Antros harmonikos generacijos mikroskopijos skydliaukės mazgų vaizdų svarbiausių statistinių parametrų įvertinimas

Evaluation of key statistical parameters of second harmonic generation microscopy images of thyroid nodules

Yaraslau Padrez[1], Lena Golubewa[1], Igor Timoshchenko[1], Danielis Rutkauskas[1]
[1]State Research Institute Center for Physical Sciences and Technology, Department of Molecular Compound Physics, Savanorių Ave. 231, LT-02300 Vilnius, Lithuania
yaraslau.padrez@ftmc.lt

Cancer is one of the leading causes of death worldwide. Its progression is often associated with the extracellular matrix (ECM) remodeling. The main protein component of ECM is collagen. As cancer develops, collagen structure, texture and content change significantly. These specific changes can serve as indicators of the disease severity and tumor metastatic potential and allow prediction of patient survival. Collagen, being a non-centrosymmetric molecule with high hyperpolarizability and, thus, effectively producing second-order non-linear signal in response to femto-second laser irradiation, represents a unique target for a label-free analysis based on second harmonic generation (SHG) microscopy.

To identify statistically significant patterns of changes in collagen structure, large amounts of microscopic data are to be analyzed. This task is often complicated by the huge number of image parameters being considered and the space of features becomes multidimensional. One way to simplify the task is to apply principal component analysis (PCA). PCA is a method which allows reducing the dimensionality of large data sets by transforming a large set of variables into a smaller one that contains most of the information in the large set.

The purpose of the present study was (i) to apply PCA for the reduction of the dimensionality of the space of features extracted from the SHG images of collagen capsules of the papillary thyroid carcinoma (PTC) nodules and (ii) to reveal the key texture features of SHG images of collagen specifically characterizing PTC.

In this study a data set of 20736 SHG images of PTC nodules were obtained by means of wide-field SHG microscopy [1]. To provide an objective analysis of a large data set, the following statistical approaches were used:

- First Order Statistics (FOS), based on the gray level intensities of the SHG images;
- Second Order Statistics (SOS), based on the analysis of the spatial relationship between gray levels of neighboring pixels reflected in Gray Level Co-occurrence Matrix (GLCM) [2];
- Higher Order Statistics (HOS), based on the analysis of the spatial distribution of gray levels of pixels with the emphasis on the groups of pixels with the same intensity reflected in a Gray Level Run Length Matrix (GLRLM) [3].

Calculated parameters included: FOS – mean, standard deviation, skewness and kurtosis; SOS – energy, contrast, correlation, homogeneity and entropy (GLCM calculation was performed for five distances between adjacent pixels: 1, 3, 6, 9, 12); HOS – short runs emphasis, long runs emphasis, gray level nonuniformity, run length nonuniformity and run percentage. Thus, 34 statistical parameters (features) were obtained for each SHG image from the experimental dataset. All features were standardized, and PCA based on covariance matrix was performed. Principal components (PCs) containing most information (variance) of initial features were calculated. To choose optimal number of PCs, a Kaiser's Rule was applied. According to it, the first four PCs describing 93% of data variance remained, thus, providing the reduction of the dimensionality of the initial space of features from 34 to 4. Being linear combinations of initial features, PCs make it possible both to reduce the dimension of features (down to those containing most of data variance), and to extract impact of initial features to maximal data variance. As a result, the most of variance in SHG images of PTC nodules is enclosed in energy, homogeneity, and correlation (SOS), short run emphasis and run percentage (HOS), (mentioned in descending order of their impact to PC1-PC4). Such data set reduction significantly simplifies any further analysis of patterns of collagen capsule changes if using these new PCs as new variables. Application of PCs as input features for K-means clustering of SHG images of PTC nodules will simplify calculations, reduce calculation time, and support estimation of characteristic parameters of capsular invasion. These estimations may be applied for explanation how each specific feature coheres with observed pathological changes and provide understanding of K-means-based reasoning.

Key words: collagen, second harmonic generation microscopy, principal component analysis.

Literature

[1] A. Dementjev, R. Rudys, R. Karpicz, D. Rutkauskas, Lith. J. Phys. **60**, 145–153 (2020).
[2] R. M. Haralick,K. Shanmugam, I. Dinstein, **6**, 610–621 (1973).
[3] M. M. Galloway, Computer Graphics and Image Processing, **4**, 2,172-179 (1975).