Analyzing the composition of remedies in ancient pharmacopeias with FCA

Agnès Braud[1], Xavier Dolques[1], Pierre Fechter[2], Nicolas Lachiche[1], Florence Le Ber[1], and Véronique Pitchon[3]

(1) Université de Strasbourg, CNRS, ENGEES, ICube UMR 7357, F67000 Strasbourg
{agnes.braud,dolques,nicolas.lachiche}@unistra.fr
florence.leber@engees.unistra.fr
(2) Université de Strasbourg, CNRS, ESBS UMR 7242, F67000 Strasbourg
p.fechter@unistra.fr
(3) Université de Strasbourg, CNRS, Archimede UMR 7044, F67000 Strasbourg
pitchon@unistra.fr

Abstract. This paper presents the collaborative work led in an interdisciplinary project on studying remedies in Arabic medieval pharmacopeia. The goal is to find new molecules in plants or combinations of active principles to substitute them for antibiotics which encounter some limits. Formal Concept Analysis is used to discover co-occurrences of ingredients. We describe the difficulties inherent to these data and the results of preliminary analyses.

1 Introduction

This paper presents the preliminary work led in an interdisciplinary project that aims at studying remedies in Arabic medieval pharmacopeia. The project gathers researchers in history, microbiology, botanic and computer science. The general goal is to discover the active ingredients in remedies and to test them experimentally on bacteriae. The extracted knowledge could be used to substitute molecules in plants or combinations of active principles for antibiotics which encounter some limits.

In this work, we focus on medieval remedies prescribed for urinary problems. We present a dataset of remedies extracted from 5 pharmacopeias. This dataset is rather small, but the modeling is not straightforward. Although the building of the database is not finished, some analyses can already be done on the composition of remedies. Formal Concept Analysis (FCA) [3] appears as particularly suitable for this task, and preliminary results we obtained using it seemed very interesting to historians and microbiologists.

Section 2 presents the process for carrying out the project and in particular data collection. Section 3 describes the data, their characteristics and the difficulties resulting from them. In Sect. 4, we present our preliminary analyses and show the benefits of using FCA in this context. Finally, Sect. 5 yields some conclusions.

RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

2 Process overview

The whole process of this research work is presented on Fig. 1.

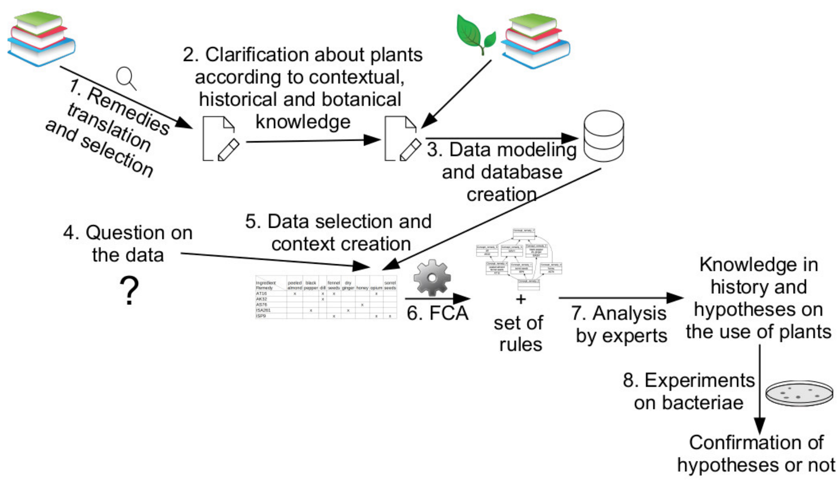

Fig. 1. Overview of the project process

The first step is to select pharmacopeias and remedies in them. This is done by an historian, who reads Arabic, on the basis of the symptoms indicated in the description of the remedy and on her knowledge of Arabic medicine. Let us note that Arabic medecine is based on the theory of Humours that considers that the body is constituted of four humours (water, fire, earth, air) having some qualities (hot, cold, dry, wet). The second step consists in clarifying some plants which are not described clearly enough in the text, with the help of a botanist. The next step is to build a database to store the remedies, whose modeling difficulties will be discussed in the next section. Then comes the formulation of questions that experts have on the data and that can be solved using FCA. The analysis of lattices and rules may bring new insights or questions both in history and on the use of plants in remedies. Some hypotheses on the use of plants stemed from this analysis will be studied in collaboration with a botanist (for the identification and characterization of therapeutic properties of plants) and a pharmacologist studying drugs based on natural substances (for the identification of the active molecules in plants), and then tested on bacteriae by microbiologists.

