CMM, Spring 2023
Breast Cancer Project

Context: As discussed at length in class, one of the most intense research areas in computational molecular medicine over the past twenty years is the prediction of cellular phenotypes, e.g., properties of cancerous growths, based on gene expression profiles. This project concerns an extremely important and largely unsolved problem in the treatment of breast cancer, namely determining whether a person diagnosed with invasive breast cancer will or will not respond to chemotherapy treatment. If a patients is sensitive to chemotherapy, it is better to apply the chemotherapy before the surgery in order to shrink the size of the tumor. Patients sensitive to chemotherapy who undergo this treatment procedure may have better prognosis. However, if the chemotherapy is not effective on the patient, surgery needs to be prioritized. Therefore, it is important to know or predict how an upcoming new patient will respond to the chemotherapy before making the treatment plan, and we could potentially answer this question with machine learning.

Prediction Problem: There are two classes labeled "treatment-sensitive" (or just "sensitive") or "treatment-resistant" (or just "resistant"). The goal then is to build a predictor (classifier) to distinguish whether a (new) patient is sensitive or resistant to the chemotherapy. In this particular study, tumors are only labeled "sensitive" if all target lesions disappear after six months of chemotherapy. This category is called (CR) for "complete response" according to the RECIST criteria (Response Evaluation Criteria In Solid Tumors). Tumors are labeled "resistant" otherwise; these are patients for whom the RECIST label is either (PD) for "progressive disease" defined by at least a 20 percent increase in the sum of diameters of target lesions, or (SD) for "stable disease."

Data: The breast cancer bulk RNA-seq dataset contains 1171 samples (patients) and 13299 transcripts. The gene expression matrix is in *BRXdata.csv*, and meta data can be found in *BRXmeta.csv* file. The phenotype of interest is the “ChemoResponse” column in the meta data. This column contains 1171 labels, with each label being either “Resistant” or “Sensitive.” What is more, the meta file also contains important clinical variables used for diagnosing breast cancer such as ER, PR, and HER2 status. Feel free to include those features into building the classifiers together with the expression profiles.

References: Three reviews:

1. https://www.mdpi.com/2072-6694/12/6/1404
2. https://www.sciencedirect.com/science/article/pii/S1526820920301567?via
3. https://www.uptodate.com/contents/general-principles-of-neoadjuvant-management-of-breast-cancer