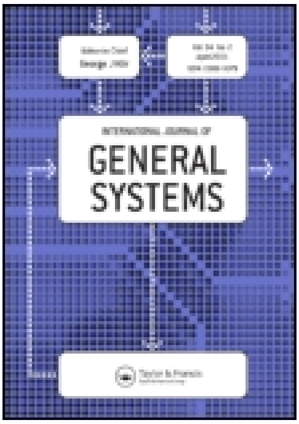

International Journal of General Systems
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ggen20

Formal concept analysis with background knowledge: a case study in paleobiological taxonomy of belemnites

Radim Belohlavek [a] , Martin Kostak [b] & Petr Osicka [a]

[a] Data Analysis and Modeling Laboratory, Department of Computer Science, Palacky University, Olomouc, Czech Republic

[b] Faculty of Science, Institute of Geology and Palaeontology, Charles University in Prague, Prague, Czech Republic

To cite this article: Radim Belohlavek , Martin Kostak & Petr Osicka (2013): Formal concept analysis with background knowledge: a case study in paleobiological taxonomy of belemnites, International Journal of General Systems, 42:4, 426-440

To link to this article: http://dx.doi.org/10.1080/03081079.2013.765079

International Journal of General Systems, 2013
Vol. 42, No. 4, 426–440, http://dx.doi.org/10.1080/03081079.2013.765079

Formal concept analysis with background knowledge: a case study in paleobiological taxonomy of belemnites

Radim Belohlavek[a]*, Martin Kostak[b] and Petr Osicka[a]

[a]*Data Analysis and Modeling Laboratory, Department of Computer Science, Palacky University, Olomouc, Czech Republic;* [b]*Faculty of Science, Institute of Geology and Palaeontology, Charles University in Prague, Prague, Czech Republic*

(*Received 10 December 2012; final version received 7 January 2013*)

We present a case study in identification of taxa in paleobiological data. Our approach utilizes formal concept analysis and is based on conceiving a taxon as a group of individuals sharing a collection of attributes. In addition to the incidence relation between individuals and their attributes, the method uses expert background knowledge regarding importance of attributes which helps to filter out correctly formed but paleobiologically irrelevant taxa. We present results of experiments carried out with belemnites—a group of extinct cephalopods which seems particularly suitable for such a purpose. We demonstrate that the methods are capable of revealing taxa and relationships among them that are relevant from a paleobiological point of view.

Keywords: formal concept analysis; background knowledge; taxonomy; paleobiology

1. Introduction and problem setting

Taxonomy, i.e. a classification scheme arranged in a hierarchical structure, is an important system-theoretic concept. Biological taxonomies and the methods of devising them are perhaps the best known and most widely studied. There exist several approaches to biological classification, with phylogenetics (cladistics) and phenetics (numerical taxonomy) being perhaps the two most important methods (Dunn & Everitt, 2004; Jardin & Sibson, 1971; Kitching, Forey, Humphries, & Williams, 1998; Sneath & Sokal, 1973). The aim of this paper is to explore the idea of utilizing formal concept analysis (FCA) in identification of taxa and devising taxonomies in paleobiology. The basic idea is to identify taxa with formal concepts which are particular groupings of objects characterized by sharing certain properties (attributes).

The overall goal of our research is to find out whether and how formal concept analysis may help identify relevant taxa, identify relationships between taxa, and reconstruct evolution from paleobiological data. In particular, we use the method developed by Belohlavek and Sklenar (2004) and Belohlavek and Vychodil (2009) which makes it possible to take into account expert's background knowledge in formation of concepts. The background knowledge concerns a relative importance of attributes and may be used for filtering out irrelevant formal concepts, thus retaining only those considered relevant by the expert. In this case study, we demonstrate that such kind of background knowledge is useful in paleobiological taxonomy. We present results obtained from data about a particular group of extinct cephalopods called belemnites, which are similar to modern squid.

---

*Corresponding author. Email: radim.belohlavek@acm.org

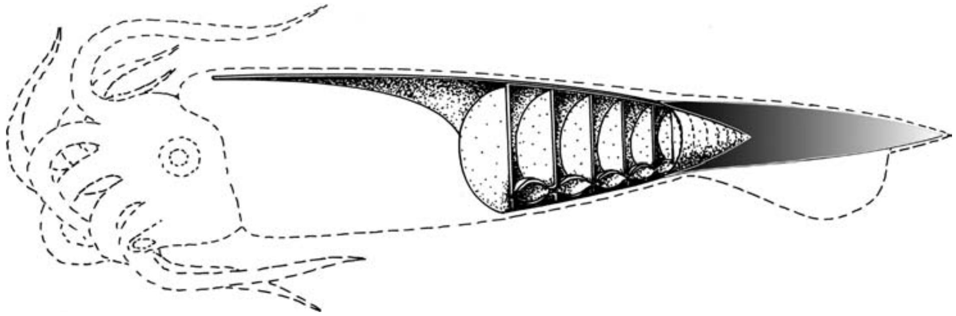

Figure 1. Schematic sketch of belemnite body with position of the internal shell. Apical grey part represents the rostrum.

Belemnites are extinct cephalopod group with no descendents (their habitus partly resembles some Recent squids). Their systematic classification, paleoecology, paleobiogeography, and stratigraphy are based on paleontological research. However, the simulation of belemnite evolution opens new options for using various mathematical models, for which the presented cephalopod group is particularly suitable. A schematic sketch of a belemnite body is shown in Figure 1. For detailed morphologic features of rostrum (i.e. solid part of the internal shell), we refer to (Christensen, 1997), (Kostak 2012), and Kostak et al. (2004).

The most important morphologic features in rostrum morphology are size, shape, and especially the part called the alveolar end and/or fracture, i.e. a space of connection between the rostrum and the phragmocone (chambered conical shell containing gas; note that this part forms a shell in the Recent *Nautilus*). From quite a large systematic group (order Belemnitida), we chose a part of family Belemnitellidae—a dominant Late Cretaceous (100–66 Ma[1]) belemnite group described below.

