Using pattern structures for analyzing ontology-based annotations of biomedical data

Adrien Coulet, Florent Domenach, Mehdi Kaytoue, Amedeo Napoli

Using pattern structures for analyzing ontology-based annotations of biomedical data

Adrien Coulet[1] **Florent Domenach**[2] **Mehdi Kaytoue**[3] **Amedeo Napoli**[1]

[1]LORIA (Université de Lorraine – CNRS – Inria Nancy Grand Est, UMR 7503), BP 239, F-54506 Vandoeuvre-lès-Nancy, France

[2]Computer Science Department, University of Nicosia, 46 Makedonitissas Av., P.O.Box 24005, 1700 Nicosia, Cyprus

[3]Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

{coulet, napoli}@loria.fr, domenach.f@unic.ac.cy, mkaytoue@liris.cnrs.fr

Abstract

Annotating data with concepts of an ontology is a common practice in the biomedical domain. Resulting annotations, *i.e.,* data-concept relationships, are useful for data integration whereas the reference ontology can guide the analysis of integrated data. Then the analysis of annotations can provide relevant knowledge units to consider for extracting and understanding possible correlations between data. Formal Concept Analysis (FCA) which builds from a binary context a concept lattice can be used for such a knowledge discovery task. However annotated biomedical data are usually not binary and a scaling procedure for using FCA is required as a preprocessing, leading to problems of expressivity, ranging from loss of information to the generation of a large number of additional binary attributes. By contrast, pattern structures offer a general FCA-based framework for building a concept lattice from complex data, *e.g.,* a set of objects with partially ordered descriptions. In this paper, we show how to instantiate this general framework when descriptions are ordered by an ontology. We illustrate our approach with the analysis of annotations of drug related documents, and we show the capabilities of the approach for knowledge discovery.

1 Introduction

Annotating data resources with the concepts of an ontology is a common practice in the biomedical domain. The resulting *annotations* are reified as links between data and concepts of a "reference ontology", and provide a support for data exchange, data integration and data analysis tasks [18]. Usually annotations can be built in three main ways, manually, automatically and semi-automatically. In manual annotation, links between data and concepts are provided by human domain experts. In automated annotation, specialized programs are parsing data for providing such links. In semi-automated annotation, specialized programs are suggesting links between data and concepts, that are subsequently validated by domain experts [17].

In the following, we are interested in the analysis of annotations of several data resources from different biomedical domains, e.g. molecular biology and medicine, w.r.t. a reference ontology. Indeed, the annotation process plays a major role in linking these different biomedical domains and understanding their relations.

In this way, this is one objective of *translational bioinformatics* to analyze molecular biology data along with clinical data for discovering correlations between them [6]. Then, hypotheses about molecular mechanisms can be proposed through the discovery of correlations between molecular data and clinical observations. Such correlations can be discovered thanks to the analysis of annotations that link both molecular and clinical data to ontology concepts.

Formal Concept Analysis (FCA) is a mathematical framework for data analysis [8], which is a candidate for our knowledge discovery task. However, some adaptations are required as annotations can be considered as complex data. Firstly, given a reference ontology, annotations are considered as pairs <*document, set of concepts*> and cannot be directly represented as a binary context. Secondly, the ontology that encompasses

the concepts used in the annotation should also be taken into account in the analysis. In FCA, several approaches exist for dealing with complex data.

A first approach is based on *scaling*, which relies on the transformation of non-binary data into binary data. Several types of scaling are known in FCA, *e.g.*, nominal, ordinal, interordinal scalings [8]. But scaling leads to several problems such as an arbitrary transformation of data, a loss of information and a potential binary attribute flooding, forbidding a comprehensive visualization of the results (see for example experiments and discussion in [10]).

Another approach is based on *pattern structures* that allows to directly analyze complex data whose descriptions are taken from a semi-lattice of descriptions [7]. Descriptions may have various types, such as numerical intervals [11], set of attributes [7] or graphs [14]. However, a partial order on descriptions is required in pattern structures. This partial order is defined according to a *similarity operator* and an associated *subsumption relation*. Pattern structures allow for the reuse of standard FCA algorithms with slight modifications, for building the pattern concept lattice and all related operations. It can be noticed that the formalism of pattern structures has gained interest in the last years due to the need for FCA to analyze large volumes of complex data.

In this paper, we present an original approach to analyze annotations based on concepts lying in a reference ontology using the formalism of pattern structures. A first requirement for using pattern structures is to define descriptions of objects, then a similarity operation with its associated subsumption relation (thus a partial ordering on descriptions). In the present case, descriptions are based on concepts lying in a reference ontology. Accordingly, the ordering of concepts in the reference ontology is used to define an original similarity operator on object descriptions and the associated subsumption relation. This is –to the best of our knowledge– the first attempts to analyze data annotations thanks to a pattern structure. Moreover, this shows the potential of pattern structures as an effective formalism for dealing with real-world data. Actually, the resulting pattern concept lattice can be used for guiding a resource annotation process, and for completing annotations that are returned by an automatic annotation tool, that can be possibly wrong or incomplete. This is particularly valuable as the work of a domain expert for correcting and completing annotations is time consuming, especially when large corpora are considered.

From now, and for avoiding any confusion, we use the term "concept" for concept lying in ontologies (represented within Description Logics or DL) and "formal concept" or "pattern concept" for concepts in FCA and pattern structures.

The paper is organized as follows. Section 2 recalls fundamental definitions used in the paper. Section 3 presents our adaptation of pattern structures to ontology-based annotations. It introduces also a concrete example about biomedical data for illustrating the approach. Section 4 details the similarity and subsumption operations on descriptions, while Section 5 provides a discussion about the analysis of annotations of biomedical data using our approach.

