International Journal of Foundations of Computer Science
Vol. 19, No. 2 (2008) 345–358

World Scientific
www.worldscientific.com

ADJUSTING ANNOTATED TAXONOMIES

TIM B. KAISER
Department of Mathematics, Darmstadt University of Technology, 64289 Darmstadt, Germany

and

STEFAN E. SCHMIDT
Institute of Algebra, Technische Universität Dresden, 01062 Dresden, Germany

and

CLIFF A. JOSLYN
Knowledge Systems Group, Pacific Northwest National Laboratory Richland, WA 99352, USA

Received 15 February 2007
Accepted 30 November 2007
Communicated by Sadok Ben Yahia and Engelbert Mephu Nguifo

We show how the concept of an *annotated ordered set* can be used to model large taxonomically structured ontologies such as the Gene Ontology. By constructing a formal context consistent with a given annotated ordered set, their concept lattice representations are derived. We develop the fundamental mathematical relations present in this formulation, in particular deriving a conceptual pre-ordering of the taxonomy, and constructing a correspondence between the annotations of an ordered set and the closure systems of its filter lattice. We study an example from the Gene Ontology to demonstrate how the introduced technique can be utilized for taxonomy review.

Keywords: Annotated ordered sets; concept lattices; knowledge representation; formal concept analysis; gene ontology; taxonomies.

1. Introduction

This work is an extension of [9].

Ontologies, taxonomies, and other semantic hierarchies are increasingly necessary for organizing large quantities of data, and recent years have seen the emergence of new large taxonomically structured ontologies such as the Gene Ontology (GO) [1][a], the UMLS Meta-Thesaurus [2], object-oriented typing hierarchies [10], and verb typing hierarchies in computational linguistics [4]. Cast as Directed Acyclic

[a]http://www.geneontology.org

Graphs (DAGs), these all entail canonical mathematical representations as *annotated ordered sets* (previously called "poset ontologies" [6]).

The size and complexity of these modern taxonomic hierarchies requires algorithmic treatement of tasks which could previously be done by hand or by inspection. These include reviewing the consistency and completeness of the underlying hierarchical structure, and the coherence of the *labeling* (the assignment of objects to ontological categories). The close similarity of the annotated ordered set representations of these taxonomies to concept lattices in Formal Concept Analysis (FCA) [5] suggests pursuing their representation within FCA, in order to gain a deeper understanding of their mathematical structure and optimize their management and analytical tractibility (see also [7]).

We begin this paper by defining annotated ordered sets, and demonstrate their appropriateness for representing the GO. Then, we define a formal context appropriate for annotated ordered sets, and thereby construct their concept lattices. We analyze the relationship between an annotated ordered set and its concept lattice representation, which includes the formulation of a correspondence between the annotations of an ordered set and the closure systems of its filter lattice. Also, we introduce the fundamental concept of an adjustment of an annotated ordered set and provide quantitative measures of the impact of adjustments. Additionally, we study an example from the GO. The paper is concluded with a discussion of future applications and extensions of the outlined approach. Throughout, we assume that the reader is knowledgable of the theory of FCA [5].

2. Taxonomic Ontologies as Annotated Ordered Sets

We use the GO as our touchstone for the general concept of an annotated ordered set. Fig. 1 (from [1]) shows a sample portion of the GO. Nodes in black represent functional categories of biological processes, basically things that proteins "do". Nodes are connected by links indicating subsumptive, "is-a" relations between categories, so that, for example, "DNA ligation" is a kind of "DNA repair". Elsewhere in the GO, nodes can also be connected by compositional, "has-part" relations, but for our purposes, we will consider the GO as singly-typed. It should be emphasized that Fig. 1 shows only a small fragment of the GO, which currently has on the order of 20,000 nodes in three disjoint taxonomies, annotated by hundreds of thousands of proteins from dozens of species.

Colored terms attached to each node indicate particular proteins in particular species which perform those functions. This assignment is called "annotation". Note that proteins can be annotated to multiple functions, for example yeast MCM2 does both "DNA initiation" and "DNA unwinding". Furthermore, an annotation to a node should be considered a simultaneous annotation to all ancestor nodes, so that yeast CDC9 does both "DNA ligation" and "DNA repair". So explicit such annotations, for example CDC9 annotation to both "DNA ligation" and "DNA recombination" in Fig. 1, are actually redundant. Finally, note the presence of multiple inheritance: "DNA ligation" is both "DNA repair" and "DNA recombination".

It is therefore appropriate to model structures such as the GO as structures called annotated ordered sets (previously referred to as poset ontologies [6]).

