Similarity Search for Multi-dimensional NMR-Spectra of Natural Products

Karina Wolfram[1], Andrea Porzel[2], and Alexander Hinneburg[1]

[1] Institute of Computer Science
Martin-Luther-University of Halle-Wittenberg, Germany
{wolfram, hinneburg}@informatik.uni-halle.de
[2] Leibniz Institute of Plant Biochemistry (IPB), Germany
aporzel@ipb-halle.de

Abstract. Searching and mining nuclear magnetic resonance (NMR)-spectra of naturally occurring products is an important task to investigate new potentially useful chemical compounds. We develop a set-based similarity function, which, however, does not sufficiently capture more abstract aspects of similarity. NMR-spectra are like documents, but consists of continuous multi-dimensional points instead of words. Probabilistic semantic indexing (PLSI) is an retrieval method, which learns hidden topics. We develop several mappings from continuous NMR-spectra to discrete text-like data. The new mappings include redundancies into the discrete data, which proofs helpful for the PLSI-model used afterwards. Our experiments show that PLSI, which is designed for text data created by humans, can effectively handle the mapped NMR-data originating from natural products. Additionally, PLSI combined with the new mappings is able to find meaningful "topics" in the NMR-data.

1 Introduction

Nuclear magnetic resonance (NMR)-spectra are an important finger printing method to investigate the chemical structure of organic compounds from plants or other tissues. Two-dimensional-NMR spectroscopy is able to capture the influences of two different atom types at the same time (e.g. ^{1}H, hydrogen and ^{13}C carbon). The result of an 2D-NMR experiment can be seen as an intensity function measured over two variables[1]. Regions of high intensity are called peaks, which contain the real information about the underlying molecular structure. The usual visualizations of 2D-NMR spectra are contour plots as shown in figure 1. An ideal peak would register as a small dot, however, due to the limited resolution available (dependent on the strength of the magnetic field) multiple peaks may appear as a single merged object with non-convex shape. In the literature peaks are noted by their two-dimensional positions without any information about the shapes of the peaks. Content-based similarity search of 2D-NMR spectra would be a valuable tool for structure investigation by comparing spectra of unknown compounds with a set of spectra, for which the structures

[1] The measurements are in parts per million (ppm).

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): PKDD 2006, LNAI 4213, pp. 650–658, 2006.

are known. While the principle is already in use for 1D-NMR spectra [5,4,1], to the best of our knowledge, no effective similarity search method is known for 2D-NMR-spectra.

Simplified, a 2D-NMR spectrum is a set of two-dimensional points. There is an analogy to text retrieval, where documents are usually represented as sets of words. Latent space models [3,2] were successfully used to model documents and thus improved the quality of text retrieval.

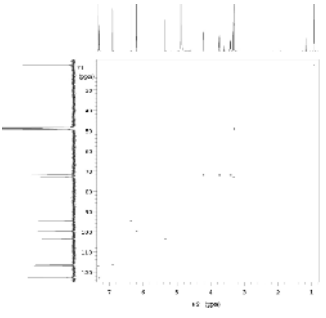

Fig. 1. 2D-NMR spectrum of quercetrin. The plots at the axes are the corresponding 1D-NMR spectra.

The contribution of this paper are methods to map 2D-NMR spectra to discrete text-like data, which can be analyzed and searched by any text retrieval method. Additionally, we propose a simple similarity function, which operates directly on the peaks of the spectra and serves as bottom line benchmark in the experimental evaluation. We demonstrate on real data that our mapping methods in combination with PLSI [3] improve the quality of similarity search of 2D-NMR spectra. Our results indicate at a larger scope that text retrieval and mining methods, designed for text data created by humans, in combination with appropriate mapping functions may yield the potential to be also successful for experimental data from naturally occurring objects. In this paper we consider exemplarily ^{1}H, ^{13}C one-bond heteronuclear shift correlation 2D-NMR spectra.

The paper is structured as follows: first, in section 2, we define 2D-NMR spectra and propose a simple similarity function. In section 3, we propose the new mapping functions for 2D-NMR spectra. In section 4, we describe our experimental evaluation and section 5 concludes the paper.

