Formal Concept Analysis Applications in Bioinformatics

SARAH ROSCOE, MINAL KHATRI, and ADAM VOSHALL, University of Nebraska-Lincoln, USA
SURINDER BATRA and SUKHWINDER KAUR, University of Nebraska Medical Center, USA
JITENDER DEOGUN, University of Nebraska-Lincoln, USA

The bioinformatics discipline seeks to solve problems in biology with computational theories and methods. Formal concept analysis (FCA) is one such theoretical model, based on partial orders. FCA allows the user to examine the structural properties of data based on which subsets of the dataset depend on each other. This article surveys the current literature related to the use of FCA for bioinformatics. The survey begins with a discussion of FCA, its hierarchical advantages, several advanced models of FCA, and lattice management strategies. It then examines how FCA has been used in bioinformatics applications, followed by future prospects of FCA in those areas. The applications addressed include gene data analysis (with next-generation sequencing), biomarkers discovery, protein-protein interaction, disease analysis (including COVID-19, cancer, and others), drug design and development, healthcare informatics, biomedical ontologies, and phylogeny. Some of the most promising prospects of FCA are identifying influential nodes in a network representing protein-protein interactions, determining critical concepts to discover biomarkers, integrating machine learning and deep learning for cancer classification, and pattern matching for next-generation sequencing.

CCS Concepts: • **Applied computing** → **Bioinformatics**; • **Information systems** → **Data mining;**

Additional Key Words and Phrases: Biomarkers discovery, biomedical ontologies, cancer classification, disease classification, drug design and development, formal concept analysis, gene expression data, healthcare informatics, next-generation sequencing data analysis, phylogeny, protein-protein interactions

ACM Reference format:
Sarah Roscoe, Minal Khatri, Adam Voshall, Surinder Batra, Sukhwinder Kaur, and Jitender Deogun. 2022. Formal Concept Analysis Applications in Bioinformatics. *ACM Comput. Surv.* 55, 8, Article 168 (December 2022), 40 pages.
https://doi.org/10.1145/3554728

1 INTRODUCTION

Computation has been used to solve several important biological problems in the past few decades because of the rise of computational power and increasing availability of large datasets. A principal example is the Human Genome Project [2], an effort in the early 2000s to produce a database of human genome sequences, or sequences of molecular structures that instruct a cell what to do in response to a situation. The human genome was fully sequenced in [124], and the database

Authors' addresses: S. Roscoe, M. Khatri, A. Voshall, and J. Deogun, School of Computing, University of Nebraska-Lincoln, Lincoln, NE, 68588-0115; email: sroscoe@huskers.unl.edu, khatri.cs16@gmail.com, s-avoshal1@huskers.unl.edu, deogun@cse.unl.edu; S. Batra and S. Kaur, University of Nebraska Medical Center, Department of Biochemistry and Molecular Biology, Omaha, NE, 68198; email: sbatra@unmc.edu.

0360-0300/2022/12-ART168 $15.00
https://doi.org/10.1145/3554728

produced by the project is still being analyzed today to determine which areas produce biological functions and identify how variations and mutations in genes can lead to onset of diseases, such as cancer or Alzheimer's.

After the success of the Human Genome Project in the early 2000s, much attention was given to how large amounts of genomic data can be analyzed. Clustering [87] is an important method of analysis, as grouping genes that have similar expression levels can reveal which ones work together in situations to perform a biological function. Clustering can be done with **formal concept analysis (FCA)**, a theoretical framework of computing that forms a hierarchical grouping of data. The hierarchy is based on partial orders, and structural properties of the data can be examined [164].

FCA was developed in the 1980s by Wille to revitalize lattice theory, part of abstract algebra. While FCA was originally developed for binary data, variations and extensions have been a popular area of research. These advanced models allow for FCA's applications on varying types of data. The models include fuzzy [26, 112], probabilistic [45, 46, 160], triadic [95, 146], or relational [31, 145, 161] concept analysis. Since its conception, FCA has benefited fields looking at the structure of its data, ontologies, and other areas where hierarchical analysis is helpful.

FCA works by first determining concepts, maximal subsets of the dataset that are marked by Trues. Then, a mathematical structure called a lattice is developed to hierarchically represent the relationships of those maximal subsets. For smaller datasets, the lattice can also be visualized to uncover implicit information about the data and how subsets of the data depend on one another. Implementations to visualize the lattice, such as LatViz and Concept Explorer [9, 166], were developed in the early 2000s. For larger datasets, the lattice may be prohibitively complex due to the exponential number of concepts. The exponential risk is one of the main limitations of using FCA. However, FCA has been used for a variety of bioinformatics applications over the last few decades.

This survey describes the FCA method, current bioinformatics applications of FCA, and future prospects of FCA in such applications.

We summarize some of the important applications here. The first ones using FCA to analyze biological data clustered gene expression data. Although expression data is measured now with high-throughput RNA sequencing [163], most published work using FCA on this topic was published before RNA sequencing, and instead used microarrays, chips that store up to tens of thousands of individual DNA strands. In studying groups of similarly expressed genes, biomarkers can be identified. Biomarkers are indicators of a future biological process. In addition to determining concepts from FCA, critical concepts may be identified and thus discern genes indicated to work together. Data representing the interactions between proteins can also be analyzed. When interaction data is modeled as a graph, its adjacency matrix can be analyzed through FCA, and significant nodes and thus significant proteins can be identified. FCA has been used with a concept-based weight function to determine the influential nodes.

Many biological datasets are very large and complex due to evolving omics technologies. The analysis of such high-throughput data is difficult for many models. Because of this, machine learning and deep learning are valuable approaches in bioinformatics, as complex data can be handled rather easily. A few studies have used these approaches in bioinformatics, in conjunction with FCA. Some studies use FCA for feature selection, or to show which features were determined most important by various classifiers. A recent prospect is the combination of FCA and deep learning methods to perform cancer classification. An advanced model of FCA, fuzzy FCA, has also been used to aid this task. The more frequent use of advanced models, such as rough FCA, relational concept analysis, and three-way (triadic) FCA, is a prospect as well. In addition, using distributive computing frameworks has been shown to greatly reduce computation time for FCA algorithms.

A survey addressing FCA's use for bioinformatics was done by Poelmans et al. in 2013 [130]. The survey focuses on FCA's use for knowledge discovery, including text mining, web mining, software mining, bioinformatics, chemistry, and medicine. Gene expression data analysis is the most frequent type of bioinformatics study cited. We discuss gene expression data analysis as well.

In 2017, another survey on the use of FCA to discover knowledge, specifically from biological data, was presented by Raza [136]. The survey addresses FCA's use to analyze gene expression data, the gene regulatory network, and the identification of fragments and compounds in the drug design and development application. We discuss these topics as well. Due to there being no recent developments in how FCA is used in enzyme classification and protein binding sites identification, we refer the reader to Raza's article for a summary of these applications.

We present a more comprehensive and up-to-date review by including a discussion of FCA, its hierarchical advantage, and various data management strategies. In addition, we address recent updates in two areas discussed by Raza (protein-protein interactions, drug design and development), cover seven new areas not addressed by Raza, and identify several areas where each application can improve in future investigations with FCA.

The contributions of this article are as follows:

(1) We survey the literature that uses FCA to study bioinformatics applications. In particular, we examine gene expression data analysis, disease classification, COVID-19-related analysis, gene regulatory network analysis, protein-protein interaction, drug design and development, next-generation sequencing, biomarkers discovery, healthcare informatics, biomedical ontologies, and phylogeny.
(2) In each application, we discuss the current trend of research and, if relevant, outline its chronological development.
(3) We identify several prospective areas where each of the above applications can benefit from future research involving FCA:
  (a) We identify several ways gene expression data analysis with FCA may be improved.
  (b) We motivate and present a new hybrid deep learning and FCA classification method using cancer image data.
  (c) Ways to detect COVID-19 treatment similarities, symptom tracking, and recommendation that individuals isolate when exposed are proposed.
  (d) Analysis of the gene regulatory network may be further developed by discovering seed genes, genes that interact strongly with the seed genes, or attempting GRN reconstruction with FCA.
  (e) The drug design and development application can use FCA to visualize the structure-activity relationship, use FCA-based social media analysis methods to determine drug side effects or efficacy, and perform more medieval pharmacy data analysis.
  (f) Next-generation sequencing opportunities using brain storm optimization, as well as pattern matching using deterministic finite automata methods, are proposed.
  (g) Protein-protein interactions can be identified with FCA by interfacing with graph theory.
  (h) Biomarkers that are discovered with the help of FCA can be experimentally validated in the lab. Critical concepts leading to biomarker discovery can also be identified using user-friendly web tools.
  (i) FCA can be used more to identify missing data and help communicate between different biomedical ontologies.
  (j) Unique solutions of currently developed phylogeny algorithms can result in a smooth transformation between FCA's concept lattice and phylogeny's median graph. Examining an organism's evolutionary development can also be done with the resulting FCA method.

Table 1. Animals and Their Characteristics Represented in a Binary FCA Context Called $\mathbb{K}_1$

$\mathbb{K}_1$	Tail	Whiskers	Fur	Feathers	Scales	Fins	Swims	Noise
Dog	X	X	X				X	X
Cat	X	X	X					X
Fish	X				X	X	X	
Parrot	X			X				X
Duck	X			X			X	X
Snake	X				X		X	X

The rest of this article is organized as follows. First, we provide an overview of how FCA works, benefits to its hierarchical structure, its notable advanced models, and methods of managing data in Section 2. Section 3 contains the survey of how FCA is used in various bioinformatics applications. Section 4 identifies future prospects in each of the applications addressed in the previous section. Finally, we conclude in Section 5.

2 FORMAL CONCEPT ANALYSIS

FCA was introduced in the 1980s by Wille and developed in tandem with Ganter [60]. It is a theoretical model for representing relationships between subsets of a dataset. These relationships are hierarchical, given that FCA was developed to revitalize lattice theory, a mathematical area based on partial orders. The structure of the data can reveal many important insights, and these insights are applied to biomarkers discovery, protein-protein interactions, co-expressed genes, and other bioinformatics applications. After the FCA model was developed, several generalizations were done to allow the use of multi-valued, probabilistic, and other non-binary data. Lattice management techniques have also been developed for managing data, including attempts to calculate only certain concepts, remove redundant objects or attributes, and simplify the lattice once it is calculated. We now describe the basics of formal concept analysis.

2.1 Basics of Formal Concept Analysis

The input to formal concept analysis is a formal context, a theoretical structure similar to a matrix. The rows of the context represent objects, and the columns represent attributes. A True (represented by an "X" in our example contexts) in a certain row and column of the context indicates that the corresponding object has the corresponding attribute. Conversely, a False (respectively, the absence of an "X") indicates that the corresponding object does not possess the corresponding attribute. Note that objects and attributes may be purely abstract.

Formally, the context is a triple $\mathbb{K} = (G, M, I)$, where G is the set of objects, M is the set of attributes, and I is the binary relation between the object and attribute sets. One simple example of a context, $\mathbb{K}_1$, is shown in Table 1. Various animals chosen (Dog, Cat, Fish, Parrot, Duck, and Snake) may or may not possess the various attributes (having Tail, Whiskers, Fur, Feathers, Scales, or Fins; able to Swim; making Noise). A naive observation of a collection of these animals results in data that indicates all the animals have a Tail; only Dog and Cat have Whiskers and Fur; Parrot and Duck have Feathers; Snake and Fish have Scales; Fish is the only animal with Fins; Dog, Fish, Duck, and Snake can Swim; and all except for Fish make Noise. This data is stored in the matrix representing the relation I between G and M. Later, in Section 2.3, we will address whether these attributes are true for all individuals of these species.

Once the context is constructed, we discover implicit information in the data by looking at its structure and hierarchy. This is achieved with a formal concept, referred to as simply "concept,"

Table 2. A Sample Concept from the Context $\mathbb{K}_1$ in Table 1

	Tail	Feathers	Noise
Parrot	X	X	X
Duck	X	X	X

Notice that for the objects Parrot and Duck, all attributes in common to both objects (Tail, Feathers, Noise) are included in the concept.

for which FCA is named. The formal concept is a subset of the matrix with Trues in all locations, according to some constraints addressed in Definitions 1 and 2.

Definition 1 (Intent). For a set $A \subseteq G$ of objects, the *intent* of A is denoted by A', where

$$A' := \{m \in M | \text{ every object } g \text{ in } A \text{ possesses every attribute } m\}.$$

Alternatively, this is the corresponding set of attributes that all objects in A possess.

Example 1. As a simple example, for the set $A = \{\text{Dog, Cat}\}$, the set A' is {Tail, Whiskers, Fur, Noise}.

Definition 2 (Extent). For a set $B \subseteq M$ of attributes, the *extent* of B is denoted by B', where

$$B' := \{g \in G | \text{ every attribute } m \text{ in } B \text{ is possessed by object } g\}.$$

Alternatively, this is the corresponding set of objects that possess all the attributes in B.

Example 2. For the set $B = \{\text{Swims, Noise}\}$, the set B' is {Dog, Duck, Snake}.

We can also obtain A'' from Example 1, which is {Dog, Cat}. Similarly, B'' from Example 2 is {Tail, Swims, Noise}.

Definition 3 (Formal Concept). A *formal concept* is a pair (A, B) of objects and attributes, respectively, such that $A' = B$ and $B' = A$. A is called the extent of the concept and B is called the intent of the concept.

Put another way, a concept is simply a pair of objects and attributes (A, B) such that every object in A possesses all attributes in B, and every attribute in B is possessed by all objects in A. Due to this definition, (A, A') from Example 1 is a concept because $A' = A'$ trivially and $A'' = A$. However, (B', B) from Example 2 is not a concept because $B'' \neq B$. We define the set of all concepts of a context as $\mathfrak{B}(G, M, I)$.

With these definitions, another example concept is shown in Table 2. This concept is constructed by choosing the set A= {Parrot, Duck}, then obtaining all attributes that satisfy A'. These attributes are B = {Tail, Feathers, Noise}. Even though Noise and Tail are both attributes that belong to objects outside the set A, the inclusion of the attribute Feathers in B restricts the object set A to only Parrot and Duck, since both objects are the only ones to share the attribute Feathers.

Because the concepts are specific subsets of the original context, some concepts may depend on each other. We demonstrate this with another valid concept in Table 3, and Definition 4.

Definition 4 (Subconcept, Superconcept). Given two concepts $C_1 = (A_1, B_1)$ and $C_2 = (A_2, B_2)$, C_1 is a *subconcept* of C_2 (denoted $C_1 \leq C_2$) iff $A_1 \subseteq A_2$ or $B_1 \supseteq B_2$. Equivalently, C_2 is said to be a *superconcept* of C_1 and $C_2 \geq C_1$.

Table 3. A Concept (A, B) Where A = {Cat, Parrot, Duck, Snake} and B = {Tail, Noise}

	Tail	Noise
Cat	X	X
Parrot	X	X
Duck	X	X
Snake	X	X

It is simple to see that every context will have two trivial concepts: the first has a potentially empty extent and an intent containing all attributes in the context. The second has an extent containing all objects in the context, and a potentially empty intent. These are valid parts of their concepts due to the possible coincidence that all objects share one or more attributes (as in the case of the attribute Tail in $\mathbb{K}_1$) or all attributes being possessed by one or more objects. For ease of repetition, we will call the all-object concept C_G and the all-attribute concept C_M.

FCA was created with lattice theory, an existing field of modern algebra, to reconnect abstract mathematics to the modeling of real-world situations and data. We now present some key definitions of lattice theory, beginning with the join and meet of a set in Definitions 5 and 6, then discuss how the definitions function in FCA.

Definition 5 (Join). A set S with a partial order $\leq$ has a *join* iff there is an element $c \in S$ such that $c \geq x$ for every $x \in S$. c is also called the supremum or least upper bound of S.

Definition 6 (Meet). A set S with a partial order $\leq$ has a *meet* iff there is an element $c \in S$ such that $c \leq x$ for every $x \in S$. c is also called the infimum or greatest upper bound of S.

The set of all concepts, $\mathfrak{B}(G, M, I)$, as well as its subsets, fulfills a pair of strict definitions involving the join and meet of its elements.

Definition 7 (Lattice). A *lattice* is a set S such that there is a join and meet for every pair of elements of S.

Since pairs of concepts in $\mathfrak{B}(G, M, I)$ have at least the general supremum C_G and the general infimum C_M, $\mathfrak{B}(G, M, I)$ is a lattice. In addition, $\mathfrak{B}(G, M, I)$ has a structural property that every one of its subsets has a join and meet. The meaning of this in lattice theory is defined in Definition 8 and its implications are discussed in the following paragraphs.

Definition 8 (Complete Lattice). A *complete lattice* S is a lattice such that every subset of S has a join and a meet.

The term for the lattice of concepts ordered by the subconcept-superconcept relation is defined in Definition 9.

Definition 9 (Concept Lattice). The set of all concepts, ordered by $\leq$, is called the *concept lattice* and is denoted by $\underline{\mathfrak{B}}(G, M, I)$.

