Unsupervised analysis of staining quality in histopathological images

Fabi Prezja[1*], Teijo Kuopio[2,3], Sami Äyrämö[1]

[1]Faculty of Information Technology, University Of Jyväskylä, Finland
[2]Department of Biological and Environmental Science, University of Jyväskylä, Finland.
[3]Department of Pathology, Central Finland Health Care District, Jyväskylä, Finland

* fabi.f.prezja@jyu.fi

Hematoxylin and eosin (H&E) staining is vital for emphasizing important histological information for pathologists. In practice, staining results vary from laboratory to laboratory, although the staining method is basically the same. We used a large staining quality control material to quantify the extent of the variation. A total of 66 laboratories were sent an unstained slide with a histologic sample from skin, colon, and kidney. The laboratories stained the slides and returned them to the organizer of the quality control round. After that we scanned the slides and used unsupervised machine learning methods to discover, model, and analyze the stainings. We aimed to cluster data without morphological features, inspect cluster distance and overlap. We cropped all tissue types into separate images, resulting in 66 x 3 = 198 images. Each image was described by three (Red, Green, Blue) color histograms (256 bins per channel). The color histograms were concatenated and transformed onto principal components by PCA, and the components (accounting for 95 % of the total variance) were input to the K-means algorithm. After grid-search between 2 – 24 K-means clusters, the K=2 and K=3 configuration exhibited the highest average silhouette score. In total all average silhouette scores were relatively low, indicating limited separability in PCA space. Our primary explanation for this result is that such techniques may provide substantial visual clues when morphological structures are present, but not in vice-versa.