2 Directly Computing Similarity

A two-dimensional NMR-spectrum of an organic compound captures many structural characteristics like rings and chains. Most important are the positions of the peaks. As the shape of a peak and its height (intensity) strongly varies over different experiments with the same compound, the representation of a spectrum includes the peak positions only. A **2D NMR-spectrum** A is defined as a set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^2$. The $|\cdot|$ function denotes the size of the spectrum $|A| = n$. A peak matches other peaks only within a certain spatial neighborhood, which is defined by the ranges α and β. A peak x from spectrum A **matches** a peak y from spectrum B, if $|x.c - y.c| < \alpha$ and $|x.h - y.h| < \beta$, where $.c$ and $.h$ denote the NMR measurements for carbon and hydrogen respectively. Note that a single peak of a spectrum can match several peaks from another spectrum. Given two spectra A and B, the subset of peaks from A which find matching partners in B is denoted as $matches(A, B) = \{x\colon x \in A, \exists y \in B\colon x \text{ matches } y\}$.

The function *matches* is not symmetric, but helps to define a symmetric similarity measure Let be A and B two spectra and $A' = matches(A, B)$ and $B' = matches(B, A)$, so the **similarity** is defined as

$$sim(A, B) = \frac{|A'| + |B'|}{|A| + |B|}$$

The measure is close to one if most peaks of both spectra are matching peaks. Otherwise the similarity drops towards zero.

3 Mapping 2D-NMR-Spectra to Document-Like Data

Like a 2D-NMR spectrum consists of a set of peaks, a document consists of many words, which typically are modeled as a set. So assuming a 2D-NMR spectrum can be transformed into a text-like object by mapping the continuous 2D peaks to discrete variables, a variety of text retrieval models can be applied. However, it is an open question, whether models designed for quite different data, namely texts created by humans, are effective on data which comes for naturally occurring compounds and thus do not include human design patterns. Because the patterns which are important to 2D-NMR spectra similarity search might be quite different from pattern found in document collections, we chose a retrieval model which is capable of learning relevant patterns from training data. Probabilistic latent semantic indexing (PLSI) introduced in [3] is a model for text retrieval with such a learning ability. For 2D-NMR spectra similarity search it is not clear, what is the best way to map the peaks of a spectrum to discrete words.

In this section we propose different methods to map the peaks of an NMR-spectrum from the continuous space of measurements to a discrete space of words. With the help of such a mapping, methods for text retrieval like PLSI can be directly applied. However, the quality of the similarity search depend on how the peaks are mapped to discrete words.

3.1 Grid-Based Mapping

First, we introduce a simple grid-based method, on which we will build more sophisticated methods. A simple grid-based method is to partition each of the both axis of the two-dimensional peak space into intervals of same size. Thus, an equidistant grid is induced in the two-dimensional peak space and a peak is mapped to exactly one grid cell it belongs to. When a grid cell is identified by a discrete integer vector consisting of the cells coordinates the mapping of a peak $x \in \mathbb{R}^2$ is formalized as $g(x) = (g_c(x.c), g_h(x.h))$ with $g_c(x.c) = \left\lfloor \frac{x.c}{w_c} \right\rfloor$, $g_h(x.h) = \left\lfloor \frac{x.h}{w_h} \right\rfloor$ The quantities w_c and w_h are the extensions of a cell in the respective dimensions, which are parameters of the mapping. The grid is

centered at the origin of the peak space. The cells of the grid act as words. The vocabulary generated by the mapped peaks consists of those grid cells which contain at least one peak. Empty grid cells are not included in the vocabulary. A word consists of a two-dimensional discrete integer vector.

Unfortunately the grid-based mapping has two disadvantages. First, close peaks may be mapped to different grid cells. This may lead to poor matching of related peaks in the discrete word space. Second, peaks of new query spectra are ignored when they are mapped to grid cells not included in the vocabulary. So some information from the query is not used for the similarity search which may weaken the performance.

3.2 Redundant Mappings

We propose three mappings which introduce certain redundancies by mapping a single peak to a set of grid cells. The redundancy in the new mappings shall compensate for the drawbacks of the simple grid-based mapping.

Shifted Grids. The first disadvantage of the simple grid-based method is that peaks which are very close in the peak space may be mapped to different grid cells, because a cell border is between them. So proximity of peaks does not guaranty that they are mapped to the same discrete cell.