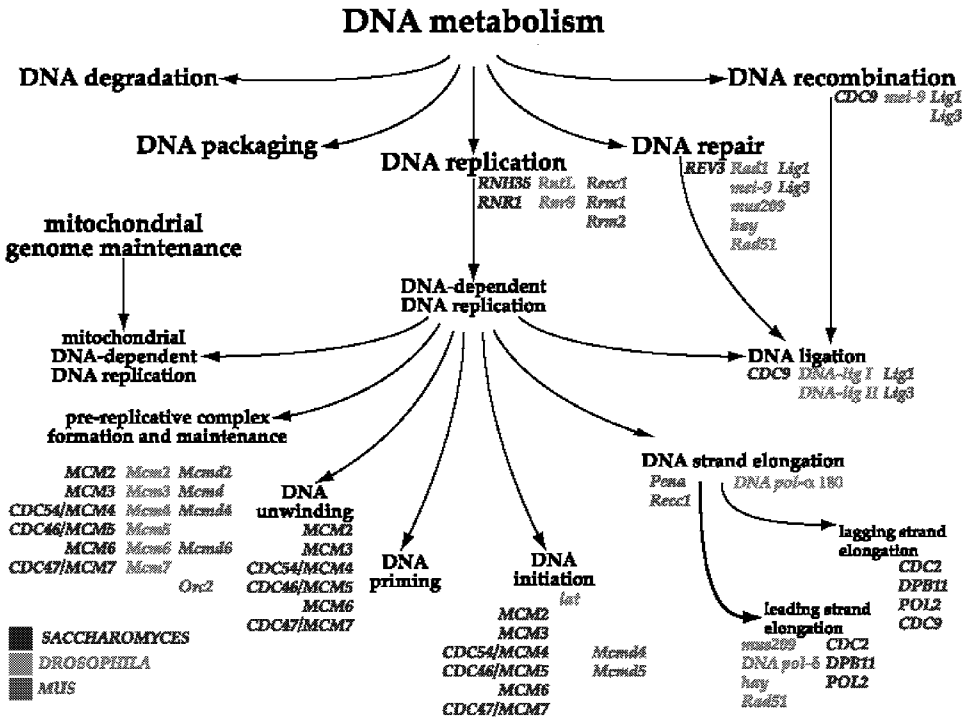

Fig. 1. A portion of the BP branch of the GO (used with permission from [1]). GO nodes in the hierarchy have genes from three species annotated below them.

Definition 1 (Annotated Ordered Set) Let $\mathcal{P} := (P, \leq_{\mathcal{P}})$ be a finite ordered set (poset), let X be a finite set of *labels*, and let $F : X \to 2^P$ be an *annotation function*. Then we call $\mathbb{O} := (\mathcal{P}, X, F)$ an *annotated ordered set* and refer to (X, F) as an *annotation* of $\mathcal{P}$. In case $\mathcal{P}$ is a (complete) lattice we call $\mathbb{O}$ an *annotated (complete) lattice* denoted $\mathbb{L}$. If $|F(x)| = 1$ for all $x \in X$, for convenience, we regard F as a map from X to P and say that $\mathbb{O}$ is *elementary*.

The annotated ordered sets form a category under the following concept of morphism. This will enable us to compare annotated ordered sets (and their *adjustments* as will be seen later) adequately.

Definition 2 (Morphism) Let $\mathbb{O}_1$ and $\mathbb{O}_2$ be annotated ordered sets where $\mathbb{O}_i := (\mathcal{P}_i, X_i, F_i)$ and $\mathcal{P}_i := (P_i, \leq_i)$. Then a pair of maps, (μ, λ), will be called a morphism from $\mathbb{O}_1$ to $\mathbb{O}_1$ if μ is an order-preserving map from $\mathcal{P}_1$ to $\mathcal{P}_2$ and λ is a map from X_1 to X_2 such that

$$\mu(F_1(x)) \subseteq F_2(\lambda(x))$$

holds for all $x \in X_1$. The morphism (μ, λ) is called *strong* if

$$\mu(F_1(x)) = F_2(\lambda(x))$$

holds for all $x \in X_1$.

As usual, the powerset of a set X will be denoted by 2^X, and for a map f from a set X to a set Y, by abuse of notation, the powerset lifting f will be given by the map 2^f from 2^X to 2^Y with $2^f(T) := \{f(t) \,|\, t \in T\}$ for all $T \in 2^X$. Now, a morphism as defined above can be captured by the following semi-commutative diagram:

$$\begin{array}{ccc} X_1 & \xrightarrow{\lambda} & X_2 \\ {\scriptstyle \mathrm{F}_1}\downarrow & & \downarrow{\scriptstyle \mathrm{F}_2} \\ 2^{P_1} & \xrightarrow{2^\mu} & 2^{P_2}. \end{array}$$

3. Concept Lattice Representations

We are now prepared to construct concept lattice representations of annotated ordered sets by deriving the appropriate formal contexts. For an ordered set $\mathcal{P} := (P, \leq_\mathcal{P})$ and node $q \in P$ we denote by $\uparrow q := \{p \in P \,|\, q \leq_\mathcal{P} p\}$ the principal filter of q and dually by $\downarrow q$ the principal ideal. In general, for $Q \subseteq P$ we define $\uparrow Q := \{p \in P \,|\, \exists q \in Q : q \leq_\mathcal{P} p\}$ and dually $\downarrow Q$. Given an annotated ordered set $\mathbb{O} := (\mathcal{P}, X, F)$ we can construct a formal context $\mathbb{K}_\mathbb{O} := (X, P, I)$ where

$$xIp :\Longleftrightarrow \downarrow p \cap F(x) \neq \emptyset$$

for $x \in X, p \in P$. Note also that

$$xIp \iff \exists q \leq_\mathcal{P} p : q \in F(x) \iff p \in \bigcup_{q \in F(x)} \uparrow q.$$

The concept lattice of $\mathbb{K}_\mathbb{O}$ will be denoted by $\underline{\mathfrak{B}}_\mathbb{O} := (\mathfrak{B}_\mathbb{O}, \leq_{\underline{\mathfrak{B}}_\mathbb{O}})$, where $\mathfrak{B}_\mathbb{O} := \mathfrak{B}(\mathbb{K}_\mathbb{O})$ is the set of formal concepts of the formal context $\mathbb{K}_\mathbb{O}$ [5]. $\underline{\mathfrak{B}}_\mathbb{O}$ is called the *concept lattice representation* of the annotated ordered set $\mathbb{O}$.

