Automated Enzyme classification by Formal Concept Analysis

François Coste, Gaëlle Garet, Agnès Groisillier, Jacques Nicolas, Thierry Tonon

Automated Enzyme classification by Formal Concept Analysis

François Coste[1], Gaëlle Garet[1], Agnès Groisillier[2], Jacques Nicolas[1], and Thierry Tonon[3]

[1] Irisa / Inria Rennes, Campus de Beaulieu, 35042 Rennes cedex, France
jacques.nicolas@inria.fr,
WWW home page: http://www.irisa.fr/dyliss

[2] Sorbonne Universités, UPMC Univ Paris 06, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, F-29688, Roscoff cedex, France

[3] CNRS, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, F-29688, Roscoff cedex, France
WWW home page: http://www.sb-roscoff.fr/umr7139.html

Abstract. Enzymes are molecules with a catalytic activity that make them essential for any biochemical reaction. High throughput genomic technics give access to the protein sequence of new enzymes found in living organisms. Guessing the enzyme functional activity from its sequence is a crucial task that can be approached by comparing the new sequences with those of already known enzymes labeled by a family class. This task is difficult because the activity is based on a combination of small sequence patterns and sequences greatly evolved over time. This paper presents a classifier based on the identification of common subsequence blocks between known and new enzymes and the search of formal concepts built on the cross product of blocks and sequences for each class. Since new enzyme families may emerge, it is important to propose a first classification of enzymes that cannot be assigned to a known family. FCA offer a nice framework to set the task as an optimization problem on the set of concepts. The classifier has been tested with success on a particular set of enzymes present in a large variety of species, the haloacid dehalogenase superfamily.

Keywords: bioinformatics, protein classification, FCA application

1 Introduction: enzyme classification

Enzymes are molecules with a catalytic activity that make them essential for any biochemical reaction. Enzymes are mainly named and classified according to the reaction they catalyze. Thus a name does not refer to a single enzyme protein but to a group of proteins from different living organisms (e.g. bacterial, plant or animal species) with the same catalytic properties. Enzymes are classified according to the report of a Nomenclature Committee appointed by the

International Union of Biochemistry and Molecular Biology[4]. Effectively, this committee assigns each enzyme a recommended name and a EC (Enzyme Commission) number representing four hierarchical levels to classify the enzyme. The first level (indicated by a number from 1 to 6) divides enzymes in six main groups (simply called classes), according to the type of chemical reaction catalyzed (e.g. 3 refers to hydrolases, which involve all hydrolytic reactions - cleavage of chemical bonds by the addition of water- and their reversal). The second and third levels provide increasing refinements on the mechanism of the reaction. The fourth level is a serial number that is assigned to inform on substrate specificity.

It is also possible to organize and classify proteins into families and super-families based on similarities between sequences and/or structure for more distant relationships between proteins. A number of studies have observed that, whilst relatives within enzyme super-families may perform different functions or transform substrates in a very different way, there is often conservation of some aspects of their chemistry/mechanisms of reactions between them. Thus, an important step when making hypotheses on the functional activity of an enzyme is to be able to determine its membership to a structural super-family and/ or family. Two classifications of protein three-dimensional structures have been developed to capture their evolutionary relationships, CATH [1] and SCOPe [2]. Both of these classifications use elementary substructures called domains, with proteins featuring one or several domains organized in various ways, and often with different functions. There is a relatively small number of super-families with respect to the number of domains (e.g. the CATH database release v3.5 contains 2626 super-families for 175536 domains) and the issue of predicting the super-family of a protein from its sequence is relatively easy since it is associated to the presence of key domains with some characteristic motifs. In contrast, the family level remains hard to predict from sequences.

In this study, given a known super-family, we consider the issue of classifying a set of new enzyme sequences (the unlabeled set) at the family level with respect to a set of sequences that have already been classified (the labeled set).

2 Coding enzymes using multiple partial local alignment

Enzyme functions can be associated to particular positions in their sequences, corresponding to amino acids involved in molecular interactions that impose a spatial structure, participate in specific binding of a substrate, or are involved in the catalytic machinery. In practice, short common words extracted from sequences of enzymes sharing a same known activity - i.e. short lists of successive amino acids - can help to point out such active sites. However, important aspects have to be considered for this task: the level of biochemical knowledge on protein elements and the divergence within protein sequences through evolution, due to point mutations, large domain rearrangements and insertion/deletions.

When dealing with protein sequences, it is important, first, to take into account the similarities due to shared physico-chemical properties between letters

[4] http://www.iubmb.org/1984

in the alphabet of the 20 standard amino acids used in proteins: some amino acid replacements that might have occurred during evolution have no impact on the function or the structure of the protein while others have. To consider this knowledge, a standard approach in machine learning consists in directly recoding the proteins on a smaller property-based alphabet, such as the hydropathy index or the Dayhoff encoding ([3], [4] and [5]). These coding schemes suffer from being a priori fixed, while the properties of an amino acid involved at one position may differ from those involved for the same amino acid at another position. The work described in this manuscript is based on a more specific data-driven approach where coding is based on the detection of local conservations shared by labeled and unlabeled sequences. The second point concerns the identification of putative domains and active sites in the enzyme sequences that relies on the detection of local similarities in the labeled set.

