Motivation

Proteins, miRNA, and metadata are largely linearly independent, and also have high correlation to ovarian cancer individually. See figure 1.

1. The feature correlation matrix has an approximate block diagonal structure. The miRNA are self-correlated, and so are the proteins and metadata, but the miRNA and proteins do not exhibit a high degree of inter-correlation.
2. Many of the miRNA (e.g., miR-200c) and proteins (e.g., HE4) have high correlation to the cancer labels. Thus, the miRNA and proteins have good synergy, and are well-suited to train a linear classification model.
3. The selection of the miRNA features was done using forward regression. More details are provided in the methods section.

Goals

The goals of this work are summarized as follows:

• Can we improve a currently deployed ovarian cancer triage test? - the current test is based on proteins and metadata. We aim to investigate whether miRNA can improve the current test.
• Testing and comparison - to do this, we compare model performance on internal and external validation sets provided by Aspria Women’s Health (AWH) and Brigham Women’s Hospital (BWH).
• Identifying patients for surgery - the patients are presenting with an ovarian mass, and the goal is to determine who should be recommended for surgery. We aim to investigate whether miRNA can improve model sensitivity, particularly among early-stage and serous cancers.

Data sets considered

• Data - We test our hypothesis on data provided by AWH, and BWH.
• Properties - The data is comprised of a training cohort, and two validation cohorts, one internal and one external. See table 1. The patients considered are women, mostly presenting with ovarian masses.

Breakdown by stage and serous vs non-serous cancers

The combined model offers the best accuracy when identifying serous and early-stage cancers.

ROC curve plots

The combined model offers the best AUC score on the internal and external validation sets. In particular, on the external set, the ROC curve offered by the combined model is always above the ROC curves compared against.

Sensitivity and specificity

The sensitivity and specificity scores are tabulated in table 2. The combined features offer a highly sensitive model, whereas the protein + metadata model is more specific.

Figure 2. TSNE plots of internal (top row) and external (bottom row) validation sets, using different feature sets.

The combined feature set (all) is most optimal for training a linear classification model.

Figure 3. ROC plots on internal and external validation sets.

Table 2. Sensitivity and specificity scores with cancer probability threshold corresponding to the point on the training ROC curve closest to (0, 1) (see figure 2).

Methods

The original triage test is based on 7 proteins (e.g., CA-125) and age and menopausal status. We introduce a panel of 180 miRNA. To reduce the miRNA dimension, we use forward regression [3, 1]. Let $\hat{\mathbf{x}} = [\mathbf{x}_1, \ldots, \mathbf{x}_k] \in \mathbb{R}^k$ be the miRNA data, and let $Y \in \{0, 1\}$ be the cancer labels. Then, we implement forward regression [3, 1]:

1. Initialize $S_0 = \emptyset$ and the number of features, $k$.
2. for every integer $j \in \{1, \ldots, k\}$ do
3. Let $\mathbf{x}_j$ be a variable maximizing $R_{S_{j-1}}^2 / \|\mathbf{x}_j\|^2$ and set $S_{j+1} = S_j \cup \{\mathbf{x}_j\}$.
4. Output $S_k$.

where

$$R_{S_k}^2 = \frac{\text{Var}(Y) - E\{Y - \hat{Y} \}^2}{\text{Var}(Y)}$$

and $Y = \sum_{j=1}^k v_j \mathbf{x}_j$, where $\mathbf{v} = [v_1, \ldots, v_k]^T$ is defined $v_j = \text{arg min}_{v} \|\mathbf{v}\|_2 \mathbf{v}^T \|\mathbf{x}_j - \mathbf{x}\|_2$. The output, $S_k$, identifies the optimal miRNA features, which are combined with the proteins and metadata. $k = 38$ was identified as the optimal $k$ using cross-validation. Then, we train a softmax function [2] to map $X$ to $Y$.

$$P(Y = j | X) = \frac{e^{x_j^T \mathbf{w}_j}}{\sum_{l=0}^{L-1} e^{x_l^T \mathbf{w}_l}}$$

where $j \in \{0, 1\}$ is the class label, $x \in \mathbb{R}^{19}$ is a patient sample $i$, one row of $X$, and the $(\mathbf{w}_j)_l$ are weights and biases to be trained. Here y denotes the class label assigned to $X$.

The class with the highest probability $P$ is then chosen for membership.

References