The concept lattice $\underline{\mathfrak{B}}(G, M, I)$ is a complete lattice, the rigorous proof of which is out of the scope of this article; however, it can be found in Theorem 3 in Ganter and Wille's book [60]. Instead, we offer a general understanding of the theorem. If $\underline{\mathfrak{B}}(G, M, I)$ was not complete, there would be some set $Y \subseteq \mathfrak{B}(G, M, I)$ for which the join or meet would not exist. Y could not be C_G or C_M, since these two concepts, always present in the concept lattice, are the join and meet of themselves: C_G

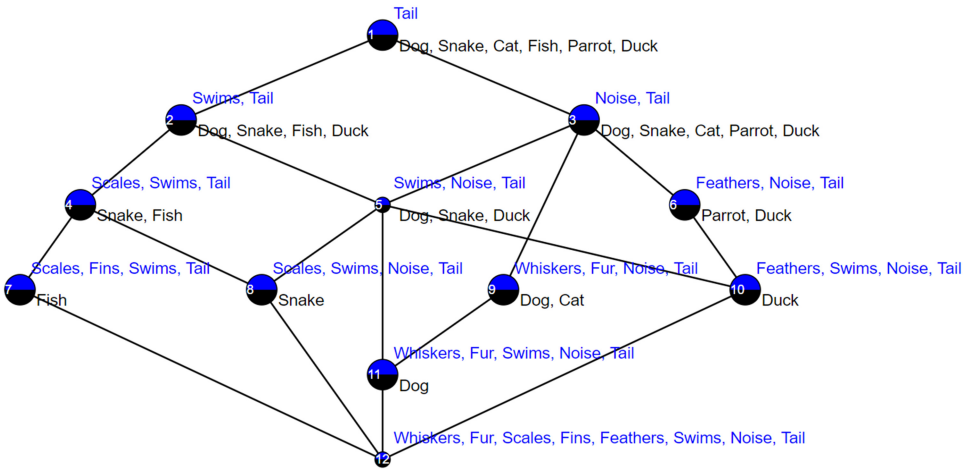

Fig. 1. The concept lattice for context $\mathbb{K}_1$.

Table 4. A Simplified Context

$\mathbb{K}_2$	*Noise*	*Scales*	*Fins*
Fish		*X*	*X*
Snake	*X*	*X*	
Parrot	*X*		

is the join of itself and C_M, while C_M is the meet of itself and C_G. It remains for Y to be a singleton concept in the lattice. However, since C_G and C_M are the respective join and meet of the entire lattice, it is impossible for Y to be a singleton. This means that because $\underline{\mathfrak{B}}(G, M, I)$ is a complete lattice, it must be *connected* in a graph theoretic sense. In this way, although the data may be disjoint (concepts apart from C_G and C_M are not sub- or superconcepts of each other), it will not be disconnected, and can therefore be analyzed accordingly.

A visualization of $\mathbb{K}_1$, made with LatViz [9], is shown in Figure 1. In this type of diagram, there are three main parts for interpretation: vertices, edges, and labels. Circles are concepts; lines represent subconcept-superconcept ordering, with a subconcept represented below the superconcept; labeling enables the reader to see which objects share certain attributes in a way that will be explained once the entire diagram is constructed. Even though one concept may be below another, physical distance is of no relevance in the diagram. We now present a more detailed example of how the concept lattice is constructed and how it may be interpreted in Example 3.

Example 3. Taking a subset of data to be $\mathbb{K}_2$ (Table 4) from the original context $\mathbb{K}_1$, we compute six concepts: C_M with an empty object set, the set of all attributes the animal Snake possesses (C_2); the set of all attributes the animal Fish possesses (C_3); the set of all objects possessing the attribute Scales (C_4); the set of all objects possessing the attribute Noise (C_5); and C_G with an empty attribute set. According to our definitions, no other concepts apart from these six are possible, as it would involve one of the object or attribute sets not equaling the equivalent intent or extent, respectively.

Now that we have computed all the concepts, we determine if some are subconcepts or superconcepts of each other. By examining the corresponding object and attribute sets, we find $C_M \leq C_G$, as expected. Additionally, $C_2 \leq C_4, C_5$, and C_G, while $C_3 \leq C_4$ and C_G. We find that C_4 and C_5 have no relation between each other because neither has a set that is a subset of the other's. C_2 and C_3

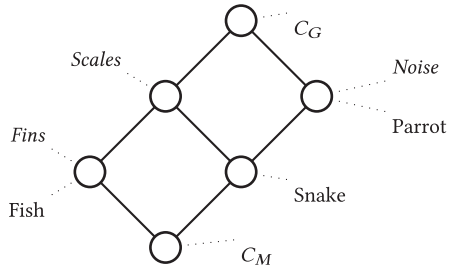

Fig. 2. Concept lattice for the context X_2 from Table 4.

are similarly disjoint. All six concepts are superconcepts of C_M. All six concepts are subconcepts of C_G. Drawing these dependencies according to the previously mentioned specifications, we obtain the concept lattice (Figure 2).

We label the objects on the concept lattice by first working from the bottom-up [59]. We label the concept with an object if it is the object's first appearance in the lattice is in that concept. Similarly, work top-down to label (in italics) the first occurrence of each attribute. When fully labeled, this allows us to work "up" the concept lattice to find an object, then see, by moving only to superconcepts of the concept where the object first appears, the attributes that object possesses. Similarly, if an attribute is found, moving only to subconcepts, we can see the objects that possess that attribute. In this way, we can see that fins are only possessed by fish in our simplified context, and that a parrot's only attribute is that it makes noise. Thus, the concept lattice allows for a more intuitive method of examining the data, rather than looking solely at the binary matrix.

2.2 Hierarchical Capability of FCA

With the concept lattice, FCA has natural hierarchical capabilities, which benefits various applications. For smaller datasets, visual inspection and interpretation of the lattice is possible, and relationships between concepts can be explored.

A seemingly natural application for the concept lattice is phylogeny, where the lattice may demonstrate an organism's evolutionary development. However, the task is not as simple. We will describe its complications in Section 3.11.

Unfortunately, biological datasets can be very large, so even calculating concepts and their relationships with FCA can be intractable. For this reason, FCA (or ideally, an advanced model of FCA) can be paired with a data mining or machine learning algorithm that reduces the dataset or performs initial analysis. Studies can use multiple algorithms, such as SVM and ANOVA, to identify important features, and then FCA analyzes the features and the algorithms that predicted them. The concept lattice is demonstrating the relationship between data mining algorithms and important features produced. It allows for a better understanding of what data mining algorithms prioritize in a dataset. This has been done in biomarkers discovery [70, 71] and healthcare informatics [120, 172].

Because the lattice is a type of Hasse diagram, it can also be used to visualize the structure-activity relationship, which visualizes molecular biological properties. Another chemistry application performs classification using the concept lattice [154], classifying uncertain amino acids. We describe these applications more in Section 3.6.

Even if the data is prohibitively complex for visualizing the concept lattice, examining the structure of and similarity between concepts can be useful. If concepts are ranked, the relationship between concepts can lead to revelations in healthcare informatics, discovery of influential nodes in protein-protein interaction data, and drug design and development applications. The concepts

calculated are frequently used to identify similarly expressed genes, in studies such as [99] and others addressed in Section 3.1. Relationships between concepts are also examined in some gene expression data applications such as [11].

2.3 Advanced Models of FCA

Although FCA was developed to use a binary context, most real-life applications may not use binary data. In bioinformatics, much of the data is numerical: gene expression data uses floating-point values to represent protein concentration for how genes react in different situations, and cancer image data may be represented in a matrix of color values.

Many advanced models of FCA have been developed to handle such data. We first highlight a few studies that use advanced models, then describe some of the most popular models in this section. Pattern structures are used to avoid information loss of binarizing gene expression data [98]. Fuzzy FCA is used to classify types of cancer in gene expression data [77] and detect adverse drug events [43]. A three-way fuzzy context is used for medical diagnoses [146]. Biclusters in gene expression data are identified with $\mathcal{K}$-FCA [67]. Compound fragment relationships are identified with FragFCA [110]. Rough set theory is used in conjunction with FCA to grade healthcare institutions [120], as well as inform medical diagnoses [156]. Finally, EFCA is used to extract sentiment analysis for healthcare informatics [6].

We now describe several popular advanced models that may be used in many bioinformatics applications: fuzzy FCA, probabilistic FCA, pattern structures, relational concept analysis, and monotone concepts. Advanced models that are specific to an application are described in the respective subsection of Section 3.

To motivate fuzzy FCA, we observe that different breeds of cats in the context $\mathbb{K}_1$ have differing amounts of fur. Some breeds (Turkish Angora, British Longhair) have long fur, and others (Sphynx, Peterbald) are hairless or almost so. Other cats (Siamese, Tabby) can fall between the two extremes. To represent this variation accurately in our context, one option is to expand the "Fur" attribute into three: Fur-Long, Fur-Short, and Fur-None. We may do a similar refinement process for the attribute "Noise," since each of the animals in $\mathbb{K}_1$ make different noises. We may expand Noise to Noise-Bark, Noise-Growl, Noise-Mew, Noise-Hiss, Noise-Squawk, and Noise-Quack. While this strategy may help clearly refine our small context and eliminate possible confusion of varying breeds, it is not necessarily scalable for the whole dataset. For contexts with an already high number of dimensions, this process could lead to an explosion of binary attributes, many of which are not relevant for objects that did not possess the attribute initially. This method of expanding attributes to suit the data is known as scaling and will be further addressed in Section 3.1.1. As an alternative, fuzzy FCA is sometimes used to accommodate variation in an object's possession of an attribute. A user-friendly explanation of Fuzzy FCA is detailed in [112], a small summary of which is given here. Instead of a binary value, fuzzy contexts' attribute values are real-valued numbers, usually in the range of $[0, 1]$, where the attribute value represents the degree of attribute in the object (e.g., amount of fur on a cat). A fuzzy subset of a context is made up of fuzzy object and attribute sets, each one specifying how relevant it is to specify the element it contains. This is often represented as a threshold to determine an object or attribute's inclusion. To compute concepts, a residuum, or fuzzy implication, determines whether the threshold is met for each value. From here the concept lattice, a residuated lattice, can be constructed. A recent survey of fuzzy FCA describes the historical development of this area, as well as recent application areas such as knowledge discovery in databases, data mining, information retrieval, and ontology engineering [119].

Another way FCA can be applied is in probabilistic applications [45, 90], the theory of which is based on rough sets. Unlike fuzzy sets, which involve the degree of an attribute manifesting in an object, rough sets are used to extract knowledge from missing or incomplete data by estimating the

missing or incomplete set's boundary regions. These regions are computed to attempt to determine the certainty of an object possessing a given attribute. Rough set theory can also help reduce data redundancy by treating objects that possess the same attributes as indiscernible and thus identical. For example, if the pair (Cat, Swims) was present in the context $\mathbb{K}_1$, the objects Cat and Dog would be indiscernible for the purposes of rough sets, and the objects could be merged to create a reduced dataset called the reduct. FCA can also be used with rough set theory to generate probabilistic association rules [160] and further discover maximal potentially useful rules in databases [47, 48].

To avoid the explosion of the number of attributes, pattern structures [58] may be used, which can capture information about numerical attributes without scaling. Pattern structures can also be connected with Fuzzy FCA [37]. A few authors have analyzed gene expression data with FCA and pattern structures [21, 78, 87, 92, 96–98]. This developing field is a promising alternative to scaling and should be further explored to assess its user-friendliness and scalability.

Relational concept analysis (RCA) is another advanced model based on FCA that uses decision logic and can handle relational data, such as graphs or relational databases. Recent advances include combining its techniques with fuzzy formal concept analysis [31] and discovering that concepts obtained through RCA are at least as expressive as ones obtained through FCA [161]. In bioinformatics, RCA has recently been used to maintain character and relationship data for a knowledge base Knomana to help determine which plants may be used as pesticides to protect crops in sub-Saharan French-speaking African countries [145]. With RCA, the authors perform conceptual classification to determine if pests or plants may keep another plant or pest away.

Additionally, there are models of FCA, such as using monotone concepts [49] that allow for a different kind of relationship between objects or attributes. Monotone concepts specifically use disjunctions among conjunctions of attributes, in addition to the traditional conjunction that current concepts employ. This allows for more information to be captured among the data. Such a method may be useful in ontologies or applications that use implication bases.

Use of these advanced models in analysis of bioinformatics data is becoming more popular and provides ease of use. The models described here provide significant prospects for analyzing bioinformatics data.

2.4 Lattice Management

Even if using an advanced model, one of the most common issues in formal concept analysis is the size of the concept lattice. When a dataset's size increases, so does the number of concepts, as well as memory constraints in computing concepts. Even if memory constraints are not an issue, the large number of concepts can lead to a cluttered structure, making data interpretation difficult. However, there are steps that can be taken to minimize such issues. While some of these steps are related to reducing the number of concepts already computed, some involve reducing the dataset before concepts are computed. These steps include algorithms to compute (select) only specific concepts, to simplify the context by removing redundant objects and/or attributes, or to simplify the concept lattice itself by eliminating concepts that capture information already represented in the lattice, a similar idea to the discussion of reducts in the previous section. A comprehensive survey of this topic was written in 2015 by [51], which addresses the categories of selection, removing redundant information, and simplifying the lattice. We mention here strategies of lattice reduction in these categories that have emerged after the survey's publication or that were not mentioned.

Selection of concepts is done while computing the initial list of all concepts. Algorithms to do so have been developed, with varying improvements on the naive algorithm. These include improved computation requirements, threshold of concept inclusion, and incremental construction of the concept lattice. ParallelGenerate and FCbO [103, 104] require no synchronization when computing concepts. In-Close2 [13] allows for an input of minimum support threshold for

a concept's object and attribute pair inclusion, which reduces the overall number of concepts and potentially allows more "significant" concepts to be included. AddIntent [158] constructs the concept lattice diagram incrementally, rather than computing all concepts and expensively building the diagram from scratch. The data mining algorithm LCM (Linear time, Closed itemset Miner) is shown to be equivalent to the CbO family of algorithms from the perspective of FCA [85]. LCM is compared to the CbO algorithm family, and it is determined that newer algorithms like FCbO and In-Close2+ may have additional benefits compared to LCM. Other improvements of In-Close include a preprint-proposed In-Close5, which uses vertical bit array storage during concept computation [171], and a high-performance parallelized algorithm FPCbO, which performs better than a similar parallelized algorithm PCbO [173]. A concept similarity measure using type-2 fuzzy sets and interordinal scaling is proposed [57]. The measure is compared with WordNet information concept similarity (*ics*) and Jaccard similarity to show that the proposed Asim similarity measure is close to a human judgment. The measure could be used by multi-valued FCA concepts to determine which concepts to select while computing the concept lattice.

The other main methods of selecting concepts include functional dependencies [93, 115], an expression of FCA's attribute implications. However, functional dependencies are often defined over numerical and categorical attributes, while in FCA, implications are more commonly defined over binary attributes. An approach for classical FCA was developed [109] and is quadratic for the number of objects in the context, and thus not scalable for large datasets. To overcome this limitation, authors in [21] present an approach for transformation by introducing partition pattern structures. These allow binarization to be applied on a many-valued formal context over attributes rather than objects. This results in reduced concept lattice size but with equivalent results that may be scalable for large datasets.

Methods to remove objects and attributes from the context include algorithms to change the specificity of binary data and to locally reduce attributes instead of globally, and theoretical methods to find the reduct of the context. The algorithms Fold [162] and Unfold [174] are used to increase the granularity (fineness) of attributes. This is a significant step in binarization of a multi-valued matrix to a formal context, but the resulting set of attributes impacts the total number of concepts, thus meriting inclusion in this discussion. Another method to reduce attributes is to use decision rules to determine local attribute reductions [134], to avoid loss of information locally. This method can help improve three-way classification. Other theoretic methods include [39], which proposes a method for constructing a simplified discernibility matrix that does not require computation of all formal concepts. Further, a fast heuristic algorithm is designed based on graph theory to greedily find the vertex cover corresponding to the attribute set, and so obtain the reduct of the formal decision context. Using the Min-Ex algorithm [128], intrinsically noisy data can be reduced by mining association rules, then extracting sets of objects that are more significant than the naive explosion of concepts through traditional means. This produces a so-called "fault-tolerant" FCA, with concepts bounded by a number of dimensions. One study [17] removes unnecessary attributes such that the meet-irreducible concepts in both the original lattice and the reduced lattice are equivalent. The infimum of elements is characterized so the reduction minimizes the number of unnecessary attributes. A concept stability measure, as introduced in [76], determines the probability that the intent will stay the same if certain objects are removed from the given concept. After deleting redundant attributes, the intensional concept stability does not change. The theorem may help verify that only redundant attributes are removed. Another attribute reduction technique is described in [170] (preprint). The algorithm improves similar attempts by using concept bit-array storage to handle "and" and "or" operations, resulting in intent reduction with minimal extent reduction.

To simplify the lattice by choosing only representative concepts from the total list of concepts, attribute clustering, theoretical extensions of FCA, and minimal generators are used.

Fuzzy attribute clustering [108] allows for more flexibility in multi-valued data, while clustering attributes with the Jaccard similarity [152] can be completed with binary data. Theoretical extensions of FCA can also assist in choosing representative concepts. Authors in [23] demonstrated that increasing uncertainty in formal fuzzy concept analysis leads to the concept lattice size being reduced. In their previous work [22], authors presented an attribute-oriented concept lattice where attributes were derived as positive and negative information from the given data. In [23] they show that the uncertainty is naturally modeled in Fuzzy FCA, which in turn leads to lattice size reduction. A fairly common approach is to identify the minimal combination of objects or attributes (called minimal generators) that distinguish the objects of one concept from other. In 2002, an incremental algorithm to mine minimal generators [129] was proposed. However, the minimal generators may still contain redundant attributes. Therefore, in 2005, a new algorithm was proposed to remove the redundancy [52]. A depth-first search approach is followed to build a depth-first search tree representing the enumeration of all subsets of the attribute set. Then, useless branches are removed from the tree. Their approach is experimentally validated on two real test datasets, UCI Mushroom and colon tumor gene expression [8]. The results show that the succinct approach can deal with high-dimensional and large real datasets. The interesting concepts in a lattice may be determined by a conceptual relevance index [82]. The index is determined by obtaining the minimal generators (attributes) for less redundant patterns. Compared to intensional stability measures, the conceptual relevance index appears to both perform faster (in polynomial space and time) and obtain more relevant information.

A program is developed to modify how the concept lattice is drawn so statistical information may be conveyed [102]. A cascading line diagram is proposed so a concept node's height corresponds to an extent's cardinality. The logarithm of the attribute support ensures the line diagram is readable with minimal horizontal lines. The resulting diagram permits both statistical and data mining measures (such as support, lift, and confidence) to be read easily.