In this project, 5 pharmacopeias have been selected:

– two versions of the pharmacopeia of Sābūr ibn Sahl (9th century). Sābūr ibn Sahl was born in Jundishāpūr (Iran) and was chief of the hospital there before moving to Iraq. A version of the pharmacopeia [6] was for the pharmacists of ‘Aḍudī’s hospital in Baghdad (this pharmacopeia will be designated by

Fig. 2. Example of pages of manuscripts - Left: Dioscorides Pedanius of Anazarbos, Kitab Disqūrīdis (Dioscorides's book), circa 1100–1200, National Library of France. Right: Al-Kindī Folio 98r (photo credit: Aya Sofia, Istanbul, Ms. N° 3603)

ISA), and a shorter one [4] for pharmacists outside of the hospital who were considered less skillful (designated by ISP);

– the pharmacopeia of Al-Kindī (9th century) [8], who studied in Bassora and Baghdad. This pharmacopeia will be designated by AK;
– the pharmacopeia of Al-Tilmīdh (11th-12th centuries) [5], who was chief of 'Aḍudī's hospital in Baghdad. This pharmacopeia will be designated by AT;
– the pharmacopeia of Al-Samarqandī [9]. This pharmacopeia will be designated by AS.

The historian has based her choice firstly on the availability of books describing them, then preferably around Baghdad because of her knowledge on this area and also because Arabic pharmacopeias have a very rich content, and on a long period for a wide historical view. A few other criteria are linked to interesting historical questions.

Pharmacopeias are available as translations and/or original manuscripts. Figure 2 shows pages of two manuscripts. Both translations and original manuscripts may contain errors, so in the first case the historian may also have an original manuscript to check, and in the second case there may be errors made by copyists so that several different manuscripts are necessary. Among several hundreds of remedies, 38 prescribed for urinary symptoms were found. Among these, we removed 3 for this work: 1 from ISA which was already in ISP with exactly the same recipe, and 2 universal remedies with more than 50 ingredients (1 in ISP, 1 in ISA).

In the following, remedy identifiers will be prefixed by the acronym of the pharmacopeia in which they appear.

3 Data characteristics

The data that have been collected are rather small, but the collection requires a lot of time and the data are actually complex. The description of a remedy is composed of a list of ingredients (see Table 1 for an example) with quantities, and a description of the preparation. Let us note that part of the ingredients are present in the recipe to treat the disease, and some others may be present to treat their side effects, but this is not indicated.

Table 1. List of ingredients of a remedy

- Armenian borax	- Ginger	- Peeled sweet almonds
- Kerman cumin	- White pepper	- Rue leaves
- Parsley	- Scammony	- Dates from Hairūn

Ingredients belong to different categories: plant, mushroom, mineral, animal (like honey or musc) and metal. The description of plants is not homogeneous:

- some plant descriptions correspond to different taxonomic rank (e.g. species for Scammony, genus for Euphorbia, several families for Fern);
- some descriptions correspond to a plant, other to a part of a plant, or even a part of a part of... a plant (e.g., Acorn/Acorn shell/Inside peel of acorn shell, or Thymus without part while just a part is probably used). Let us notice that several parts of the same plant can be used in a remedy;
- some ingredients have some transformations associated (e.g., Peeled and grilled bitter almonds);
- some ingredients are a mix of ingredients (e.g., Persian za'atar);
- a few ingredients have an origin (e.g., Kerman cumin) and it is an important information since it usually reflects a high quality;
- some are still to determine (e.g., Idrūmaghmū).