2 Background definitions

2.1 Formal Concept Analysis

We recall here the standard FCA notations and we refer readers to [8] for details and proofs. A *formal context* (G, M, I) is defined as a set G of objects, a set M of attributes, and a binary relation $I \subseteq G \times M$. $(g, m) \in I$ means that "the object g is related with the attribute m through the relation I". Two derivation operators can be defined on sets of objects and sets of attributes as follows, $\forall A \subseteq G, B \subseteq M$:

$$A' = \{m \in M : \forall g \in A, (g, m) \in I\}$$

$$B' = \{g \in G : \forall m \in B, (g, m) \in I\}$$

The two operators $(\cdot)'$ define a Galois connection between the power set of objects $\wp(G)$ and the power set of attributes $\wp(M)$. A pair $(A, B), A \subseteq G, B \subseteq M$, is a *formal concept* iff $A' = B$ and $B' = A$. A is called the *extent* and B the *intent* of the formal concept. The set of all formal concepts, ordered by inclusion of extents (or dually by inclusion of intents), *i.e.*, $(A_1, B_1) \leq (A_2, B_2)$ iff $A_1 \subseteq A_2$ (or dually $B_2 \subseteq B_1$), forms a complete lattice [4], called *concept lattice*.

2.2 Pattern structures

A pattern structure can be understood as a generalization of a formal context to analyze complex data [7] : an object has a description lying in a semi-lattice where an "intersection" (or meet) is defined. This intersection allows for characterizing the similarity of two descriptions, i.e. what they do have in common.

Formally, let G be a set of objects, let $(\mathcal{D}, \sqcap)$ be a meet-semi-lattice of object descriptions and let $\delta : G \longrightarrow \mathcal{D}$ be a mapping associating each object with its description. $(G, (\mathcal{D}, \sqcap), \delta)$ is called a pattern structure. Elements of $\mathcal{D}$ are called descriptions or patterns and are ordered by a subsumption relation $\sqsubseteq$ such as $\forall c, d \in \mathcal{D}, c \sqsubseteq d \Longleftrightarrow c \sqcap d = c$. A pattern structure $(G, (\mathcal{D}, \sqcap), \delta)$

gives rise to two derivation operators denoted by $(\cdot)^{\square}$:

$$A^{\square} = \prod_{g \in A} \delta(g) \qquad for\ A \subseteq G$$
$$d^{\square} = \{g \in G | d \sqsubseteq \delta(g)\} \qquad for\ d \in (\mathcal{D}, \sqcap).$$

These operators form a Galois connection between the power set of objects $\wp(G)$ and $(\mathcal{D}, \sqcap)$. Pattern concepts of $(G, (\mathcal{D}, \sqcap), \delta)$ are pairs of the form (A, d), $A \subseteq G$, $d \in (\mathcal{D}, \sqcap)$, such that $A^{\square} = d$ and $A = d^{\square}$. For a pattern concept (A, d), d is the pattern intent and is the common description to all objects in A, the pattern extent. When partially ordered by $(A_1, d_1) \leq (A_2, d_2) \Leftrightarrow A_1 \subseteq A_2 \ \ (\Leftrightarrow d_2 \sqsubseteq d_1)$, the set of all pattern concepts forms a complete lattice called pattern concept lattice. The operator $(\cdot)^{\square\square}$ is a closure operator and pattern intents are closed patterns. Pattern structures have been applied to numerical intervals [11] and to graphs [14].

2.3 $\mathcal{EL}$ ontologies

Ontologies considered in this work are DL ontologies *i.e.*, are based on a set of concepts, relations and individuals represented within Description Logic (DL) [2]. Concepts can be either *atomic* or *defined*. In the first case, their description is reduced to a label and in the second case their description is a DL axiom that includes constructors such as conjunction and existential quantification.

The $\mathcal{EL}$ DL allows for conjunction ($\wedge$) and existential restriction ($\exists r.c$) [1] in the definitions of concepts [1]. This simple DL is sufficient for our purpose, together with transitive roles and general concept inclusion axioms *i.e.*, axioms of the form $C \preccurlyeq D$ where C, D can be either atomic or defined concepts. Moreover, the least common subsumer (lcs) of two concepts in $\mathcal{EL}$ always exists and can be computed in polynomial time, provided that there is no cycle in concept definitions, *i.e.*, the definition of a concept c_i does not make reference to c_i itself [3].

For avoiding any confusion and making a clear distinction between the DL formalism and the pattern structure formalism, we use the classical logical notations for the $\mathcal{EL}$ DL, thus $\wedge$ for conjunction and $\preccurlyeq$ for subsumption, while we keep $\sqcap$ for the similarity operator and $\sqsubseteq$ for the subsumption relation in pattern structures.

In the following, we consider a reference ontology denoted by $\mathcal{O}$ based on the $\mathcal{EL}$ DL. $\mathcal{O}$ is composed of :

- $C(\mathcal{O})$ denotes a set of *concepts* and $R(\mathcal{O})$ denotes a set of binary relations,
- concepts c_i in $C(\mathcal{O})$ are partially ordered thanks to a subsumption relation $\preccurlyeq$, where $c_1 \preccurlyeq c_2$ means that concept c_1 is a sub-concept of c_2 and that every individual that is an instance of c_1 is an instance of c_2,
- A is a set of axioms that describe defined concepts.

3 Problem statement

3.1 The UMLS Semantic Network and semantic types

The UMLS (Unified Medical Language System) is composed of two main components : a set of ontologies of the biomedical domain (such as SNOMED CT, ICD-10, MeSH) and the UMLS Semantic Network [5]. For sake of simplicity, we use a single data resource, DrugBank [2] [12] and a single ontology, the NCI (National Cancer Institute) Thesaurus [19], which belongs to the UMLS. Thus annotations that illustrate our study rely on links between DrugBank and the NCI Thesaurus.