Querying the lattice for a concept may be done with the method proposed in [117]. Natural language processing allows for a user to input partial or incorrect information and still obtain a relevant concept. Being able to query the concept lattice may be useful for bioinformatics applications. For example, one could query similarly expressed genes or pair a querying algorithm with an incremental lattice construction algorithm to allow for robust data management.

Apart from algorithmic improvements to make the concept lattice more manageable, scalable algorithms are very important for big data tasks such as analyzing bioinformatics data. Apache Spark, a distributed framework for machine learning algorithms, was used to develop one such scalable algorithm for determining FCA concepts and calculating the FCA concept lattice [41]. Apache Spark uses in-memory computation, cache memory, and canonical tests to calculate concepts and their relationships in an iterative manner without unnecessary repetition. The method is also fault tolerant. The method results in much faster computation times for large datasets.

A scalable algorithm for determining the implication base of a formal context has also been developed [42]. An implication base is similar to functional dependencies. Minimal generators are determined via a confidence metric to help build the implication base. Using the Apache Spark framework, execution time is greatly improved as opposed to algorithms not executed on distributive frameworks. While Apache Spark is optimal for this application, using similar distributive frameworks such as Hadoop [1] and MapReduce [44] can also provide significant computational improvement as opposed to solely algorithmic improvements.

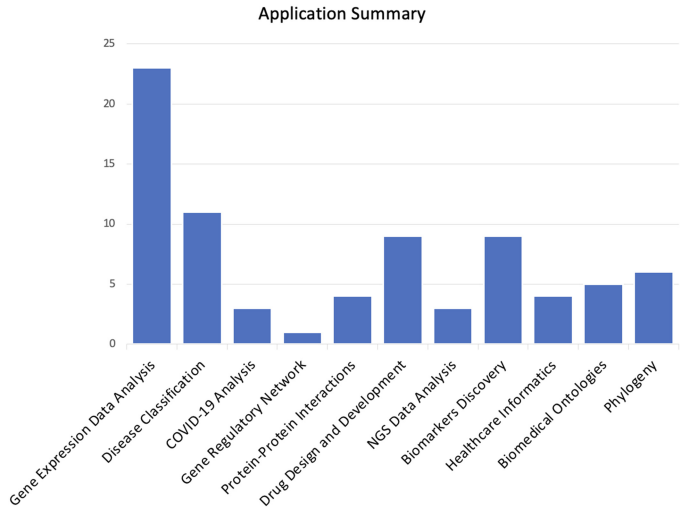

Fig. 3. A visual summary of the number of applications presented in Section 3.

Each of these methods can assist the user in constructing a manageable and interpretable concept lattice, thus aiding in analysis. This concludes our discussion of how FCA works and FCA-specific data management strategies.

3 BIOINFORMATICS APPLICATIONS

Formal concept analysis has been used as an approach for knowledge discovery in various bioinformatics fields. A comprehensive survey of the use of FCA for knowledge discovery was done by [130], and a further study focusing on the use of FCA for knowledge discovery in biological data was done by [136]. Our survey includes gene expression data analysis and expands on some topics addressed in the previous surveys. We also include several applications not addressed in the previous works. This section addresses the following applications in bioinformatics: gene expression data analysis, cancer classification, COVID-19-related analysis, gene regulatory network analysis, protein-protein interaction, drug design and development, next-generation sequencing, biomarkers discovery, healthcare informatics, biomedical ontologies, and phylogeny. The distribution of how many papers were found for these areas is in Figure 3.

3.1 Analyzing Gene Expression Data with Formal Concept Analysis

We describe general applications of FCA for gene expression data here. If the gene expression data in a study is used for a more specific area of bioinformatics, such as cancer classification, biomarkers discovery, or the gene regulatory network, we describe the study in a later section.

The structure of FCA, specifically how concepts are computed, lends itself to computing groups of genes that are expressed similarly. We now examine how gene expression data can be discretized to work with classical FCA.

3.1.1 Discretization of Gene Expression Matrices. Before examining the applications in detail, how the gene expression data is discretized must be discussed. Several common techniques are used to discretize gene expression data. Each study's technique will be discussed in Section 3.1.2, but we briefly explain them here. A summary of these techniques is in Table 5.

Table 5. Summary of Discretization Methods Used for Gene Expression Data

Study	Year	Discretization Method
[30, 128, 139]	2003, 2005, 2007	1 if expressed past threshold, 0 if not
[28]	2005	Strong expression (positive or negative) registered
[40, 56, 99, 123, 131]	2005, 2008, 2019	Interordinal scaling
[147, 148]	2006, 2007	Rough set theory permutations
[99]	2008	Conceptual scaling
[96]	2009	Interordinal scaling with pattern structures
[81]	2014	Binarize via cluster membership
[79, 80, 157]	2016, 2018	3-valued matrix discretization

Table 6. A Sample Gene Expression Matrix, with the Original Values in Two Situations s_1 and s_2 on the Left

	s_1	s_2	$s_1, \geq 1$	$s_1, \geq 2$	$s_1, \geq 3$	$s_1, \geq 4$	$s_2, \geq 0.5$	$s_2, \geq 0.75$	$s_2, \geq 1$
g_1	5.3	0.7	X	X	X	X	X		
g_2	3.36	1.2	X	X	X		X	X	X
g_3	2.5	0.25	X	X					
g_4	0.43	0.96					X	X	

The matrix discretized with interordinal scaling is on the right.

The easiest method of binarization is measuring whether a gene is **expressed or not (E/NE)**; this is determined by assigning 1 if the value exceeds a threshold for each gene [139]. This method is also used to capture whether a gene is overexpressed in a situation [30, 128]. Strong expression, whether positive or negative, can be captured similarly [28]. A variant of this is first transforming the matrix to a 3-valued matrix (often $-1, 0, 1$) based on each row's value varying from the average. From there, the matrix is discretized to a binary matrix by doing a similar computation [79, 80, 157].

Some of these methods, especially overexpression or E/NE, risk losing the strength of expression by setting a threshold. To retain expression strength, interordinal scaling is used [99, 131]. In the algorithm MICROBLAST, Potter [131] uses interordinal scaling to discretize the gene expression matrix, which is done by comparing the expression level of each gene to an inequality. If the expression level does fulfill an inequality, it is said to possess that attribute. An example of this transformation is shown in Table 6. One potential issue is that if an attribute exceeds a certain threshold, then the gene will possess all the relevant attributes, not just the closest one. This issue is addressed in [99].

The intervals chosen in interordinal scaling correspond to a scale proposed by [40, 123]. Measuring the similarity of concepts discretized with interordinal scaling was done by [56]. Interordinal scaling can also be combined with pattern structures [96].

The effects of various binarization techniques on discovered positively correlated expression groups are explored by [96, 97, 99]. In the first study [99], a binarization method called *conceptual scaling*, which converts the value of each expression level to an interval, is proposed. If an expression level does correspond to an interval, it is said to possess that attribute. This is a better method than Potter's interordinal scaling, because if two expression levels fulfill a certain inequality, there is a better chance that attributes do not overlap. An example of the original dataset (the left part of Table 6) converted to binary data with conceptual scaling is shown in Table 7. A potential downside to this method is that the number and size of intervals should be determined by an expert to avoid losing too much precision.

Table 7. Gene Expression Matrix Discretized with Conceptual Scaling Proposed by Kaytoue-Uberall et al. [99]

	$s_1, (4, \infty)$	$s_1, (3, 3.99)$	$s_1, (2, 2, 99)$	$s_1, (0, 0.99)$	$s_2, (1, 1.99)$	$s_2, (0.5, 0.99)$	$s_2, (0, 0.499)$
g_1	X					X	
g_2		X			X		
g_3			X				X
g_4				X		X	

Table 8. Summary of Papers That Analyze Gene Expression Data

Study	Year	Purpose
[139]	2003	Examine Galois concepts
[28]	2005	Propose D-Miner algorithm; find transcription factors
[131]	2005	First to use FCA with microarray data
[30]	2007	Identify positively correlated expression groups
[40]	2008	Implement [131], test multiple experiments of mouse lung tissue data
[96]	2009	Investigate effects of discretization
[97]	2011	Extract concepts with pattern structures
[13, 15, 16]	2011, 2013, 2018	Calculate concepts in developing mouse embryo tissue data
[81]	2014	Combine multiple experiment data
[67]	2016	Determine over- or under-expression; interface with gene ontology
[157]	2016	Identify negatively correlated expressed genes in time series data
[79]	2018	Cluster positively correlated expressed genes with BiFCA+
[80]	2018	Identify negative biclusters
[84]	2021	Discover and eliminate missing genotypes in ischemic stroke data due to machine error
[32]	2021	Cluster with evolutionary algorithms using multidimensional formal concepts

In the second study, the effects of binarization on algorithms Norris, CbO, and NextClosure are examined. In particular, two methods of adapting real-valued data to a form FCA can use are inspected. These two methods, interordinal scaling and pattern structures, are used to reduce the size and complexity of the concept lattice in experiments with *Laccaria bicolor*, a fungus that is found on tree roots [96]. With data in a gene expression matrix, the numerical data is represented as attributes in terms of scales. For example, if a gene g_i has value w_j in situation s_k, the attribute for g_i is represented as the one-valued attribute $s_k \leq w_j$. This is done for all values each gene may have in any situation, which accordingly leads to a very large context. To restrict attributes, the inequalities may be restricted (i.e., not included in the context) to satisfy some maximum value. As this method can result in an explosion of attributes, interval pattern structures are also explored as an alternative to binarization, reducing the risk of data loss.

Another way to binarize is to use cluster membership when clusters of genes are computed as a preprocessing step [81]. Groups of genes can also be discretized using rough set theory, where all permutations of objects are paired together [147, 148]. The attribute for the new paired object is measured by how the second attribute's value compares to (by $<$ or $\geq$) the first. Classification results using this method are reasonable. However, this method results in quadratic complexity, as discussed in [21], so using a similar process with pattern structures is proposed instead, as it is more scalable and feasible for lattice construction. However, this method has been used to analyze the effect of discretization, and not on the biological impact of such calculations.

Whatever method of discretization is used, it should be directly related to the purpose of the study. As discretization has inherent data loss, care must be taken as to which method is used.

3.1.2 FCA Analysis of Gene Expression Data. We now address how gene expression data is analyzed with FCA. A summary of the papers is found in Table 8.

Two authors initially examine the feasibility of clustering gene expression data with similar methods to FCA. In a 2003 study, [139] examines Galois concepts in discovering co-regulated genes. Results of different discretization methods are also examined, which shows insight into its importance. A 2004 survey explored the feasibility of clustering gene expression data [87]. Although FCA is not explicitly mentioned, the survey describes how clustering (including biclustering, a similar method to FCA) can be used to gain information from a microarray gene expression matrix. In particular, clustering methods specifically used for gene expression data as well as class validation and the reliability of produced clusters are discussed. Other early attempts to perform gene expression analysis with FCA include finding transcription factors, which are proteins that regulate whether a gene is expressed or not. Using the algorithm D-Miner, formal concepts with Galois operators are mined [28]. The concepts then reveal the transcription factors.

Potter, in his doctoral thesis, is one of the first to use FCA with microarray data [131]. The flawed version of interordinal scaling is used. Potter then creates the microBLAST lattice from the biological lattice, which is simply the concept lattice of the discretized biological data. No mention is made of correlated groups of genes or other specific applications to gene expression data.

Potter's seminal work is then implemented in Choi et al. [40]. The authors approach the problem of binarization in a method similar to interordinal scaling, by dividing the number line according to the "gaps" in expressed numerical values. Then the intervals are recursively partitioned into subintervals until a desired partitioning is reached. This allows any given gene in the context to possess the one-valued attribute of being in its interval. Concept lattices are constructed for each separate experiment and then compared according to a defined distance. The distance Choi et al. propose is a measure of which genes each vertex (concept) in the lattice shares with another. This allows for comparison and analysis of different experiments with mouse lung tissue gene expression data. Using multiple experiment data is further explored in [81], which proposes that multiple experiments can be beneficial to analysis because the risk of biases in individual experiments is reduced, and combining data can result in a higher confidence of results. Binarization is done by first partitioning and then consensus clustering the genes. The context is then constructed with genes as the objects, and attributes are whether each gene belongs to each cluster. Analysis is done with a time seriesdata set of fission yeast, and the biological significance of these results is discussed. Ischemic stroke data of patients versus **single nucleotide polymorphisms (SNPs)** is biclustered using FCA [84]. The data was gathered from multiple microarray experiments. The biclusters then reveal whether a patient is missing an SNP value. This allows for the discovery and elimination of missing genotypes due to machine error.

Identifying the positively correlated expression groups of a dataset is the explicit purpose of [30]. Discretization is done by overexpression: if a gene's expression level in some situation is over a set threshold, the relationship is 1, and 0 otherwise. The extraction of concepts is done with D-Miner [28], an algorithm designed to extract transcription factors, which works through local pattern discovery. The intent of these extracted concepts is then examined to find the positively correlated expression groups.

After investigating the effects of discretization, [96] performs experiments on a tree fungus dataset. Results show the role and function of genes that are expressed similarly. Concept extraction with interval pattern structures is further explored in [97]. By modifying FCA to allow for pattern structures, concept extraction from such a context can show each gene's positively correlated expression.

In order to select biclusters of genes that work together to perform biological functions, a variation on FCA, $\mathcal{K}$-FCA, is developed [67], along with a web tool WebGeneKFCA [55] that provides a variety of visualization techniques to aid the user in determining gene under- or over-expression,

as well as provide an interface with gene ontology. $\mathcal{K}$-FCA provides an analysis method similar to Fuzzy FCA that reduces the need to binarize the entire dataset. Lattice reduction is also addressed; a proposed solution is to "filter" concepts by judging if an intent contains too many or too few genes, according to some threshold that is again determined by an expert. This allows the lattice to be somewhat simplified and allows the "critical" concepts to be computed and displayed.

An algorithm BiFCA+ was developed [79] to cluster gene expression data into biclusters based on FCA, which corresponds to groups of positively correlated expressed genes. The algorithm is based on biclustering data that is binarized to a three-state matrix based on expression variance in conditions before transforming into a two-state matrix based on the average value per gene. The set of biclusters is reduced by the Bond correlation if there is significant overlap between samples.

Negative correlation between genes is examined [80, 157]. In [157], algorithm NCFCA finds negatively correlated genes in cell cycle time series data. The matrix is discretized by transforming the expression value to 1, 0, or −1 if the value in the time series data increased, did not change, or decreased, respectively. Then it is binarized based on the row's average value deviation. The concept lattice is filtered by limiting the minimum number of genes in each subset.

Another algorithm, NBF, was proposed to find negative correlations, which the author refers to by negative biclusters [80]. Discretization is similar to [157], except when going from the 3-state to the 2-state matrix, two binary matrices are produced, one representing a positive average change per row, and another representing a negative change. Experimental results are obtained from yeast cell-cycle, human B-cell lymphoma, and Alzheimer gene expression databases.

In addition to determining positively or negatively correlated groups of gene expression, FCA helps identify genes that perform biological functions. Using developing mouse embryo tissue gene expression data like [40], the makeup of genes in components of tissues is found [13, 15, 16]. The algorithm In-Close 2 was developed [13] and tested on the mouse tissue data to see which concepts were detected. Binarization in all three studies was done by allowing any expression level in a gene. Mouse tissue components were visualized with the acquired knowledge of the genes expressed in each component. These studies could be further explored with a method of binarization that does not reduce the strength of each gene's expression.

Multi-dimensional formal contexts are used to perform multimodal clustering with evolutionary algorithms [32]. Approaches of multidimensional FCA are presented for gene expression data encoding and genetic clustering. The evolutionary algorithms are tested on a myocardial infarction dataset.

This concludes our survey of FCA applications using gene expression data.

3.2 Disease Classification with FCA

Disease diagnosis and prediction are important areas of bioinformatics. In particular, FCA has been used to classify (diagnose) or predict the presence of various diseases. These diseases include renal diseases, heart disease, tuberculosis, diabetes, and different types of cancer. Section 3.2.1 details how FCA is used to classify various types of diseases, followed by Section 3.2.2, which describes how FCA has been used to classify cancer data. The papers discussed in this section are summarized in Table 9.

3.2.1 Classification of Various Diseases. There have been several studies that perform classification or prediction on various types of disease.

Testing for disease similarity has been done using FCA [100]. FCA is formulated as a graph. Subgraphs and spanning trees are used to discover concepts. Data of up- or down-regulated genes in renal disease biopsy tissue is analyzed. Inspecting the concept lattice determines relationships between renal diseases, and disease dependence can be inferred.

Table 9. Summary of Papers in the Disease Classification Application

Study	Year	Purpose
[156]	2011	Diagnose heart disease from patient historical health data
[100]	2012	Test for disease similarity
[19]	2014	Interpretation of brain tumor images; FCA interfaces with descriptive logic
[155]	2017	Improve DNC classification algorithm; feature selection feeds into FCA classifier
[146]	2018	Diagnose multiple diseases from medical data with a three-way fuzzy context
[140]	2018	Classify histopathological images; FCA is used in a four-step sampling method
[151]	2019	Predict tuberculosis from symptoms with CUR decomposition context reduction
[142]	2019	Classify diabetes dataset; FCA adds explainability to deep learning model
[77]	2020	Classify type of cancer using fuzzy FCA method
[53, 54]	2021	Transform decision trees into a concept lattice; tested on cancer, heart disease, mammographic, and diabetes datasets

FCA is used as a framework to diagnose whether a patient has heart disease [156]. Suitable decision rules are mined with the help of rough set theory. The dataset used is historical health data of a patient's record of chest pain, blood pressure, cholesterol, blood sugar, electrocardiography, and maximum heart rate. Classification is performed to determine whether a patient has hypertensive heart disease.

