Professor Witold Danikiewicz
Institute of Organic Chemistry PAS
ul. Kasprzaka 44/52
01-224 Warsaw
Poland

Warsaw, 20-09-2022

R E V I E W

of the PhD thesis of Michał Aleksander Ciach entitled "Algorithms for computational mass spectrometry based on the optimal transport theory" performed at the Faculty of Mathematics, Informatics and Mechanics of the University of Warsaw under the supervision of Professor Anna Gambin and at the Data Science Institute at Haselt University (Belgium) under the supervision of Professor Dirk Valkenborg

Qualitative and quantitative analysis of complex mixtures of organic compounds, mostly of biological origin, is currently one of the fastest growing sectors of analytical chemistry. The fields of proteomics and metabolomics are based on the results of analyzes of mixtures of proteins and peptides (proteomics) or small molecule compounds (metabolomics). The most widely used spectral methods in these analyzes are nuclear magnetic resonance (NMR) and mass spectrometry (MS), the latter most often combined with one of the chromatographic separation techniques: gas chromatography (GC-MS) for the separation of volatile compounds, i.e. low-molecular and non-polar or high performance liquid chromatography (HPLC-MS) to separate other types of compounds. The NMR method gives much greater certainty in identifying the components of a mixture, and also allows the determination of their quantitative proportions, but requires relatively large amounts of material, and also fails in the case of more complex mixtures. Therefore, it requires preliminary, preparative separation of the mixture under investigation by means of chromatographic techniques. In addition, it is characterized by a low dynamic range, i.e. it can be used to determine the quantitative proportions of compounds whose contents in the mixture do not differ by more than one order of magnitude. Mass spectrometry, on the other hand, is characterized by a very high sensitivity and a wide dynamic range (even up to 3 orders of magnitude), but - apart from the molecular weight and possibly the summary formula of the tested compound - it provides only limited information about the structure of the compound. However, due to the combination with one of the chromatographic techniques, it does not require preliminary separation of the mixture.

The main problem currently faced by researchers analyzing complex mixtures is the amount of data that needs to be interpreted. Therefore, bioinformatics, i.e. a field of science that uses computer algorithms, among others to support the analysis of complex mixtures of natural compounds, is currently developing intensively. The HPLC-MS technique can provide many gigabytes of data from a single measurement, therefore the automation of their interpretation is very important. The doctoral dissertation of M.Sc. Michał Ciach describes an attempt to use advanced mathematical algorithms for such automation.. He decided to use the theory of optimal transport and the concept of Wasserstein distance to analyze the mass spectra of complex mixtures of organic compounds. In particular, his aim was to develop algorithms

enabling the identification of ions from specific compounds in these mixtures. This is a very complex problem because the appearance of the spectrum essentially depends on the ionization method used. I am afraid that M.Sc. Ciach did not appreciate this problem, because he does not comment on it anywhere in his dissertation - I will return to this issue later in the review. Nevertheless, if certain conditions for recording spectra are met, the methodology proposed by the Author may turn out to be very useful. As a chemist, I am not able to substantively assess the formal correctness of the mathematical and IT apparatus used, which is why in my review I will focus primarily on assessing the usefulness of the methodology developed by M.Sc. Ciach in mass spectrometry.

However, before I proceed to the assessment of the scientific value, I must deal with the formal aspect of the reviewed dissertation. Even a cursory reading of the thesis indicates that it has more in common with the collection of scientific publications than a "classic" doctoral dissertation, as evidenced by the division into subsections, characteristic of scientific works rather than doctoral dissertations. In Chapter 1, the Author lists the scientific publications in which the results obtained during the research under the doctorate are described. The comparison of the content of these works with the relevant fragments of the dissertation clearly shows that very extensive fragments of the text and practically all the drawings from these works were transferred to the dissertation in an unchanged form. I would have no objections if M.Sc. Ciach chose the form of a doctoral dissertation permitted by Polish law as a collection of publications preceded by an introduction introducing the reader to the issues discussed in the dissertation. Of course, then the declarations of the co-authors of the publication would be necessary. However, M.Sc. Ciach chose a so far unknown to me the "hybrid" form, that is, incorporating significant fragments of the publication into the content of the dissertation. The plural number used throughout the dissertation, generally accepted in multi-author publications, but not in doctorates - at least in those with which I have dealt so far - unfortunately does not help the reader to find out what the Author did and what the co-authors of the publication did. In order to clarify my doubts, I turned to M.Sc. Ciach through his co-promoter, prof. Anna Gambin, to provide a statement describing the actual role of the Author in the creation of the publications on which the dissertation is based. I received such a declaration and on the basis of it I could say that his contribution was decisive or at least very significant, which completely dispelled my doubts.