These goals can be achieved by looking for optimal multiple alignment of sequences. In fact, an alignment does not only provides a recoding of sequences, it also keep track of the chaining of elements since the matching edges between elements in the alignment are constrained not to cross We have extended the standard alignment search by loosening the constraints on admissible alignments in two ways: the alignment is local (involving only substrings) and it is partial (involving only sequences subsets instead of the whole set of sequences as in classical alignment). Altogether, this leads to a partial local multiple alignment (PLMA) of the sequences. Each short strongly conserved region in the PLMA (also called block) will form one of the characters for recoding the sequences. At this stage, it is important to note that the new sequences to be assigned, the unlabeled sequences, need to be also encoded and are aligned together with the sequences of known class, the labeled sequences. The computation of PLMA has been introduced as the first step performed in Protomata-Learner ([6]), a grammatical inference program aiming at learning finite state automata for the characterization of protein family sequence sets. But whereas the choice of the alignment parameters is important in Protomata-Learner to tune the desired level of generalization, we have only used a basic default set of parameters in this study to represent each sequence by the sequence of blocks it is involved in.

3 Class assignment from formal concept analysis

3.1 Formalization of the classification problem

Once each protein sequence have been converted in a boolean sequence or vector of block presences, it remains to assign each unlabeled sequence to a class. This is either a known family class or a new class that has not been observed in the labeled set but gains some evidence from the concurrent presence of specific blocks in the unlabeled set.

A natural approach for such an assignment task is to build a classification of all sequences with respect to the binary attributes (block presence) and to decide the class of unlabeled sequences from their place among the labeled sequences in the labeled tree. This requires to define a similarity measure on the set of binary attributes, and to set a threshold to discriminate the meaningful clusters. Problems quickly arise when trying to follow this approach: the number of attributes may greatly vary from one superfamily to the other and from one sequence to the other within a same (super)family. A decision taken on statistical arguments is not fully satisfactory because it is hard to fix universal values for the necessary parameters and ultimately a biologist has to check the assignments on the basis of the argumentation logics, his own knowledge, and further biochemical characterization of the sequence(s) of interest.

We have thus decided to use a FCA approach to solve this issue. The proposition of formal concept analysis as a first step for a supervised classification task is a common application that can be found in the literature for various topics. The vast majority of related papers have used concept lattices built on a learning set of labeled objects to produce a classifier that is used in a second step to assign new objects with unknown class . The concept lattice provides a nice ordering for the search of rules. This search may be pruned by transforming the initial lattice (closed label lattice, [7]). The most efficient way of generating the classifier is to derive the rules or decision nodes directly from selected concepts in the lattice. For instance, it is possible to build a decision tree from the lattice [8]. A more complex procedure is possible via the computation of concept intersections in the lattice [9]. In all these studies, the set of unlabeled objects is used only once the classifier is built. In contrast, the work of [10] considers the lattice built on both the labeled and the unlabeled set to focus the search on links between known and unknown objects. Then scores are calculated on concepts to estimate the plausibility that a concept represent a set of neighbors (objects belonging to a same class). The class label of objects is thus taken into account through scoring. The method selects first the most discriminating concept for each unlabeled object and classify it with respect to the neighbors class.

In our study, the attributes are blocks and there are two kinds of objects, the labeled and unlabeled enzyme sequences. The idea is then to introduce the knowledge on the classes of labeled instances directly in the formal context. This can be solved simply by adding the class values as new objects. Each time a block b is observed in a sequence of class c, the pair (b, c) is added to the formal context. Including the classes in the context as objects allows to have the right semantics for the binary relation: it reflects the presence of each block in each enzyme and each class. In practice, it is only necessary to produce concepts having at least one unlabeled sequence in the object set, otherwise it it is not useful for sequence labeling. The size of the relation remains sufficiently small in this context to produce the whole lattice of formal concepts for this relation. The assignment procedure is based on the exploitation of the lattice.

In a general setting, let A be the attribute set, C the class set, L the labeled set of objects and U the unlabeled set of objects. Let $\mathcal{I}$ denote the binary relation over $(L + U + C) \times A$ and $\mathcal{B}((L + U + C), A, \mathcal{I})$ the concept lattice.

The problem is to find a minimal extension N of C and an argumentation allowing to assign classes of $N \cup C$ to elements of U on the basis of $\mathcal{B}((L + U + C), A, \mathcal{I})$.

For this purpose, we propose an iterative scheme where each unlabeled sequence is assigned in turn by looking for its compatible class assignments. A *compatible class assignment* is defined as a class that belongs to some concepts sharing a maximal set of blocks with the unlabeled sequence. Maximality is defined here with respect to set inclusion. Each class assignment may be associated to a concept that we call *attribute-compatible concept* (see definition 2).