A three-way fuzzy context is used to diagnose medical data [146]. Similar concepts can also be identified with a Euclidian distance measure. Using a neutrosophic graph representation of the concept lattice, the diseases viral fever, malaria, typhoid, gastritis, and stenocardia are diagnosed.

CUR decomposition helps with formal context reduction [151]. The method is tested on a tuberculosis symptoms dataset. Predictions based on decision attributes are performed to determine if the disease is present based on symptoms.

Machine learning (ML) models are developed to perform classification tasks in many domains. The ML models are often considered black boxes because they hide the internal logic from the users. As a result, they are often not used in real-world applications. Authors in [142] attempt to overcome this limitation by using a formal concept analysis framework to explain the logic behind the classification performed by the ML models. The proposed method builds a concept lattice for each value of the target class variable. The new instances are classified using subset matching between the concept lattice's intents and the concept lattice formed by the new instance. If a matching intent is found in any lattice, the new instance is assigned the corresponding class; otherwise, it is given a label of "not determined." This helps the user find the attributes responsible for the classification of an instance into a given class. The approach is tested on a diabetes dataset with 11 features and 690 samples. The results show that on the diabetes dataset, the FCA model performs better than the ML model. The results also suggest that for larger datasets, FCA might *not* perform better than the ML model. However, this research's goal is not to perform better than the ML model, but instead add explainability to the classification task performed by ML models. The proposed approach is domain independent and can also be applied to problems other than classification. A slightly similar method to [142] (though one that does not explicitly use FCA) proposes definitions and methods to determine the most predictive and discriminative features in an instance of supervised classification [24]. Using a Pareto front (which uses partial order definitions), a lattice structure shows the predictive and discriminant features in a diabetes database.

3.2.2 Cancer Classification. Cancer is a disease in which cells grow uncontrollably [5]. If untreated, cancerous cells may become metastatic and grow beyond their areas to other parts of the body. Uncontrolled cell growth can happen for multiple reasons, such as failure to receive signals for the cell to die, growth despite having signals to not grow, and tricking or evading the immune system, which is supposed to stop the growth of cancerous cells. There may also be cases of abnormal cell growth, but if the cells do not grow uncontrollably, the growth cells may be called benign. Certain types of cancer may be genetic, meaning that multiple people in the same family may get the same disease. It is estimated that almost 40% of people in the United States will be diagnosed with cancer during their lifetimes [4]. Because of this, it is important to assess ways to lower the incidence of the disease, diagnose it early, or determine ways to stop or remove the growth of cells once cancer is present.

Computer science is best poised to address the diagnosis problem. Research typically takes the form of cancer classification, which takes cancer data as input and classifies a portion of the dataset, typically as cancerous vs. benign, or cancerous vs. normal cell. This has been the focus of grand challenges such as BACH to perform analysis using machine learning and deep learning on breast cancer images [18]. Now the various ways FCA performs classification on cancer datasets are described.

An image interpretation framework based on descriptive logic for knowledge representation and an FCA framework for reasoning is introduced in [19]. An image is viewed as an observation, and an interpretation is the best explanation based on the background knowledge about the image's context. A background knowledge base represented in descriptive logic is constructed using expert knowledge of the domain and the spatial relations in the image, extracted using image processing algorithms. The interpretation is obtained by applying two reasoning operators on the concept lattice of background knowledge. The framework is tested on pathological brain tumor image interpretation. The prior pathological knowledge on the brain and the spatial information on the brain image is formalized using descriptive logic language, and then the presence of a tumor in the image is identified by the lattice reasoning. This approach interfaces descriptive logic with FCA to solve a non-standard inference problem in descriptive logic.

The classification algorithm DNC was improved [155] by extracting nominal attributes. The study presents a feature selection method to improve performance of a FCA-based classifier. A new feature selection method based on the Shannon entropy is proposed, which improves classification results. Tests are done on a UCI breast cancer dataset with 0% error rate. The study shows that performing feature selection prior to FCA classification provides better generalization and specialization relationships in the data. The prior feature selection also removes redundancy.

Breast cancer is a leading cause of death in women all over the world [7]. Automating cancer image classification can save millions of lives, and much effort has been made in this domain. Authors in [140] present a breast cancer histopathological image classifier in which each pixel in an image is classified as cancerous or benign. This helps to identify the cancerous regions in an image. However, this pixel classification is time-consuming for large datasets. To overcome this limitation, authors developed a unique four-step sampling method based on formal concept analysis to improve the efficiency of a machine learning classifier on the breast cancer image data: (1) A formal context is generated using a similarity measure on the pairwise objects in the original dataset. (2) An object-pattern table is generated based on the formal context. The pattern is the set of all possible combinations of attributes in the context. The objects from the original data may be mapped to one or more patterns due to pairwise comparison in step 1. (3) A hypercontext reduction algorithm is applied to the object-pattern table to find the most representative patterns. (4) A conceptual sampling algorithm is applied to find the objects with the maximum total

Table 10. Description of Papers in the COVID-19 Application

Study	Year	Purpose
[34]	2021 (preprint)	Perform knowledge discovery on phase 3 or higher COVID-19 vaccines
[74]	2021	Determine patients with similar symptoms
[165]	2021	Identify close contacts of a person infected with COVID-19

number of patterns. The obtained sample is used by the machine learning algorithm to perform classification. The approach is experimentally validated on a breast cancer tissue dataset and the results show that the performance was improved in terms of sample size and quality.

FCA was used to identify patients' type of cancer (the options being breast cancer, kidney renal clear cell carcinoma, colon adenocarcinoma, lung adenocarcinoma, and prostate adenocarcinoma) from gene expression values [77]. FCA is one of the four data mining methods used to determine the results, using a fuzzy FCA classification method from [69].

An ensemble of decision trees can be transformed into a concept lattice to perform classification [54]. The decision tree premises and their classification target have a partial order the same as the FCA concept lattice, so a decision lattice can be built out of the existing classification rules. The experiments are tested on a variety of health and bioinformatics classification datasets, including breast cancer, heart disease, and diabetes. The decision semilattice is further developed in [53]. An ensemble of decision trees is transformed into a decision semilattice, which has the same predictions.

3.3 COVID-19 Analysis

With the recent advent of COVID-19, FCA has been used to analyze various aspects of the disease and its treatments. The summary of papers in this section is found in Table 10.

In a preprint [34], knowledge representation is done on COVID-19 vaccines that are phase 3 or higher. Attributes include the vaccine type (such as mRNA, inactivated virus, DNA, or protein subunit), the dosing schedule, the drug's safety, and its immunogenicity. The information is obtained for 19 vaccine candidates and is stored in a binary table. The main attribute categories are split into various binary sub-attributes. Multiple concept lattices are constructed for analysis. Attribute implications for the concept lattices identify important common attributes such as reactogenicity. Attributes of similar vaccines can also be examined. The dual vaccine concept lattice, with attributes and objects switched, shows the affinity of the various phase 3 vaccines.

Another application of knowledge discovery for COVID-19 is done by [74]. Knowledge discovery with FCA is performed on patient symptoms from COVID-19. The context is made up of patients (objects) and symptoms (attributes). The concept lattice calculation algorithm is incremental, allowing for new records to be added to the context. Once the concept lattice is obtained, the concepts may be queried using SQL to find patients who have the same symptoms.

FCA is also used to detect which close contacts of a person infected with COVID-19 should quarantine and seek testing [165]. FCA contains graph information in an adjacency matrix representation. γ-quasi cliques are described and used, which allow for a less strict definition of cliques for the close contact problem. In an experiment, the graph is constructed from 1 day of GPS data from a town in the United Kingdom. One user is chosen to have COVID-19 results, and a list of people in their network who should get tested and quarantine is constructed.

3.4 Gene Regulatory Network Analysis

The **gene regulatory network (GRN)** is a group of genes that interact to perform and regulate cell function. The main contribution to the GRN application was done by Gebert et al. in [61], summarized in Table 11. Unknown gene members of the regulatory network are detected.

Table 11. Summary of Paper Examining Gene Regulatory Network Analysis

Study	Year	Purpose
[61]	2008	Identify seed genes with template matching in FCA

Table 12. Summary of Papers in the Protein-protein Interactions Application

Study	Year	Purpose
[101]	2014	Predict gold standard domain-domain interactions from protein-protein interaction data
[153]	2017	Detect influential nodes by adding a weight calculation to attributes
[75]	2017	Evaluate similarity between molecular structures
[83]	2020	Identify significant objects from a graph's adjacency matrix

The proposed method is influential because similarly regulated genes do not necessarily appear in the same graph cluster. Gene interaction data is organized into a graph, with a subset of genes designated as seed genes. An edge between two genes represents any interaction between them. After the graph is constructed, the Kendall correlation is used to determine gene template matching. A formal concept represents whether a gene matches a given template, and the concept lattice is calculated. Concepts with either sub- or superconcept relation to those having seed genes are of interest. The genes in the sub- or superconcepts are considered candidates for the regulatory network. The candidate list is condensed with the shifted noise-robust Kendall correlation coefficient. As a result, genes that strongly interact with the seed genes are identified. In testing the method, a tuberculosis dataset is used to repair an SOS gene network.

3.5 Protein-protein Interactions

Studying protein-protein interactions can convey valuable information for drug design. A summary of the papers in this area is in Table 12.

In 2014, one of the first papers to use FCA to analyze **protein-protein interaction (PPI)** data was [101]. PPI data is used to predict gold standard domain-domain interactions. The formal context is represented as the protein IDs as objects and the domains as attributes. Once formal concepts and the lattice are calculated, promiscuous domains are found near the top of the lattice. The concepts are scored, and the domain interactions inferred from the protein interactions are shown to be reliable when domains are promiscuous.

Discovering influential nodes in a graph or network can lead to identifying essential proteins in protein-protein interactions [68]. Influential nodes can be detected using FCA by adding a weight calculation to each attribute [153]. After representing a graph's adjacency matrix as a formal concept, the weight measure is calculated in two steps. First, determine which concepts have the given attribute. Then, sum each concept's ratio of the number of objects over the number of attributes. Once each attribute has a weight, the attributes can be ranked, where a high weight signifies an influential node. In a similar method, significant objects may be identified using FCA [83]. A graph's adjacency matrix is the formal context. This method results in formal concepts that correspond to cliques in the graph. The cross-face centrality is calculated for the nodes in the graph, which identifies the influential vertices.

Another method for identifying interactions with FCA is evaluating similarity between graphs [75]. The process can be used to identify similar molecular structures in drugs. A graph is represented as a modified adjacency matrix in FCA, and the concept lattice for each graph is built. Then the similarity of concept lattices for different graphs is computed.

Table 13. Summary of Papers in the Drug Design and Development Application

Study	Year	Purpose
[110]	2008	Combine fragments to identify diverse molecules with FCA concepts
[138]	2011	Compare 15 structure-activity relationship models with FCA
[141]	2013	Identify correlating mutagenic and carcinogenic compounds, and non-mutagenic and non-carcinogenic compounds
[118]	2018	Visualize structure-activity relationship with Hasse diagram
[91]	2019	Summarize FCA's use in knowledge discovery of databases; apply with antibacterial drug discovery pattern mining
[137]	2020	Ontologically identify chemical element
[154]	2020	Classify uncertain hydrophobic-polar amino acids
[43]	2021	Detect adverse drug effects with social media analysis
[33]	2021	Identify medieval remedies and ingredients

3.6 Drug Design and Development

Drug design is a useful area of bioinformatics, as the development of new drugs can help people suffering from various conditions. FCA has been used in drug design processes by visualizing or representing the structure-activity relationship, identifying relationships in chemical compounds, and identifying molecular compounds and elements. These methods are described in the following paragraphs, and summarized in Table 13.

FCA has many potential applications in chemistry, which are summarized in a 2017 literature review [135]. One application is structure-activity relationships. Structure-activity relationships are valuable in drug design, and are used for discovering the biological properties in molecules. FCA compares to this relationship in several ways. FCA was compared to 15 quantitative structure-activity relationship models to find concepts from a set of 95 amines [138]. The study found that FCA and the concept lattice may be very useful for this task. The structure-activity relationship may be visualized in a Hasse diagram [118], the same kind of diagram used by the FCA concept lattice. This indicates that the FCA concept lattice may enable further analysis of structure-activity relationships.

A 2019 thesis [91] describes how FCA can be used for knowledge discovery in databases and how FCA can cluster complex data. An application of these methods is shown by using pattern mining for antibacterial drug discovery.

We now describe various studies that use FCA to identify compounds and elements or classify molecules. In 2008, a variation of FCA, FragFCA, was developed to examine the relationship between fragments found in active compounds [110]. The signature fragments can be identified, and the fragments can be searched interactively. Formal concepts containing fragments result in combinations to help identify diverse molecules. A similar study compares molecular compound selectivity and extract compounds with similar selectivity compared to existing drugs or drug candidates [111]. FCA can be used to identify correlating mutagenic and carcinogenic compounds, as well as non-mutagenic and non-carcinogenic compounds [141]. A chemical element can be ontologically identified by using concepts from FCA, by using chemical attribute relations and the atomic number [137]. Amino acids are classified in [154], which visualizes the concept lattice for different categories of amino acids and uses the lattice to classify uncertain hydrophobic-polar amino acids based on existing concepts. The understanding of protein folding can be aided by such a study.

Detecting adverse drug effects on social media and comparing them with documented side effects in a reputable database is done by [43]. Tweets about three drugs and their side effects are

Table 14. Summary of Papers Performing NGS Data Analysis with FCA

Study	Year	Purpose
[116]	2020	Compare granular counting method for multi-reads to FCA
[143]	2021	Identify biclusters of bacterial strains from long read data
[38]	2021	Assist brain storm optimization algorithm with clustering

chosen, as are academic papers about the drugs' side effects. After text mining papers and tweets, fuzzy formal concept analysis is used to keep track of two contexts. One context is whether the symptoms discussed in the tweet are reliably described according to the drug effects database. The other is whether a paper mentions the side effects. Fuzzy FCA is used in this case for correlation evaluation. What could be expanded upon is how the FCA implementation is used. It is well known that with a large dataset like the one used in the study (20,000 tweets and 4,000 papers), FCA may run very slowly or encounter memory constraints. A more detailed explanation of how the fuzzy FCA model creates the correlation is recommended.

An interdisciplinary study on medieval pharmacy remedies has been conducted [33]. Remedies and their ingredients are obtained from manuscripts and translated, and the remedies (objects) and their ingredients (attributes) form a context. Remedies versus ingredients from plants also form a context. The goals are to discover which ingredients are used most frequently, and which ingredients are frequently used together. The concepts with the largest extents are calculated to discover which ingredients appear most frequently. For the second research question, association rules are determined from concepts in the lattice. The dataset used is small due to the effort of determining and translating remedies and ingredients from primary texts. However, this type of application may be uniquely suited for FCA, as it is a relatively small dataset that reduces the likelihood of memory concerns.

3.7 NGS Data Analysis

Next-generation sequencing (NGS) allows for very fast genome sequencing [25]. FCA has been mentioned or used in three recent NGS studies, which are described here and summarized in Table 14.

Granular counting of uncertain data can help count multi-reads in RNA-Seq output [116]. While the study does not use FCA, the method is compared to FCA. The probabilistic method proposed is more complex than using FCA with binary data, but an advanced probabilistic model of FCA may allow a similar method to be performed on RNA-Seq data.

FCA is used in an NGS study that identifies bacterial strains from long read data [143]. Formal concept analysis is used to identify biclusters of the strains and their genes. The biclusters are then put into a classification tree to determine the closeness of the strains. Brain storm optimization, a type of algorithm aimed at imitating swarm intelligence, has been improved with the addition of FCA [38]. The HBSO algorithm attempts to perform the k-means algorithm. FCA assists HBSO by clustering individuals already determined to be part of a population. The method helps to rectify one of the original algorithm's shortcomings. Using brain storm optimization algorithms is optimal for multiple sequence alignment compared to the genetic algorithm [86], and has potential for the NGS data analysis application.

3.8 Biomarkers Discovery

Biomarkers are features in a database that may predict whether an individual will be healthy or develop a disease in the future [150]. Here we describe how biclustering with FCA has been used

Table 15. Summary of Papers in the Biomarkers Discovery Application

Study	Year	Purpose
[121]	2008	Distinguish genes involved in tumor-based or metastatic breast cancer; uses microarray data
[10–12]	2012, 2013, 2015	Identify breast cancer biomarker candidates from hyper- or hypomethelated genes
[107]	2015	Identify genes expressed in lung cancer tumors
[70, 71]	2016, 2019	Identify predictive biomarkers in metabolomic data; FCA validates knowledge discovery
[73]	2020	Discover biomarkers with Knowledge Extraction and Management software
[113]	2021 (preprint)	Discover significant biclusters in lung adenocarcinoma PSG genes

to identify biomarkers in genomic and metabolic data. A summary of the papers can be found in Table 15.

Breast cancer biomarkers have been discovered using FCA classification in microarray data [121]. Specifically, it is determined which genes are associated with tumor-based breast cancer and which are associated with metastatic breast cancer. Discretization is done by setting a threshold for each gene. One attribute represents the experimental value being below or equal to the threshold, and another represents the value being above. Noise is reduced by a heuristic that sets a minimum and maximum number of genes allowed in an intent. A training and validation context is constructed to identify and verify the biomarkers, respectively. FCA performs four points better than a MATLAB decision tree algorithm. While FCA's accuracy is only 85%, the method identifies 29 valid biomarkers, using an average of 3.8 genes.

A series of studies identify breast cancer biomarker candidates from hypermethelated or hypomethelated genes [10–12]. FCA finds genes that have inhibited [12] or uninhibited [10] expression in breast cancer subtypes. Objects are the inhibited genes. Attributes are whether that gene corresponds to a breast cancer subtype. The concepts reveal which groups of inhibited genes correspond to cancer subtypes. The genes found as a result of these studies were then verified in the lab [94], so a similar approach may determine other biologically significant results.