In case $\mathbb{O}$ forms an annotated complete lattice and $(A, B) \in \mathfrak{B}_\mathbb{O}$ is a formal concept in $\mathfrak{B}_\mathbb{O}$, we observe that $A = B^I$ is the set of all $x \in X$ such that $\bigwedge B$ is an upper bound of $F(x)$. Also, for convenience, for a node $p \in P$ denote $p^I := \{p\}^I \subseteq X$.

We can define a new relation on P induced by the concept lattice $\underline{\mathfrak{B}}_\mathbb{O}$. Let $\mu_\mathbb{O}$ be the map which assigns to every element p in P its attribute concept (p^I, p^{II}). The range of $\mu_\mathbb{O}$ is denoted by $P_\mathbb{O}$. We say p is *conceptually* less or equal than q if and only if $\mu_\mathbb{O}(p) \leq_{\underline{\mathfrak{B}}_\mathbb{O}} \mu_\mathbb{O}(q)$, denoted by $p \sqsubseteq_\mathbb{O} q$. We call $\sqsubseteq_\mathbb{O}$ the *conceptual pre-order* of $\mathbb{O}$. In general, the relation $\sqsubseteq_\mathbb{O}$ is not an order since for different $p, q \in P$ the corresponding attribute concepts $\mu_\mathbb{O}(p)$ and $\mu_\mathbb{O}(p)$ can match. Two annotations $(X, F_1), (X, F_2)$ of $\mathcal{P}$ are called *annotationally equivalent* if their conceptual pre-orders coincide.

Definition 3 (Adjustment) For an annotated ordered set $\mathbb{O} := (\mathcal{P}, X, F)$, the *adjustment* of $\mathbb{O}$ is given by the annotated ordered set $\mathbf{Ad}(\mathbb{O}) := (\mathcal{P}_\mathbb{O}, X, F_\mathbb{O})$ where

$$F_\mathbb{O}(x) := \{\mu_\mathbb{O}(p) \,|\, x \in p^I\} \text{ for all } x \in X$$

and $\mathcal{P}_\mathbb{O} := (P_\mathbb{O}, \leq_\mathbb{O})$ with $\leq_\mathbb{O} := \leq_{\mathbb{B}_\mathbb{O}} \cap P_\mathbb{O} \times P_\mathbb{O}$.

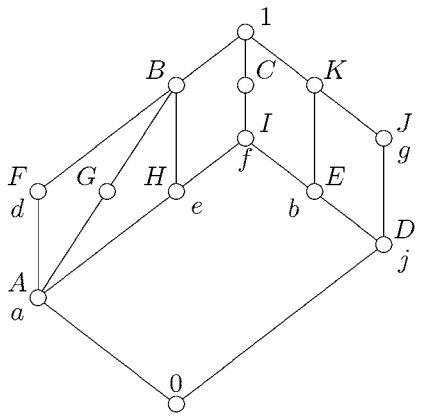

Fig. 2. Example of an annotated lattice.

	0	A	B	C	D	E	F	G	H	I	J	K	1
a		×	×	×			×	×	×				×
b				×		×				×		×	×
d			×				×						×
e			×	×					×	×			×
f				×						×			×
g											×	×	×
j				×	×	×				×	×	×	×

Fig. 3. Context for the annotated lattice in Fig. 2.

Proposition 1 Let $\mathbb{O} := (\mathcal{P}, X, F)$ be an annotated ordered set. Then $(\mu_{\mathbb{O}}, id_X)$ is a morphism from $\mathbb{O}$ to its adjustment $\mathbf{Ad}(\mathbb{O})$. □

Let us agree to call an annotated ordered set *adjusted* if it is isomorphic to its adjustment. Though an annotated ordered set is not necessarily adjusted, every adjustment is.

In Sections 4, 5, and 7 we will explore the relationship between the original ordered set $\mathcal{P}$ and the constructed conceptual pre-order $(P, \sqsubseteq_{\mathbb{O}})$ and hint at potential applications arising from this comparison.

An example for an elementary annotated lattice $\mathbb{L} := (\mathcal{P}, X, F)$ is given in Fig. 2 where $\mathcal{P} := (\{A, B, \ldots, K, 0, 1\}, \leq_{\mathcal{P}}), X = \{a, b, d, e, f, g, j\}$, and F and $\leq_{\mathcal{P}}$ are defined as illustrated. Fig. 3 shows the formal context $\mathbb{K}_{\mathbb{L}}$, and Fig. 4 the resulting concept lattice $\underline{\mathfrak{B}}_{\mathbb{O}}$, which in this case also is the adjustment of $\mathbb{L}$.

4. Mathematical Properties of Concept Lattice Representations

In the first part of this section we will analyze how the order of the annotated ordered set and the order of its concept lattice are related. In the second part we will investigate how the concept lattice representations, derived from a given ordered set using different annotations, can be classified.