Definition 1. (compatible class assignment) *Given a concept lattice $B = \mathcal{B}((L+U+C), A, \mathcal{I})$ and an element u of U, a compatible class assignment is an element $c \in C$ such that there exists a concept $(\{u, c\} \cup X, Y)$ in B, $X \subset L+U+C$, and no Y is larger among the possible concepts.*

3.2 Supervised classification

Our method tries to maximize the specificity of the classification decisions and proposes several quality levels for a class assignment towards this end.

At level 1, it checks if some blocks that are specific of a class (i.e. they are present in sequences belonging to a single class) are also present in the current unlabeled sequence. These blocks are called *characteristic blocks* and are assigned the highest quality value since they do not lead to any ambiguity if present alone. It corresponds to build a characteristic partition that splits L in subsets $L_i, i = 1, m$ corresponding to a common class value for each subset and A in $m+1$ possibly empty subsets $A_i, i = 0, m$ corresponding to attributes only present in elements of L_i, with $A_0 = A \setminus \cup_{i=1}^{n} A_i$.

If there exists a single compatible class assignment c using only characteristic blocks, the sequence is classified at level 1, with label c.

If there are several compatible class assignment c using only characteristic blocks, the sequence has an ambiguous classification and if it cannot be classified at the next level, it is said ambiguous and all its possible classes are displayed.

For sequences that have not been classified at level 1, the method checks at level 2 if some concepts are attribute-compatible with respect to the current unlabeled sequence, irrespective of the specificity of its blocks.

If there exists a single compatible class assignment c, the sequence is classified, with label c.

If there are several compatible class assignment c , the sequence is said ambiguous and all its possible classes are displayed.

The remaining cases are when no concept is compatible with the unlabeled sequence. It means either that the sequence has no block in common with another sequence and it remains unclassified, or that it is a member of a new family never observed before that use blocks found only in unlabeled sequences.

For instance, figures 1(a),1(b) and 1(c) represent partial local multiple alignments and in each figure, colored sequences (e.g. $s1$ and $s2$) are labeled sequences while black sequences (e.g. $s3$) are unlabeled and waiting for class assignment. On figure 1(a) the unlabeled sequence $s3$ gets only one compatible class corresponding to the orange concept and can thus be unambiguously classified. However, on figure 1(b) there are two compatible concepts (orange and green), and the unlabeled sequence class assignment is ambiguous. Figure 1(c), provides an example of an unlabeled sequence, $s3$, that remains unclassified because the multiple alignment has found no common block with an other sequence.. On the same picture a new family is formed with a characteristic concept involving only unlabeled sequences : $\{s4, s5, s6\} \times \{Block1, Block2, Block3\}$. The purpose of the next subsection is to detail the search of such new classes.

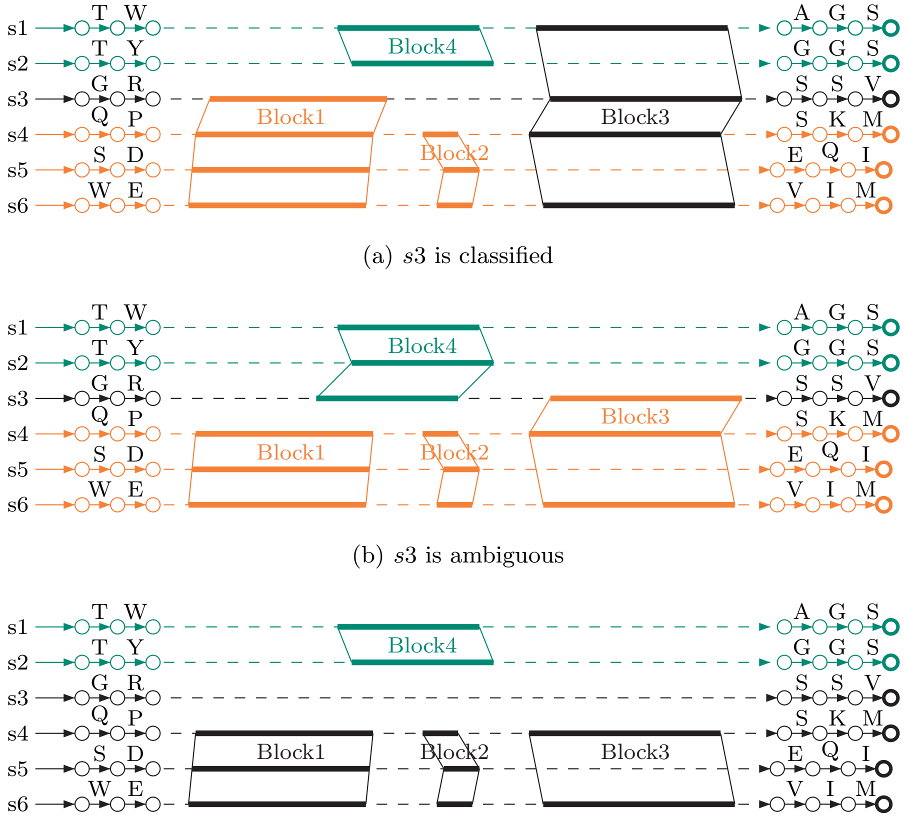
3.3 Unsupervised classification

In terms of FCA, a new family can be characterized like for other families by an associated concept that gathers the sequences of this family and the blocks that form a signature of this family. These blocks are characteristic of unlabeled sequences as is the case for level 1 classification, but this time it is an unsupervised task since the set of classes N is unknown.