FCA is used for knowledge discovery in lung cancer tumors [107]. After feature selection is done by a T-test, FCA is run on a dataset with both tumor and normal tissue samples. Then, clusters in the lattice diagram are analyzed. All tumor samples are contained in one cluster, indicating FCA is beneficial for this purpose. The attributes in the cluster are then examined, and the genes involved in each attribute are discussed as to how they may be affecting the tissue sample. LYVE1 is identified as a gene whose abnormal expression may be targeted in treatment.

Two studies identify predictive biomarkers by using FCA for knowledge discovery in metabolomic data [70, 71]. RF, ANOVA, and SVM are used to determine biological patterns and select the most predictive features. The classifiers and predictive features are represented using a cross table (formal context). Visualization of the context is done, and interpretation sees the role of each classifier. The method shows how the different classification processes may decide on overlapping or independent features, and so FCA serves as a validation step in the process of knowledge discovery.

Biomarkers discovery is completed with FCA with the Knowledge Extraction and Management software [73]. Genomic data is collected from patients in the clinical trial of a drug aiming to combat mild to moderate Alzheimer's disease. FCA concepts represent association rules, and support, confidence, and lift metrics are calculated. The biomarkers discovered may help identify patients that can be targeted with a specific therapy to treat the disease.

Significant biclusters are discovered with FCA in lung adenocarcinoma PSG genes in a 2021 preprint [113]. The significant biclusters identified imply that the PSG+ genes are statistically significant. A Python package, COINCIDENCETEST, is developed as well to determine the significance of binary clusters for similar tasks.

Table 16. Summary of Papers in the Healthcare Informatics Application

Study	Year	Purpose
[36]	2013	Analyze patients' sequence data for diagnostic information and cancer care assumptions
[172]	2020	Classify risk of there being an epidemic
[120]	2021	Explore attributes to reverse-engineer hospital ranking
[6]	2021	Judge healthcare quality with sentiment analysis on patients' social media

3.9 Healthcare Informatics

FCA can provide useful information in healthcare informatics settings by performing analysis of sequence data or attribute exploration. Papers addressing these topics are summarized in Table 16.

Cancer sequence data describing patients' hospitalizations, reason for hospitalizations, and any procedures underwent is analyzed in [36]. A new method of analyzing sequence data is proposed. Pattern structures and their projections are used to analyze assumptions about cancer care as well as diagnostic information.

A hybrid classification method was developed using SVM, K-Means, and FCA to classify whether there is risk of an epidemic [172]. Data from five recent epidemics (SARS, MERS, Ebola, H1N1, and H5N1) is used to test the algorithm. FCA's role in this application is feature extraction and data reduction by grouping objects in one concept together.

FCA is used for attribute exploration in reverse-engineering hospital ranking information [120]. Given scoring data for a hospital, rough set theory may be used to find equivalence classes among the data. After the LEM2 algorithm produces decision rules, a formal context is constructed per decision class for which attributes correspond to the objects in support of the rule. The concept lattice then shows relationship information about the formal concepts.

Patient sentiment analysis is performed using EFCA to judge the quality of healthcare [6]. On social media, patients express opinions on their care. One challenge to existing methods is identifying aspects to analyze in sentiment analysis, as opinions on healthcare do not always translate directly to parts of speech. After social media posts are gathered, EFCA is used to determine which parts of speech are discussing an aspect of care. FCA is used for feature reduction and selection, to identify the most important aspects. Finally, SVM classifies whether each feature identified is indeed an aspect. EFCA is tested on six medical specialties. The reliability of the method is found to be better than segregational, conditional random field, and convolutional neural network methods.

3.10 Biomedical Ontologies

Ontologies have been used to represent data based on abstraction of human thoughts. As formal concepts and their relationships form an integral part of how FCA works, the framework is uniquely suited to provide knowledge discovery for applications such as ontologies. Previous discussion of FCA's use in ontologies for knowledge discovery can be found in [130]. Ontologies have been used to support mental health decisions [50], personalize medical treatment [149], and combine with FCA for data mining [29]. A summary of the papers in this section can be found in Table 17.

Cluster analysis is performed on colon cancer gene data [27]. Genes from the dataset are objects, and the expression profiles, gene ontology terms, and knowledge bases are attributes. The concepts produced are examined with enrichment analysis and gene ontology to determine the biological reasons the genes are grouped together in the concepts.

Various methods of interoperability (the healthcare system's ability to communicate information and data) in healthcare are described in [35]. The attempts at interoperability that are relevant to this survey are focused on communicating semantic meaning. FCA has been able to structurally

Table 17. Summary of Papers Examining Biomedical Ontologies

Study	Year	Purpose
[27]	2014	Determine biological reasons for colon cancer gene concept grouping
[167]	2018	Identify similar terms in different ontologies
[125]	2018	Develop ontology for chronic tropical diseases
[106]	2020	Perform ontology matching with DeepFCA
[168]	2021	Find missing concepts in biomedical terminology database

Table 18. Summary of Papers Examining the Use of FCA for Phylogeny

Study	Year	Purpose
[132]	2013	Propose algorithms to transform concept lattice to median graph
[64]	2018	Address error in [132]; propose new algorithm
[62]	2020	Transform context into distributive lattice
[17]	2021	Prove attribute reduction method in distributive concept lattice
[63]	2022	Construct context of a median lattice
[65]	2022	Develop algorithm to minimize distributive $\vee$ semilattice

analyze the SNOMED-CT medical ontology [89] and has been used to analyze its semantic completeness [88]. More recently, FCA has been used for ontology matching, which identifies similar items in different ontologies [167].

An ontology using FCA was developed for chronic tropical diseases [125]. The method was implemented with Protege-OWL. Tuberculosis is included in this ontology, along with leprosy, meningitis, malaria, ebola, and others.

FCA has been used in conjunction with deep learning in [106]. The study combines FCA with representation learning to perform biomedical ontology matching with the algorithm DeepFCA. Formal concepts based on tokens and entity names are computed, and the concept lattice is constructed. Compared with other representation learning systems, the algorithm performs better in F1-measure, though there is still room for improvement. This forms a basis for other algorithms using FCA and deep learning for knowledge discovery.

FCA can also help find missing concepts in a biomedical terminology database [168]. The concept lattice, with terms formed by longest common substrings, can help the user identify the position where a term may be missing, and the missing term can be inferred from its position. The method's success is validated with an automatic process, and manual review is needed to verify that the terms identified are indeed missing.

3.11 Phylogeny

Phylogeny, the study of evolutionary development of organisms [3], is a possible area of interest in FCA. A summary of papers described in this section can be found in Table 18. One of the main benefits of using FCA for such an application is the concept lattice, which may be used as a visual structure to analyze evolutionary development.

FCA's potential in this area began with [132]. The study explains the link between the concept lattice and phylogeny's use of median networks to represent genetic changes to species. Algorithms are proposed to reduce a concept lattice into a median network. One of the main differences between median graphs and the concept lattice, however, is that the concept lattice is not necessarily distributive: if any pair of elements has a lower bound, then any three elements must have a lower bound as well. This is further addressed in [64]. In [64], another algorithm is proposed

that transforms the concept lattice into a median graph. In particular, the authors point out an algorithm error in [132] relating to the transformation. The transformation of the concept lattice into a distributive lattice requires more formalism, and an algorithm to do so is presented. In [62], a method to transform a context into one that meets the distributive lattice conditions is described. A method for attribute reduction in a distributive concept lattice is proved in [17]. The efforts to construct a distributive semilattice is expanded on in [63], which describes algorithms to construct a median graph and construct a context of a median lattice, so FCA can be used for phylogeny analysis. A remaining issue is minimality of the median lattice, but [65] develops an algorithm to minimize the distributive $\vee$ semilattice. However, there may not be unique solutions to this algorithm.

4 FUTURE PROSPECTS OF FCA IN BIOINFORMATICS

There have been many recent developments in the applications mentioned in this survey. In this section, we describe how the methods used invite new methods of research. First, we will summarize the prospects of FCA, and then we will examine the prospects individually by category.

Biclustering gene expression or metabolomic data with FCA can reveal genes that work together to perform a biological function. This information can be used to discover biomarkers or identify similarly expressed genes.

Disease gene expression datasets, particularly ones for cancer, are often analyzed by FCA classification algorithms [20, 72, 114, 127, 133]. However, with very large datasets, as well as more complex data types, such as histopathological images, it is recommended that FCA interface with a machine learning or deep learning framework. In light of this, we have proposed a new hybrid deep learning and formal concept analysis image classification technique.

Combining FCA with a weight calculation or computing with a graph theoretic metric can result in detection of a graph's influential nodes. This can be helpful for determining seed genes of a gene regulatory network or discovering significant protein-protein interactions.

Next-generation sequencing has been used to identify bacterial strains from long reads. There are also prospects to use algorithms from other areas of computer science with FCA to perform this task. One is a **brain storm optimization (BSO)** algorithm, which imitates swarm intelligence, in conjunction with FCA. Because BSO is optimal for performing multiple sequence alignment, a BSO method with FCA could have great success in aligning multiple sequences. Another prospect of this area is pattern matching, which combines FCA and deterministic finite automata. Both prospects are not currently used for bioinformatics data analysis but could have great success.

A primary asset of FCA is its hierarchical structure. Once the concept lattice is calculated, objects and attributes are not just clustered, but organized in the subconcept-superconcept relationship. This has helped determine promiscuous domain interactions [101], as well as inspired the current trend to perform phylogenetic analysis with the concept lattice transforming to a median graph.

Now, we describe these prospects in more detail.

4.1 Gene Expression Data

FCA has been used to identify groups of genes expressed together (either positively or negatively). There are several ways this type of analysis can be improved.

A user-friendly tool to integrate gene expression data analysis using fuzzy FCA, pattern structures, or rough set theory. While many studies address the benefits to using variations of FCA, increasing use of these advanced models can help mitigate naive binarization data loss. Developing tools like WebGeneKFCA can also provide user-friendly options to perform this analysis.

Negatively correlated gene expression. While much work has been done in the past 15 years to identify positively expressed correlation groups, the identification of groups of genes that are negatively co-expressed is a more recent area. Groups of genes negatively expressed may provide a fuller explanation of gene expression in situations, especially in datasets related to diseases. This area could also be expanded by incorporating a variation of FCA to handle positive and negative data, such as JSM [105].

Use of the concept lattice in analysis. Most gene expression studies only calculate concepts. The concept lattice is useful by providing hierarchical analysis of the relationships between concepts. If computing concepts is done only as a clustering method, it is inefficient, and a separate, more efficient, clustering method is advised.

4.2 Disease Classification

We now motivate a new problem in the area of disease classification. The problem uses FCA and deep learning to perform classification on cancer image data.

Hybrid FCA with machine or deep learning. Some work has already combined FCA with **machine learning (ML)** or **deep learning (DL)** [106, 142, 172]. In the future, we aim to extend this work to an image classification problem with a large and high-dimensional dataset. Convolutional neural networks are the state-of-the-art technique in image classification problems. However, FCA has the following advantages over CNNs: (1) FCA requires one single pass over the dataset to create a lattice, unlike CNN, which is an iterative training process. (2) FCA can incorporate any change in input data by adding a new concept or extending the intent and extent of an existing concept in the lattice. This supports transfer learning, as FCA does not require retraining from scratch like DNN. (3) FCA creates a lattice for a visual representation of data. It helps experts to identify useful patterns. Also, experts can add new concepts to represent the information not discovered from the data or remove concepts that give inaccurate information. (4) FCA discovers the relationship between the feature set and the target variable during the classification task. Combining these two computing techniques can improve the computational power of FCA. Although FCA has benefits over ML models, as mentioned earlier, FCA can still be computationally expensive depending on how concepts are computed. Using another technique to reduce the dataset or otherwise improve FCA can be beneficial. In addition, FCA has most often been used for knowledge discovery once ML algorithms extract relevant information from the dataset. This can help provide information about the relationship between extracted features. In particular, the concept lattice, when used for classification, may provide information that would not otherwise be accessible when the two strategies are combined.

4.3 COVID-19 Analysis and Healthcare Informatics

FCA has identified similarities between COVID-19 vaccines, determined hospital patients who may have the disease, and identified close contacts of a supposed infected individual [34, 74, 165]. Extrapolating, we see the following possibilities.

Treatment similarities. With new treatments and vaccines being developed and implemented in the community, a similar study to [34] may provide insight in treatments aiming to reduce the likelihood of severe illness and hospitalization.

Symptom tracking and isolation recommendation. In colleges or assisted living or correctional facilities, officials may wish to mitigate the likelihood of COVID-19 spreading among their members. An approach similar to [74] may help officials track symptoms and recommend treatment or isolation guidelines to those affected. If close contacts can be identified with device data as in [165], this may also help curb the spread of the virus.

4.4 Gene Regulatory Network Analysis

As [61] is one of the only studies to work with the gene regulatory network, we describe two of many open prospects.

Discovering seed genes or genes interacting strongly. Discovering nodes in the network and analyzing the network are both areas for further research. There is potential for overlap in this area with other applications addressed in this survey. Methods like those used in identifying protein-protein interaction could be used to find seed genes as influential nodes in the graph. A similar approach could help find genes interacting strongly with seed genes. As FCA has also been used for social media analysis such as [14, 144], using a social media approach could aid this topic as well.

GRN reconstruction. Reconstruction of the network is also an open problem that, to our knowledge, has not yet been attempted using FCA. A probabilistic graphical model is one way to perform reconstruction [169]. FCA could be used in a probabilistic advanced model to accomplish the task.

4.5 Drug Design and Development

FCA has been used to analyze **structure-activity relationships (SARs)** and perform social media analysis for detecting adverse drug effects and identifying compounds, in both modern settings and medieval pharmacy data. We describe how FCA may be used in each of these areas to produce further interesting results.

Structure-activity relationship and FCA concept lattice. Further work can be done to compare FCA to the SAR. Representing the Hasse diagram visualization of an SAR could lead to new insights of the relationship between biological properties in molecules and further develop FCA as useful to chemical applications.

Social media for drug side effect analysis. As [43] discovers adverse drug effects through social media analysis, this approach could be replicated to determine adverse effects for more types of drugs. It could also be used to determine drug side effects or efficacy in real time.

Identifying compounds with biomarkers discovery methods. Being able to identify compounds with FCA has great potential. Further classifying types of compounds can be performed. An entire FCA pipeline could be built to discover biomarkers from genomic or metabolic data, and the biomarkers information can be used in drug development to develop treatments for the identified targets.

Medieval pharmacy analysis. After analyzing medieval pharmacy data [33], more studies analyzing similarly historical data could produce valuable information. Because of the curse of dimensionality, small datasets like the one in [33] (due to the difficulty in obtaining it) are uniquely suited to be analyzed with FCA. This approach could be combined with other data mining methods to integrate findings from the historical data with modern drug discoveries.

4.6 Next-generation Sequencing

FCA can help identify bacterial strains from long reads [143]. There have been two recent studies incorporating FCA where a similar method could be used to perform multiple sequence alignment.

Brain storm optimization for sequence alignment. Brain storm optimization is a type of algorithm aimed at imitating swarm intelligence. Although the HBSO algorithm that uses FCA [38] is not currently used for sequencing, the method could be tested on alignment data, which may lead to an efficient multiple sequence alignment algorithm.

Pattern matching for sequence alignment. Pattern matching may be done with the FCA concept lattice [159]. Deterministic and non-deterministic finite automatas can be converted into

the concept lattice for pattern matching. An incremental approach can be leveraged so the entire lattice does not need to be constructed at once. While no bioinformatics applications are mentioned, an implementation of this method may benefit RNA sequence alignment.

4.7 Protein-protein Interaction

Most work to identify protein-protein interactions using FCA does so by determining influential nodes in a graph. Discerning the similarity between graphs [75] has also been done. These two approaches inform how interactions may be determined. A few possibilities to extend this application are described below.

Association rule mining biclustering. Association rule mining (unrelated to FCA) has been used to discover protein-protein interactions related to COVID-19 [66, 126]. Calculating maximal biclusters may lead to discovering novel interactions, and this information could assist in drug development to treat those infected. Formal concept analysis could contribute to this method by using concept calculation to bicluster human-protein interaction data, as in [66]. Other methods could be identifying significant objects [122] to determine protein-protein interactions, similar to [83].

Compound similarity analysis using graphs. In [75], graph similarity is calculated to determine compound similarity. Similarity between a family of renal diseases has also been investigated using gene expression data [100]. Although the datasets are different, a similar approach may be made to identify strong relationships between compounds involved in interactions. Additionally, given that [100] formulates FCA using graph theory, these methods could be combined or inform one another to provide valuable information about a protein-protein interaction graph.

4.8 Biomarkers Discovery

Methods using FCA to predict biomarkers primarily identify genes that occur frequently in concepts or identify significant biclusters in a gene expression dataset. Here are two ways these methods may be extended.

Identify critical concepts with $\mathcal{K}$-FCA. Critical concepts can be identified in the $\mathcal{K}$-FCA approach [67]. The software could be used to discover biomarkers with critical concepts.

Experimental validation of gene discovery. FCA analysis of gene expression data has revealed potential genes of interest that may be biomarkers for diseases. Several have been experimentally validated in [94]. Further experimental validation can show whether FCA's methods assist in discovering genes that contribute to cancerous behavior in cells.

4.9 Biomedical Ontologies

Ontologies can provide valuable enrichment to gene expression data analysis, or vice versa [27]. Studies have also shown that FCA can help identify similarities in different ontologies or identify missing terms.

Identification of missing data. Missing terms in a biomedical ontology can be identified [168], which leads to a better understanding of the limits of a given ontology. If deep learning is used, as in [106], this task may be easier.

Communication between ontologies. FCA has been shown to benefit ontologies by identifying similar terms in different ontologies [167]. This may be helpful in areas where ontological concepts and their relationships can have different placements in different ontologies, such as medical knowledge.