Data modeling requires expert knowledge to determine which information is important to capture; for example, should Acorn and Inside peel of acorn shell be comparable? Moreover, for a work on active principles, having a precise description of plants is important as different species may have different properties. The exact species may be implicit for the author, who is used to a specific one or to the local one. A more precise information can be inferred on the species, either by checking the translation, searching for contextual information in books, or relying on illustrations present in the book. In this last case, the botanist may help as well as with knowledge on the parts of plants traditionally used. At the end, the discussions on data modeling have raised new questions. For example, after a discussion on the problem of representing acorn and the different parts used, the historian has decided to study further in books the use of oak.

All these points make that a relational database is not suitable to store these data. A graph database offers more flexibility to represent them and to adapt the representation to new cases. We have thus chosen this type of database. For the same reasons, building a context for FCA requires some choices among the information.

Currently, the data on which we work are the data issued from the translation of texts. The work to clarify species and parts of plants is in progress, led by the historian and the botanist, and will take some time, but it is still possible to start some analyses.

4 Preliminary analyses

For our preliminary analyses on the data, we have worked on two questions asked by the microbiologist: "which ingredients appear the most often?" (Q1) and "which ingredients often appear together?" (Q2). As shown in previous works [2], FCA-based approaches are very suitable to solve this type of questions.

We have worked both on the ingredients as they appear in the recipes, and on these ingredients with the ones coming from plants replaced just by the corresponding plant (without part and transformations, such as Oak for Acorn shell or Almond tree for Peeled sweet almonds). The plant can be at the level of species, genus, family... Let us note R the set of remedies, I the set of ingredients without modification and IP the set of ingredients with the ingredients coming from plants replaced by the plant, $contains \subseteq R \times I$ and $containsP \subseteq R \times IP$ two incidence relations describing the composition of remedies. We have thus built one context $CI = (R, I, contains)$ and another context $CIP = (R, IP, containsP)$. Let us recall that we work on 35 remedies concerning urinary symptoms. They are distributed in the 5 pharmacopeias as follows: 3 in AK, 5 in AS, 7 in AT, 6 in ISA and 14 in ISP. Their recipes contain between 2 and 21 ingredients giving 170 different ingredient descriptions in CI, and 144 attributes in CIP. The lattice obtained from CI ($\mathcal{L}_{CI}$) contains 138 concepts, the one obtained from CIP ($\mathcal{L}_{CIP}$) 151.

Starting with question Q1 ("which ingredients appear the most often?"), we study the lattices and in particular the concepts with the biggest extents. Table 2 shows some of the concepts obtained in $\mathcal{L}_{CI}$ and the ones in $\mathcal{L}_{CIP}$ corresponding to the same ingredients but considering the plant. The extent is presented so that the different pharmacopeias are separated. Concept ({*Celery seeds*}, {*AK114*, *AT16*, *AT20*, *AT72*, *ISA137*, *ISA162*, *ISP100*, *ISP16*, *ISP185*, *ISP186*, *ISP78*, *ISP9*}) from $\mathcal{L}_{CI}$ reveals that Celery seeds is the ingredient that appears the most often (12 times) in remedies. Then Indian nard appears 9 times, Saffron and Black pepper 7 times. Of course, it is possible to know which ingredients appear most often by computing statistics in the database, but the extension of the concept also shows that Celery seeds appears in 4 out of 5 pharmacopeias. Similarly, lattice $\mathcal{L}_{CIP}$ reveals Celery as appearing the most often in remedies. There is however a little change with Oak appearing with 8 occurrences, while Acorn appeared only 6 times in $\mathcal{L}_{CI}$. As can be seen in Table 2, this is due to the use of different parts of oak with different transformations (Acorn in 6 remedies, Acorn shell and Grilled inside peel of acorn shell in remedy AT137, Grilled acorn shell in AT80). There are thus 3 concepts introducing oak parts in $\mathcal{L}_{CI}$. Moreover, among the 6 remedies in the extension of the concept introducing Acorn shell, 2 are from ISA pharmacopeia and none from ISP while they are from the same author. This may be due to the fact that ISP was for pharmacists outside of the hospital who where considered less skillful than those from the hospital, and acorn shell is toxic and thus should be used carefully.