This problem is related to biclustering [11]. However, the goal of biclustering consists in simultaneous partitioning of the set of objects and attributes. In our case, it is not realistic to expect a partition of both sets. The objects (sequences) share numerous attributes (blocks) and frequently, it is the way they are combined which allow to distinguish different clusters. The issue of object clustering from a formal context is treated in paper [12]. Authors propose a two-step procedures where formal concepts are enlarged to approximate concepts during the first step and then merged in a second step when they overlap sufficiently. This approach draws on the concept lattice as we do in order to find clusters but it shares some common drawbacks with biclustering with respect to our application domain. A partition of objects is useful but not necessary in our case and furthermore, it is not easy to tune the parameters associated to the method to get meaningful approximate concepts. In [13], the idea of using the set of formal concepts is further elaborated and no need for thresholds is longer required. Instead of starting from the object×attributes concept lattice, the authors propose to consider the lattice built on the object×concepts context in order to build the object clusters. It seems an interesting idea that could be experimented on the protein classification task. However, the interpretation of clusters becomes more difficult and it is an important preoccupation for the biologist to master the decision process. Another related aspect of all these method is their heuristic nature. Concept analysis is an exact method and it seems somewhat unfortunate losing this property in the classification task.

We decided to keep on the idea of associating a concept to each class. We also looked for an exact search of the concepts without parameter tuning, a requirement that implies a neat specification of the target concepts. The issue of deciding the occurrence of new families in N is non trivial due to the conjunction of two difficulties that have to be taken into consideration:

- A given set of sequences participates to a number of concepts. A subset of concepts has to be extracted that covers the set of sequences;



(a) $s3$ is classified

(b) $s3$ is ambiguous

(c) $s3$ is unclassified and $\{s4, s5, s6\}$ forms a new family

Fig. 1. Examples of partial local multiple alignments with labeled (colored) and unlabeled (black) sequences

- The set of new families is not necessary a partition: although it should be avoided as much as possible, a given sequence that has evolved to get a bifunctional capacity could belong to two different families.

We have set this issue as the following optimization problem : find an optimal cover of the new family sequences by the set of concepts including characteristic new blocks -only present in unlabeled sequences-. Optimality depends on three criteria of decreasing priority:

1. minimize the number of ambiguous sequences in the concepts (i.e. get closer to a partition);
2. minimize the size of N (i.e. parsimonious hypothesis with a minimum number of necessary new families);
3. maximize globally the support of the new families in terms of number of characteristic blocks.

These criteria are coded within a set of logical constraints using Answer Set programming [14]. Optimal concepts are produced by a dedicated solver through a conflict-driven constrained enumeration of admissible solutions [15]. This way, exact solutions can be produced.

Another important aspect of the quality of a classification decision is its support with respect to existing labeled sequences. A compatible concept can be associated to each decision. This concept gets a support in terms of its number of blocks. Another measure is the support in terms of labeled sequences. However, the compatible concept is not the best one with respect to this measure. It may exist a concept in the lattice, called seq-compatible concept, with a larger sequence support:

Definition 2. (attribute-compatible and object-compatible concept) *Given a concept lattice* $B = \mathcal{B}((L+U+C), A, \mathcal{I})$*,* $u \in U$*, and* $c \in C$*, the attribute-compatible concept and object-compatible concept are concepts* $BC(u,c) = (\{u,c\} \cup X, Y_{max})$ *and* $BC(u,c) = (\{u,c\} \cup X_{max}, Y)$ *of* B*, where* $Y_{max} = max\{Y \subset A : (\{u,c\} \cup X, Y) \in B, X \subset L+U+C\}$ *and* $X_{max} = max\{X \subset L+U+C : (\{u,c\} \cup X, Y) \in B, Y \subset A\}$*.*

This way, each class assignment may be scored by the number of blocks of its attribute-compatible concept and the number of sequences of its object-compatible concept.