Instead of mapping a peak to a single grid cell, we propose to map it to a set of overlapping grid cells. This is achieved by several shifted grids of the same granularity. In addition to the base grid some grids are shifted into the three directions $(1,0)(0,1)(1,1)$. One grid is shifted in each of the directions by half of the extent of a cell. In general, there may be $k-1$ grids shifted by fractions of $1/k, 2/k, \ldots, k-1/k$ of the extent of a cell in each direction respectively. For the mapping of the peaks to words which consist of cells from the different grids, two additional dimensions are needed to distinguish (a) the $k-1$ grids in each direction and (b) the directions themselves. The third coordinate represents the fraction by which a cell is shifted and the fourth one represents the directions by the following coding: value 0 is (0,0), 1 is (1,0), 2 is (0,1) and 3 is (1,1). So each peak is mapped to a finite set of four-dimensional integer vectors. The mapping of a peak $x \in \mathbb{R}^2$ is

$$s(x) = \{(g_c(x.c), g_h(x.h), 0, 0)\} \cup \bigcup_{i=1}^{k-1} \big\{(g_c(x.c + i/k \cdot w_c), g_h(x.h), i, 1),$$
$$(g_c(x.c), g_h(x.h + i/k \cdot w_h), i, 2), (g_c(x.c + i/k \cdot w_c), g_h(x.h + i/k \cdot w_h), i, 3)\big\}$$

Thus, a single peak is mapped to $3(k-1)+1$ words. A nice property of the mapping is that there exists at least one grid cell for every pair of matching peaks both peaks are mapped to.

Different Resolutions. The second disadvantage of the simple grid-based mapping comes from the fact that empty grid cells (not occupied by at least

one peak from the set of training spectra) do not contribute to the representation to be learned for similarity search. So peaks of new query spectra mapped to those empty cells are ignored. That effect can be diminished by making the grid cells larger. However, this is counterproductive for the precision of the similarity search due to the coarser resolution. Thus, there are two contradicting goals, namely (a) to have a fine resolution to handle subtle aspects in the data and (b) to cover at the same time the whole peak space by a coarse resolution grid so that no peaks of a new query spectrum have to be ignored.

Instead of finding a tradeoff for a single grid, both goals can be served by combining simple grids with different resolutions. Given l different resolutions $\{(w_c^{(1)}, w_h^{(1)}), \ldots, (w_c^{(l)}, w_h^{(l)})\}$ a peak is mapped to l grid cells of different sizes. In order to distinguish between the different grids an additional discrete dimension is needed. So the mapping function is

$$r(x) = \bigcup_{i=1}^{l} \{(g_c^{(i)}(x), g_h^{(i)}(x), i)\}$$

with $g_c^{(i)}$ and $g_h^{(i)}$ use $w_c^{(i)}$ and $w_h^{(i)}$ respectively. Note that a hierarchical, quadtree like partitioning is a special case of the proposed mapping function with $w_c^{(i)} = 2^{i-1} w_c$ and $w_h^{(i)} = 2^{i-1} w_h$.

Combining shifted Grids with different Resolutions. Both methods are designed to compensate for different drawbacks of the simple grid mapping. So it is natural to combine both mappings. The parameters of such a mapping are the number of shifts k, the number of different grid cell sizes l and the actual sizes $\{(w_c^{(1)}, w_h^{(1)}), \ldots, (w_c^{(l)}, w_h^{(l)})\}$. Beside the two coordinates for the grid cells, additional discrete dimensions are needed for the shift, the direction and the grid resolution. Using the the definitions from above the mapping function of the combined mapping of a peak is

$$c(x) = \bigcup_{i=1}^{l} \{(g_c^{(i)}(x.c), g_h^{(i)}(x.h), 0, 0, i)\} \cup \bigcup_{j=1}^{k-1} \Big\{ (g_c^{(i)}(x.c + i/k \cdot w_c^{(i)}), g_h^{(i)}(x.h), j, 1, i),$$
$$(g_c^{(i)}(x.c), g_h^{(i)}(x.h + i/k w_h^{(i)}), j, 2, i), (g_c^{(i)}(x.c + i/k w_c^{(i)}), g_h^{(i)}(x.h + i/k \cdot w_h^{(i)}), j, 3, i) \Big\}$$

Thus a single peak is mapped to $l(3(k-1)+1)$ words. In the next section all mappings are compared with respect to the effectiveness for similarity search.

4 Evaluation and Results

The data used are mostly secondary metabolites of plants and fungi. The substances cover a representative area of naturally occurring compounds The database includes about 587 spectra, each has about 3 to 35 peaks. The total number of

Group	#Spectra	#Peaks
Pregnans	11	17–26
Anthrquinones	8	3–6
Aconitanes	8	22–26
Triterpenes	17	24–31
Flavonoids	18	5–8
Isoflavonoids	16	5–7
Aflatoxins	8	8–10
Steroids	12	16–23
Cardenolides	15	18–25
Coumarins	19	3–8

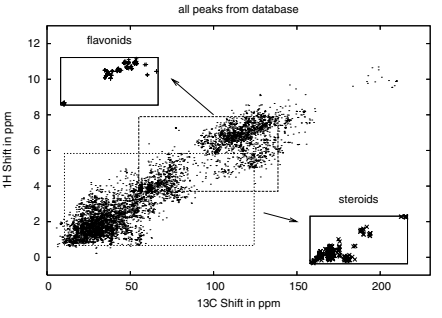

Fig. 2. Left: Groups with number of spectra and range of peaks, Right: Distribution of the peaks of all spectra with the distribution within the groups of flavonoids and steroids

peaks is 7029. Ten small groups of chemically similar compounds are included in the database for controlled experiments. The groups with the number of spectra and number of peaks are listed in figure 2 left. The peak space with all peaks in the database is shown in figure 2 right.