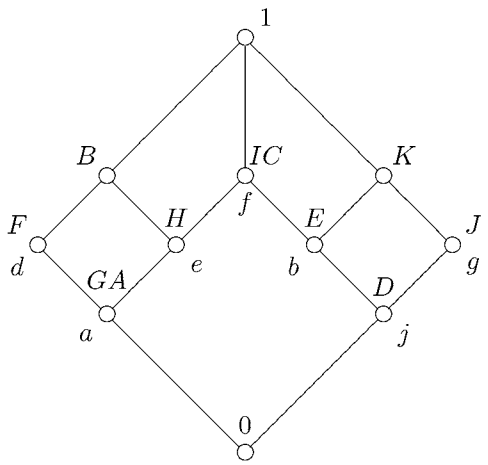

Fig. 4. Concept lattice representation of the annotated lattice in Fig 2.

4.1. *Annotated ordered sets and their conceptual pre-order*

The following proposition connects an annotated ordered set with its concept lattice representation. In the following, let $\mathbb{O} := (\mathcal{P}, X, F)$ be an annotated ordered set. We define $\mu : P \to \mathfrak{B}_{\mathbb{O}}$ via $\mu(p) = (p^I, p^{II})$ for all $p \in P$, that is, it maps each poset node to its *attribute concept* in $\mathfrak{B}_{\mathbb{O}}$. Note, that μ is the composition of $\mu_{\mathbb{O}}$ with the canonical embedding of $\mathcal{P}_{\mathbb{O}}$ into the concept lattice representation of $\mathbb{O}$.

Proposition 2 Let $\mathbb{O}$ be an annotated ordered set. Then μ constitutes an order-homomorphism between $\mathcal{P}$ and the concept lattice representation of $\mathbb{O}$.

Proof. Let $p, q \in P$. Then we have $p \leq_{\mathcal{P}} q \Longleftrightarrow \downarrow p \subseteq \downarrow q$ which implies

$$p^I = \{x \in X \mid \downarrow p \cap F(x) \neq \emptyset\} \subseteq \{x \in X \mid \downarrow q \cap F(x) \neq \emptyset\} = q^I.$$

Since the last statement is equivalent to $\mu(p) \leq_{\mathfrak{B}_{\mathbb{O}}} \mu(q)$ this asserts the proposition. □

By definition of $\leq_{\mathbb{O}}$, it follows that $p \leq_{\mathcal{P}} q \implies p \sqsubseteq_{\mathbb{O}} q$. Clearly, the converse is wrong, since in general $\sqsubseteq_{\mathbb{O}}$ is only a pre-order. Even for the factor order associated with the pre-order, the converse implication does not hold as is verified by the example shown in Figure 5, where $c \sqsubseteq_{\mathbb{O}} a$, but c and a are non-comparable in $\mathcal{P}$. But for adjusted annotated ordered sets we have a nice situation.

Proposition 3 Let $\mathbb{O}$ be an adjusted annotated ordered set. Then μ constitutes an order-reflecting embedding of $\mathcal{P}$ into the concept lattice representation of $\mathbb{O}$.

For elementary annotated complete lattices we find the following connection which goes further than the results for annotated ordered sets.

Proposition 4 Let $\mathbb{L} = (\mathcal{P}, X, F)$ be an elementary annotated complete lattice. The concept lattice of $\mathbb{L}$ is order-embedded into $\mathcal{P}$ via the map $\varphi : \mathfrak{B}_{\mathbb{L}} \to P$ where $(A, B) \mapsto \bigwedge B$.

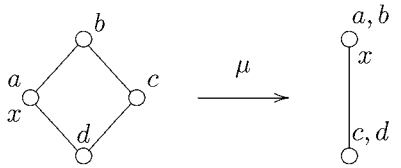

Fig. 5. Counter-example to μ inducing an order isomorphism.

Proof. For all $\mathfrak{c}_1, \mathfrak{c}_2 \in \mathfrak{B}_{\mathbb{L}}$, we have to show that $\mathfrak{c}_1 \leq_{\underline{\mathfrak{B}}} \mathfrak{c}_2$ holds if and only if $\varphi(\mathfrak{c}_1) \leq_{\mathcal{P}} \varphi(\mathfrak{c}_2)$. Let $\mathfrak{c}_1 = (A, B)$ and $\mathfrak{c}_2 = (C, D)$ be concepts in $\mathfrak{B}_{\mathbb{L}}$.
"$\Rightarrow$": Assume $(A, B) \leq_{\underline{\mathfrak{B}}} (C, D)$. This is equivalent to $D \subseteq B$ which implies $\bigwedge B \leq_{\mathcal{P}} \bigwedge D$.
"$\Leftarrow$": Assume $\bigwedge B \leq_{\mathcal{P}} \bigwedge D$. Since $\mathbb{L}$ is elementary, B^I is the set of all labels $x \in X$ such that $\bigwedge B$ is an upper bound of $F(x)$ it follows that $B^I \subseteq D^I$ and therefore we have $(B^I, B) \leq_{\underline{\mathfrak{B}}} (D^I, D)$ as required. □

For elementary annotated lattices the previously introduced mappings μ and φ combine in a surprising way.

Theorem 1 Let $\mathbb{L} := (\mathcal{P}, X, F)$ be an elementary annotated complete lattice. Then (φ, μ) forms a residuated pair between the concept lattice representation of $\mathbb{L}$ and $\mathcal{P}$. In particular, φ is an injective $\bigvee$-morphism and μ is a surjective $\bigwedge$-morphism.