4 An experiment with the HaloAcid Dehalogenase enzyme superfamily (HAD)

The haloacid dehalogenase superfamily (HAD) is a large superfamily (120193 sequences reported; http://pfam.sanger.ac.uk/clan/CL0137) of ubiquitous enzymes present in all three superkingdoms of life. The numbers of sequences differ between organisms, from around twenty in the bacteria *Escherichia coli* [16] to between 150-200 in the benchmark biological models *Arabidopsis* thaliana and *Homo sapiens* [17]. HADs are involved in a variety of cellular processes and serve as the predominant catalysts of metabolic phosphate ester hydrolysis [18]. Enzymes in this superfamily are related by their ability to form covalent enzyme-substrate intermediates via a conserved aspartic acid site. These enzymes catalyze enzymatic cleavage, by nucleophilic substitution, of carbon-halogen bonds (C-halogen), and also feature a variety of hydrolytic activities including phosphatase (CO-P), phosphonatase (C-P) and phosphoglucomutase (CO-P hydrolysis and intramolecular phosphoryl transfer) reactions. The figure 2 provides an example of HAD structure from a bacteria.

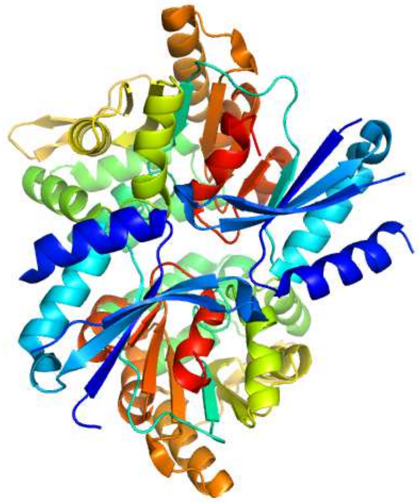


Fig. 2. 3D structure of HAD hydrolase T0658 from *Salmonella enterica*

All structurally characterized superfamily members share a conserved alpha/beta-core domain, termed the "HAD-like" fold by SCOP. HAD superfamily enzymes usually function as homodimers (i.e., a complex made of two identical proteins).The core domain is similar to the 'Rossmann-fold' with a six stranded parallel β-sheet, flanked by five α-helices. The typical fold of HAD phosphatases contains three additional structural signatures that allow the enzyme to adopt distinct conformational states and that contribute to substrate specificity: the squiggle, flap, and cap domains. Effectively, most superfamily members have a cap domain, and its site of incorporation within the sequence is one of the parameter supporting the enzymatic diversity within the HAD superfamily [19]. Dehalogenases have received an increased interest in the last decade since they have the potential to be used in both industrial and pharmaceutical applications, in addition to bioremediation processes [20].

For this experiment, we have worked on the following datasets :

1. Sequences from various organisms extracted from the supplementary data of article [19]. 34 families, 3 sequences in each family (102 sequences);
2. Sequences from *E. coli* extracted from [16] 23 sequences;
3. Sequences from *H. sapiens* extracted from [17] 40 sequences ;
4. Sequences from *A. thaliana* extracted from the TAIR database by identifying proteins containing a HAD domain and sequences identified by bibliographic analysis [19] 153 sequences, including 23 unlabeled sequences.

In all the study, dataset 1 is used as the labeled set and contains sequences labeled with a family class. The three remaining datasets have been used as unlabeled sets, and the sequence family prediction made by FCA have been compared to the classification existing for some of the sequences contained in these datasets in order to assess the performance of our analysis. Indeed, many sequences from E. coli, Homo sapiens and Arabidopsis thaliana have been biochemically characterized and/or have been considered for in silico/in vivo structural analysis, and this provided experimental results on their classification.

Figure 3 shows the complete lattice obtained on the smallest context corresponding to the E. coli unlabeled dataset. This line diagram has been drawn using the software erca (Eclipse's Relational Concept Analysis [5]) and a reduced labeling. The top concept 0 contains all blocks and no sequence or class. The bottom concept 4 contains all sequences and classes and no block. The edges going to concept 9 and others were slightly intertwined and we have used a blue color to better distinguish them. The concepts having at least one unlabeled sequence in the figure are colored in sea green. These concepts contain the set of blocks of the unlabeled sequences, a maximal subset of which has to be used for classification.

Assignment results are summarized in table 1.

		E. coli	*H. sapiens*	*A. thaliana*
Classified (%)	**True**	**61**	**65**	**56**
	False	9	3	6
Ambiguous (%)	True	17	18	18
	False	13	3	8
Unclassified (%)	True	0	8	8
	False	0	3	5
Total		100	100	100

Table 1. Percentage by species of sequences correctly/wrongly assigned

The row "Classified" refers to sequences with only one predicted compatible class. The row "Ambiguous" refers to sequences with several compatible classes.