The UMLS Semantic Network provides a set of broad subject categories, or *semantic types*, that is used as a high level classification for concepts of UMLS ontologies [15]. An overview of the 133 semantic types is available at http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html. Semantic types are organized as a tree denoted hereafter as $\mathcal{ST}_{tree}$. For example, some semantic types are more general than others such as "Organism", which is more general than "Human" or "Anatomical Structure", which is more general than "Tissue".

Every concept of a UMLS ontology is mapped to one or more semantic types (*i.e.*, to a non-empty set of semantic types). In addition, the hierarchy of $\mathcal{ST}_{tree}$ can be used to map a concept c_1 to the set of semantic types that are ancestors of the semantic types of c_1. For example, if the concept c_1 has for semantic type "Disease or Syndrome", it can be mapped to "Pathologic Function" and "Biologic Function" too (as the laters are ancestors of the former in $\mathcal{ST}_{tree}$). Accordingly, we are using the hierarchy $\mathcal{ST}_{tree}$ to dispose of the full set of semantic types that can be mapped to each concept. Figure 1 illustrates the mappings of some concepts of the NCI Thesaurus with their semantic types.

In our approach, a selection of semantic types chosen by the analyst will be used as upper level classes for concepts annotating biomedical documents.

1. In addition we used a different operator to distinguish the DL subsumption ($C \preccurlyeq D$) from the partial ordering on pattern concepts ($(A_1, d_1) \leq (A_2, d_2)$) described in 2.2.

2. Publicly available at http://www.drugbank.ca/

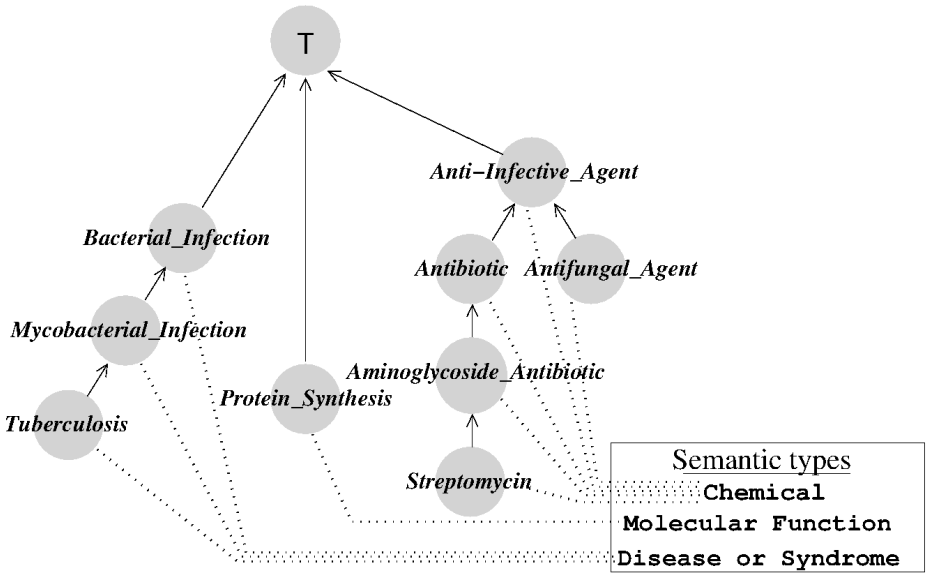

FIGURE 1 – Detail of the NCI Thesaurus with associated semantic types from the UMLS. Nodes are concepts of the ontology, arrows represent subsumption relationships ($\leqslant$). Doted lines map each concept to its semantic type as defined in the UMLS Semantic Network.

3.2 Building a pattern structure for biomedical annotations

In this work, we are interested in the discovery of associations between sets of concepts annotating biomedical documents. This knowledge discovery method should take into account domain knowledge, *i.e.*, the NCI Thesaurus and semantic types. For example, an expert may be interested in a drug-disease association, *e.g.*, Antibiotic-Inflammation, checking whether the association is frequent and searching for a potential associated molecular mechanism.

For analyzing annotations it may be worth to distinguish concepts thanks to domains of interests (kinds of points of view). For example, a domain expert may group concepts according to their membership to distinct portions of an ontology to separate concepts about diseases from concepts about drugs. Accordingly, we consider in this work that the domain expert defines a set of dimensions $\mathcal{ST} = \{\mathtt{st}_1, \mathtt{st}_2, ..., \mathtt{st}_k\}$ where each $\mathtt{st}_i$ is a semantic type. Then a biomedical document will be annotated w.r.t. $\mathcal{ST}$ dimensions. More precisely, given a biomedical document g, the annotation of g w.r.t. the reference ontology $\mathcal{O}$ and $\mathcal{ST}$ dimensions is a pair $(g, \langle \mathtt{ST}_1(g), \mathtt{ST}_2(g), ..., \mathtt{ST}_k(g)\rangle)$ where $\mathtt{ST}_i(g)$ is the set of concepts annotating g for the dimension $\mathtt{st}_i$ of $\mathcal{ST}$ (possibly some of the $\mathtt{ST}_i(g)$ can be empty).