Table 2. Examples of concepts in $\mathcal{L}_{CI}$ and $\mathcal{L}_{CIP}$

<table>
<tr><th colspan="2">Concepts of $\mathcal{L}_{CI}$</th><th colspan="2">Concepts of $\mathcal{L}_{CIP}$</th></tr>
<tr><th>Intent</th><th>Extent (cardinality)</th><th>Intent</th><th>Extent (cardinality)</th></tr>
<tr><td>Celery seeds</td><td>AK114,
AT16, AT20, AT72,
ISA137, ISA162,
ISP100, ISP16, ISP185,
ISP186, ISP78, ISP9 (12)</td><td>Celery</td><td>AK114,
AT16, AT20, AT72,
ISA137, ISA162,
ISP100, ISP16, ISP185,
ISP186, ISP78, ISP9 (12)</td></tr>
<tr><td>Indian nard</td><td>AT137, AT72, AT80,
ISA137, ISA162, ISA261,
ISP16, ISP27, ISP78 (9)</td><td>Indian nard</td><td>AT137, AT72, AT80,
ISA137, ISA162, ISA261,
ISP16, ISP27, ISP78 (9)</td></tr>
<tr><td>Saffron</td><td>AT16,
ISP16, ISP186, ISP27,
ISP6, ISP70, ISP9 (7)</td><td>Saffron</td><td>AT16,
ISP16, ISP186, ISP27,
ISP6, ISP70, ISP9 (7)</td></tr>
<tr><td>Black pepper</td><td>AS78,
ISA137, ISA261,
ISP47, ISP54, ISP6,
ISP70 (7)</td><td>Black pepper</td><td>AS78,
ISA137, ISA261,
ISP47, ISP54, ISP6,
ISP70 (7)</td></tr>
<tr><td>Acorn</td><td>AS124b, ASN39,
AT137, AT81
ISA125, ISA231 (6)</td><td rowspan="3">Oak</td><td rowspan="3">AS124b, ASN39,
AT137, AT377, AT80,
AT81,
ISA125, ISA231 (8)</td></tr>
<tr><td>Cypress, Olibanum bark, Olibanum, Lavender, Cumin, Acorn shell, Grilled inside peel of acorn shell, Indian nard</td><td>AT137 (1)</td></tr>
<tr><td>Cypress, Olibanum bark, Olibanum, Lavender, Grilled acorn shell, Indian nard</td><td>AT80 (1)</td></tr>
</table>

For question Q2 (”which ingredients often appear together?”), association rules are a good starting point. We computed those with a minimal support of 3 remedies and a minimal confidence of 80%. For context CI, we obtained 15 rules. Among these rules, 10 were implication rules. Three of these implication rules are given in Table 3.

One appeared with a support of 7: *Celery seeds* → *Fennel seeds*, another one with a support of 6: *Opium* → *Saffron*. With a support of 4, we had 2

Table 3. Examples of implication rules obtained on context CI

Rule	**Support**
Celery seeds → Fennel seeds	7
Opium → Saffron	6
Asarabacca, Indian nard → Celery seeds	4

implication rules involving 3 ingredients such as: *Asarabacca*, *Indian nard* → *Celery seeds*. Context CIP leads to more rules: 24, among which 20 implication rules. If we consider the first rule (*Celery seeds* → *Fennel seeds*), we then explore the subconcepts of the concept introducing Celery seeds to see which ingredients other than Fennel seeds also appear with it. Figure 3 shows an extract of this part of the lattice. This analysis is interesting for microbiologists, in order to make hypotheses on the respective roles of the ingredients, and which ones play similar roles. Finally, experts were really interested by the results obtained with FCA on this first dataset.