4.10 Phylogeny

Encouraging progress has been made for using the concept lattice to study evolutionary development. The beginning attempts were to transform the concept lattice into a median graph, though

the transformation was not done correctly. In recent attempts, an algorithm has been developed to rectify the transformation. We describe this area's prospects.

Identify unique solutions. Identifying unique solutions to an algorithm calculating the minimal distributive $\vee$ semilattice is an immediate potential area of future research.

Experiments on phylogenetic data. Once the median graph is obtained from a formal context, performing analysis on an organism's evolutionary development can show the effect of the algorithmic development in this area.

5 DISCUSSION AND CONCLUSIONS

FCA is a framework that allows for analysis and visualization of data. Relationships within the dataset can be analyzed with formal concepts, an abstraction of human thought. In a dataset, concepts are maximal subsets of the data, and concepts are ordered by set inclusion. That partial order can be visualized in the concept lattice, a powerful asset to knowledge discovery. This survey has described how FCA provides a framework for analyzing recent applications in bioinformatics. In this section, we describe the prospects, challenges, and limitations of using FCA for such tasks. First, we address the FCA strategies that are commonly used in many studies. Next, we address some ways in which bioinformatics studies have helped the advancement in FCA formalisms. Finally, we discuss the challenges and limitations of using FCA for biomedical data, then conclude.

There are many FCA strategies that are commonly used in studies examining biological data. These include cluster analysis of concepts, querying the constructed concept lattice, and using a similarity measure or performing pattern matching. Similarly, using decision and association rules, feature selection or extraction, and performing graph analysis with FCA are also commonly used strategies.

Cluster analysis is performed in cancer classification and drug design and development. This strategy aids the determination of whether symptoms, gene profiles, or protein responses may be significant. Querying the concept lattice lets the user glean information from the data. For example, querying is used in COVID-19 analysis, gene template matching, protein-protein interactions, identifying drug fragment relationships, and determining biological information from gene concepts. Utilizing similarity measures and pattern matching is useful for drug design and development, as well as healthcare informatics. It has been used to find groups of data that correspond with one another or match patterns that appear in the dataset.

While FCA methods have been used to identify missing elements or discover common elements in two ontologies, similar methods may be used in (for example) reconciling datasets from different experiments or determining the classification of an as-yet-unseen object. Association and decision rules are used to classify diseases, discover biomarkers, and reverse-engineer healthcare informatics data. These kinds of rules represent implicit relationships found in the dataset, and the implications of such rules can be used to perform classification. Feature selection and extraction have been used in health-related problems, such as disease classification and healthcare informatics. In these situations, FCA is used to identify important features after machine learning algorithms perform initial analysis to reduce the dataset.

Finally, work in COVID-19 analysis, disease classification, and protein-protein interaction has used graph analytic methods. A graph of protein interactions can be analyzed to see if it has influential nodes or cliques. Cliques are obtained in COVID-19 studies, to identify close contacts of an individual. Graph similarity is used to examine disease similarity.

It is evident that many FCA techniques can be used in disparate applications. However, some methods are confined to certain applications because of commonly used data formats. For example, techniques such as pattern mining and decision rules are more likely to use sequence data.

Therefore, it is more likely that disease classification or healthcare informatics applications will take advantage of these methods.

Many studies that use FCA to analyze bioinformatics data have advanced the formalism of FCA. Some researchers realized that their problems cannot be solved using existing techniques, so new FCA models were designed to solve those problems. These models include EFCA [6], FragFCA [110], and $\mathcal{K}$-FCA [67]. Other researchers found that their biological data was in graph format, and FCA was used in new ways to perform analysis on those graphs. For example, [75] developed an FCA method to evaluate similarity between graphs, [165] determined γ-quasi cliques from a graph using FCA, and [100] discovered renal disease similarity using both graph theory and FCA methods. Since graphs are used to model many types of data, from gene networks and social media networks to compiler algorithms, such developments introduce the possibility of using FCA for many other domains of analysis.

One of the principal challenges of FCA for bioinformatics applications is the cost and time to calculate the concept lattice. When the naive algorithm is used to accomplish this task, the calculation is exponential. However, if an algorithm such as In-Close or its derivations [13] are used, the calculation may be more manageable. The main limitation of FCA is related to its exponential risk. When FCA is used on large data, if binarization of a multivalued data set is attempted, loss of data is a result. Otherwise, it can be almost impossible to calculate the concept lattice. The challenges and limitations are multiplied greatly when FCA is used to analyze biological data, which are typically high-dimensional. The limitation can be mitigated by using an advanced model of FCA or augmenting the method with a machine learning algorithm. Using a distributive computing framework such as Apache Spark can also greatly reduce computation time.

With this review, we have shown that formal concept analysis has been useful in analyzing various types of biological data for many important and current bioinformatics applications. First, we described how the FCA formalism works, including the concept lattice, and the advantages of FCA's hierarchical structure. Then we discussed several advanced models of FCA, and how the concept lattice can become manageable. Next, we described 11 applications that use FCA for analysis. We have also identified many promising future prospects where FCA may further advance these applications. When examining how far the state of FCA research is from real-world practicality, we acknowledge that with the exponential risk, there are still theoretical difficulties, but progress is being made in several directions. The In-Close algorithm and its recent improvements [171, 173] can refine the process of concept and lattice calculation. Distributive computing frameworks can also bring about practicality, as computation time can be greatly reduced for very large datasets. FCA has also been used in COVID-19 analysis [34, 74, 165], which is an emerging and useful area of study since 2020. Since methods from other areas of FCA research are applicable to bioinformatics, and theoretical advancements are still being made, FCA continues to be relevant and useful for bioinformatics applications.

REFERENCES

[1] Apache Hadoop. n.d. https://hadoop.apache.org/.
[2] A Brief Guide to Genomics. n.d. https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics.
[3] Phylogeny - an Overview | ScienceDirect Topics. n.d. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/phylogeny.
[4] National Cancer Institute. 2015. Cancer Statistics. https://www.cancer.gov/about-cancer/understanding/statistics.
[5] National Cancer Institute. 2007. https://www.cancer.gov/about-cancer/understanding/what-is-cancer.
[6] Zohair Ahmed, Junwen Duan, FangXiang Wu, and Jianxin Wang. 2021. EFCA: An extended formal concept analysis method for aspect extraction in healthcare informatics. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'21)*. 1241–1244. https://doi.org/10.1109/BIBM52615.2021.9669754
[7] Mehmet Fatih Akay. 2009. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications* 36, 2, Part 2 (March 2009), 3240–3247. https://doi.org/10.1016/j.eswa.2008.01.009

[8] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS* 96, 12 (June 1999), 6745–6750. https://doi.org/10.1073/pnas.96.12.6745

[9] Amedeo Napoli and Thi Nhu Nguyen Le. n.d. Lattice Editor. https://latviz.loria.fr/about.html.

[10] I. I. Amin, S. K. Kassim, A. e Hassanien, and H. A. Hefny. 2013. Using formal concept analysis for mining hyomethylated genes among breast cancer tumors subtypes. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI'13)*. 521–526. https://doi.org/10.1109/ICACCI.2013.6637226

[11] Islam Ibrahim Amin, Aboul Ella Hassanien, Samar K. Kassim, and Hesham A. Hefny. 2015. Big DNA methylation data analysis and Visualizing in a Common Form of Breast Cancer. In *Big Data in Complex Systems: Challenges and Opportunities*, Aboul Ella Hassanien, Ahmad Taher Azar, Vaclav Snasael, Janusz Kacprzyk, and Jemal H. Abawajy (Eds.). Springer International Publishing, Cham, 375–392. https://doi.org/10.1007/978-3-319-11056-1_13

[12] I. I. Amin, S. K. Kassim, A. e Hassanien, and H. A. Hefny. 2012. Formal concept analysis for mining hypermethylated genes in breast cancer tumor subtypes. In *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA'12)*. 764–769. https://doi.org/10.1109/ISDA.2012.6416633

[13] Simon Andrews. 2011. In-Close2, a high performance formal concept miner. In *Conceptual Structures for Discovering Knowledge (Lecture Notes in Computer Science)*, Simon Andrews, Simon Polovina, Richard Hill, and Babak Akhgar (Eds.). Springer, Berlin, 50–62.

[14] Simon Andrews, Helen Gibson, Konstantinos Domdouzis, and Babak Akhgar. 2016. Creating corroborated crisis reports from social media data through formal concept analysis. *J. Intell. Inf. Syst.* 47, 2 (Oct. 2016), 287–312. https://doi.org/10.1007/s10844-016-0404-9

[15] Simon Andrews and Kenneth McLeod. 2013. Gene co-expression in mouse embryo tissues. *International Journal of Intelligent Information Technologies (IJIIT)* 9, 4 (Oct. 2013), 55–68. https://doi.org/10.4018/ijiit.2013100104

[16] Simon Andrews and Kenneth McLeod. 2018. A visual analytics technique for exploring gene expression in the developing mouse embryo. In *Graph-based Representation and Reasoning (Lecture Notes in Computer Science)*, Peter Chapman, Dominik Endres, and Nathalie Pernelle (Eds.). Springer International Publishing, 137–151.

[17] Roberto G. Aragón, Jesús Medina, and Eloísa Ramírez-Poussa. 2021. Identifying non-sublattice equivalence classes induced by an attribute reduction in FCA. *Mathematics* 9, 5 (Jan. 2021), 565. https://doi.org/10.3390/math9050565

[18] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, Gerardo Fernandez, Jack Zeineh, Matthias Kohl, Christoph Walz, Florian Ludwig, Stefan Braunewell, Maximilian Baust, Quoc Dang Vu, Minh Nguyen Nhat To, Eal Kim, Jin Tae Kwak, Sameh Galal, Veronica Sanchez-Freire, Nadia Brancati, Maria Frucci, Daniel Riccio, Yaqi Wang, Lingling Sun, Kaiqiang Ma, Jiannan Fang, Ismael Kone, Lahsen Boulmane, Aurélio Campilho, Catarina Eloy, António Polónia, and Paulo Aguiar. 2019. BACH: Grand challenge on breast cancer histology images. *Medical Image Analysis* 56 (Aug. 2019), 122–139. https://doi.org/10.1016/j.media.2019.05.010

[19] J. Atif, C. Hudelot, and I. Bloch. 2014. Explanatory reasoning for image understanding using formal concept analysis and description logics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, 5 (May 2014), 552–570. https://doi.org/10.1109/TSMC.2013.2280440

[20] Hayfa Azibi, Nida Meddouri, and Mondher Maddouri. 2020. Survey on formal concept analysis based supervised classification techniques. In *Frontiers in Artificial Intelligence and Applications*, Antonio J. Tallón-Ballesteros and Chi-Hua Chen (Eds.). IOS Press. https://doi.org/10.3233/FAIA200762

[21] Jaume Baixeries, Mehdi Kaytoue, and Amedeo Napoli. 2014. Characterizing functional dependencies in formal concept analysis with pattern structures. *Annals of Mathematics and Artificial Intelligence* 72, 1–2 (Oct. 2014), 129–149. https://doi.org/10.1007/s10472-014-9400-3

[22] Eduard Bartl and Jan Konecny. 2016. L-concept analysis with positive and negative attributes. *Information Sciences* 360 (Sept. 2016), 96–111. https://doi.org/10.1016/j.ins.2016.04.012

[23] Eduard Bartl and Jan Konecny. 2019. L-concept lattices with positive and negative attributes: Modeling uncertainty and reduction of size. *Information Sciences* 472 (Jan. 2019), 163–179. https://doi.org/10.1016/j.ins.2018.08.057

[24] Alexandre Bazin, Miguel Couceiro, Marie-Dominique Devignes, and Amedeo Napoli. 2021. An approach to identifying the most predictive and discriminant features in supervised classification problems. In *Graph-based Representation and Reasoning (Lecture Notes in Computer Science)*, Tanya Braun, Marcel Gehrke, Tom Hanika, and Nathalie Hernandez (Eds.). Springer International Publishing, Cham, 48–56. https://doi.org/10.1007/978-3-030-86982-3_4

[25] Sam Behjati and Patrick S. Tarpey. 2013. What is next generation sequencing? *Archives of Disease in Childhood - Education and Practice* 98, 6 (Dec. 2013), 236–238. https://doi.org/10.1136/archdischild-2013-304340

[26] Radim Belohlavek and Vilem Vychodil. 2005. What is a fuzzy concept lattice? In *Proc. Concept Lattices and their Applications*, 12.

[27] Sidahmed Benabderrahmane. 2014. Formal concept analysis and knowledge integration for highlighting statistically enriched functions from microarrays data. In *International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO'14)*. 12.

[28] Jérémy Besson, Céline Robardet, Jean-François Boulicaut, and Sophie Rome. 2005. Constraint-based concept mining and its application to microarray data analysis. *Intelligent Data Analysis* 9, 1 (Jan. 2005), 59–82. https://doi.org/10.3233/IDA-2005-9105

[29] Sanjiv K. Bhatia and Jitender S. Deogun. 2014. Data Mining Tools: Formal Concept Analysis and Rough Sets. https://www.igi-global.com/chapter/data-mining-tools/www.igi-global.com/chapter/data-mining-tools/107268. 655–663 pages. https://doi.org/10.4018/978-1-4666-5202-6.ch060

[30] Sylvain Blachon, Ruggero G. Pensa, Jérémy Besson, Céline Robardet, Jean-François Boulicaut, and Olivier Gandrillon. 2007. Clustering formal concepts to discover biologically relevant knowledge from gene expression data. *In Silico Biology* 7, 4,5 (Jan. 2007), 467–483.

[31] Stefania Boffa, Petra Murinová, and Vilém Novák. 2021. A proposal to extend relational concept analysis with fuzzy scaling quantifiers. *Knowledge-Based Systems* 231 (Nov. 2021), 107452. https://doi.org/10.1016/j.knosys.2021.107452

[32] Mikhail Bogatyrev et al. 2021. Multimodal clustering with evolutionary algorithms. In *FCA4AI'21*. 71–85.

[33] Agnes Braud, Xavier Dolques, Pierre Fechter, Nicolas Lachiche, Florence Le Ber, and Veronique Pitchon. 2021. Analyzing the composition of remedies in ancient pharmacopeias with FCA. In *RealDataFCA'21*.

[34] Javier Burgos-Salcedo. 2021. A Comparative Analysis of Clinical Stage 3 COVID-19 Vaccines Using Knowledge Representation. (March 2021), 2021.03.07.21253082 pages. https://doi.org/10.1101/2021.03.07.21253082

[35] Rashmi Burse, Michela Bertolotto, Dympna O'Sullivan, and Gavin McArdle. 2021. Chapter 4 - Semantic interoperability: The future of healthcare. In *Web Semantics*, Sarika Jain, Vishal Jain, and Valentina Emilia Balas (Eds.). Academic Press, 31–53. https://doi.org/10.1016/B978-0-12-822468-7.00018-3

[36] Aleksey Buzmakov, Elias Egho, Nicolas Jay, Sergei O. Kuznetsov, Amedeo Napoli, and Chedy Raïssi. 2013. The representation of sequential patterns and their projections within formal concept analysis. In *Workshop Notes for LML (PKDD)*.

[37] Aleksey Buzmakov and Amedeo Napoli. 2016. How fuzzy FCA and pattern structures are connected. In *5th Workshop "What Can FCA Do for Artificial Intelligence?" (FCA4AI'16)*.

[38] Fengrong Chang, Lianbo Ma, Yan Song, and Aoshuang Dong. 2021. Brain storm optimization algorithm based on formal concept analysis. In *Advances in Swarm Intelligence (Lecture Notes in Computer Science)*, Ying Tan and Yuhui Shi (Eds.). Springer International Publishing, Cham, 479–491. https://doi.org/10.1007/978-3-030-78743-1_43

[39] Jinkun Chen, Jusheng Mi, Bin Xie, and Yaojin Lin. 2019. A fast attribute reduction method for large formal decision contexts. *International Journal of Approximate Reasoning* 106 (March 2019), 1–17. https://doi.org/10.1016/j.ijar.2018.12.002

[40] V. Choi, Y. Huang, V. Lam, D. Potter, R. Laubenbacher, and K. Duca. 2008. Using formal concept analysis for microarray data comparison. *J. Bioinform. Comput. Biol.* 6, 1 (Feb. 2008), 65–75. https://doi.org/10.1142/S021972000800328X

[41] Raghavendra K. Chunduri and Aswani Kumar Cherukuri. 2019. Scalable formal concept analysis algorithms for large datasets using spark. *J. Ambient Intell. Human Comput.* 10, 11 (Nov. 2019), 4283–4303. https://doi.org/10.1007/s12652-018-1105-8

[42] Raghavendra Kumar Chunduri and Aswani Kumar Cherukuri. 2021. Scalable algorithm for generation of attribute implication base using FP-growth and spark. *Soft. Comput.* 25, 14 (July 2021), 9219–9240. https://doi.org/10.1007/s00500-021-05844-9

[43] Michela De Rosa, Giuseppe Fenza, Alessandro Gallo, Mariacristina Gallo, and Vincenzo Loia. 2021. Pharmacovigilance in the era of social media: Discovering adverse drug events cross-relating Twitter and Pubmed. *Future Generation Computer Systems* 114 (Jan. 2021), 394–402. https://doi.org/10.1016/j.future.2020.08.020

[44] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. *Commun. ACM* 51, 1 (Jan. 2008), 107–113. https://doi.org/10.1145/1327452.1327492

[45] Alexander Demin, Denis Ponomaryov, and Evgeny Vityaev. 2012. Probabilistic concepts in formal contexts. In *Perspectives of Systems Informatics (Lecture Notes in Computer Science)*, Edmund Clarke, Irina Virbitskaite, and Andrei Voronkov (Eds.). Springer, Berlin, 394–410. https://doi.org/10.1007/978-3-642-29709-0_33

[46] Jitender Deogun et al. 2003. Probability logic modeling of knowledge discovery in databases. In *Foundations of Intelligent Systems (Lecture Notes in Computer Science)*, Ning Zhong, Zbigniew W. Raś, Shusaku Tsumoto, and Einoshin Suzuki (Eds.). Springer, Berlin, 402–407. https://doi.org/10.1007/978-3-540-39592-8_56

[47] Jitender Deogun and Liying Jiang. 2005. SARM — Succinct association rule mining: An approach to enhance association mining. In *Foundations of Intelligent Systems (Lecture Notes in Computer Science)*, Mohand-Said Hacid, Neil V. Murray, Zbigniew W. Raś, and Shusaku Tsumoto (Eds.). Springer, Berlin, 121–130. https://doi.org/10.1007/11425274_13

[48] Jitender Deogun, Liying Jiang, and Vijay V. Raghavan. 2004. Discovering maximal potentially useful association rules based on probability logic. In *Rough Sets and Current Trends in Computing (Lecture Notes in Computer Science)*, Shusaku Tsumoto, Roman Słowiński, Jan Komorowski, and Jerzy W. Grzymała-Busse (Eds.). Springer, Berlin, 274–284. https://doi.org/10.1007/978-3-540-25929-9_32

[49] Jitender S. Deogun and Jamil Saquer. 2004. Monotone concepts for formal concept analysis. *Discrete Applied Mathematics* 144, 1 (Nov. 2004), 70–78. https://doi.org/10.1016/j.dam.2004.05.001

[50] Jitender S. Deogun and William Spaulding. 2010. Conceptual development of mental health ontologies. In *Advances in Intelligent Information Systems*, Janusz Kacprzyk, Zbigniew W. Ras, and Li-Shiang Tsay (Eds.). Vol. 265. Springer, Berlin, 299–333. https://doi.org/10.1007/978-3-642-05183-8_13

[51] Sérgio M. Dias and Newton J. Vieira. 2015. Concept lattices reduction: Definition, analysis and classification. *Expert Systems with Applications* 42, 20 (Nov. 2015), 7084–7097. https://doi.org/10.1016/j.eswa.2015.04.044

[52] Guozhu Dong, Chunyu Jiang, Jian Pei, Jinyan Li, and Limsoon Wong. 2005. Mining succinct systems of minimal generators of formal concepts. In *Database Systems for Advanced Applications (Lecture Notes in Computer Science)*, Lizhu Zhou, Beng Chin Ooi, and Xiaofeng Meng (Eds.). Springer, Berlin, 175–187.