Proof. Firstly, we deduce from Proposition 4 that φ is injective. For residuated pairs this implies the surjectivity of the second map. It remains to show that (φ, μ) forms a residuated pair.
Since $\mathbb{L}$ is elementary, F can be regarded as a map from X to P and then the incidence relation I of $\mathbb{K}_{\mathbb{L}}$ is defined via xIp if and only if $F(x) \leq_{\mathcal{P}} p$; therefore $x^I = {\uparrow} F(x)$ for all $x \in X$. In the following let (A, B) be an arbitrary concept in $\mathfrak{B}_{\mathbb{L}}$. We derive $B = A^I = \bigcap_{a \in A} x^I = \bigcap_{a \in A} {\uparrow} F(x) = {\uparrow} \bigvee F(A)$; hence, we receive $\varphi(A, B) = \bigwedge B = \bigvee F(A)$. We conclude the proof as follows:

$$\begin{aligned} \varphi(A,B) \leq_{\mathcal{P}} p \quad &\Longleftrightarrow \quad \bigvee F(A) \leq_{\mathcal{P}} p \quad \Longleftrightarrow \quad p \in {\uparrow} \bigvee F(A) = A^I \\ &\Longleftrightarrow \quad A \subseteq p^I \quad \Longleftrightarrow \quad (A,B) \leq_{\underline{\mathfrak{B}}} \mu(p) \end{aligned}$$

□

As a consequence of our theorem we know that φ embeds the concept lattice representation of an elementary annotated complete lattice into its underlying lattice as a kernel system. This fact applies to the example from Figure 2 and is visualized in Fig. 6.

Though, in general, the concept lattice representations of elementary annotated ordered sets cannot be embedded into their underlying ordered set, it is feasible to embed them into a well-known extension of the former. For a subset Q of an ordered set $\mathcal{P}$, we will use the notation $Q^{\downarrow}$ for the set of all lower bounds of Q in $\mathcal{P}$.

Theorem 2 Let $\mathbb{O} := (\mathcal{P}, X, F)$ be an elementary annotated ordered set and let $\mathbb{K} := (P, P, \leq_{\mathcal{P}})$. Then the map $\mathfrak{B}_{\mathbb{O}} \to \mathfrak{B}(\mathbb{K})$ with $(A, B) \mapsto (B^{\downarrow}, B)$ forms a $\bigvee$-embedding of the concept lattice representation of $\mathbb{O}$ into the Dedekind- MacNeille completion of $\mathcal{P}$.

Proof. Firstly, we refer to Theorem 4 in [5], p.48, for details regarding the Dedekind-MacNeille completion.
Since $\mathbb{O}$ is elementary, $x^I = \uparrow F(x)$ is an intent not only of $\mathbb{K}_{\mathbb{O}}$ but also of $\mathbb{K}$ for every $x \in X$; trivially, p^I is an extent of $\mathbb{K}_{\mathbb{O}}$ for every $p \in P$. By Definition 69 in [5], p. 185, this means that I is a bond from $\mathbb{K}_{\mathbb{O}}$ to $\mathbb{K}$. Now Corollary 112 in [5], p. 256, implies that the map φ_I from $\underline{\mathfrak{B}}_{\mathbb{O}}$ to $\underline{\mathfrak{B}}(\mathbb{K})$ with $\varphi_I(A, B) = (A^{I\downarrow}, A^I) = (B^{\downarrow}, B)$ is a $\bigvee$-morphism, which clearly is injective. □

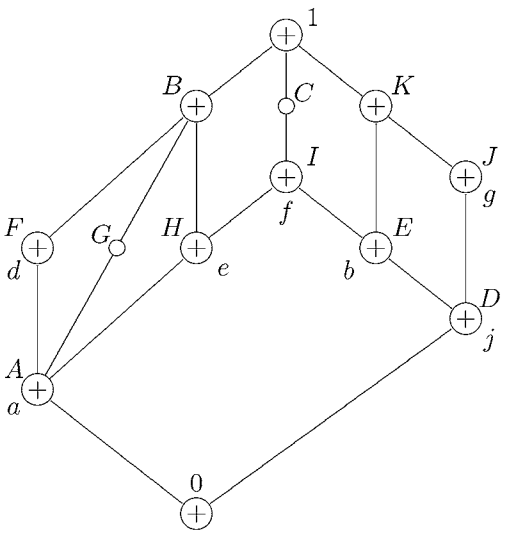

Fig. 6. The concept lattice representation from Fig. 4 embedded as kernel system in its annotated lattice from Fig. 2

4.2. *Classifying the annotations of an ordered set*

We start with giving two rather extreme examples for different concept lattices derived from the same ordered set via different annotations. In the following, it is more convenient to regard an annotated ordered set $\mathbb{O} := (\mathcal{P}, X, F)$ as a formal context with an ordered set of attributes. Since the annotation function $F : X \to 2^P$ set-theoretically is a relation $F \subseteq X \times P$, the formal context (X, P, F) together with the ordered set $\mathcal{P} = (P, \leq_{\mathcal{P}})$ yields another way of looking at an annotated ordered set. It is obvious, that the formal context $\mathbb{K}_{\mathbb{O}}$ is equal to $(X, P, F \circ \leq_{\mathcal{P}})$, where $\circ$ denotes the relational product. We recall Theorem 4 from [5] which states that for an ordered set $\mathcal{P}$ its Dedekind-MacNeille completion is isomorphic to the concept lattice of the formal context $(P, P, \leq_{\mathcal{P}})$. Now it is easy to see, that the *identical* labelling function $F_{id} : P \to 2^P$ with $p \mapsto \{p\}$ yields an annotated ordered