For example, let us consider the document DB01082 (gathering data about Streptomycin) in the DrugBank database. Figure 2 shows this document and an annotation relating three concepts of the NCI Thesaurus (here the reference ontology $\mathcal{O}$). Moreover, let us consider $\mathcal{ST}$ dimensions as $\mathcal{ST}$ = {`Disease or Syndrome`, `Bacterium`, `Molecular Function`, `Chemical`}. Then the annotation of DB01082 can be read as :

$$(\text{DB01082}, \langle \{Tuberculosis\},\ \{\},\ \{Protein\_Synthesis\},\ \{Streptomycin\}\rangle)$$

Now we have everything for defining the pattern structure $(G, (\mathcal{D}, \sqcap), \delta)$ for analyzing annotations of biomedical documents :

- $G = \{g_1, g_2, ..., g_n\}$ is a set of annotated biomedical documents ;
- $\mathcal{O}$ is the reference ontology, *i.e.*, the NCI Thesaurus ;
- $\mathcal{ST} = \{\mathtt{st}_1, \mathtt{st}_2, ..., \mathtt{st}_k\}$ is a subset of semantic types of the UMLS Semantic Network that defines the dimensions of the annotation vector ;
- $\mathcal{D} = \mathcal{P}(\mathtt{st}_1) \times \mathcal{P}(\mathtt{st}_2) \times ... \times \mathcal{P}(\mathtt{st}_k)$ where $\mathcal{P}(\mathtt{st}_i)$ is the power set of the set of concepts of semantic type $\mathtt{st}_i$. As a product of complete lattices, $\mathcal{D}$ is also a complete lattice (and thus a semi-lattice). Elements of $\mathcal{D}$ are named hereafter *ontological patterns* ;
- $\delta : G \to \mathcal{D}$ is a mapping associating a document $g_i \in G$ with a description in $\mathcal{D}$ or more precisely a vector in $\mathcal{D}$,

$$\delta(g_i) = \langle \mathtt{ST}_1(g_i), \mathtt{ST}_2(g_i), ..., \mathtt{ST}_k(g_i)\rangle$$

where $\mathtt{ST}_j(g_i)$ is the set of concepts of semantic type $\mathtt{st}_j$ annotating g_i.

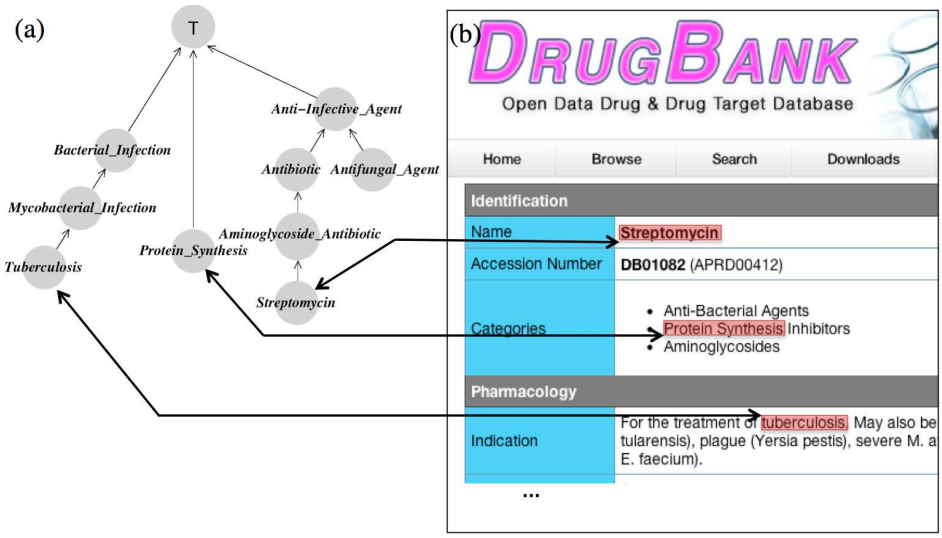

FIGURE 2 – (a) The left part of the Figure shows the NCI Thesaurus ontology ; (b) the right part is an excerpt of the document DB01082 of DrugBank related to the Streptomycin drug. Bold arrows connecting (a) and (b) represent the annotation of DB01082.

Table 1 gives an example of this pattern structure. The fourth row of the table shows the annotation of the document DB01082 (about Streptomycin). The different columns are filled with the concepts annotating DB01082 w.r.t. the semantic type provided in the header of each column.

Now, it remains to define the similarity operation $\sqcap$ between two descriptions $\delta(g_1)$ and $\delta(g_2)$:

$$\delta(g_1) = \langle \mathtt{ST}_1(g_1), \mathtt{ST}_2(g_1), ..., \mathtt{ST}_k(g_1) \rangle$$

$$\delta(g_2) = \langle \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_2) \rangle$$

$$\delta(g_1) \sqcap \delta(g_2) = \langle \mathtt{ST}_1(g_1) \sqcap \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_1) \sqcap \mathtt{ST}_2(g_2),$$

$$..., \mathtt{ST}_k(g_1) \sqcap \mathtt{ST}_k(g_2) \rangle$$

where $\mathtt{ST}_1(g_1) \sqcap \mathtt{ST}_1(g_2)$ is a light notation for $\langle \mathtt{ST}_1(g_1) \rangle \sqcap \langle \mathtt{ST}_1(g_2) \rangle$. $\mathtt{ST}_1(g_1) \sqcap \mathtt{ST}_1(g_2)$ is the *convex hull* in $\mathcal{O}$ of all concepts in $\mathtt{ST}_1(g_1)$ and $\mathtt{ST}_1(g_2)$. The definition of the convex hull is made precise in the next section.

