1D10

Information-theoretical Data Analysis Approaches to Raman Micro-spectroscopic Images

Khalifa Mohammad Helal[1], Harsono Cahyadi[4], J. Nicholas Taylor[2, 3], Akira Okajima[5], Yasuaki Kumamoto[5], Hideo Tanaka[5], Yoshinori Harada[3, 5], Tamiki Komatsuzaki[1, 2, 3]
[1]Graduate School of Life Science, Hokkaido University, Japan; [2]Research Institute for Electronic Science, Hokkaido University, Japan; [3]JST/CREST; [4]Dept. of Methodologies for Medical Research, Kyoto Prefectural Univ. of Medicine, Japan; [5]Dept. of Pathology and Cell Regulation, Kyoto Prefectural Univ. of Medicine, Japan

[Abstract] Raman micro-spectroscopic imaging is a non-invasive and label-free live cell imaging technique where the images contain spatial and spectral information of molecular vibrations, e.g., in single cells. The spectral features are unique to specific bio-molecules such as cytochrome C, lipids, etc. As Raman images contain much information but signals in raw spectra are weak, development of useful data processing methods are needed. Here, we develop information-theoretical data analysis approaches to Raman micro-spectroscopic images. Data obtained from two different disease types of rat hepatic steatosis were analyzed. Due to background contamination and large noise fluctuations in raw spectra, we preprocessed the data with several data analysis tools such as recursive polynomial fitting, singular value decomposition, etc., to enhance discriminating features of them. An information-theoretical soft clustering method using rate-distortion theory is used to classify the disease types. We show how Raman imaging is useful for differentiating cellular conditions.

Keywords: Raman imaging, Preprocessing, Clustering, Information theory, Rate-distortion theory.

[Introduction] Raman imaging is expected to provide a new means to diagnose diseases and differentiate the types of diseases even in the case that morphologies in single cells may be unaltered. Raman imaging gives structural and chemical information about not only specific molecules but also the whole sample being analyzed, so that one does not need to identify marker proteins. The data are hyperspectral Raman images where each pixel represents a spectrum containing rich molecular information. Multivariate statistical methods are very useful for processing of Raman and IR spectral data because of their ability to analyze the vast spectral distribution and discriminate between spectra of different samples that show only very minor changes [1, 2]. Preprocessing of the Raman data is required to reduce effects of unwanted signals [1, 3]. It is thus needed to establish a theoretical and algorithmic platform to differentiate the underlying cell types, the stage of disease, and prediction of disease by referencing spectral differences in terms of molecular fingerprints buried in Raman signal.

[Brief concept of clustering analysis in Raman shift feature space] Figure 1 provides a sketch of spatiotemporal Raman spectra in a high dimension where each point corresponds to

a Raman spectrum after application of appropriate preprocessing. To quantify differences among the spectra, we choose the Manhattan (l_1) distance metric, which is known to provide the best discrimination between the different points in high dimensional data spaces [4]. The Manhattan distance between spectra S_i and S_j is $d_{ij} = \sum_w |S_i(w) - S_j(w)|$, where w is the spectral dimension (wavenumber). The information-theoretical rate-distortion theory [5, 6] is implemented as a clustering scheme, enabling the low S/N ratio of Raman spectra to be better considered in comparison to a standard clustering scheme. This theory clusters a set of spectra into a smaller number of groups via minimization of the functional $\mathcal{F} = I(C;S) + \beta\langle d \rangle$ with respect to the conditional probability $p(C_k|S_i)$, through an iterative calculation. Here C denotes a set of clusters to be obtained, S denotes a given set of spectra, $I(C;S)$ is the mutual information between C and S, $\langle d \rangle$ is the mean distortion among all spectra within each cluster and averaged over all clusters, and β is a Lagrange multiplier. From this feature space, cell types, conditions, and time evolution, in terms of Raman fingerprints of molecules in the system, can be classified. In this presentation, we show our recent analysis on liver tissues of different disease types of rat hepatic steatosis. The identification of disease types of rat hepatic steatosis so far relies mainly on the morphological information of tissues. We discuss how Raman spectra can differentiate and predict disease-related changes in biomolecular composition of the liver tissue.

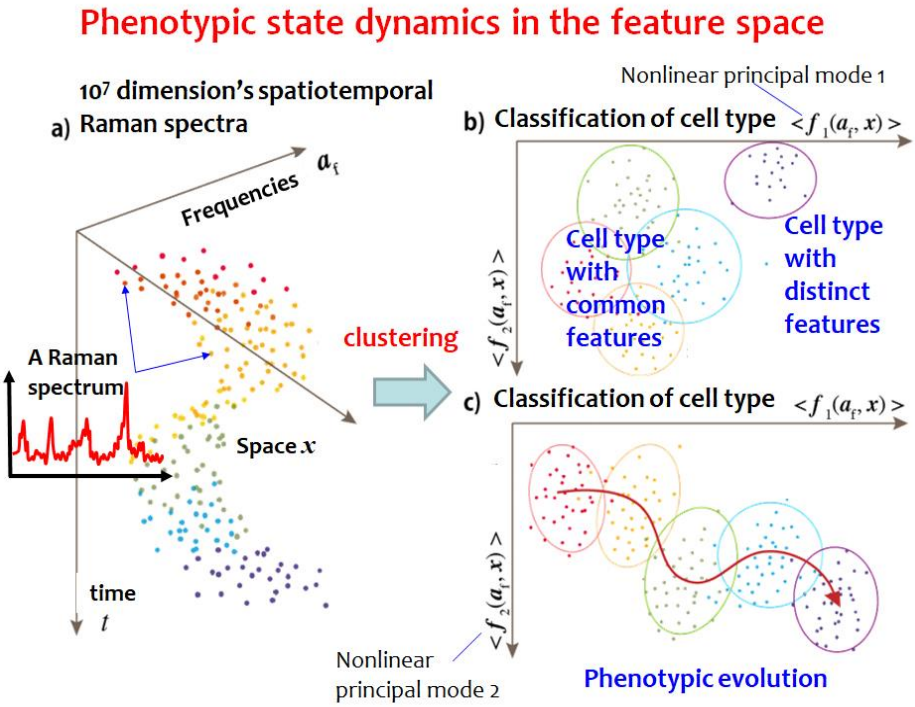

Figure 1

[References]

[1] Gautam *et al. EPJ Tech. Instrumentation.* (2015) 2:8
[2] O'Connell *et al. Appl Spectrosc.* (2010) 64:1109–21
[3] T. Bocklitz *et al. Anal. Chim Acta.* (2011) 704:47–56.
[4] C. Aggarwal, A. Hinneburg and D. Keim. On the surprising behavior of distance metrics in high dimensional space. In *Proceedings of 8th International Conference on Database Theory (ICDT),* (2001).
[5] C. Shannon. A Mathematical Theory of Communication. *Bell Sys. Tech. J.* **27**, 379-423, 623-656 (1948)
[6] J. N. Taylor *et al. Sci. Rep.* (2015) 5:9174