set $\mathbb{O}_{id} := (\mathcal{P}, P, F_{id})$ which is isomorphic to the Dedekind-MacNeille completion of $\mathcal{P}$, because $F_{id} \circ \leq_{\mathcal{P}} = \leq_{\mathcal{P}}$ which yields $\mathbb{K}_{\mathbb{O}} = (P, P, \leq_{\mathcal{P}})$. On the other hand – as complicated as $\mathcal{P}$ might be – if the labelling function is constant with $F_P(x) = P$ for any label $x \in X$ we get a formal context with $I = X \times P$. That means, the concept lattice representation shrinks the ordered set into a single element.

To get a more comprehensive description of the interplay of the annotations of an ordered set $\mathcal{P}$ and its concept lattice representations we will use the *filter lattice* of an ordered set $\mathcal{P}$ – which is defined as $\mathcal{F}(\mathcal{P}) := (\{F \subseteq P \mid \uparrow F = F\}, \subseteq)$ – as a framing structure.

Theorem 3 The annotations of an ordered set $\mathcal{P} = (P, \leq_{\mathcal{P}})$ are, up to annotational equivalence, in one-to-one correspondence to the closure systems in the filter lattice of $\mathcal{P}$.

Proof. Let $x \in X$ be a label. The object intent $x^{F \circ \leq_{\mathcal{P}}}$ of x in $(X, P, F \circ \leq_{\mathcal{P}})$ is of the form $\{p \in P \mid \exists q \in x^F : q \leq_{\mathcal{P}} p\}$ which is equal to the filter $\uparrow x^F$ in $\mathcal{P}$. Since the intents of all concepts of a concept lattice are exactly the meets of the object intents, the intents of the concepts of $\mathfrak{B}(X, P, F \circ \leq_{\mathcal{P}})$ are exactly the meets of filters of the form $x^{F \circ \leq_{\mathcal{P}}}$ with $x \in X$, and therefore, form a closure system in the filter lattice of $\mathcal{P}$.

Let us assume that $\mathcal{X} \subseteq 2^P$ is a closure system in the filter lattice of $\mathcal{P}$. We consider the formal context $(\mathcal{X}, P, \ni)$. For $X \in \mathcal{X}$, we get $X^{\ni} = \{p \in P \mid p \in X\} = X$. Therefore the intents of the associated concept lattice constitute exactly the closure system $\mathcal{X}$. And since in our situation $\ni \circ \leq_{\mathcal{P}}$ is equal to $\ni$, an annotation $(\mathcal{X}, \ni)$ corresponding to $\mathcal{X}$ is found. □

The above theorems say that the cosmos of possible structures which can be produced via annotating an ordered set and forming its concept lattice are restricted to closure systems in the filter lattice of the original ordered set – and also exhaust them.

5. Application to the Gene Ontology

In this section we apply our proposed technique to the GO cutout depicted in Figure 1. The given diagram can be seen as an annotated ordered set where the underlying ordered set consists of the functional categories of biological processes (as e.g. *DNA replication*) and the order is given by the arrows. The set of labels consists of the proteins and the annotation function maps a protein to a function category if it is listed at the respective function category node. Clearly, this annotated ordered set can not be interpreted as an annotated lattice, since infima and suprema do not exist for any subset of nodes, e.g. the infimum over all function categories is not present. Also the annotation function attaches some proteins to several nodes as it is the case for *Lig1* and *Lig3* who are attached to the functional categories of *DNA ligation*, *DNA recombination*, and *DNA repair*. Figure 7 shows a diagram of the concept lattice representation of this annotated ordered set where we have omitted function categories where there is no protein attached to the nodes or to some subnode. Figure 8 shows the conceptual pre-ordering of the functions derived from the concept lattice representation (in this case it is an order).