2. Formal concept analysis with attribute priorities

Formal concept analysis is a method used for data analysis with applications in various domains (Carpineto & Romano, 2004; Ganter & Wille, 1999). The input to FCA consists of a data table describing a set X of objects and a set Y of attributes. The table specifies which objects have which attributes. The main aim of FCA is to reveal from the data a hierarchically organized set of particular clusters, called formal concepts, and a small set of particular attribute dependencies, called attribute implications. FCA aims at formalizing and utilizing a traditional theory of concepts, in which a concept is understood as an entity consisting of its extent (collection of all objects to which the concept applies) and its intent (collection of all attributes to which are characteristic for the concept). For example, the extent of the concept *dog* consists of all dogs while its intent consists of attributes, such as *barks*, *has a tail*, *has four limbs* and so on. The information extracted from data in FCA is well comprehensible by users because FCA employs notions which humans use in ordinary life. This is an important feature of FCA which makes it appealing for data analysis purpose.

In the basic setting of FCA, attributes are assumed to be binary, i.e. a given object either has or does not have a given attribute. A data table with binary attributes is represented by a triplet $\langle X, Y, I\rangle$, called a *formal context*, which consists of the above-mentioned sets X and Y of objects and attributes, and a binary relation I between X and Y

(incidence relation, to-have relation). Thus, $I \subseteq X \times Y$, $\langle x, y\rangle \in I$ indicates that object x has attribute y, and $\langle x, y\rangle \notin I$ indicates that x does not have y. Objects $x \in X$ correspond to table rows, attributes $y \in Y$ correspond to table columns, and I is represented by ×s in the table entries—× indicates that the corresponding object has the corresponding attribute while a blank indicates the opposite.

A formal concept of $\langle X, Y, I\rangle$ is any pair $\langle A, B\rangle$ of sets $A \subseteq X$ (*extent*, set of objects to which the concept applies) and $B \subseteq Y$ (*intent*, set of attributes characterizing the concept) such that B is just the set of attributes shared by all objects from A, and A is the set of all objects sharing all attributes from B. In symbols, this can be written as $A^{\uparrow} = B$ and $B^{\downarrow} = A$, where

$$A^{\uparrow} = \{y \in Y \mid \text{for each}\, x \in A : \langle x, y\rangle \in I\},$$
$$B^{\downarrow} = \{x \in X \mid \text{for each}\, y \in B : \langle x, y\rangle \in I\}.$$

These arrow operators form a Galois connection between X and Y and play an important role in FCA. The set of all formal concepts of $\langle X, Y, I\rangle$ is denoted by $\mathcal{B}(X, Y, I)$. That is,

$$\mathcal{B}(X, Y, I) = \{\langle A, B\rangle \mid A^{\uparrow} = B, B^{\downarrow} = A\}.$$

Under a partial order $\leq$, defined for $\langle A_1, B_1\rangle, \langle A_2, B_2\rangle \in \mathcal{B}(X, Y, I)$ by

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle \;\text{ iff }\; A_1 \subseteq A_2\, (\text{iff}\, B_2 \subseteq B_1),$$

$\mathcal{B}(X, Y, I)$ happens to be a complete lattice, the so-called *concept lattice* of $\langle X, Y, I\rangle$ (Ganter & Wille, 1999). The partial order $\leq$ represents the subconcept–superconcept (generalization–specialization) relationship between formal concepts. That is, $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$ means that formal concept $\langle A_2, B_2\rangle$ (representing e.g. *mammal*) is more general than $\langle A_1, B_1\rangle$ (e.g. *dog*). Note also that there exist efficient algorithms for computing $\mathcal{B}(X, Y, I)$.

The concept lattice is visualized by a labelled line diagram (or Hasse diagram) (Ganter & Wille, 1999). The diagram consists of nodes, some of which are connected by lines. The formal concepts are represented by nodes, the subconcept–superconcept relationship is represented by the lines, and a particular way of labelling the nodes by object and attribute names is used (Ganter & Wille, 1999). In particular, every node (formal concept) is connected to all of its direct predecessors (formal concepts which are more specific) and direct successors (formal concepts which are more general). The diagram is easily understood by users and provides a hierarchical view on the data.

In various situations, the expert finds certain formal concepts in the concept lattice interesting and relevant while others not. One reason for this is that in addition to the input data, i.e. the relation between the objects and the attributes, the expert may have further knowledge regarding the objects and attributes. According to this knowledge, the expert may regard some formal concepts interesting (relevant) and the others not. In (Belohlavek & Sklenar, 2004; Belohlavek & Vychodil, 2009), we developed an approach to handling one particular type of a background knowledge which we utilize in this paper. The idea is that the attributes may not be equally important in concept formation.

To give a simple example, consider the following attributes of books: *hardbound*, *paperback*, *engineering*, *science*, and *philosophy*. In a reasonable taxonomy of books, we naturally consider the attributes *engineering*, *science*, and *philosophy* more important than *hardbound* or *paperback*. As a consequence, we would consider a formal concept characterized by the attribute *hardbound* (all books which are hardbound) not natural (irrelevant, uninteresting). On the other hand, the formal concept characterized by *engineering* (books on engineering), the formal concept characterized by *science*, and perhaps also the formal concept characterized by *engineering* and *paperback* (paperback books on engineering) would be considered natural. Such a background knowledge may be represented by the so-called *attribute-dependency formulas* (AD formulas) (Belohlavek & Sklenar, 2004; Belohlavek & Vychodil, 2009). In our example, the AD formula representing our background knowledge is

$$\textit{hbound} \sqcup \textit{pback} \sqsubseteq \textit{engineering} \sqcup \textit{science} \sqcup \textit{philosophy},$$

but in general, the background knowledge is represented by a set of AD formulas.