3.3 The similarity between descriptions

Given an ontology $\mathcal{O}$, and two concepts c_1 and c_2, the least common subsumer, denoted by $\mathtt{lcs}(\{c_1, c_2\})$, is the most specific concept subsuming both c_1 and c_2 w.r.t. the ontology $\mathcal{O}$. Here $\mathcal{O}$ is an $\mathcal{EL}$ ontology where no cycle appears in concept definitions, thus the `lcs` of two concepts of $\mathcal{O}$ always exists [3]. Indeed, the existence of the `lcs` is guaranteed as soon as $\mathcal{O}$ has a join semi-lattice structure. The `lcs` operation can be defined (recursively) for a set of concepts $C_n = \{c_1, c_2, ..., c_n\}$ as follows :

$$\forall n \in \mathbb{N},\ \mathtt{lcs}(C_n) = \mathtt{lcs}(\{\mathtt{lcs}(C_{n-1}), c_n\})$$

For example, the `lcs` of *Streptomycin* and $Antifungal\_Agent$ is $Anti-Infective\_Agent$ (see Figure 2).

The `lcs` itself could be used to define a similarity operation between two descriptions. But, an objective here is to complete annotations of documents as much as possible. Thus, the convex hull operation appears to be a better similarity operation. Moreover, if one concept was missed by the annotation process but is available in the ontology, it can be retrieved within the convex hull of the initial set of annotating concepts.

The *convex hull* of the set of concepts $\{c_1, c_2\}$, denoted by $\mathtt{CVX}(\{c_1, c_2\})$, is defined as a set of concepts $\{x_1, x_2, ..., x_n\}$ verifying :

- $x_i \leqslant \mathtt{lcs}(\{c_1, c_2\})$,
- $(x_i \geqslant c_1$ and $x_i \wedge c_1 \equiv c_1)$ or $(x_i \geqslant c_2$ and $x_i \wedge c_2 \equiv c_2)$,
- $x_i \not\equiv \top$

For example, $\mathtt{CVX}(Streptomycin, Antifungal\_Agent)$ = $\{Aminoglycoside\_Antibiotic$, $Anti-Infective\_Agent$, $Antibiotic$, $Antifungal\_Agent$, $Streptomycin\}$.

As the `lcs` operation, the convex hull operation can be generalized (recursively) to a set of concepts $C_p = \{c_1, c_2, ..., c_p\}$:

$$\forall p \in \mathbb{N},\ \mathtt{CVX}(C_p) = \mathtt{CVX}(\{\mathtt{CVX}(C_{p-1}), c_p\})$$

TABLE 1 – A pattern structure where objects are DrugBank documents and attributes are semantic types. Each document is annotated with a set of concepts of the NCI Thesaurus (the reference ontology) having distinct semantic types. The document DB01082 of DrugBank is annotated with three concepts, including the concept *Tuberculosis* of semantic type `Disease or Syndrome`.

G \ $\mathcal{ST}$	Disease or Syndrome	Bacterium	Molecular Function	Chemical
Drug1	{Tuberculosis, Bacterial_Infection}	{}	{Protein_Synthesis}	{Antibiotic, Antifungal_Agent}
Drug2	{Bacterial_Infection}	{}	{Protein_Synthesis}	{}
Drug3	{Tuberculosis, Bacterial_Infection}	{}	{}	{Anti-Infective_Agent}
DB01082	{Tuberculosis}	{}	{Protein_Synthesis}	{Streptomycin}
Drug5	{Tuberculosis, Bacterial_Infection}	{}	{}	{Antibiotic, Antifungal_Agent}

We use the expression "convex hull" by analogy with the Euclidean geometry. In Euclidean geometry, a convex hull of a set of points is the minimal convex set that can be formed by these points. In our case, the convex hull of a set of concepts is the minimal set of concepts including the initial concepts, their least common subsumer and all concepts in between.

The *similarity operation* on descriptions applies to two vectors having the same dimensions and returns a vector where the components are filled with the convex hull of the union of the two initial sets of concepts. Formally we have :

$$\delta(g_1) = \langle \mathtt{ST}_1(g_1), \mathtt{ST}_2(g_1), ..., \mathtt{ST}_k(g_1) \rangle$$

$$\delta(g_2) = \langle \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_2) \rangle$$

$$\delta(g_1) \sqcap \delta(g_2) = \langle \mathtt{ST}_1(g_1) \sqcap \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_1) \sqcap \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_1) \sqcap \mathtt{ST}_k(g_2) \rangle$$

where

$$\mathtt{ST}_i(g_1) \sqcap \mathtt{ST}_i(g_2) = \mathtt{CVX}(\mathtt{ST}_i(g_1) \cup \mathtt{ST}_i(g_2)).$$

It can be noticed that the definition of the similarity operation on concepts can be likened to the the definition of the similarity operation for numerical intervals as the convex hull of two intervals (see for example [11]). Moreover, similarly as for intervals we have the following property :

$$\delta(g_1) \sqcap \delta(g_2) = \delta(g_1) \text{ iff } \delta(g_1) \sqsubseteq \delta(g_2)$$

As an illustration let us consider the two objects "Drug1" and "DB01082" and their descriptions δ(Drug1) and δ(DB01082) given in the Table 1. Their meet is

$$\delta(\text{Drug1}) \sqcap \delta(\text{DB01082}) = \langle \{Bacterial\_Infection, Mycobacterial\_Infection, Tuberculosis\}, \{\}, \{Protein\_Synthesis\}, \{Aminoglycoside\_Antibiotic, Anti-Infective\_Agent, Antibiotic, Antifungal\_Agent, Streptomycin\} \rangle.$$

The meet semi-lattice of pattern elements (actually of convex hulls) defined by the similarity operation is given in Figure 3. This semi-lattice is associated with the context of Table 1 and the order defined by the NCI Thesaurus given in Figure 2.