4.1 Comparison

The different methods for similarity search of 2D-NMR-spectra are compared using recall-precision curves. The search quality is high, when both – recall and precision – are high. So the upper curves are the best.

First, the direct similarity function is tested. Each spectrum from the ten groups is used as a query while the rest of the respective group should be found as answers. The plot in figure 3a shows averages over all queries. The size of the matching neighborhood is varied over $\alpha = 4, 6, 8, 10$ and $\beta = 0.4, 0.6, 0.8, 1.0$ respectively. As expected, the search quality is low. In fact on average, it fails to deliver a spectrum from the answer set in the top ranks which is indicated by the hill-like shape of the curves.

Next, a series of experiments is conducted using our proposed mapping functions in combination with PLSI. All curves are averages from cross validation over all groups. As the groups are very small the leave-one-out testing scheme is employed. The results for the simple grid-based mapping are shown in figure 3b. The sizes of the grid cells are varied over $w_c = 4, 6, 8, 10$ and $w_h = 0.4, 0.6, 0.8, 1.0$ respectively. The results are already much better than those for the direct similarity function. Small sizes give the best results. The use of shifted grids improves the performance substantially over simple grids, as shown in figure 3c,d. The plots show the experiments for $k = 2, 3$. The quality of $w_c = 4$ and $w_h = 0.4$ with $k = 2$ and $k = 3$ are almost identical. However, the vocabulary for $k = 2$ is much smaller, so the model has much less parameters to train. In practise, the smaller model with $k = 2$ shifts is favored.

Also the mapping based on grids with different grid cell sizes are assessed. Due to lack of space, only the results from combinations of $w_c^{(1)} = 4, w_h^{(1)} = 0.4$ with

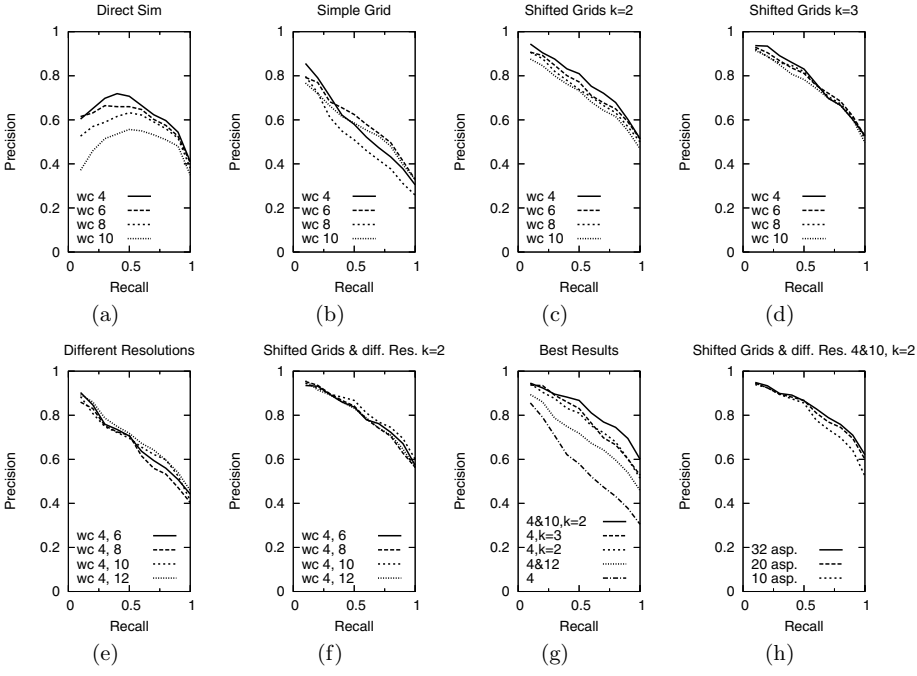

Fig. 3. Average recall-precision curves from leave-one-out cross validation experiments

other sizes are reported, because those performed best among all combinations. Figure 3e shows that also the mapping based on different grid cell sizes outperforms the simple grid-based mapping. But the improvement is not as much as for shifted grids. The set of resolutions $\{(w_c^{(1)} = 4, w_h^{(1)} = 0.4), (w_c^{(2)} = 12, w_h^{(2)} = 1.2)\}$ performs best.