An *AD formula* over a set Y of attributes is an expression

$$D_1 \sqsubseteq D_2,$$

where $D_1, D_2 \subseteq Y$. A *formal concept* $\langle A, B\rangle$ *satisfies the AD formula* $D_1 \sqsubseteq D_2$, in symbols

$$\langle A, B\rangle \models D_1 \sqsubseteq D_2,$$

if $D_1 \cap B \neq \emptyset$ implies $D_2 \cap B \neq \emptyset$. That is, $\langle A, B\rangle \models D_1 \sqsubseteq D_2$ means that whenever the concept contains an attribute from D_1, it has to contain already an attribute from D_2. The rationale is that in forming reasonable concepts, one should proceed by adding first the more important attributes, i.e. those from D_2 such as *engineering*, and only then one may add the less important attributes, i.e. those from D_1 such as *paperback*. As a result, the formal concept characterized by *hardbound* is not compatible with the above background knowledge because in this case, $B = \{\textit{hardbound}\}$, and thus for the above AD formula $D_1 \sqsubseteq D_2$ with $D_1 = \{\textit{hbound}, \textit{pback}\}$ and $D_2 = \{\textit{engineering}, \textit{science}, \textit{philosophy}\}$, we have $D_1 \cap B \neq \emptyset$ but $D_2 \cap B = \emptyset$. Intuitively, $\langle A, B\rangle$ contains a less important attribute but does not contain any of the prescribed more important attributes.

A *formal concept satisfies a set T of AD formulas*, in symbols $\langle A, B\rangle \models T$, if $\langle A, B\rangle$ satisfies each AD formula $D_1 \sqsubseteq D_2$ from T. The set of all formal concepts of $\langle X, Y, I\rangle$ compatible with T is denoted by $\mathcal{B}_T(X, Y, I)$, i.e.

$$\mathcal{B}_T(X, Y, I) = \{\langle A, B\rangle \in \mathcal{B}(X, Y, I) \mid \langle A, B\rangle \,\text{satisfies every}\, D_1 \sqsubseteq D_2 \,\text{in}\, T\}.$$

$\mathcal{B}_T(X, Y, I)$ thus comprises a part of the whole concept lattice $\mathcal{B}(X, Y, I)$, and is considered the set of all natural (relevant, interesting) formal concepts in the data $\langle X, Y, I\rangle$ given the background knowledge represented by T.

3. Belemnite evolution in time and space

Coleoid cephalopods played an important role in the Cretaceous ecosystem in both the Northern and Southern hemispheres. Inside this diverse group, especially belemnites were a common part of nectonic assemblages (most of free swimmers) in marine shallower water ecosystems. However, the internal shell characteristic is strongly different in both non-related cephalopod groups. In belemnites, the internal shell is composed of three major components—the most commonly preserved calcitic rostrum, aragonitic phragmoconus, and dorsal proostracum (dorsal sheet, probably built by organic matter). As no soft body parts are preserved for taxonomic distinction, belemnite taxonomy is mainly based on the following morphological characteristics of the belemnite rostrum:

- structure of the alveolar end,
- shape and size of the rostrum,
- internal structures at alveolar end,

- external characteristics of the rostrum,
- structure of apex.

Rostrum, a hydrodynamic and also hydrostatic organ, shows gradual changes in evolutionary history.

The morphologic changes in time and space are well documented especially in early belemnitellids (Kostak, 2004). Relatively few morphologic features in belemnite rostrum (Christensen, 1997; Kostak, 2004), clear taxonomy, and new systematic and stratigraphic revision (see below) make it possible to use formal methods for the identification of taxa.

Belemnites show an interesting model of migration patterns and provincialism in Jurassic and Cretaceous coleoids. We have used the Upper Cretaceous belemnites (Belemnitellidae). Their evolution centre lay in the SE parts of the East European (Russian) Platform and mass migrations affected shallow seas in Central and NW Europe, Siberian areas, and North America during the Cenomanian (100.5–93.9 Ma) through the Coniacian (89.8–86.3 Ma). They show marked provincialism (high number of endemic taxa) in the North European, the North American, and the East European Provinces (Christensen, 1997; Kostak et al., 2004). The last belemnites became extinct at the K/T boundary (Cretaceous/Tertiary boundary, ca. 65.6 Ma). The fall in belemnitellid diversity is related to the reduction of shallow sea areas, eustasy (i.e. sea-level changes), and the radiation (evolutionary expansion) of new vertebrate groups, i.e. Neoselachii (‘modern’ sharks) and Teleostei (bony fishes). Our analysis, selected results of which are provided below, focuses on their evolution, particularly on the frequency of morphologic changes, endemism (species living in restricted geographical areas), and species and generic relationships.

We chose the Upper Cretaceous belemnites from the Cenomanian–Coniacian interval for the following reasons:

- Systematic revision has been done recently.
- Clear stratigraphic position.
- The evolutionary lineage from one ancestor is well documented.
- Recognized migration patterns and spreading succession in time and space.
- Quite simple morphology of preserved parts, with clear changes in time.

We have analysed 26 species belonging to three genera in the Cenomanian–Coniacian interval (i.e. 97–87 Ma). This phase of belemnite evolution shows the highest frequency in morphologic changes, and we suppose this to be suitable for our purpose.

4. Analysis and results

Dataset

The data (i.e. a formal context $\langle X, Y, I\rangle$) used in our experiment consists of 26 belemnite species (objects of X) that are specified in Table 1 and 36 rostrum characteristics (attributes of Y) that are specified in Table 2. The data is displayed in Table 3.

Attribute priorities and AD formulas

The corresponding concept lattice $\mathcal{B}(X, Y, I)$ consists of 616 formal concepts and is depicted in Figure 2. From the paleontological point of view, it contains both formal concepts that seem natural and interesting and those that are not natural. To eliminate the latter, we used AD formulas (see Section 2), because a paleontologist considers some attributes more important than others in his reasoning about possible taxa. The use of AD

Table 1. Objects: selected belemnite species.