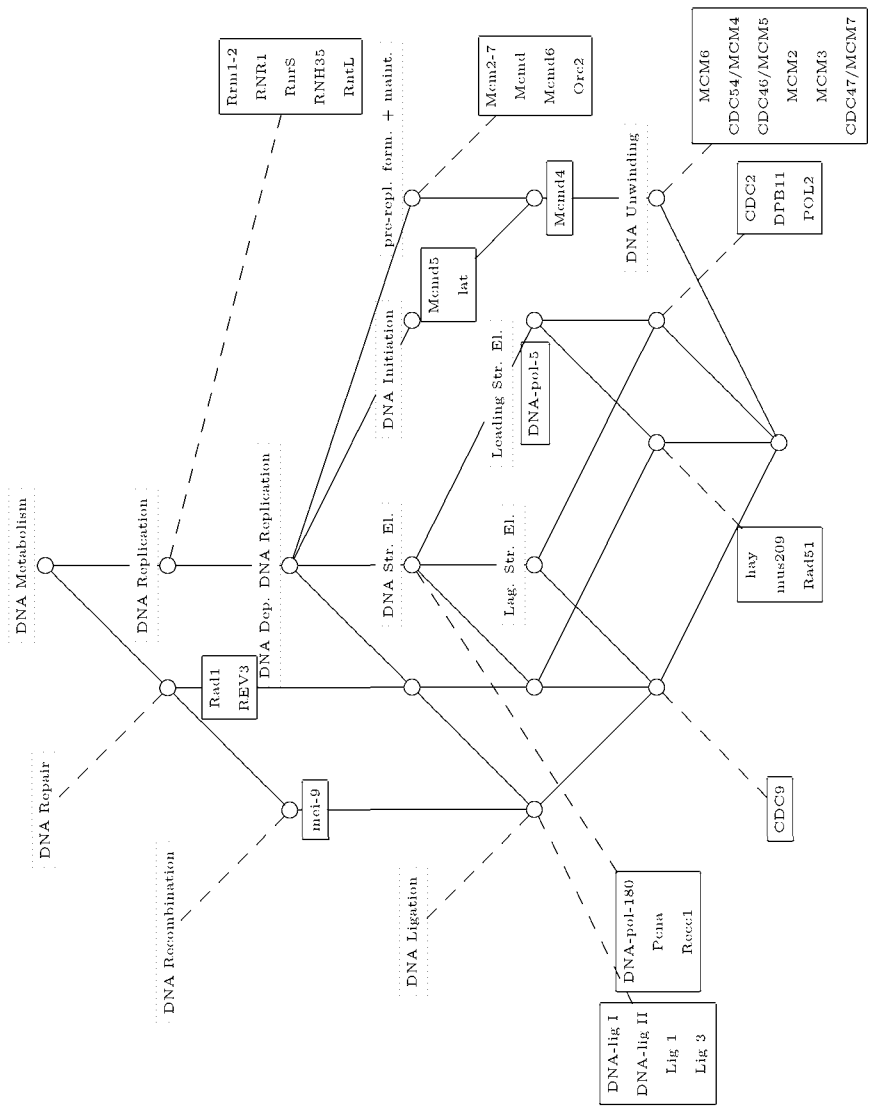

Fig. 7. Concept lattice for the GO cutout depicted in Fig. 1.

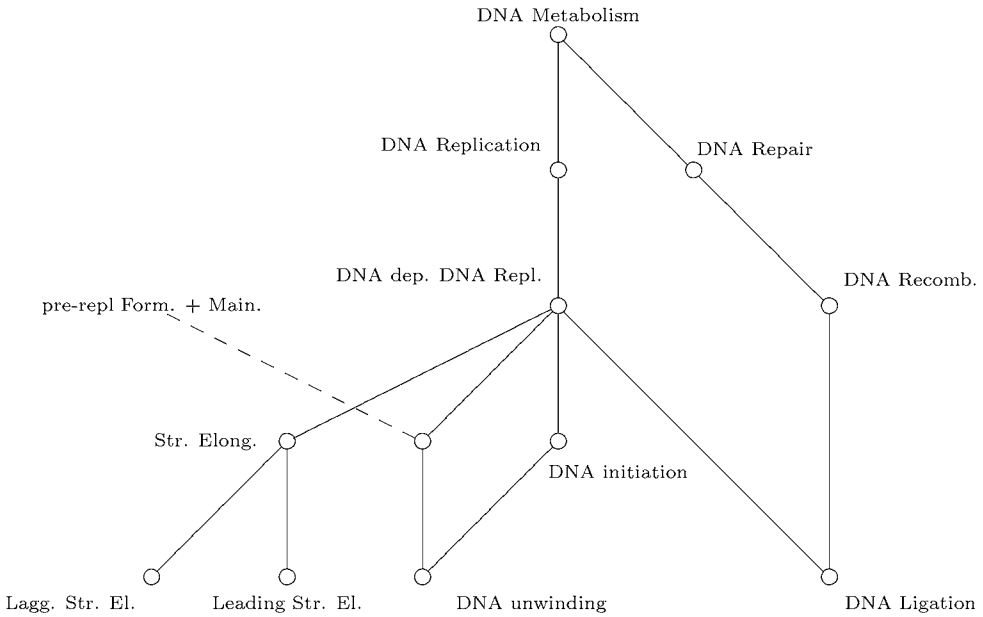

Fig. 8. Function categories ordered conceptually.

We want to point out some interesting differences between the annotated ordered set and its concept lattice. Conceptually, the function category *DNA Recombination* is less than *DNA Repair* while in the GO the two nodes are not comparable. This change occurs because *DNA Repair* "inherits" the proteins from *DNA Ligation* which yields a superset of proteins annotated to *DNA Repair* compared to *DNA Recombination*. Since the design of the function category ordering of the GO differs from the conceptual pre-ordering the question arises if some proteins exist but are not present in the GO which justify the non-comparability or if the ordering should be redesigned.

If we focus our attention on the protein *CDC9* we see that it is annotated to two quite horizontally distinct nodes in the GO, *Lagging strand elongation* and *DNA ligation*. In the concept lattice representation, the new object concept node for *CDC9* thus ties together these two GO nodes through the intermediate concept shown there, the *CDC9* object concept atom on the left. Now, we could ask the question if there is a meaningful label for this node and if it should eventually be introduced in the GO.

6. Soundness and Completeness of Annotated Ordered Sets

As described in the last section, there can be interesting differences between annotated ordered sets and their adjustments. In this section we distinguish two aspects

- *soundness*
- *completeness*

for measuring those differences.