M.Sc. Ciach's doctoral dissertation is written in English. It consists of seven chapters, the results obtained by the Author are described in chapters 3-6, and chapter 7 contains a summary and conclusions. This part of the work does not raise my objections, which I cannot say about the first two chapters. Chapter 1, entitled "Introduction," was intended to be an introduction to the dissertation. Unfortunately, it leaves a huge dissatisfaction. Since M.Sc. Ciach's dissertation concerns the analysis of mass spectra, much more should be written about this analytical method in the context of how the measurement techniques used affect the appearance of these spectra, because without knowledge about it, the entire analysis may not make sense. In particular, I mean the influence of various ionization methods on the appearance of the spectrum. For example, electron ionization (EI), as a high-energy method, provides spectra in which fragment ions are always present, and the molecular ion may be missing (and often is). On the other hand, the electrospray technique and related methods most often used for the analysis of polar, non-volatile compounds and often also with large and very large

molecules usually yield spectra in which only pseudo-molecular ions of the $[M + H]^+$, $[M + Na]^+$, $[M - H]^-$ type appear, and many more. Especially in the case of spectra of small molecules with not very high basicity - and this is the case in, for example, metabolomics - more than one type of pseudo-molecular ion is most often observed. While developing his method of mass spectra analysis, M.Sc. Ciach should be aware of this, because it is obvious that the simultaneous presence of e.g. $[M + H]^+$ and $[M + Na]^+$ ions in the spectrum significantly influences the assignment of the peaks in the spectrum to the appropriate compounds. I did not find information on this subject in the reviewed dissertation, and it is of key importance from the point of view of the practical applications of this method.

There is also a lack of information that I gave in the introduction to this review, i.e. the use of chromatographic separation methods of complex mixtures as an introduction to their analysis with spectral methods. If the author of the dissertation mentioned that in the case of complex mixtures, some components will always remain chromatographically unseparated and in such situations his method would be applicable, the dissertation would be placed in its proper context. For I cannot imagine (although perhaps my imagination is limited ...) that the method described by M.Sc. Ciach could be applied to the analysis of a mixture of several hundred or even only a few dozen ingredients, which is typical for samples of biological origin.

I omit in this review a number of errors and inaccuracies regarding mass spectrometry contained in this chapter - I have marked them in the form of comments in the PDF file of the dissertation, which I gave to the Author. Of course, I understand that mass spectrometry is not the subject of studies by M.Sc. Ciach, but it would be appropriate to read a good, contemporary textbook on mass spectrometry, and not refer to the classic, but already very archaic monograph by McLafferty and Tureček, or consult this part of the dissertation with a specialist from fields of MS.

Chapter 1, which follows, is in fact an extended summary of the dissertation. In this part, the author also lists the publications, printed and sent for printing, which contain the results of his research. By the way, at present the names of all the co-authors of the publication are given in the bibliography, and this must certainly be done in the case of unprinted works. Chapter 1 closes with a section entitled "Acknowledgment of scientific collaboration" in which M.Sc. Ciach thanks his colleagues. However, this is not a formal statement about the contribution of the work of the co-authors of the publication, but - as I wrote above - I received such a statement.

Chapter 2, entitled "Current approaches to analysis of spectral data" was intended to show - as the title indicates - the current state of knowledge in the field of the use of mathematical methods for spectral analysis, by implication - mass spectra. In fact, however, it repeats almost the entire content of the publication mentioned in the bibliography under number 2. This work concerns the analysis of nuclear magnetic resonance spectra. It does use the methodology based on the theory of optimal transport, which is the main subject of the dissertation, but this chapter gives the impression of being artificially attached. I would rather expect a discussion on known methods of mathematical analysis of mass spectra. This information does not appear until the beginning of the next three chapters.

The substantive part of the dissertation is included in chapters 3 - 6. Chapters 3 - 5 are based on the content of publication numbers 5 and 6 in the bibliography, but it should be emphasized that they contain much more information. In Chapter 3, M.Sc. Ciach describes the

spectral comparison algorithms based on the Wasserstein distance. This is an introduction to the more advanced methods described in the next two chapters. As I have already written, I am not able to evaluate the presented methodology from a mathematical point of view, and this chapter still lacks real-life examples. which would allow the assessment from the point of view of usefulness in mass spectrometry. In the next chapter, the Author describes the application of algorithms based on Wasserstein distance to deconvolution of mass spectra, which, in order to avoid terminological ambiguities, he describes as linear regression of mass spectra. Chapter 6 describes the methodology of linear regression of mass spectra, taking into account their imperfections, such as noise and signals from contamination of the tested sample. M.Sc. Ciach shows in this chapter. that the methodology proposed by him allows for effective identification of "foreign" peaks in spectra without losing the information needed to identify chemical compounds, the presence of which is expected in the tested sample. Importantly, the methodology can be used both for the analysis of centroid and profile spectra, which makes it possible to apply it to the analysis of other types of spectra, e.g. NMR. Undoubtedly, it is an interesting proposal addressed to researchers involved in the analysis of complex mixtures of organic compounds.