[5] https://code.google.com/p/erca/

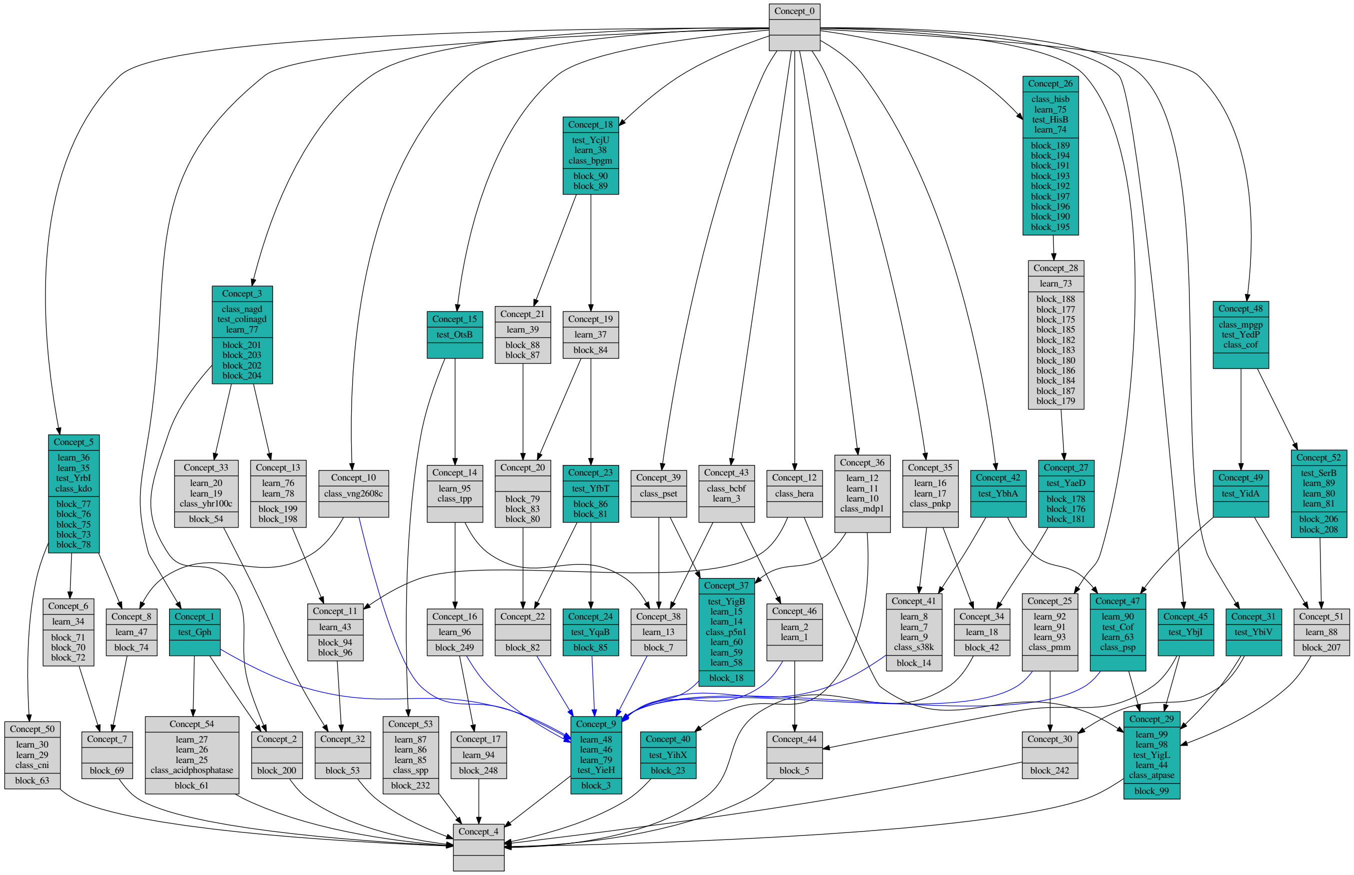
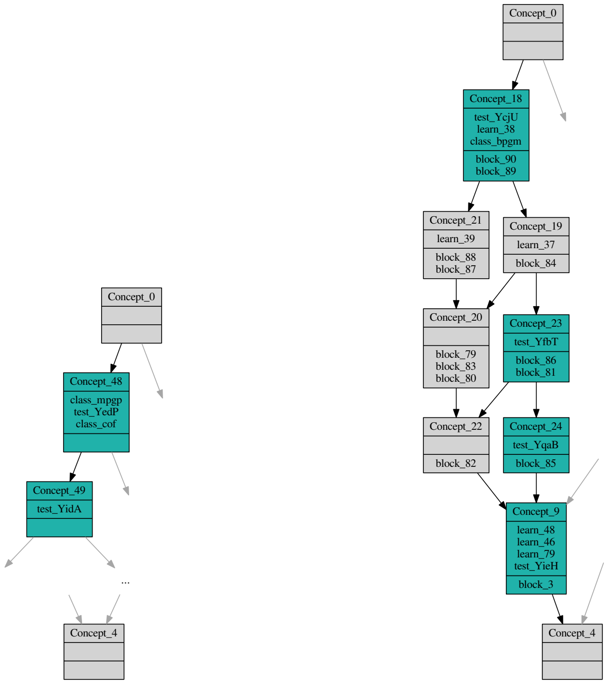


Fig. 3. Hasse diagram from lattice blocks x sequences/classes obtained on experiment with E. coli as unlabeled dataset



(a) YedP and YidA are ambiguous with two possible class labels, mpgp and cof

(b) YfbT and YcjU can be classified and assigned uniquely with the class bpgm

Fig. 4. Different kinds of assignment decisions

The classification is assumed to be correct (true) if one of these compatible classes is the good one. The percentage of correctly/wrongly assigned sequences is given.