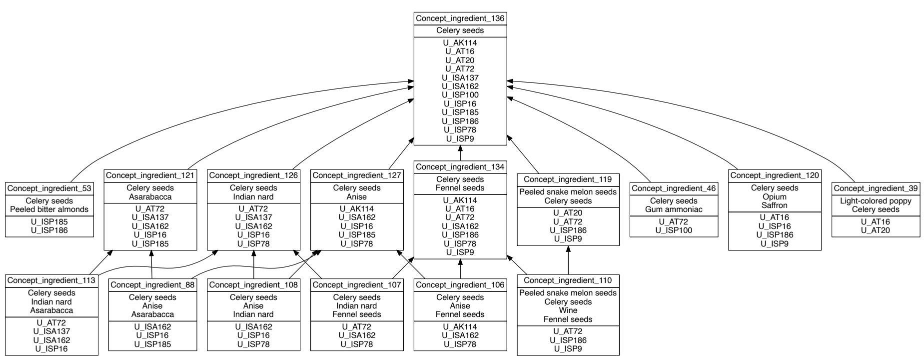

Fig. 3. Extract of lattice $\mathcal{L}_{CI}$ extracted from context CI

The work in [1] is based on the same objective of finding co-occurring ingredients but in a different pharmacopeia and for skin, mouth, or eye infections. It uses community detection in a network of ingredients. FCA is a more direct process that gives a complete view of the combinations of ingredients. Moreover the expert can see the remedies corresponding to each combination and exploit association rules. An FCA-based approach has been used in a similar project focusing on the identification of local plants for addressing sanitary problems in Sub-Saharan Africa [10, 7]. In this project, data represent plants or part of plants, and how, and in which country, they are used to treat animal or plant diseases. They are collected from various contemporary texts and concern contemporary practices, unlike our data for which we thus have more uncertainty.

5 Conclusion

This project highlights several interesting points about the work on such real data. First, the necessity and richness of adopting an interdisciplinary approach. Indeed, interdisciplinarity is required at each step of the process. Data preparation requires historians, botanists and computer scientists. The questioning on data leading to the analyses realized by computer scientists comes from historians and microbiologists. Finally, the results are discussed with historians and microbiologists, who will also perform experimental tests of interesting hypotheses. The questioning evolves all along the process with the discussions and new results. Second point, FCA appears as particularly suitable for expert needs, and even the simple analyses of lattices and rules presented in this paper have revealed useful insights.

The next steps of the project will aim at enriching the database, both by clarifying ingredients and by integrating other data like quantities for the ingredients, symptoms for which remedies are prescribed, preparation of the remedies, remedies for other kind of diseases, and by adding information on plants such as toxicity. More complex analyses will then be possible.

Currently, the amount of data the historian needs to work is rather small, however if the amount of data increases we may have to search for techniques to facilitate the exploration of the results.

References

1. Connelly, E., del Genio, C.I., Harrison, F.: Data mining a medieval medical text reveals patterns in ingredient choice that reflect biological activity against infectious agents **11**(1) (2020)
2. Couceiro, M., Napoli, A.: Elements About Exploratory, Knowledge-Based, Hybrid, and Explainable Knowledge Discovery. In: ICFCA 2019, Frankfurt, Germany. LNAI, vol. 11511, pp. 3–16. Springer (Jun 2019)
3. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag (1999)
4. Kahl, O.: The Small dispensatory: Sābūr ibn Sahl. Brill, Leiden (2003)
5. Kahl, O.: The dispensatory of Ibn at-Tilmīdh, Arabic text, English translation, study and glossary. Brill, Leiden (2007)
6. Kahl, O.: Sābūr ibn Sahl's Dispensatory in the Recension of the 'Aḍudī Hospital. Brill, Leiden; Boston (2009)
7. Keip, P., Gutierrez, A., Huchard, M., Le Ber, F., Sarter, S., Silvie, P., Martin, P.: Effects of Input Data Formalisation in Relational Concept Analysis for a Data Model with a Ternary Relation. In: ICFCA 2019, Frankfurt, Germany. LNCS, vol. 11511, pp. 191–207. Springer (Jun 2019)
8. Levey, M.: The Medical Formulary or Aqrābādhīn of Al-Kindī. Univ. of Pennsylvania Press, Philadelphia (1966)
9. Levey, M., al Khaledy, N.: Chemistry in the Medical Formulary of Al-Samarqandī. Univ. of Pennsylvania Press, Philadelphia (1967)
10. Silvie, P.J., Martin, P., Huchard, M., Keip, P., Gutierrez, A., Sarter, S.: Prototyping a Knowledge-Based System to Identify Botanical Extracts for Plant Health in Sub-Saharan Africa. Plants **10**(5) (2021)