Belemnite species	*Label*
Actinocamax verus antefragilis	1
Praeactinocamax primus	2
Praeactinocamax plenus	3
P. plenus cf. *strehlensis*	4
Praeactinocamax triangulus	5
P. aff. *triangulus*	6
Praeactinocamax aff. *sozhensis*	7
Praeactinocamax contractus	8
Praeactinocamax planus	9
Praeactinocamax coronatus	10
Praeactinocamax matesovae	11
Praeactinocamax medwedicicus	12
P. sp. 1	13
P. sp. 2	14
Praeactinocamax strehlensis	15
Praeactinocamax bohemicus	16
P. aff. *bohemicus*	17
Praeactinocamax cobbani	18
Praeactinocamax manitobensis	19
P. cf. *manitobensis*	20
Praeactinocamax sternbergi	21
Praeactinocamax walkeri	22
G. intermedius	23
G. surensis	24
Goniocamax christenseni	25
Goniocamax lundgreni	26

formulas to make his background knowledge explicit and to take it into account in the subsequent analysis is therefore a natural solution.

Based on expert's opinion, we partitioned the set Y of attributes into seven groups shown in Table 2 and partially ordered them as depicted in Figure 3. Then, for every two groups $D_1, D_2 \subseteq Y$ of the seven groups, we add to the set T of our AD formulas the AD formula $D_1 \sqsubseteq D_2$. For example, since the group Others B, consisting of β, γ, δ, ϵ, and η, is a direct predecessor of Others A, consisting of y, z, α, we add

$$\beta \sqcup \gamma \sqcup \delta \sqcup \epsilon \sqcup \eta \sqsubseteq y \sqcup z \sqcup \alpha$$

to T.

The concept lattice

The constrained concept lattice $\mathcal{B}_T(X, Y, I)$, which consists of only those formal concepts in $\mathcal{B}(X, Y, I)$ that comply with the background knowledge represented by T, contains 121 formal concepts and is depicted in Figure 4. The concepts are labeled by numbers which we use below when interpreting the formal concepts and taxa and their relationships.

Interpretation

Table 4 describes selected formal concepts of the constrained concept lattice $\mathcal{B}_T(X, Y, I)$, namely the most interesting concepts from the paleontological viewpoint. We selected

Table 2. Attributes and their groups.

Group/name	*Label*
Size	
Large guards (more 65 mm)	a
Medium guards (65–80 mm)	b
Small guards (less 65 mm)	c
Cigar shape in dorsoventral (DV) view	d
Lanceolate in DV view	e
Slightly lanceolate in DV view	f
Subcylindrical in DV view	g
Conical in DV view	h
Cigar shape in lateral (L) view	i
Lanceolate in L view	j
Slightly lanceolate in L view	k
Subcylindrical in L view	l
Conical in L view	m
Flattening	
Laterally flattened	n
Dorsally flattened	o
Ventrally flattened	p
Alveolar fracture (AF), pseudoalveolus	
High conical AF	q
Low conical AF	r
Shallow pseudoalveolus (less 3 mm)	s
Deep pseudoalveolus (more 3 mm)	t
Cross section	
Oval cross section of AF	u
Oval to triangular cross section of AF	v
Triangular cross section of AF	w
Circular cross section of AF	x
Others A	
Bottom of ventral fissure	y
Dorsolateral furrows	z
Ventral furrow	α
Others B	
Dorsolateral depressions	β
Granulation of the whole guard	γ
Striation of the whole guard	δ
Mucro	ϵ
Vascular imprints frequent	η
Others C	
Conellae	θ
Granulation of a part of the guard	ι
Striation of a part of the guard	κ
Vascular imprints rare	λ

these 28 concepts from among the 121 concepts of $\mathcal{B}_T(X, Y, I)$ manually based on our paleontological expertise. The other concepts in $\mathcal{B}_T(X, Y, I)$ represent meaningful taxa as well but the selected concepts are particularly interesting. For every concept, we list its number (the numbers coincide with the labels in Figure 4), its extent (the collection of objects covered by the concept), and its intent (the collection of attributes characterizing the concept). Formal concepts may naturally be interpreted as taxa. Namely, for a given formal concept, its extent represents the collection of species covered by the

Table 3. Formal context of belemnites and their attributes.

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*	α	β	γ	δ	ϵ	η	θ	ι	κ	λ
1			×	×					×					×			×							×			×									
2	×	×	×	×	×	×	×			×	×	×				×		×			×						×	×			×	×				×
3	×	×	×		×	×	×				×	×				×		×			×	×					×	×			×	×		×	×	×
4		×			×					×						×		×			×	×						×							×	×
5	×	×	×		×						×	×				×				×			×				×	×						×	×	×
6	×					×						×				×				×		×					×	×				×				×
7	×	×	×		×						×	×				×	×				×						×	×			×					
8		×	×		×						×	×				×		×			×	×					×	×				×		×	×	×
9		×			×						×	×				×		×			×						×	×			×			×	×	×
10		×					×						×			×				×	×						×									×
11		×			×							×				×			×		×						×	×			×			×	×	
12		×		×			×					×				×				×		×					×	×								×
13		×	×			×	×					×				×				×		×						×				×				
14		×			×							×				×				×	×	×						×				×				×
15		×	×		×					×	×					×		×			×	×														×
16		×				×	×					×	×			×		×				×	×					×	×	×		×				×
17		×				×	×					×				×		×			×	×						×								×
18		×			×	×					×	×				×		×			×	×					×	×		×		×			×	×
19	×	×				×					×	×				×		×	×			×	×				×	×	×	×						×
20	×					×										×		×	×		×							×	×	×		×				×
21	×	×				×					×					×		×			×	×					×	×	×	×		×		×	×	×
22		×				×	×				×	×				×		×			×	×					×	×				×			×	×
23		×	×		×	×					×	×				×				×		×					×	×			×			×	×	
24			×	×	×	×						×				×			×	×		×					×	×				×		×	×	
25		×				×						×			×	×				×	×				×	×	×	×			×			×	×	×
26		×				×						×			×	×				×	×				×	×	×	×			×			×	×	×