These first results are encouraging. More than 50% of sequences are correctly classified into the 34 possible families, new families detected by the method are effective and sequences not belonging to the superfamily remain unclassified.

For a fraction of unlabeled sequences, their right classification is actually unknown (datasets *H. sapiens* and *A. thaliana*). Yet, it is possible to look for possible class assignments. Table 2 give the percentage of such sequences that could be classified by our method. It shows that most of these unknown sequences could be assigned to one or several classes.

	H. sapiens	*A. thaliana*
Classified (%)	50	54
Ambiguous (%)	50	21
Unclassified (%)	0	25
Total	100	100

Table 2. Percentages of unknown sequences in datasets assigned to one, several or none of the classes

The percentages of sequences belonging to new families and of unclassified sequences are also given. Unclassified sequences are sequences that can neither been assigned to a known class nor be assigned to a new family cluster.

For the three datasets, *E. coli*, *H. sapiens* and *A. thaliana*, we find 0, 2 and 11 new subfamilies respectively.

In the case of the *H. sapiens* dataset, sequences predicted to belong to new families are described in the reference paper [17]. As a matter of fact, these families are not present in the labeled set and are thus correctly predicted as new.

In the case of *A. thaliana* dataset, it is difficult to know if predicted new families are real because of the number of uncertain sequences. However, we have detected after some bibliographical study that 11 unclassified sequences seem to have been wrongly assigned to the HAD superfamily by the TAIR query.

The specificity of the detection of new families has been tested too. For each known family in the labeled set, a new labeled set has been built that contain all sequences from the initial labeled set except the sequences belonging to this family. The unlabeled set has been built with the *E. coli* dataset plus the sequences of the selected family (3 sequences). The goal was to retrieve all the family sequences in the unlabeled set family detected as a new family by our method. We have computed the percentage of retrieved sequences for all families. The results are shown in table 3. Note that some families are not present in *E. coli* and this is indicated by the column label "new family alone". For the others (column new family + *E. coli*), the number of *E. coli* sequences belonging to the family is given within brackets.

new family alone	% of retrieved sequences	new family + *E. coli*	% of retrieved sequences
EYA	100	NagD (+1)	100
SPSC	100	Cof (+6)	44
ATPase	0	PSP (+1)	75
PNKP	100	HisB (+2)	100
deoxy	0	BPGM (+6)	67
s38K	100	Sdt1p (+4)	43
HerA	0	TPP (+1)	100
PMM	100	KDO (+1)	100
Yhr100c	100	MPGP (+1)	67
CNI	67		
Enolase	100		
BCBF	100		
LPIN	100		
PseT	100		
P5N1	100		
AcidPhosphatase	100		
Phosphonatase	100		
VNG2608C	0		
SPP	100		
CNII	100		
MDP1	100		
dehr	0		
Zr25	100		
CTD	100		

Table 3. Percentage of retrieved sequences within a new family

On the 34 subfamilies present in the labeled set, the unlabeled has been convincing for 27 of them.

These already good results could be refined by finding a largest labeled subset retrieved as a new family. This could be a direct measure of the specificity of the method.

5 Conclusion

We have described a classification method based on a concept lattice including both a set of already classified objects and a set of objects to be classified. It has been applied to enzyme sequences, a group of key proteins involved in many biochemical processes and with a high potential for the discovery of new functional molecules. Our results are encouraging and show our classification method is sensitive and specific. More than half of the unlabeled sequences are correctly classified with respect to the current knowledge for 34 subfamilies and ambiguous sequences represent only one third of sequences, two thirds of them having the correct class assignment. Moreover, each classification decision may be clearly explained and related to known sequences or particular positions in the sequence corresponding to blocks. Ambiguity could be even reduced in practice by looking for sequences that are inherently ambiguous because they are made for instance of two fragments of two proteins of different class. Such potential proteins, which we call chimera, could be automatically extracted during classification.

Another aspect of this work is the unsupervised classification problem for objects with attributes that are characteristic of unlabeled objects. We have suggested a model for solving this problem as an optimization issue taking into account ambiguity, parsimony (number of needed new classes) and intent (number of attributes).

To our knowledge, it is the first time that this issue is properly formalized in bioinformatics. We have implemented all the specifications in this paper in the framework of answer set programming, a form of declarative programming adapted to combinatorial problems [14]. Once all constraints are expressed as logical formulas, a grounder transform them in a (large) set of boolean formulas and a solver looks for possible models of this set (the answers), which give access to the solutions of the initial problem. We have used the solver Clasp developed in Potsdam University [15].