6.1. *Soundness*

To measure conceptual soundness, firstly, one can investigate the degree of collapse, that is, to calculate how many elements of the initial taxonomy are identified. The more elements get identified the larger is the deviation from the initial taxonomy to its adjustment. We propose to use the following supermodular isotone normalized function on the lattice of equivalence relations on a set P, given by,

$$\sigma_0 : \mathrm{Eq}(P) \longrightarrow \mathbb{R},\ \theta \mapsto \frac{|\theta| - |P|}{|P \times P| - |P|}.$$

In our setting of an annotated ordered set $\mathbb{O} := (\mathcal{P}, X, F)$ we define the *degree of collapse* as

$$s_0(\mathbb{O}) := \sigma_0(ker(\mu_{\mathbb{O}})), \text{ where } \ker(\mu_{\mathbb{O}}) := \{(p,q) \in P \times P \mid \mu_{\mathbb{O}}(p) = \mu_{\mathbb{O}}(q)\}.$$

Applying this to our example of the GO we receive 0 as a degree of collapse, that is, nothing is collapsed.

Secondly, one can investigate the degree of missing links, that is, all pairs of nodes which are not comparable in the annotated ordered set but become related in the conceptual pre-order. In our setting of an annotated ordered set $\mathbb{O} := (\mathcal{P}, X, F)$ we define the normalized *degree of missing links* as

$$s_1(\mathbb{O}) := \frac{|\sqsubseteq_{\mathbb{O}}| - |\leq_{\mathcal{P}}|}{|P \times P| - |\leq_{\mathcal{P}}|}.$$

Applying this to our example of the GO we receive

$$s_1 = \frac{45 - 41}{144 - 41} = \frac{4}{103} \approx 0.039,$$

that is, since four links are added in the adjustment, the degree of missing links turns out to be approximately 0.039.

6.2. *Completeness*

To measure the completeness of an annotated ordered set we propose to take into account the concepts of the concept lattice representation which are not attribute concepts and different from the smallest concept. Those concepts might be considered as proposals for new nodes in the taxonomy, formally for new elements of P. We define the *degree of potentially missing taxonomy nodes* as

$$c_0(\mathbb{O}) := \frac{|\mathfrak{B}_{\mathbb{O}}| - |\mu(P)| - 1}{|P|}$$

Applying this to our example of the GO we receive

$$c_0 = \frac{19 - 12 - 1}{12} = 0.5$$

as a degree of potentially missing taxonomy nodes, that is, for every two nodes present in the annotated taxonomy a new node is present in the concept lattice representation.

7. Discussion

As a main application area of our technique we see the task of review or refinement of taxonomic ontologies as insinuated in the last sections.

It should be noted that the formal properties of the GO are just now beginning to be explored. Joslyn *et al.* [6] have done preliminary measurements of its poset properties, including height, width, and ranks. And while we've noted that the GO is not specifically lower-bounded, if a lower bound is asserted, then it can be questioned how many pairs of nodes do not have unique meets and joins, and thus how close it comes to our idealized annotated lattice. This is something we have addressed specifically elsewhere [7], including proposing a method to measure this degree of lattice-ness based on the FCA reconstruction of the (un-annotated) GO.

We see future work in this line of research in evolving measures and tools to make the technique operable for large taxonomies (see also [7, 8]). This would involve the design of expert systems, which support a semi-automated review or reengineering process of taxonomic ontologies.

References

1. The Gene Ontology Consortium: *Gene Ontology: Tool For the Unification of Biology*, *Nature Genetics*, v. **25**:1, pp. 25-29 (2000)

2. Bodenreider, Olivier; Mitchell, Joyce A; and McCray, Alexa T: *Evaluation of the UMLS As a Terminology and Knowledge Resource for Biomedical Informatics*, in: *AMIA 2002 Annual Symposium*, pp. 61-65 (2002)

3. B. A. Davey, H. A. Priestly: *Introduction to Lattices and Order*, Cambridge University Press (1990)

4. Davis, Anthony R: *Types and Constraints for Lexical Semantics and Linking*, Cambridge UP (2000)

5. B. Ganter, R. Wille: *Formal Concept Analysis*, Mathematical Foundations. Springer, Berlin Heidelberg New York (1999)

6. C.A. Joslyn, S.M. Mniszewski, A. Fulmer, and G.G. Heaton: *The Gene Ontology Categorizer*, Bioinformatics, v. 20:s1, pp.169-177 (2004)

7. C.A. Joslyn, D.D.G. Gessler, S.E. Schmidt, and K.M. Verspoor: *Distributed Representations of Bio-Ontologies for Semantic Web Services*,

in Joint BioLINK and 9th Bio-Ontologies Meeting (JBB06) (2006) http://bio-ontologies.man.ac.uk/download/Joslyn1EtAlDistributed.pdf

8. Joslyn, Cliff; Gessler, DDG; and Verspoor, KM: *Knowledge Integration in Open Worlds: Utilizing the Mathematics of Hierarchical Structure*, in: Proc. 2007 Int. Conf. of Semantic Computing, Irvine, CA (2007)

9. T.B. Kaiser, S.E. Schmidt, C.A. Joslyn: *Concept Lattice Representations of Annotated Taxonomies*, to appear in CLA 06 conference post-proceedings, LNAI Springer, Berlin Heidelberg New York (2007)

10. Knoblock, B. Todd, and J. Rehof: *Type Elaboration and Subtype Completion for Java Bytecode*, in: *Proc. 27th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (2000)