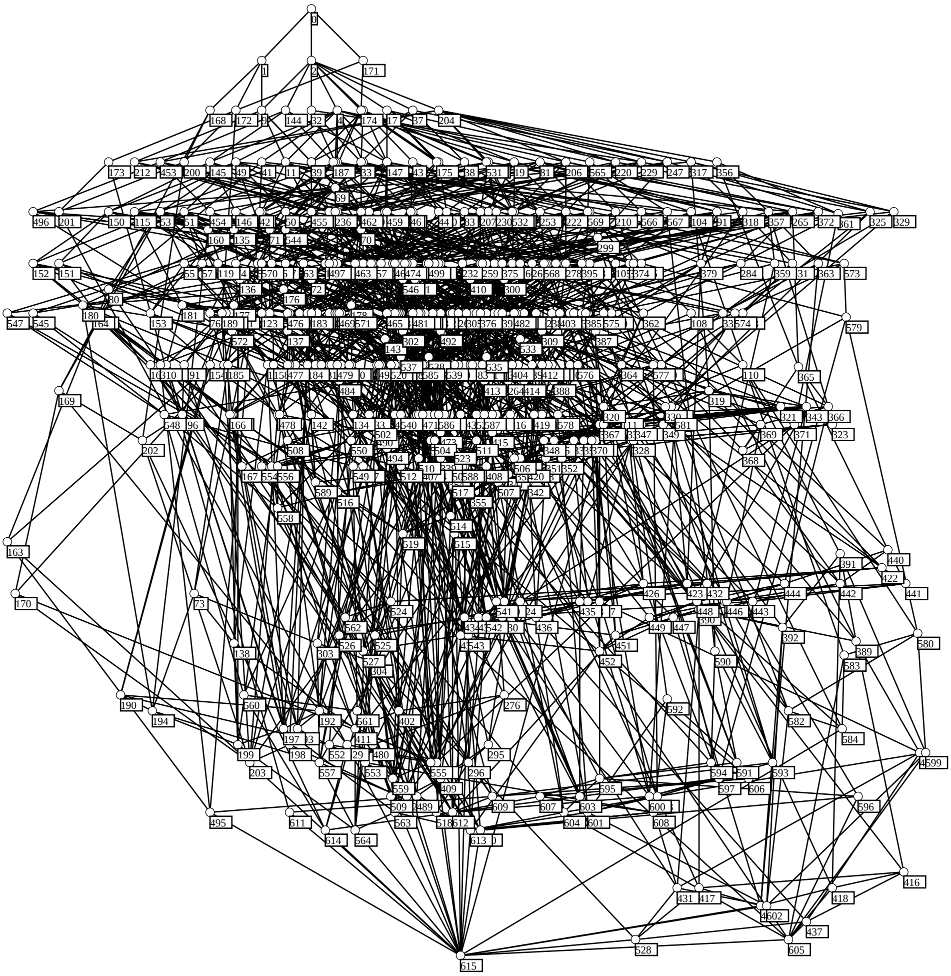

Figure 2. Hasse diagram of unconstrained concept lattice containing 616 concepts.

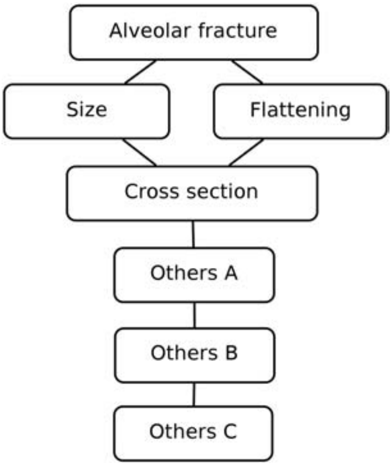

Figure 3. Hasse diagram of the ordered set of attribute groups.