The next step will consist in testing the robustness of the method on species that are very evolutionary distant compared to the other organisms for which test sets were considered. To this aim, we have selected the brown alga *Ectocarpus siliculosus*, for which the genome sequence has been recently published [21]. We will test if the best in silico assignment within classes correlates with potential substrate specificity. To this aim, a number of algal sequences will also be biochemically characterized.

References

[1] Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R., Yeats, C., Thornton, J.M., Orengo, C.A.: New functional families (funfams) in cath to improve the mapping of conserved functional sites to 3d structures. Nucleic Acids Research **41** (2013) D490–D498

[2] Fox, N.K., Brenner, S.E., Chandonia, J.M.: SCOPe: Structural Classification of Proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Research **42** (2014) D304–D309

[3] Yokomori, T., Ishida, N., Kobayashi, S.: Learning local languages and its application to protein α-chain identification. In: HICSS (5). (1994) 113–122

[4] Peris, P., López, D., Campos, M., Sempere, J.M.: Protein motif prediction by grammatical inference. In Sakakibara, Y., Kobayashi, S., Sato, K., Nishino, T., Tomita, E., eds.: ICGI. Volume 4201 of Lecture Notes in Computer Science., Springer (2006) 175–187

[5] Peris, P., López, D., Campos, M.: Igtm: An algorithm to predict transmembrane domains and topology in proteins. BMC Bioinformatics **9** (2008)

[6] Kerbellec, G.: Apprentissage d'automates modélisant des familles de séquences protéiques. PhD thesis, Université Rennes 1 (2008)

[7] Wang, J., Liang, J., Qian, Y.: Closed-label concept lattice based rule extraction approach. In Huang, D.S., Gan, Y., Premaratne, P., Han, K., eds.: ICIC (3). Volume 6840 of Lecture Notes in Computer Science., Springer (2011) 690–698

[8] Kovacs, L.: Generating decision tree from lattice for classification. In: 7th International Conference on Applied Informatics. Volume 2. (2007) 377–384

[9] Sahami, M.: Learning classification rules using lattices. In Lavrac, N., Wrobel, S., eds.: ECML. Volume 912 of Lecture Notes in Computer Science., Springer (1995) 343–346

[10] Ikeda, M., Yamamoto, A.: Classification by Selecting Plausible Formal Concepts in a Concept Lattice. In: Workshop on Formal Concept Analysis meets Information Retrieval (FCAIR2013). (2013) 22–35

[11] Busygin, S., Prokopyev, O., Pardalos, P.M.: Biclustering in data mining. Comput. Oper. Res. **35** (2008) 2964–2987

[12] Gaume, B., Navarro, E., Prade, H.: Clustering bipartite graphs in terms of approximate formal concepts and sub-contexts. International Journal of Computational Intelligence Systems **6** (2013) 1125–1142

[13] Navarro, E., Prade, H., Gaume, B.: Clustering sets of objects using concepts-objects bipartite graphs. In Hüllermeier, E., Link, S., Fober, T., Seeger, B., eds.: SUM. Volume 7520 of Lecture Notes in Computer Science., Springer (2012) 420–432

[14] Brewka, G., Eiter, T., Truszczyński, M.: Answer set programming at a glance. Commun. ACM **54** (2011) 92–103

[15] Gebser, M., Kaufmann, B., Schaub, T.: Conflict-driven answer set solving: From theory to practice. Artif. Intell. **187** (2012) 52–89

[16] Kuznetsova, E., Proudfoot, M., Gonzalez, C.F., Brown, G., Omelchenko, M.V., Borozan, I., Carmel, L., Wolf, Y.I., Mori, H., Savchenko, A.V., Arrowsmith, C.H., Koonin, E.V., Edwards, A.M., Yakunin, A.F.: Genome-wide Analysis of Substrate Specificities of the *Escherichia coli* Haloacid Dehalogenase-like Phosphatase Family. Journal of Biological Chemistry **281** (2006) 36149–36161

[17] Seifried, A., Schultz, J., Gohla, A.: Human HAD phosphatases: structure, mechanism, and roles in health and disease. FEBS Journal **280** (2013) 549–571

[18] Koonin, E.V., Tatusov, R.L.: Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity: Application of an iterative approach to database search. Journal of Molecular Biology **244** (1994) 125–132

[19] Burroughs, A.M., Allen, K.N., Dunaway-Mariano, D., Aravind, L.: Evolutionary Genomics of the HAD Superfamily: Understanding the Structural Adaptations and Catalytic Diversity in a Superfamily of Phosphoesterases and Allied Enzymes. Journal of Molecular Biology **361** (2006) 1003 – 1034

[20] Janssen, D.B.: Biocatalysis by dehalogenating enzymes. Volume 61 of Advances in Applied Microbiology. Academic Press (2007) 233 – 252

[21] Mark Cock, J., Sterck, L., Rouzé, P., Scornet, D., Allen, A., Amoutzias, G., Anthouard, V., Artiguenave, F., Aury, J., Badger, J.: The Ectocarpus genome and the independent evolution of multicellularity in brown algae. Nature (2010) 617–621