Dually, it is also possible to define a *join operation* on descriptions, making $(\mathcal{D}, \sqcap, \sqcup)$ a complete lattice. This operation is not necessary for the definition of pattern structures but exists in our case because of the property of $\mathcal{D}$, the space of descriptions. The join of two descriptions $\delta(g_1)$ and $\delta(g_2)$ is defined as follows :

$$\delta(g_1) \sqcup \delta(g_2) = \langle \mathtt{ST}_1(g_1) \sqcup \mathtt{ST}_1(g_2), \mathtt{ST}_2(g_1) \sqcup \mathtt{ST}_2(g_2), ..., \mathtt{ST}_k(g_1) \sqcup \mathtt{ST}_k(g_2) \rangle$$

where

$$\mathtt{ST}_i(g_1) \sqcup \mathtt{ST}_i(g_2) = \mathtt{CVX}(\mathtt{ST}_i(g_1)) \cap \mathtt{CVX}(\mathtt{ST}_i(g_2)).$$

Actually, the result of the join operation is the set of common concepts in the two convex hulls of $\mathtt{ST}_i(g_1)$ and $\mathtt{ST}_i(g_2)$.

For example, the join of the descriptions of "Drug1" and "DB01082" is :

$$\delta(\text{Drug1}) \sqcup \delta(\text{DB01082}) = \langle \{Tuberculosis\}, \{Protein\_Synthesis\}, \{\} \rangle.$$

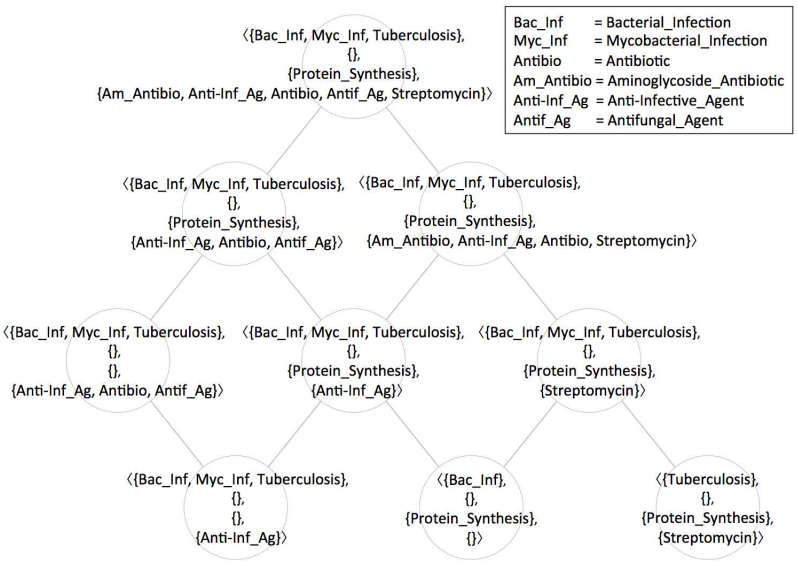

FIGURE 3 – The meet semi-lattice of convex hulls associated with the context represented in Table 1 and the NCI Thesaurus. To enlighten the semi-lattice, we used abbreviations that are clarified in the upper right frame.

The intersection of two convex hulls may be empty as shown in the above example. However, it can be noticed that even if $\delta(g_1)$ and $\delta(g_2)$ may have no common element, they can still have a join as illustrates the following example. Suppose that we have only one dimension and let us consider the reference ontology in Figure 2 :

$$\delta(g_1) = \langle\{Bacterial\_Infection, Tuberculosis\}\rangle$$

$$\delta(g_2) = \langle\{Mycobacterial\_Infection\}\rangle.$$

Actually, the results of the meet and join operations on these two descriptions are :

$$\delta(g_1) \sqcap \delta(g_2) = \langle\{Bacterial\_Infection, Mycobacterial\_ Infection, Tuberculosis\}\rangle$$

$$\delta(g_1) \sqcup \delta(g_2) = \langle\{Mycobacterial\_Infection\}\rangle.$$

In addition, we remark that we do not have $\delta(g_1) \sqcap \delta(g_2) = \delta(g_1)$ as $\delta(g_1)$ is not a convex hull and thus we do not have either $\delta(g_1) \sqsubseteq \delta(g_2)$.

3.4 Computing Pattern Structures with CloseByOne

In FCA, an efficient way of computing closed formal concepts that are the basic bricks of concept lattices is the algorithm CloseByOne [13, 16]. To adapt CloseByOne to the general case of pattern structures, one has to replace the original Galois connection, usually denoted by $(\cdot)'$, with the derivation operator denoted by $(\cdot)^\square$. Below, we give the basic pseudo-code of the algorithm CloseByOne (Algorithms 1 and 2) for computing patterns. In addition to the new derivation operator, one must replace the intersection of standard FCA with the similarity operation on patterns ($\sqcap$, line 5 of Algorithm 2) that is adapted to the nature of patterns. This adaptation of CloseByOne does not affect termination, correctness and complexity of the algorithm.

A simple implementation of Algorithms 1 and 2 is proposed at `github.com/coulet/OntologyPatternIcfca/`.

Alg. 1 CloseByOne.

1: $L = \varnothing$
2: **for each** $g \in G$
3: process$(\{g\}, g, (g^{\square\square}, g^\square))$
4: L is the concept set.

Alg. 2 process$(A, g, (C, D))$ with $C = A^{\square\square}$ and $D = A^\square$ and $<$ the lexical order on object names.

if $\{h | h \in C \backslash A \text{ and } h < g\} = \varnothing$ **then**
2: $L = L \cup \{(C, D)\}$
for each $f \in \{h | h \in G \backslash C \text{ and } g < h\}$
4: $Z = C \cup \{f\}$
$Y = D \sqcap \{f^\square\}$
6: $X = Y^\square$
process$(Z, f, (X, Y))$
8: **end if**

4 Analyzing annotations of biomedical data

We illustrate our approach with the analysis of annotations of DrugBank documents with the ontology "NCI Thesaurus". These annotations are provided by the NCBO (National Center for Biomedical Ontology) Resource Index presented hereafter.