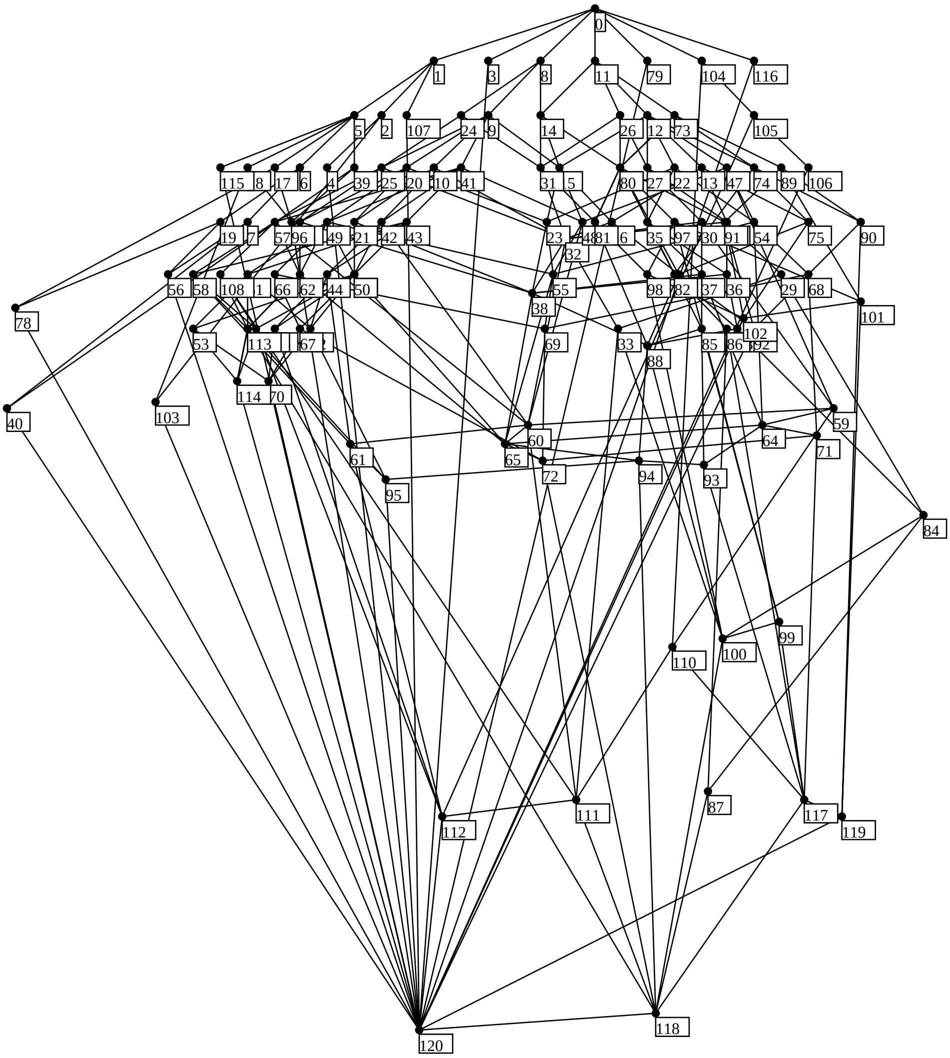

Figure 4. Hasse diagram of ordered set of 121 concepts resulting from the size reduction via AD formulas.

corresponding taxon and its intent represents the attributes characterizing the taxon. For example, the taxon represented by concept No. 2 covers species 23 and 24, i.e. *Goniocamax intermedius* and *Goniocamax surensis*, and may be characterized by the presence of attributes c, e, f, l, p, t, v, z, β, λ, and ϵ. Furthermore, the concept ordering represents the subtaxon–supertaxon relationship. For instance, the taxon represented by concept No. 4 is a subtaxon of that represented by concept No. 2 because the extent {24} of concept No. 4 is included in the extent {23, 24} of concept No. 2; dually, the intent of concept No. 2 is included in the intent of concept No. 4. Clearly, a meaningful analysis of the results and a useful interpretation of formal concepts as taxa require paleontological expertise. For the present data, such analysis and interpretation of the most important formal concepts of the constrained concept lattice are provided in Appendix.

Table 4. Interesting formal concepts.

Concept	*Extent*	*Intent*
3	1	c, d, i, n, q, x, z
2	23, 24	c, e, f, l, p, t, v, z, β, λ, ϵ
4	24	c, d, e, f, l, p, s, t, v, z, β, κ, λ, ϵ
103	23	b, c, e, f, k, l, p, t, v, z, β, δ, λ, ϵ
8	3, 4, 8, 15, 16, 17, 18, 19, 21, 22	b, p, r, v, α
79	11	b, e, l, p, s, u, z, β, δ, λ, ϵ
11	2, 3, 4, 5, 9, 15, 17, 18, 21, 22	b, p, r, v, β, κ, α
104	2, 3, 20, 21	a, f, p, r, u, β, κ, α
116	7	a, b, c, e, k, l, p, q, u, z, β, δ
17	12, 14	b, l, p, t, v, β, α
115	5	a, b, c, e, k, l, p, t, w, z, β, λ, ϵ, α
18	12, 23	b, l, p, t, v, z, β
78	14	b, e, l, p, t, u, v, β, κ, α
40	10	b, g, m, p, t, u, z, α
114	19	a, b, f, k, l, p, r, s, v, w, z, β, γ, ι, α
70	16	b, f, g, l, m, p, r, v, w, β, γ, ι, κ, α
95	18	b, e, f, k, l, p, r, u, v, z, β, ι, κ, ϵ, α
112	21	a, b, f, k, p, r, u, v, z, β, γ, ι, κ, λ, ϵ, α
117	2, 3	a, b, c, e, f, g, k, l, p, r, u, z, β, δ, κ, α
119	2	a, b, c, d, e, f, g, j, k, l, p, r, u, z, β, δ, κ, α
118	3	a, b, c, e, f, g, k, l, p, r, u, v, z, β, δ, κ, λ, ϵ, α
87	3, 9	b, e, k, l, p, r, u, z, β, δ, λ, ϵ, α
84	3, 8, 9	b, e, k, l, p, r, u, z, β, λ, ϵ, α
102	15	b, c, e, j, k, p, r, u, v, α
101	2, 15	b, c, e, j, k, p, r, u, α
90	2, 4	b, e, j, p, r, u, β, α
5	5, 10, 12, 13, 14, 23, 25, 26	b, p, t
56	25, 26	b, f, l, o, p, t, u, θ, y, z, β, δ, λ, ϵ, α

5. Conclusions

The presented case study demonstrates that FCA with background knowledge makes it possible to reveal natural taxa and interesting relationships in paleobiological data. Importantly, our study demonstrates that formal concepts may naturally represent paleobiological taxa. Namely, the intent of a particular formal concept represents a collection of attributes characterizing the corresponding taxon, whereas the extent represents a collection of individuals (species in our case) covered by the taxon. From a paleobiological viewpoint, the major conclusions of the present analysis, described in detail in Appendix, may be summarized as follows:

1. The analysis suggests that the belemnite taxonomic group studied (family Belemnitellidae) is polyphyletic (i.e. there is more than one common ancestor). This suggestion contradicts the present opinion according to which this family is monophyletic (i.e. origins from one common ancestor).
2. The analysis indicates the existence of three morphologic trends going to independent, but morphologically similar species in one belemnite genus (the so-called parallel evolution). Up to now, only two parallel morphological lineages were considered inside this genus.
3. The analysis revealed the genus origin (speciation of genus *Goniocamax*) and derivation of endemic taxa (both more and less connected with allopatric speciation); see Appendix for details.