In the last, sixth chapter of the dissertation, M.Sc. Ciach describes the application of the optimal transport theory to the analysis of images obtained by means of surface imaging by mass spectrometry (MSI - Mass Spectrometry Imaging). The algorithms developed by the Author are used here to improve the uniqueness of the identification of ions from selected organic compounds on the surface of the test sample, which is most often a properly prepared cross-section of plant or animal tissue. The most common methods of ionization are Secondary Ion Mass Spectrometry (SIMS) and Matrix-Assisted Laser Desorption / Ionization (MALDI). Both methods are characterized by a large scatter of the relative intensities of the peaks between consecutive measurements, which creates major interpretation problems, especially in the case of the presence of numerous compounds in the sample. The methodology described by the Author allows for a significant improvement in the quality of the obtained images, which is very important in biochemical and biomedical research. Therefore, I rate this part of the dissertation the highest by far. I believe that MSI image analysis is the field in which the application of algorithms based on the optimal transport theory can bring the most benefits. According to the information provided in the dissertation, a publication on this subject, prepared in cooperation with the team of prof. Olga Vitek from Northeastern University is under review. I do not have access to the manuscript of this work, but when reading chapter 6 it is difficult to resist the impression that it is simply this publication, as evidenced by the division into subsections, characteristic of scientific papers rather than doctoral dissertations. Of course, this does not diminish the substantive value, but it would be appropriate to write about it in the dissertation and define the M.Sc. Ciach contribution to its creation.

The last, seventh chapter of the dissertation contains a concise summary and conclusions. The doctoral student emphasizes that the methodology of spectrum analysis developed by him, using the theory of optimal signal transport between spectra: experimental and theoretical spectra of mixture components, can be used not only to analyze mass spectra of mixtures of many compounds, but also spectra obtained by means of other spectral methods, including nuclear magnetic resonance. The main mathematical tool is the Wasserstein distance, which quantifies the difference between two spectra as the minimum distance on the *m/z* axis

(for mass spectra) needed to match their signals. The final part of the dissertation is a list of cited literature containing 101 items.

Summing up the review of the doctoral dissertation by Michał Ciach, I must say that it evokes mixed feelings. On the one hand, I have no doubts that the developed methodology of spectrum analysis makes a significant contribution to this important research topic and may find numerous practical applications. The fact that some of the results were published in high-ranked international journals confirms that the methodology is mathematically correct, which I, as a chemist, cannot state. Nevertheless, from the point of view of mass spectrometry, the work has quite a few shortcomings. First of all, the Author treats such problems quite nonchalantly, such as the possibility of the presence of various pseudo-molecular ions from the same compound in the spectrum (e.g. $[M + H]^+$ and $[M + Na]^+$ ions) and fragmentation ions. It is not uncommon for the fragmentation ion of compound A to be identical to the molecular (or pseudo-molecular) ion of compound B. Another problem concerns the accuracy with which the values of *m/z* ions are reported. Sometimes they are nominal masses, rounded to whole numbers, and sometimes values obtained as a result of recording high resolution spectra. Therefore, I believe that before the presented methodology becomes a tool for chemists, further research will be necessary in close cooperation with specialists in the field of mass spectrometry. As far as the non-substantive evaluation is concerned, it is worth emphasizing the very careful preparation of diagrams and figures and even professional typographic layout. The only significant mistake I found concerns the incorrect notation of the averagine formula on page 46. It is worth noting that M.Sc. Ciach uses the English language, in which the dissertation is written, very efficiently.

On the other hand, however, the work in the form in which I received it raises many formal reservations regarding the use of fragments of the publication and the actual contribution of the Author to the results described in it - I have already written about them above. However, they were explained in the Author's statement that I received, so the dissertation, including the statement, is now fully acceptable to me.

In conclusion, I state that the reviewed doctoral dissertation of M.Sc. Michał Ciach meets the requirements set out in the Act on Higher Education and Science of July 20, 2018 (as amended) and therefore I refer to the Scientific Council of Mathematics and Informatics Disciplines of the University of Warsaw with application for admission of Michał Ciach to the next stages of the procedure leading to receiving of the PhD title.

Witold Aleksander Danikiewicz

Elektronicznie podpisany przez Witold Aleksander Danikiewicz
Data: 2022.09.20 10:30:03 +02'00'