A STATISTICAL FRAMEWORK FOR BIOMARKER IDENTIFICATION USING HR-MAS 2D NMR SPECTROSCOPY

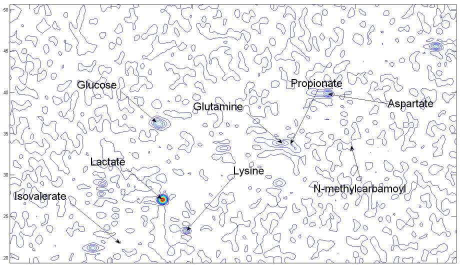
A. BELGHITH[1], C. COLLET[2], K. ELBAYED[3], L. RUMBACH[4], I. NAMER[5], and J-P. ARMSPACH[6]

[1]University of Strasbourg, LSIIT - CNRS UMR 7005, Strasbourg, Alsace, France, [2]University of Strasbourg, LSIIT - CNRS UMR 7005, France, [3]University of Strasbourg, Institut de Chimie, [4]Neurology Department CHU Minjoz Besancon -France, [5]University of Strasbourg, LINC - CNRS FRE 3289 - France, [6]University of Strasbourg, LINC - CNRS FRE 3289, France

INTRODUCTION: Metabolomics is an exponentially growing field of 'omics' research concerned with the comparison, identification and quantification of large numbers of metabolites in biological system. This emergent science of metabolomics brings increasing promise to identify biomarker diseases including biochemical changes in disease and to predict human reaction towards treatments. In this context, the 2D High Resolution Magic Angle Spinning (HR-MAS) Nuclear Magnetic Resonance (NMR) spectroscopy has emerged as an ideal platform for studying metabolites of biopsies. Indeed, the 2D HR-MAS NMR spectroscopy is considered as a revolutionized tool for the study of chemistry and biochemistry and has become arguably the single most widely used technique to elucidate the relationships between clinically relevant cell processes and specific metabolites in order to identify diseases. In this study, we particularly focus on the 2D $H^1 - C^{13}$ Heteronuclear Single Quantum Coherence (HSQC) NMR spectrum analysis.

The metabonomic analysis requires comparison of metabolite profiles obtained from multiple replicates of samples exposed to different experimental conditions. What adds difficulty to automating this analysis process is that each peak of a given metabolite (*i.e.* a metabolite is presented as a set of peaks with specified locations) can be shifted slightly from one sample to the next. The primary causes of chemical shifts in peak positions are variations in the pH and the temperature of the sample. In this study, we propose a new scheme to detect and align simultaneously peaks in different spectra. The peak detection and alignment result will be then used to identify biomarkers present in the biopsy. The method was validated on real HSQC spectra.

METHODS: In this study, each peak will be parameterized by its positions, its amplitude and its shape. These characteristics are theoretically invariable for the same metabolite but, in practice, the noise induces peak shape differences and peak locations variations. To eliminate these effects, one needs to propose an alignment step: to do that, we proposed a new vision of the alignment problem. This approach assumes that any modification of the peak location is an imprecision which is added to the spectra. It is important to note that the notion of imprecision and uncertainty are distinct ones and, unfortunately, both terms are mostly confused. Indeed, the notion of the uncertain means that a datum can be precise, but its realization is not sure. This uncertainty can be sufficiently modeled by the Bayesian theory. On the other hand, the notion of imprecision comes from the fact that we do not have enough knowledge on the datum, thus we describe it with vague terms but its realization is sure. In order to model and handle both imprecision and uncertainty, the use of the evidence theory which is a generalization of the classic Bayesian theory may be chosen [1]. Since the bottleneck for any method based on evidence theory is the definition of the mass function, we have proposed a new approach to model the mass function based on the fuzzy quantification of the imprecision degree.



Once peaks are detected and aligned, we address the problem of biomarker identification. In this study, the biomarker identification is obtained by comparing 2D NMR spectral patterns in the NMR spectrum of the biopsy with specific library coding reference spectra of pure metabolites denoted the *corpus*. In the literature, methods addressing the biomarker identification problem exploit only the location of peaks for metabolite identification [2]. To this end, they use threshold methods to accommodate the chemical shift differences between the different observed NMR spectra and the corpus. Nevertheless, the choice of the thresholds may strongly affect the robustness of the annotation method. To overcome this problem, we will use the fuzzy set theory which is appropriate to handle fuzzy situations [3]. The proposed method is more flexible and general. It allows the representation of both fuzziness and uncertainty within the inference process. The interest of such approach is in its capacity to take into account information brought by the measures and the *a priori* information that we have. The adequacy between the estimated parameters and the model is described thanks to a cost function to maximize.

RESULTS: We now deal with the problem of result validation on real HR-MAS 2D spectra. Our data base contains five 2D spectra from healthy colorectal tissues biopsy and five other spectra from cancerous colorectal tissues biopsy. Since no ground truth is available in this case to evaluate the quality of the proposed method, the peaks detection and alignment results and the biomarker identification results were manually examined by an expert investigator to validate the generated results. Note that the used corpus contains forty referenced metabolites given by the physicians. The manual inspection displayed that all metabolites were correctly detected with no error for the ten spectra. This means that our method is a robust unsupervised alternative to manual metabolite annotation requiring high expert time. The figure illustrates an example of metabolite (biomarker) identification result on 2D healthy colorectal tissues biopsy spectrum.

CONCLUSION AND DISCUSSION: In this paper we present a new statistical framework for biomarkers identification of human tissues using HR-MAS 2D NMR spectroscopy. The majors challenges for automatic metabolites identification are i) the spectral complexity inherent in many tissues can lead to a large number of peaks confined to a relatively chemical shift range ; ii) the handling of chemical shift changes introduced by the variation of pH and temperature. The use of evidence theory for peaks detection and alignment and the fuzzy set theory for biomarker identification increase the efficiency of our scheme. This method was validated on real spectra with the collaboration of NMR experts.

The proposed method offers not only a powerful automated tool for peaks detection and alignment but also a parametric representation of the NMR 2D spectrum which will be used to annotate spectra and to compare metabolite profiles obtained from different biopsies. The MATLAB (The Mathworks Inc.) implementation of the algorithm is available upon request.

REFERNCES: [1] Shafer G: A mathematical theory of evidence 1976. [2] J. Xia, T.C. Bjorndahl, P. Tang, D.S.Wishart, “MetaboMiner – semi-automated identification of metabolites from 2 D NMR spectra of complex biofluids,” *BMC bioinformatics*, vol. 9, no. 1, pp.507, 2008. [3] E. Waltz, J. Llinas, *Multisensor data fusion*, Artech House Boston, London, 1990.