4.1 A repository of annotations : The NCBO Resource Index

The NCBO Resource Index is a repository of annotations automatically populated by a Natural Language Processing tool [9]. This tool parses the textual content of several biomedical databases (*e.g.*, DrugBank, OMIM, ClinicalTrial.gov) searching for occurrences of terms refering to concepts of ontologies. When the name of a concept c_i is found in a document g_i, an annotation *i.e.*, a pair (g_i, c_i), is created and stored. On December 18th, 2012, the NCBO Resource Index contained annotations for 34 databases with concepts of 280 ontologies of the BioPortal [20]. The Resource Index can be queried either by a Web user interface [3] or by a REST Web service [4]. We used the latter to build sets of annotations.

4.2 DrugBank annotations with the NCI Thesaurus

DrugBank is a publicly available database that contains data about drugs, their indications and their molecular targets. The database is organized into documents, or entries, where each document compiles data about one drug. Data in DrugBank are for the main part made of texts is natural language. Figure 2 (b) presents the document of DrugBank in concern with Streptomycin.

Annotations considered in the following relate DrugBank documents and concepts of the "NCI Thesaurus" ontology. The NCI Thesaurus is a broad domain ontology and consequently its annotations may concern either clinics or molecular biology data that can be conjointly explored in translational bioinformatics. Moreover, the NCI Thesaurus is an $\mathcal{EL}$ ontology, thus a `lcs` always exists and its processing is tractable. We used the version 12.04 of the NCI Thesaurus encoded in OWL and available on the NCBO Bioportal [5].

3. Available at `http://bioportal.bioontology.org/resource_index`

4. Documented at `http://www.bioontology.org/wiki/index.php/Resource_Index_REST_Web_Service_User_Guide`

5. NCI Thesaurus 12.04 : `bioportal.bioontology.org/ontologies/1032`

4.3 Interpretation

We propose in Table 1 a context including annotations of five DrugBank documents based on concepts to the NCI Thesaurus. Concepts may have four distinct semantic types ($|\mathcal{ST}| = 4$) : `Disease or Syndrome`, `Bacterium`, `Molecular Function` and `Chemical`. The meet-semi-lattice of patterns associated with such annotations is depicted in Figure 3 and the corresponding pattern concept lattice is given in Figure 4. Both sets of formal concepts in the semi-lattice and in the pattern concept lattice were obtained thanks to the adapted implementation of CloseByOne (see subsection 3.4).

Now we propose an analysis of the resulting concept lattice shown in Figure 4. Consider that one of our objectives is to repair and complete the annotations associated with biomedical documents. The top formal concept in the lattice has the "largest extent", *i.e.*, the set of all the objects, and the "smallest intent", actually the largest convex hull for the annotations.

Let us consider the two formal concepts in the upper left part of the concept lattice, the first called here $c_{\#15}$ has an extent containing "drug1" and "drug5" and the second called here $c_{\#5}$ has an extent containing only "drug5". The `Chemical` semantic type of both concepts is $\{Anti-Infective\_Agent, Antibiotic, Antifungal\_Agent\}$. The `Disease or Syndrome` dimension ("DoS") in both concepts is $\{Bacterial\_Infection, Mycobacterial\_Infection, Tuberculosis\}$ as in the top concept. However, the `Molecular Function` dimension ("MF") is the same for the top concept and $c_{\#15}$, *i.e.*, $\{Protein\_Synthesis\}$; while it is redefined and empty in $c_{\#5}$. This can be interpreted as follows :

- The value of `Chemical` in both $c_{\#15}$ and $c_{\#5}$ is completed (as a convex hull) with *Anti-Infective_Agent* and is the correct annotation to be associated to documents "drug1" and "drug5" for the `Chemical` dimension. This shows how the final pattern concept lattice can effectively complete the original annotation process (especially when this process is automated).
- The same remark applies to the `Disease or Syndrome` dimension, which is also completed (as a convex hull). The concept lattice provides once again the complete annotation for both concepts $c_{\#15}$ and $c_{\#5}$.

Thus, even on this small and toy example, it is possible to understand and verify the usefulness and potential of the approach : the resulting pattern concept lattice yielded by the "ontological pattern structure" provides the means for completing the initial annotations in a way that respects the reference ontology.

Finally, we experimented the pattern structure ap-

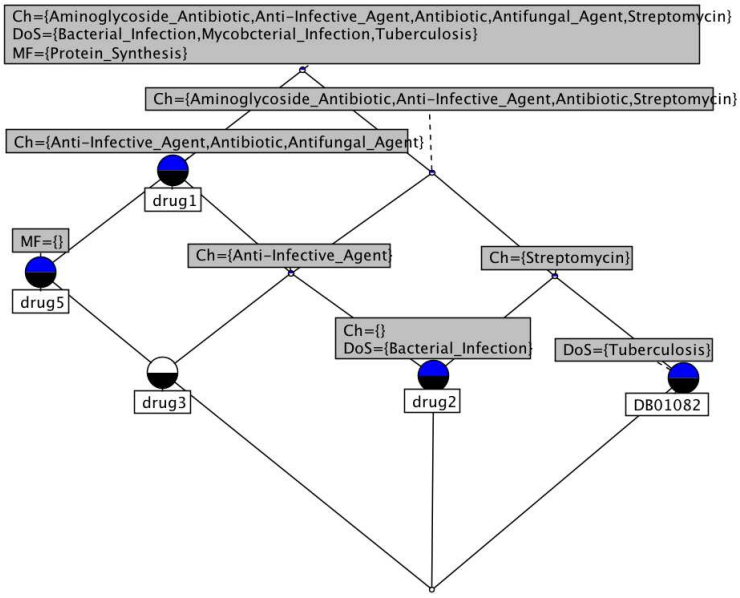

FIGURE 4 – The pattern concept lattice corresponding to the pattern structure given in Table 1 and on the NCI Thesaurus. The top concept has the intent with the larger descriptions and consequently its extent includes all the documents (objects). Traversing the lattice downward, the concepts present more specialized extents and more general intents w.r.t. the subsumption relation on descriptions "Ch", "DoS" and "MF" are respectively abbreviations for the semantic types `Chemical`, `Disease or Syndrome` and `Molecular Function`.