Although conclusion 3 has formerly been suggested and finally confirmed also by phylogenetic analysis before our study, conclusions 1 and 2 represent quite a new point of view on early belemnitellid evolution. In this respect, these conclusions stimulate further paleontological investigations.

In addition, we plan to focus on the following topics: further analyses of the present data and applications of the presented method to further fossil groups; employment of triadic concept analysis (Lehmann & Wille, 1995), which enables to analyse data describing which individuals have which attributes under which conditions; extension of the analyses to data with graded (fuzzy) attributes to enable to include attributes that come naturally in degrees in addition to the yes/no attributes considered so far (see e.g. (Belohlavek & Osicka, 2012) and (Belohlavek & Vychodil, 2012) for the methods). A long-term goal is to study from the formal concept analysis viewpoint how paleontologists form taxa and taxonomies and seek ways to formalize of this process to a reasonable extent, see also (Priss, 2003).

Acknowledgements

The research of R. Belohlavek and P. Osicka was supported by Grant No. P103/10/1056 of the Czech Science Foundation. The research of M. Kostak was supported by Grant No. MSM 0021620855 (Vyzkumny zamer). A preliminary version of this paper was presented at the 20th European Meeting on Cybernetics and Systems Research in 2010.

Note

[1] *Mega-annum*, i.e. one million years

Notes on contributors

Radim Belohlavek received Ph.D. Degree in Computer Science from the Technical University of Ostrava (Czech Republic) in 1998, Ph.D. Degree in Mathematics from Palacky University in 2001, and D.Sc. Degree in Informatics and Cybernetics from the Academy of Sciences of the Czech Republic in 2008. He is a Professor of Computer Science at Palacky University, Olomouc. Dr Belohlavek's academic interests are in the areas of uncertainty and information, fuzzy logic and fuzzy sets, applied algebra and logic, and data analysis and formal concept analysis. He published two books, *Fuzzy Relational Systems: Foundations and Principles* (Kluwer, 2002) and *Fuzzy Equational Logic* (Springer, 2005, with Vilem Vychodil), and over 150 papers in conference proceedings and journals. Dr Belohlavek is a Senior Member of IEEE (Institute of Electrical and Electronics Engineers), and a Member of ACM (Association for Computing Machinery), AMS (American Mathematical Society), and is a member of editorial boards of several international journals.

Martin Kostak received Ph.D. degree in Geological Science from the Charles University in Prague. Since 2009, he is an Associate Professor of Palaeontology and Geology and Director of the Institute of Geology and Palaeontology at the Faculty of Science, Charles University in Prague. Dr Kostak's academic interests are concentrated on the Cretaceous belemnites (and other extinct coleoid cephalopods), integrated stratigraphy of the Jurassic/Cretaceous boundary interval, and paleoecology, biostratigraphy, and eventostratigraphy of the Cretaceous basins. He published six books and monographs, including *Ancient World of Fossils* (in Czech, German, Italian, and Dutch) and *Cenomanian Through the Lowermost Coniacian*

Belemnitellidae Pavlow (Belemnitida, Coleoidea) of the East European Province, and over 30 papers in journals and conference proceedings.

Petr Osicka received Ph.D. degree in Computer Science from Palacky University, Olomouc (Czech Republic) in 2012. He is an assistant at Palacky University, Olomouc. Dr Osicka's professional interests include fuzzy logic and fuzzy sets, relational data analysis, and algorithms. He authored several papers in journals and conference proceedings. Petr Osicka is a Member of ACM (Association for Computing Machinery).

References

Belohlavek, R., & Osicka, P. (2012). Triadic concept lattices of data with graded attributes. *International Journal of General Systems*, *41*(2), 93–108.

Belohlavek, R., & Sklenar, V. (2004). Formal concept analysis constrained by attribute-dependency formulas. *Lecture notes in computer science* (Proc. ICFCA 2005), Vol. 3403 (pp. 176–191). Berlin: Springer.

Belohlavek, R., & Vychodil, V. (2009). Formal concept analysis with background knowledge: Attribute priorities. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, *39*(4), 399–409.

Belohlavek, R., & Vychodil, V. (2012). Formal concept analysis and linguistic hedges. *International Journal of General Systems*, *41*(5), 503–532.

Carpineto, C., & Romano, G. (2004). *Concept data analysis: Theory and applications*. Chichester: John Wiley & Sons.

Christensen, W. K. (1997). The Late Cretaceous belemnite family Belemnitellidae: Taxonomy and evolutionary history. *Bulletin of the Geological Society of Denmark*, *44*, 59–88.

Dunn, G., & Everitt, B. S. (2004). *An introduction to mathematical taxonomy*. Mineola: Dover.

Ganter, B., & Wille, R. (1999). *Formal concept analysis: Mathematical foundations*. Berlin: Springer.

Jardin, N., & Sibson, R. (1971). *Mathematical taxonomy*. Chichester: John Wiley & Sons.

Kitching, I., Forey, P., Humphries, C., & Williams, D. (1998). *Cladistics: Theory and practice of parsimony analysis*. Oxford: Oxford University Press.

Kostak, M. (2004). Cenomanian through the lowermost Coniacian Belemnitellidae Pavlow (Belemnitida, Coleoidea) of the East European Province. *Geolines*, *18*, 59–109.

Kostak, M. (2012). On the Turonian origin of the Goniocamax-Belemnitella stock (Cephalopoda, Coleoidea). *Géobios*, *45*(1), 79–85.

Kostak, M., Čech, S., Ekrt, B., Mazuch, M., Wiese, F., Voigt, S., & Wood, C. J. (2004). Belemnites of the Bohemian Cretaceous Basin in a global context. *Acta Geologica Polonica*, *54*(4), 511–533.

Lehmann, F., & Wille, R. (1995). A triadic approach to formal concept analysis. *Lecture Notes in Computer Science*, *954*, 32–43.