Last, experiments are performed with the combination of the previous two mappings, namely a combination of shifted grids with those of different resolutions. The performance results are shown in figure 3f which indicates that the best combination, namely the resolution set $\{(w_c^{(1)} = 4, w_h^{(1)} = 0.4), (w_c^{(2)} = 10, w_h^{(2)} = 1.0)\}$ with $k = 2$ shifts, outperforms both previous mappings. This is more clearly seen in figure 3g which compares the best performing settings from the above experiments. In summery, the mappings based on shifted grids and those with different resolutions perform significantly better than the simple grid-based mapping. Finally, the combination of shifted grids and grids with different resolutions is even better than the individual mappings.

The last point is the number of hidden aspects. For the experiments reported so far, the PLSI model is used with 20 hidden aspects. Also different numbers of aspects are tested using the best combination of mappings. Figure 3h shows that the performance with 10 aspects drops a bit The increase in the numbers of aspects from 20 to 32 is only marginally reflected in increase of search performance. So 20 is a reasonable number of aspects for the given data. In

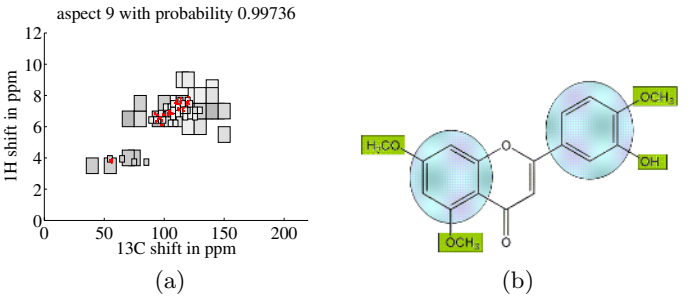

Fig. 4. (a) Main aspect of the flavonoid group which includes the region of aromatic rings (upper right cluster) and the region for oxygen substituents (lower left cluster). The gray shades indicate the strength of the association between grid cell and aspect. (b) An example of an flavonoid (3'-Hydroxy-5,7,4'-trimethoxyflavone) where the aromatic rings and the oxygen substituents (methoxy groups in this case) are marked.

conclusion, the results prove experimentally that the PLSI model, designed for text retrieval, is indeed effective for similarity search of 2D-NMR spectra from naturally occurring compounds.

4.2 Analysis of the latent Aspects

We analyzed the latent aspects learned by the PLSI model using the mapping based on the combination of shifted grids with different resolutions. The grid cells (words) with high probability for a given aspect are plotted together to describe the aspects meaning. Some aspects specialized on certain regions in the peak space which are typical for distinct molecule fragments like aromatic rings or alkane skeletons. However, also more subtle details of the data are captured by the aspect model. For example, the main aspect for the group of flavonoids specializes not only on the region for aromatic rings which are the main part of flavonoids. It also includes a smaller region which indicates oxygen substitution. A closer inspection of the database revealed that indeed many of the included flavonoids do have several oxygen substituents. The main aspect for flavonoids with the respective peak distribution of the flavonoid group is shown in figure 4a. We believe a detailed analysis of the aspects found by the model may help to investigate unknown structures of new substances when their NMR-spectra are included in the training set.

5 Conclusion

We proposed redundant mappings from continuous 2D-NMR spectra to discrete text-like data which can be processed by any text retrieval method. We demonstrated experimentally the effectiveness of the our mappings in combination with PLSI. Further analysis revealed that the aspects found by PLSI are chemically relevant. In future research we will study more recent text models like LDA [2] in combination with our mapping methods.

References

1. A. S. Barros and D. N. Rutledge. Segmented principal component transform-principal component analysis. *Chemometrics & Intelligent Laboratory Systems*, 78:125–137, 2005.
2. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
3. T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, 1999.
4. P. Krishnan, N. J. Kruger, and R. G. Ratcliffe. Metabolite fingerprinting and profiling in plants using nmr. *Journal of Experimental Botany*, 56:255–265, 2005.
5. C. Steinbeck, S. Krause, and S. Kuhn. Nmrshiftdb-constructing a free chemical information system with open-source components. *J. chem. inf. & comp. sci.*, 43: 1733 –1739, 2003.