
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
NAME
FIRSTNAME
MIDDLENAME
LASTNAME
SUFFIXNAME
ADDRESS
CITY
STATE
COUNTRY
ZIP
CHEMICAL
PK,FK1 PN
PK NAME
SMILE
IN_TITLE
IN_ABS
IN_CLAIMS
IN_BODY
PATENT
PK PN
PK CDLABEL
FK3 COUNTRY
FK3 KIND
PUBLISHED
FILED
PRD
APN
TITLE
FK1 ASSIGNEE_CODE
FK2 ASSIGNMENT_TYPE
MSTATUS
CLAIMS
EXEMP_CLAIMS
EXAMINER_PRIMARY
EXAMINER_ASSISTANT
US_TECH_CENTER
PN_RAW
APN_RAW
INSERTED
LAST_MODIFIED
DESIGNATED_COUNTRY
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
TYPE
COUNTRIES
ASSIGNEE
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
NAME
ADDRESS
CITY
STATE
COUNTRY
ZIP
ASSIGNEE_CODE
ASSIGNMENT_TYPE
ABSTRACT
PK,FK1 PN
PK,FK1 CDLABEL
PK LANGUAGE
ABSTRACT
LEGAL_AGENT
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
LREP_RAW
CLAIMS
PK,FK1 PN
PK,FK1 CDLABEL
CLAIM_RAW
PRIORITY
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
APN
PRD
APN_RAW
CLASS_IPC
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
IPC
EDITION
IS_MAIN
IS_CURRENT
IPC_RAW
CLASS_ECLA
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
ECLA
COUNTRY
SCHEME
ECLA_RAW
CLASS_US
PK,FK1 PN
PK,FK1 CDLABEL
PK NUM
IS_MAIN
IS_CURRENT
USCLASS_RAW
DESCRIPTION
PK,FK1 PN
PK,FK1 CDLABEL
DESC_RAW