Naidin, D. P. (1964). *Upper Cretaceous belemnites from the Russian Platform and adjacent areas. Actinocamax, Gonioteuthis, Belemnellocamax*. Moskva: Izdatelstvo Moskovskogo Gosudarstvennogo Universiteta (in Russian) .

Priss, U. (2003). Formalizing botanical taxonomies. *Lecture notes in artifical inteligence* (Proc. ICCS 2003), Vol. 2746 (pp. 309–322). Berlin: Springer.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. New York: W.H. Freeman.

Appendix: Paleontological interpretation of the results

In this part, we provide a detailed summary of paleontological interpretation of the results presented in Section 4. In particular, we provide an interpretation of the formal concepts in Table 4 and their relationships. These are the concepts in the constrained concept lattice $\mathcal{B}_T(X, Y, I)$ that are particularly interesting from a paleobiological point of view. For more paleobiological details and for the figures of the species mentioned, we refer to the literature cited in this section.

The Upper Cretaceous belemnites (family Belemnitellidae, i.e. belemnitellids in this paper) are represented by 7 genera *Actinocamax*, *Praeactinocamax*, *Goniocamax*, *Gonioteuthis*, *Belemnella*, *Belemnellocamax*, *Belemnitella*, and problematic genera *Belemnocamax* and *Fusiteuthis*. Although the phylogenic lineage going from *Goniocamax* to younger *Gonioteuthis*, *Belemnella*, and *Belemnitella* is quite clear and is connected with progressive calcification of the alveolar part, the relationships between earlier genera like the origin of the Late Cretaceous belemnites are still unclear, see (Christensen, 1997) and (Kostak, 2012) where one can find additional paleontological details and taxa figures.

The concept analysis strictly derived and separated genus *Actinocamax* from other belemnitellids (concept No. 3). We have used one species (subspecies) of *Actinocamax verus antefragilis* (the stratigraphic range of this species is very long and falls into the period of rapid evolution of another belemnite taxa). This and later species of *Actinocamax* are very similar to each other—they can be distinguished only at a subspecies level (note that the status of subspecies in paleontological taxonomy seems to be highly controversial). The concept analysis of this genus showed no relationship to other belemnitellid taxa. This should be interpreted like an extreme derivation from belemnitellids (size, shape of the rostrum, and alveolar fracture) and/or by presence of another belemnoid ancestors. In this respect, the *Actinocamax* evolutionary lineage should be excluded from belemnitellids as their ancestor probably does not belong to taxa related to earlier genus *Praeactinocamax*. Species of *Praeactinocamax* were commonly distributed in Euroasia and North America from the Cenomanian through Coniacian (the last one is known from the Santonian of Greenland). This genus occurred a few million years ago before the earliest *Actinocamax* and the first species of this genus—*P. primus*—is generally considered (Christensen, 1997; Naidin, 1964) to be also the earliest species of belemnitellids. If this opinion is true, *P. primus* and its straight descendent *P. plenus* (primus/plenus group) are initial taxa for all younger belemnitellids (i.e. monophyletic group), and this must be clearly detected in concept analysis. If we look at concept No. 11, we can see attributes common to another taxa of this genus. Original morphologic features in *primus/plenus* group (also concepts Nos. 117 and 118) are present especially in the European species.

Of a high importance is concept No. 8. showing attributes common to both conservative morphological features in the East European species (concept No. 14, related to concept No. 11) but also to relatively distinct Turonian North American (*P. manitobensis*, concept No. 114, *P. cobbani* and *P. sternbergi*—concepts Nos. 61, 95, and 112) and the Late Turonian species from Central and NW Europe (*P. bohemicus*, concept No. 70). In this respect, we observe at least two different evolutionary lineages—the expression of allopatric speciation (geographic division of population that prevents genetic interchange between them, i.e. a reproduction barrier and origin of new species) in different geographic areas.

High endemic species with unknown or poorly documented phylogenetic relationship are *P. sozhensis* (concept No. 116) and *P. matesovae* (concept No. 79).

Another belemnite lineage is also evolved from the earliest species of *Praeactinocamax* by forming a deeper pseudoalveolus (a space between the phragmoconus and the rostrum). This evolutionary innovation is partly observable already in some specimens of *P. plenus* population. So, the origin of this lineage took place probably in the Late Cenomanian (ca. 94 Ma). Typical morphologic features inside this morphologic lineage are summarized in concept No. 1. (see attributes in concepts Nos. 1 and 5). This evolutionary lineage shows markedly higher number of rare and endemic species (*P. medwedicicus*, *P. coronatus*, *P. aff. triangulus*, *P.* sp. 1 and 2, etc.). The iteration (short-term divergence of phylogenetic lineages) could be understood also as an expression of allopatric speciation in the East European Cretaceous sea. This lineage is also important by the origin of the earliest species of genus *Goniocamax* (*G. intermedius* and *G. surensis*—concepts Nos. 2, 4, and 103) and advanced *G. christenseni* and *G. lundgreni* (concept No. 56). The last-mentioned species couple is also considered to be an ancestor of younger *Belemnitella* stock (including also genus *Gonioteuthis*). The transition between *Goniocamax* and *Belemnitella* is gradual and relatively clear.

In summary, the following are the major conclusions drawn from our analysis. Separated evolutionary lineage in belemnite genus *Actinocamax* (which is probably derived from an ancestor different from the other species of family Belemnitellidae) indicates a polyphyletic origin of belemnitellids. The morphologic changes in the earliest genus (*Praeactinocamax*) within the family Belemnitellidae should be explained by allopatric speciation, including also iteration with endemic taxa. This is a typical expression of the parallel and gradual evolution. Intrageneric variability has clearly been demonstrated. Three evolutionary lineages within the genus *Praeactinocamax* lineages were detected using formal concept analysis. The origin of genus *Goniocamax* and its derivation from *Praeactinocamax* has been confirmed by our method.