[53] Egor Dudyrev and Sergei Kuznetsov. 2021. Summation of decision trees. In *"What Can FCA Do for Artificial Intelligence?"* 99–104.

[54] Egor Dudyrev and Sergei O. Kuznetsov. 2021. Decision concept lattice vs. decision trees and random forests. In *International Conference on Formal Concept Analysis (Lecture Notes in Computer Science)*, Agnès Braud, Aleksey Buzmakov, Tom Hanika, and Florence Le Ber (Eds.). Springer International Publishing, Cham, 252–260. https://doi.org/10.1007/978-3-030-77867-5_16

[55] Jose Maria Fernandez-Calabozo et al. 2012. WebGeneKFCA: An on-line conceptual analysis tool for genomic expression data. In *The 9th International Conference on Concept Lattices and Their Applications*. 345–350.

[56] Anna Formica. 2019. Similarity reasoning in formal concept analysis: From one- to many-valued contexts. *Knowl. Inf. Syst.* 60, 2 (Aug. 2019), 715–739. https://doi.org/10.1007/s10115-018-1252-4

[57] Anna Formica. 2021. Concept similarity in formal concept analysis with many-valued contexts. *Computing and Informatics* 40, 3 (Nov. 2021), 469–488. https://doi.org/10.31577/cai_2021_3_469

[58] Bernhard Ganter and Sergei O. Kuznetsov. 2001. Pattern structures and their projections. In *Conceptual Structures: Broadening the Base (Lecture Notes in Computer Science)*, Harry S. Delugach and Gerd Stumme (Eds.). Springer, Berlin, 129–142. https://doi.org/10.1007/3-540-44583-8_10

[59] Bernhard Ganter, Sebastian Rudolph, and Gerd Stumme. 2019. Explaining data with formal concept analysis. In *Reasoning Web. Explainable Artificial Intelligence: 15th International Summer School 2019, Tutorial Lectures*, Markus Krötzsch and Daria Stepanova (Eds.). Springer International Publishing, Cham, 153–195. https://doi.org/10.1007/978-3-030-31423-1_5

[60] Bernhard Ganter and Rudolf Wille. 2012. *Formal Concept Analysis: Mathematical Foundations*. Springer Science & Business Media.

[61] Jutta Gebert, Susanne Motameny, Ulrich Faigle, Christian V. Forst, and Rainer Schrader. 2008. Identifying genes of gene regulatory networks using formal concept analysis. *Journal of Computational Biology* 15, 2 (March 2008), 185–194. https://doi.org/10.1089/cmb.2007.0107

[62] Alain Gély, Miguel Couceiro, Laurent Miclet, and Amedeo Napoli. 2020. Steps in the representation of concept lattices and median graphs. In *15th International Conference on Concept Lattices and Their Applications (CLA'20)*. 1.

[63] Alain Gély, Miguel Couceiro, Laurent Miclet, and Amedeo Napoli. 2022. A study of algorithms relating distributive lattices, median graphs, and formal concept analysis. *International Journal of Approximate Reasoning* 142 (March 2022), 370–382. https://doi.org/10.1016/j.ijar.2021.12.011

[64] Alain Gély, Miguel Couceiro, and Amedeo Napoli. 2018. Steps towards achieving distributivity in formal concept analysis. In *The 14th International Conference on Concept Lattices and Their Applications*. 291.

[65] Alain Gély, Miguel Couceiro, and Amedeo Napoli. 2022. Towards distributivity in FCA for phylogenetic data. In *Complex Data Analysis with Formal Concept Analysis*.

[66] Moumita Ghosh, Pritam Sil, Anirban Roy, Rohmatul Fajriyah, and Kartick Chandra Mondal. 2021. Finding prediction of interaction between SARS-CoV-2 and human protein: A data-driven approach. *J. Inst. Eng. India Ser. B* 102, 6 (Dec. 2021), 1293–1302. https://doi.org/10.1007/s40031-021-00569-7

[67] Jose M. González-Calabozo, Francisco J. Valverde-Albacete, and Carmen Peláez-Moreno. 2016. Interactive knowledge discovery and data mining on genomic expression data with numeric formal concept analysis. *BMC Bioinformatics* 17, 1 (Sept. 2016), 374. https://doi.org/10.1186/s12859-016-1234-z

[68] Sathyanarayanan Gopalakrishnan, Supriya Sridharan, and Swaminathan Venkatraman. 2021. Centrality measures in finding influential nodes for the big-data network. In *Handbook of Smart Materials, Technologies, and Devices*, Chaudhery Mustansar Hussain and Paolo Di Sia (Eds.). Springer International Publishing, Cham, 1–17. https://doi.org/10.1007/978-3-030-58675-1_103-1

[69] Niruktha Roy Gotoor. 2019. Image classification using fuzzy FCA. *Embargoed Master's Theses*.

[70] Dhouha Grissa, Blandine Comte, Mélanie Pétéra, Estelle Pujos-Guillot, and Amedeo Napoli. 2019. A hybrid and exploratory approach to knowledge discovery in metabolomic data. *Discrete Applied Mathematics* 273 (Feb. 2020), 103–116. https://doi.org/10.1016/j.dam.2018.11.025

[71] Dhouha Grissa, Blandine Comte, Estelle Pujos-Guillot, and Amedeo Napoli. 2016. A hybrid knowledge discovery approach for mining predictive biomarkers in metabolomic data. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science)*, Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken (Eds.). Springer International Publishing, Cham, 572–587. https://doi.org/10.1007/978-3-319-46128-1_36

[72] Anamika Gupta, Naveen Kumar, and Vasudha Bhatnagar. 2005. Incremental classification rules based on association rules using formal concept analysis. In *Machine Learning and Data Mining in Pattern Recognition (Lecture Notes in Computer Science)*, Petra Perner and Atsushi Imiya (Eds.). Springer, Berlin, 11–20. https://doi.org/10.1007/11510888_2

[73] Harald Hampel et al. 2020. A precision medicine framework using artificial intelligence for the identification and confirmation of genomic biomarkers of response to an Alzheimer's disease therapy: Analysis of the blarcamesine (ANAVEX2-73) phase 2a clinical study. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 6, 1 (2020), e12013. https://doi.org/10.1002/trc2.12013

[74] Fei Hao and Doo-Soon Park. 2021. CoNavigator: A framework of FCA-based novel coronavirus COVID-19 domain knowledge navigation. *Human-centric Computing and Information Sciences* 6 (Feb. 2021), 12. https://doi.org/10.22967/HCIS.2021.11.006

[75] Fei Hao, Dae-Soo Sim, Doo-Soon Park, and Hyung-Seok Seo. 2017. Similarity evaluation between graphs: A formal concept analysis approach. *Journal of Information Processing Systems* 13, 5 (2017), 1158–1167. https://doi.org/10.3745/JIPS.04.0048

[76] Fei Hao, Erhe Yang, Lantian Guo, Aziz Nasridinov, and Doo-Soon Park. 2021. On invariance of concept stability for attribute reduction in concept lattice. In *Advances in Computer Science and Ubiquitous Computing (Lecture Notes in Electrical Engineering)*, James J. Park, Simon James Fong, Yi Pan, and Yunsick Sung (Eds.). Springer, Singapore, 101–106. https://doi.org/10.1007/978-981-15-9343-7_14

[77] Y. He, R. Bockmon, M. Modey, and S. Roscoe. 2020. Classification of cancer types based on gene expression data. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'20)*. 2175–2182. https://doi.org/10.1109/BIBM49941.2020.9313559

[78] Rui Henriques et al. 2015. A structured view on pattern mining-based biclustering. *Pattern Recognition* 48, 12 (Dec. 2015), 3941–3958. https://doi.org/10.1016/j.patcog.2015.06.018

[79] Amina Houari, Wassim Ayadi, and Sadok Ben Yahia. 2018. A new FCA-based method for identifying biclusters in gene expression data. *Int. J. Mach. Learn. & Cyber.* 9, 11 (Nov. 2018), 1879–1893. https://doi.org/10.1007/s13042-018-0794-9

[80] A. Houari, W. Ayadi, and S. Ben Yahia. 2018. NBF: An FCA-based algorithm to identify negative correlation biclusters of DNA microarray data. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA'18)*. 1003–1010. https://doi.org/10.1109/AINA.2018.00146

[81] Anna Hristoskova, Veselka Boeva, and Elena Tsiporkova. 2014. A formal concept analysis approach to consensus clustering of multi-experiment expression data. *BMC Bioinformatics* 15, 1 (May 2014), 151. https://doi.org/10.1186/1471-2105-15-151

[82] Mohamed-Hamza Ibrahim et al. 2021. Conceptual relevance index for identifying actionable formal concepts. In *Graph-based Representation and Reasoning (Lecture Notes in Computer Science)*. Springer International Publishing, Cham, 119–126. https://doi.org/10.1007/978-3-030-86982-3_9

[83] Mohamed Hamza Ibrahim, Rokia Missaoui, and Jean Vaillancourt. 2020. Cross-face centrality: A new measure for identifying key nodes in networks based on formal concept analysis. *IEEE Access* 8 (2020), 206901–206913. https://doi.org/10.1109/ACCESS.2020.3038306

[84] Dmitry I. Ignatov et al. 2021. Object-attribute biclustering for elimination of missing genotypes in ischemic stroke genome-wide data. In *Recent Trends in Analysis of Images, Social Networks and Texts (Communications in Computer and Information Science)*. Springer International Publishing, Cham, 185–204. https://doi.org/10.1007/978-3-030-71214-3_16

[85] Radek Janostik, Jan Konecny, and Petr Krajča. 2022. LCM from FCA point of view: A CBO-style algorithm with speed-up features. *International Journal of Approximate Reasoning* 142 (March 2022), 64–80. https://doi.org/10.1016/j.ijar.2021.11.005

[86] Pujari Jeevana Jyothi, Karteeka Pavan K, S. M. Raiyyan, and T. Rajasekhar. 2020. *MSA: An Application of Brain Storm Optimization Algorithm*. SSRN Scholarly Paper ID 3646174. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3646174

[87] D. Jiang, C. Tang, and A. Zhang. 2004. Cluster Analysis for Gene Expression Data: A Survey. *IEEE Transactions on Knowledge & Data Engineering* 16, 11 (2004), 1370–1386.

[88] Guoqian Jiang and Christopher G. Chute. 2009. Auditing the semantic completeness of SNOMED CT using formal concept analysis. *Journal of the American Medical Informatics Association* 16, 1 (Jan. 2009), 89–102. https://doi.org/10.1197/jamia.M2541

[89] Guoqian Jiang, Katsuhiko Ogasawara, Akira Endoh, and Tsunetaro Sakurai. 2003. Context-based ontology building support in clinical domains using formal concept analysis. *International Journal of Medical Informatics* 71, 1 (Aug. 2003), 71–81. https://doi.org/10.1016/S1386-5056(03)00092-3

[90] Liying Jiang. 2006. *New Data Mining Models Based on Formal Concept Analysis and Probability Logic*. Ph.D. Dissertation. The University of Nebraska - Lincoln, Nebraska.
[91] Nyoman Juniarta. 2019. *Mining Complex Data and Biclustering Using Formal Concept Analysis*. Ph.D. Dissertation. Université de Lorraine.
[92] Nyoman Juniarta, Miguel Couceiro, and Amedeo Napoli. 2019. A unified approach to biclustering based on formal concept analysis and interval pattern structure. In *22nd International Conference on Discovery Science*.
[93] Jaume Baixeries Juvilla. 2007. *Lattice Characterization of Armstrong and Symmetric Dependencies*. Ph.D. Dissertation. Universitat Politècnica de Catalunya (UPC).
[94] Samar K. Kassim, Hanan H. Shehata, Marwa M. Abou-Alhussein, Maha M. Sallam, and Islam Ibrahim Amin. 2017. Laboratory validation of formal concept analysis of the methylation status of microarray-detected genes in primary breast cancer. *Tumour Biol.* 39, 6 (June 2017), 1010428317698390. https://doi.org/10.1177/1010428317698390
[95] Mehdi Kaytoue et al. 2014. Biclustering meets triadic concept analysis. *Ann. Math. Artif. Intell.* 70, 1 (Feb. 2014), 55–79. https://doi.org/10.1007/s10472-013-9379-1
[96] Mehdi Kaytoue, Sébastien Duplessis, Sergei O. Kuznetsov, and Amedeo Napoli. 2009. Two FCA-based methods for mining gene expression data. In *Formal Concept Analysis (Lecture Notes in Computer Science)*, Sébastien Ferré and Sebastian Rudolph (Eds.). Springer, Berlin, 251–266.
[97] Mehdi Kaytoue, Sergei O. Kuznetsov, and Amedeo Napoli. 2011. Biclustering numerical data in formal concept analysis. In *Formal Concept Analysis (Lecture Notes in Computer Science)*, Petko Valtchev and Robert Jäschke (Eds.). Springer, Berlin, 135–150.
[98] Mehdi Kaytoue, Sergei O. Kuznetsov, Amedeo Napoli, and Sébastien Duplessis. 2011. Mining gene expression data with pattern structures in formal concept analysis. *Information Sciences* 181, 10 (May 2011), 1989–2001. https://doi.org/10.1016/j.ins.2010.07.007
[99] Mehdi Kaytoue-Uberall, Sébastien Duplessis, and Amedeo Napoli. 2008. Using formal concept analysis for the extraction of groups of co-expressed genes. In *Modelling, Computation and Optimization in Information Systems and Management Sciences (Communications in Computer and Information Science)*, Hoai An Le Thi, Pascal Bouvry, and Tao Pham Dinh (Eds.). Springer, Berlin, 439–449.
[100] Benjamin J. Keller et al. 2012. Formal concept analysis of disease similarity. *AMIA Jt. Summits Transl. Sci. Proc.* 2012 (March 2012), 42–51.
[101] Susan Khor. 2014. Inferring domain-domain interactions from protein-protein interactions with formal concept analysis. *PLOS ONE* 9, 2 (Feb. 2014), e88943. https://doi.org/10.1371/journal.pone.0088943
[102] Jana Klimpke and Sebastian Rudolph. 2021. Visualization of statistical information in concept lattice diagrams. In *Formal Concept Analysis (Lecture Notes in Computer Science)*, Agnès Braud, Aleksey Buzmakov, Tom Hanika, and Florence Le Ber (Eds.). Springer International Publishing, Cham, 208–223. https://doi.org/10.1007/978-3-030-77867-5_13
[103] Petr Krajca, Jan Outrata, and Vilem Vychodil. 2008. Parallel recursive algorithm for FCA. In *Proceedings of the 6th International Conference on Concept Lattices and Their Applications (CLA'08)*. 83–94.
[104] Petr Krajca, Jan Outrata, and Vilem Vychodil. 2010. Advances in algorithms based on CbO. In *Proceedings of the 7th International Conference on Concept Lattices and Their Applications (CLA'12)*, Vol. 672. 325–337.
[105] Sergei O. Kuznetsov. 2004. Machine learning and formal concept analysis. In *Concept Lattices (Lecture Notes in Computer Science)*, Peter Eklund (Ed.). Springer, Berlin, 287–312. https://doi.org/10.1007/978-3-540-24651-0_25
[106] Guoxuan Li. 2020. DeepFCA: Matching biomedical ontologies using formal concept analysis embedding techniques. In *Proceedings of the 4th International Conference on Medical and Health Informatics (ICMHI'20)*. ACM, New York, NY, 259–265. https://doi.org/10.1145/3418094.3418121
[107] Y. Li et al. 2015. Cancer gene expression data attribute partial ordered representation and knowledge discovery. In *2015 Fifth International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC'15)*. 861–865. https://doi.org/10.1109/IMCCC.2015.188
[108] Valentín Liñeiro-Barea, Jesús Medina, and Inmaculada Medina-Bulo. 2018. Generating fuzzy attribute rules via fuzzy formal concept analysis. In *Interactions between Computational Intelligence and Mathematics*, László T. Kóczy and Jesús Medina (Eds.). Springer International Publishing, Cham, 105–119. https://doi.org/10.1007/978-3-319-74681-4_7
[109] Stéphane Lopes, Jean-Marc Petit, and Lotfi Lakhal. 2002. Functional and approximate dependency mining: database and FCA points of view. *Journal of Experimental & Theoretical Artificial Intelligence* 14, 2–3 (April 2002), 93–114. https://doi.org/10.1080/09528130210164143
[110] Eugen Lounkine, Jens Auer, and Jürgen Bajorath. 2008. Formal concept analysis for the identification of molecular fragment combinations specific for active and highly potent compounds. *J. Med. Chem.* 51, 17 (Sept. 2008), 5342–5348. https://doi.org/10.1021/jm800515r
[111] Eugen Lounkine, Dagmar Stumpfe, and Jürgen Bajorath. 2009. Molecular formal concept analysis for compound selectivity profiling in biologically annotated databases. *J. Chem. Inf. Model.* 49, 6 (June 2009), 1359–1368. https://doi.org/10.1021/ci900095v

[112] Juraj Macko. 2013. User-friendly fuzzy FCA. In *Formal Concept Analysis (Lecture Notes in Computer Science)*, Peggy Cellier, Felix Distel, and Bernhard Ganter (Eds.). Springer, Berlin, 156–171. https://doi.org/10.1007/978-3-642-38317-5_10

[113] James Mathews et al. 2021. An exact test for significance of clusters in binary data. *arXiv:2109.13876 [math, stat]* (Sept. 2021). arXiv:math, stat/2109.13876

[114] Nida Meddouri and Mondher Meddouri. 2008. Classification methods based on formal concept analysis. *Proceedings of the 6th International Conference of on Concept Lattices and Their Applications*. 9–16.