proach on a larger real-world context. We selected 25 drugs of DrugBank out of 173 drugs returned by the query "antibiotic" and we retain the annotations provided by the NCBO Resource Index associated with 4 distinct semantic types. After 4.4 hours, we obtained 204,801 closed concepts on a computer with two Intel Core 2 Extreme X7900 CPUs and 4GiB of memory. The resulting concept lattice is rather large and the analysis of formal concepts with a domain expert is in progress. We think that the results of the analysis will be in accordance with the analysis presented just above for the toy example.

5 Conclusion and Perspectives

Pattern structures provide an original and effective approach within FCA to analyze complex data such as ontology-based annotations of biomedical documents. In this paper, we propose a framework based on pattern structures for dealing with annotations which are made with concepts represented within an $\mathcal{EL}$ ontology. Then we propose a pattern structure providing a classification of biomedical documents according to their annotations and the semantic types of the concepts within the annotations. The resulting concept lattice can be used for analyzing and completing the original annotations.

This work shows that pattern structures are an effective means for dealing with real-world and complex data. In the present case, more experiments remain to be done as well as a thorough study of the various pattern structures that can be associated to an annotation process depending on one or several ontologies.

Acknowledgement

This work was supported in part by the funding agency Campus France (Zenon PHC project 24855NG) and by the Research Promotion Foundation of Cyprus (project KD4CD DIAKRATIKES/KYGA/0310).

Références

[1] Franz Baader, Sebastian Brandt, and Carsten Lutz. Pushing the el envelope. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *IJCAI*, pages 364–369. Professional Book Center, 2005.

[2] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook : Theory, Implementation, and Applications.* Cambridge University Press, 2003.

[3] Franz Baader, Ralf Küsters, and Ralf Molitor. Computing least common subsumers in description logics with existential restrictions. In *IJCAI*, pages 96–103, 1999.

[4] Marc Barbut and Bernard Monjardet, editors. *Ordres et classification : Algèbre et combinatoire (tome II)*. Hachette, Paris, 1970.

[5] Olivier Bodenreider. The Unified Medical Language System (UMLS) : integrating biomedical terminology. *Nucleic Acids Research*, 32(Database-Issue) :267–270, 2004.

[6] Atul J. Butte. Viewpoint paper : Translational bioinformatics : Coming of age. *JAMIA*, 15(6) :709–714, 2008.

[7] Bernhard Ganter and Sergei O. Kuznetsov. Pattern Structures and Their Projections. In *ICCS*, volume 2120 of *LNCS*, pages 129–142. Springer, 2001.

[8] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis*. Springer, mathematical foundations edition, 1999.

[9] Clément Jonquet, Paea LePendu, Sean M. Falconer, Adrien Coulet, Natalya Fridman Noy, Mark A. Musen, and Nigam H. Shah. NCBO Resource Index : Ontology-based search and mining of biomedical resources. *J. Web Sem.*, 9(3) :316–324, 2011.

[10] Mehdi Kaytoue, Sergei O. Kuznetsov, and Amedeo Napoli. Revisiting numerical pattern mining with formal concept analysis. In *IJCAI*, pages 1342–1347, 2011.

[11] Mehdi Kaytoue, Sergei O. Kuznetsov, Amedeo Napoli, and Sébastien Duplessis. Mining gene expression data with pattern structures in formal concept analysis. *Inf. Sci.*, 181(10) :1989–2001, 2011.

[12] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, Anchi Guo, and David S. Wishart. DrugBank 3.0 : a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, 39(Database-Issue) :1035–1041, 2011.

[13] Sergei O. Kuznetsov. A fast algorithm for computing all intersections of objects in a finite semilattice. *Automatic Documentation and Mathematical Linguistics*, 27(5) :400–412, 2004.

[14] Sergei O. Kuznetsov and Mikhail V. Samokhin. Learning closed sets of labeled graphs for chemical applications. In Stefan Kramer and Bernhard Pfahringer, editors, *ILP*, volume 3625 of *Lecture Notes in Computer Science*, pages 190–208. Springer, 2005.

[15] Alexa T. McCray. An upper level ontology for the biomedical domain. *Comp. Funct. Genom.*, 4 :80–84, 2003.

[16] Jan Outrata and Vilém Vychodil. Fast algorithm for computing fixpoints of galois connections induced by object-attribute relational data. *Inf. Sci.*, 185(1) :114–127, 2012.

[17] Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7), 2009.

[18] Daniel L. Rubin, Nigam Shah, and Natalya Fridman Noy. Biomedical ontologies : a functional perspective. *Briefings in Bioinformatics*, 9(1) :75–90, 2008.

[19] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI Thesaurus : A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1) :30–43, 2007.

[20] Patricia L. Whetzel, Natalya Fridman Noy, Nigam H. Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. BioPortal : enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(Web-Server-Issue) :541–545, 2011.