[115] Raoul Medina and Lhouari Nourine. 2010. Conditional functional dependencies: An FCA point of view. In *Formal Concept Analysis (Lecture Notes in Computer Science)*, Léonard Kwuida and Barış Sertkaya (Eds.). Springer, Berlin, 161–176.

[116] C. Mencar and W. Pedrycz. 2020. Granular counting of uncertain data. *Fuzzy Sets and Systems* 387 (May 2020), 108–126. https://doi.org/10.1016/j.fss.2019.04.018

[117] Marek Menšík, Adam Albert, and Tomáš Michalovský. 2021. Using FCA and concept explications for finding an appropriate concept. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*. 49–60.

[118] Jean-Philippe Métivier, Bertrand Cuissart, Ronan Bureau, and Alban Lepailleur. 2018. The pharmacophore network: A computational method for exploring structure–activity relationships from a large chemical data set. *J. Med. Chem.* 61, 8 (April 2018), 3551–3564. https://doi.org/10.1021/acs.jmedchem.7b01890

[119] Alwersh Mohammed. 2021. Integration of FCA with fuzzy logic: A survey. *Multidiszciplináris Tudományok* 11, 5 (2021), 373–385. https://doi.org/10.35925/j.multi.2021.5.41

[120] Arati Mohapatro et al. 2021. A knowledge evocation model in grading healthcare institutions using rough set and formal concept analysis. In *Advances in Distributed Computing and Machine Learning (Lecture Notes in Networks and Systems)*. Springer, Singapore, 327–334. https://doi.org/10.1007/978-981-15-4218-3_32

[121] Susanne Motameny, Beatrix Versmold, and Rita Schmutzler. 2008. Formal concept analysis for the identification of combinatorial biomarkers in breast cancer. In *Formal Concept Analysis (Lecture Notes in Computer Science)*, Raoul Medina and Sergei Obiedkov (Eds.). Springer, Berlin, 229–240.

[122] Dariusz Mrozek. 2018. *Scalable Big Data Analytics for Protein Bioinformatics: Efficient Computational Solutions for Protein Structures*. Computational Biology, Vol. 28. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-98839-9

[123] Amedeo Napoli. 2005. Chapter 41 - A smooth introduction to symbolic methods for knowledge discovery. In *Handbook of Categorization in Cognitive Science*, Henri Cohen and Claire Lefebvre (Eds.). Elsevier Science Ltd., Oxford, 913–933. https://doi.org/10.1016/B978-008044612-7/50096-2

[124] Sergey Nurk et al. 2022. The complete sequence of a human genome. *Science* 376, 6588 (April 2022), 44–53. https://doi.org/10.1126/science.abj6987

[125] Segun Olatinwo et al. 2018. An ontology-based system for chronic tropical diseases using the protégé-OWL tool. *Songklanakarin J. Sci. Technol.* 40, 6 (Dec. 2018). 1386–1395.

[126] Debasmita Pal and Kartick Chandra Mondal. 2022. Predicting novel interactions from HIV-1-human PPI data integrated with protein signatures and GO annotations. *International Journal of Bioinformatics Research and Applications* 17, 6 (Jan 2022), 537–559.

[127] P. Pattaraintakorn, V. Boonjing, and J. Tadrat. 2008. A new case-based classifier system using rough formal concept analysis. In *2008 3rd International Conference on Convergence and Hybrid Information Technology*, Vol. 2. 645–650. https://doi.org/10.1109/ICCIT.2008.343

[128] Ruggero G. Pensa and Jean-François Boulicaut. 2005. Towards fault-tolerant formal concept analysis. In *Advances in Artificial Intelligence (Lecture Notes in Computer Science) (AI*AI'05)*, Stefania Bandini and Sara Manzoni (Eds.). Springer, Berlin, 212–223.

[129] John L. Pfaltz and Christopher M. Taylo. 2002. Closed set mining of biological data. In *2nd Workshop on Data Mining in Bioinformatics (BIOKDD'02)*.

[130] Jonas Poelmans, Dmitry I. Ignatov, Sergei O. Kuznetsov, and Guido Dedene. 2013. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications* 40, 16 (Nov. 2013), 6538–6560. https://doi.org/10.1016/j.eswa.2013.05.009

[131] Dustin Potter. 2005. *A Combinatorial Approach to Scientific Exploration of Gene Expression Data: An Integrative Method Using Formal Concept Analysis for the Comparative Analysis of Microarray Data*. Ph.D. Dissertation. Virginia Polytechnic Institute and State University.

[132] Uta Priss. 2013. Representing median networks with concept lattices. In *Conceptual Structures for STEM Research and Education (Lecture Notes in Computer Science)*, Heather D. Pfeiffer, Dmitry I. Ignatov, Jonas Poelmans, and Nagarjuna Gadiraju (Eds.). Springer, Berlin, 311–321.

[133] Olga Prokasheva, Alina Onishchenko, and Sergey Gurov. 2013. Classification methods based on formal concept analysis. In *Formal Concept Analysis Meets Information Retrieval (FCAIR'13)*. 103–112.

[134] Keyun Qin, Hong Lin, and Yuting Jiang. 2019. Local attribute reductions of formal contexts. *Int. J. Mach. Learn. Cyber.* (April 2019). https://doi.org/10.1007/s13042-019-00956-z

[135] Nancy Y. Quintero and Guillermo Restrepo. 2017. Formal concept analysis applications in chemistry: From radionuclides and molecular structure to toxicity and diagnosis. In *Partial Order Concepts in Applied Sciences*. Springer International Publishing, Cham, 207–217. https://doi.org/10.1007/978-3-319-45421-4_14

[136] Khalid Raza. 2017. Formal concept analysis for knowledge discovery from biological data. *IJDMB* 18, 4 (2017), 281. https://doi.org/10.1504/IJDMB.2017.10009312 arXiv:1506.00366

[137] Guillermo Restrepo. 2020. A formal approach to the conceptual development of chemical element. In *What Is A Chemical Element?* Oxford University Press, New York. https://doi.org/10.1093/oso/9780190933784.003.0013

[138] Guillermo Restrepo, Subhash C. Basak, and Denise Mills. 2011. Comparison of QSARs and characterization of structural basis of bioactivity using partial order theory and formal concept analysis: A case study with mutagenicity. *Curr. Comput. Aided Drug Des.* 7, 2 (June 2011), 109–121. https://doi.org/10.2174/157340911795677639

[139] Francois Rioult. 2003. Mining concepts from large SAGE gene expression matrices. In *KDID*.

[140] Tooba Salahuddin, Fatima Haouari, Fahad Islam, Rahma Ali, Sara Al-Rasbi, Nada Aboueata, Eman Rezk, and Ali Jaoua. 2018. Breast cancer image classification using pattern-based hyper conceptual sampling method. *Informatics in Medicine Unlocked* 13 (Jan. 2018), 176–185. https://doi.org/10.1016/j.imu.2018.07.002

[141] Mostafa A. Salama, Aboul Ella Hassanien, and Adel M. Alimi. 2013. Formal concept analysis approach for comparison between mutagenicity and carcinogenicity in cheminformatics. In *13th International Conference on Hybrid Intelligent Systems (HIS'13)*. 267–272. https://doi.org/10.1109/HIS.2013.6920494

[142] Amit Sangroya, C. Anantaram, Mrinal Rawat, and Mouli Rastogi. 2019. Using formal concept analysis to explain black box deep learning classification models. *What Can FCA do for Artificial Intelligence? (FCA4AI'19)*, 19–26.

[143] Grégoire Siekaniec, Emeline Roux, Téo Lemane, Eric Guédon, and Jacques Nicolas. 2021. Identification of isolated or mixed strains from long reads: A challenge met on Streptococcus thermophilus using a MinION sequencer. *Microb. Genom.* 7, 11 (Nov. 2021), 000654. https://doi.org/10.1099/mgen.0.000654

[144] Paula Silva et al. 2017. Formal concept analysis applied to professional social networks analysis. In *Proceedings of the 19th International Conference on Enterprise Information Systems*. SCITEPRESS - Science and Technology Publications, Porto, Portugal, 123–134. https://doi.org/10.5220/0006333401230134

[145] Pierre J. Silvie, Pierre Martin, Marianne Huchard, Priscilla Keip, Alain Gutierrez, and Samira Sarter. 2021. Prototyping a knowledge-based system to identify botanical extracts for plant health in sub-saharan africa. *Plants* 10, 5 (May 2021), 896. https://doi.org/10.3390/plants10050896

[146] Prem Kumar Singh. 2018. Medical diagnoses using three-way fuzzy concept lattice and their euclidean distance. *Comp. Appl. Math.* 37, 3 (July 2018), 3283–3306. https://doi.org/10.1007/s40314-017-0513-2

[147] D. Slezak and J. Wroblewski. 2006. Rough discretization of gene expression data. In *2006 International Conference on Hybrid Information Technology*, Vol. 2. 265–267. https://doi.org/10.1109/ICHIT.2006.253621

[148] Dominik Ślęzak and Jakub Wróblewski. 2007. Roughfication of numeric decision tables: The case study of gene expression data. In *Rough Sets and Knowledge Technology (Lecture Notes in Computer Science)*. Springer, Berlin, 316–323.

[149] William Spaulding and Jitender Deogun. 2011. A pathway to personalization of integrated treatment: Informatics and decision science in psychiatric rehabilitation. *Schizophrenia Bulletin* 37, Suppl_2 (Sept. 2011), S129–S137. https://doi.org/10.1093/schbul/sbr080

[150] Kyle Strimbu and Jorge A. Tavel. 2010. What are biomarkers? *Curr. Opin. HIV AIDS* 5, 6 (Nov. 2010), 463–466. https://doi.org/10.1097/COH.0b013e32833ed177

[151] K. Sumangali and Ch. Aswani Kumar. 2019. Knowledge reduction in formal contexts through CUR matrix decomposition. *Cybernetics and Systems* 50, 5 (July 2019), 465–496. https://doi.org/10.1080/01969722.2019.1602300

[152] K. Sumangali and Aswani Kumar Ch. 2019. Concept lattice simplification in formal concept analysis using attribute clustering. *J. Ambient Intell. Human Comput.* 10, 6 (June 2019), 2327–2343. https://doi.org/10.1007/s12652-018-0831-2

[153] Zejun Sun et al. 2017. Identifying influential nodes in complex networks based on weighted formal concept analysis. *IEEE Access* 5 (2017), 3777–3789. https://doi.org/10.1109/ACCESS.2017.2679038

[154] Adrian-Sorin Telcian, Daniela-Maria Cristea, and Ioan Sima. 2020. Formal concept analysis for amino acids classification and visualization. *Acta Universitatis Sapientiae, Informatica* 12, 1 (July 2020), 22–38. https://doi.org/10.2478/ausi-2020-0002

[155] Marwa Trabelsi, Nida Meddouri, and Mondher Maddouri. 2017. A new feature selection method for nominal classifier based on formal concept analysis. *Procedia Computer Science* 112 (Jan. 2017), 186–194. https://doi.org/10.1016/j.procs.2017.08.227

[156] B. K. Tripathy, Debi Acharjya, and V. Cynthya. 2011. A framework for intelligent medical diagnosis using rough set with formal concept analysis. *International Journal of Artificial Intelligence & Applications (IJAIA)* 2 (May 2011). https://doi.org/10.5121/ijaia.2011.2204

[157] X. Tu, Y. Wang, M. Zhang, and J. Wu. 2016. Using formal concept analysis to identify negative correlations in gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13, 2 (March 2016), 380–391. https://doi.org/10.1109/TCBB.2015.2443805

[158] Dean van der Merwe, Sergei Obiedkov, and Derrick Kourie. 2004. AddIntent: A new incremental algorithm for constructing concept lattices. In *Concept Lattices (Lecture Notes in Computer Science)*, Peter Eklund (Ed.). Springer, Berlin, 372–385. https://doi.org/10.1007/978-3-540-24651-0_31

[159] Frederick Johannes Venter. 2021. *Formal Concept Analysis Applied to Pattern Matching and Automata.* Thesis. Stellenbosch: Stellenbosch University.

[160] E. Vityaev et al. 2012. Probabilistic generalization of formal concepts. *Program. Comput. Soft.* 38, 5 (Sept. 2012), 219–230. https://doi.org/10.1134/S0361768812050076

[161] Mickaël Wajnberg, Petko Valtchev, Mario Lezoche, Alexandre Blondin-Massé, and Hervé Panetto. 2021. FCA went (multi-)relational, but does it make any difference? In *9th Workshop"What Can FCA Do for Artificial Intelligence?" Colocated with 30th International Joint Conference on Artificial Intelligence (IJCAI'21) (CEUR Workhop Proceedings)*, Vol. 2972. Montréal, Canada, 27–38.

[162] Y. Wan and L. Zou. 2019. An efficient algorithm for decreasing the granularity levels of attributes in formal concept analysis. *IEEE Access* 7 (2019), 11029–11040. https://doi.org/10.1109/ACCESS.2019.2892016

[163] Zhong Wang, Mark Gerstein, and Michael Snyder. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 1 (Jan. 2009), 57–63. https://doi.org/10.1038/nrg2484

[164] Rudolf Wille. 2009. Restructuring lattic theory: An approach based on hierarchies of concepts. In *Formal Concept Analysis (Lecture Notes in Computer Science)*, Sébastien Ferré and Sebastian Rudolph (Eds.). Springer, Berlin, 314–339.

[165] Yixuan Yang, Fei Hao, Doo-Soon Park, Sony Peng, Hyejung Lee, and Makara Mao. 2021. Modelling prevention and control strategies for COVID-19 propagation with patient contact networks. *Human-centric Computing and Information Sciences* 11, 45 (Dec. 2021), 15. https://doi.org/10.22967/HCIS.2021.11.045

[166] S. A. Yevtushenko. 2000. System of data analysis. In *Proc. 7th National Conference on Artificial Intelligence (KII'00)*, 127–134.

[167] Mengyi Zhao et al. 2018. Matching biomedical ontologies based on formal concept analysis. *Journal of Biomedical Semantics* 9, 1 (March 2018), 11. https://doi.org/10.1186/s13326-018-0178-9

[168] Fengbo Zheng, Rashmie Abeysinghe, and Licong Cui. 2021. Identification of missing concepts in biomedical terminologies using sequence-based formal concept analysis. *BMC Med. Inform. Decis. Mak.* 21, 7 (Nov. 2021), 234. https://doi.org/10.1186/s12911-021-01592-w

[169] Guangyong Zheng and Tao Huang. 2018. The reconstruction and analysis of gene regulatory networks. In *Computational Systems Biology: Methods and Protocols*, Tao Huang (Ed.). Springer, New York, NY, 137–154. https://doi.org/10.1007/978-1-4939-7717-8_8

[170] Jianqin Zhou et al. 2021. Concept and attribute reduction based on rectangle theory of formal concept. *arXiv:2111.00005 [cs]* (Oct. 2021). arXiv:cs/2111.00005

[171] Jianqin Zhou et al. 2021. A new algorithm based on extent bit-array for computing formal concepts. *arXiv:2111.00003 [cs]* (Oct. 2021). arXiv:cs/2111.00003

[172] S. Zidi. 2020. SVM and formal concept analysis for epidemics detection. In *2020 International Conference on Computing and Information Technology (ICCIT-1441'20)*. 1–6. https://doi.org/10.1109/ICCIT-144147971.2020.9213801

[173] Ligeng Zou et al. 2022. A new parallel algorithm for computing formal concepts based on two parallel stages. *Information Sciences* 586 (March 2022), 514–524. https://doi.org/10.1016/j.ins.2021.12.008

[174] Ligeng Zou, Zuping Zhang, and Jun Long. 2016. An efficient algorithm for increasing the granularity levels of attributes in formal concept analysis. *Expert Systems with Applications* 46 (March 2016), 224–235. https://doi.org/10.1016/j.eswa.2015.10.026

Received 12 October 2021; revised 25 June 2022; accepted 18 July 2022