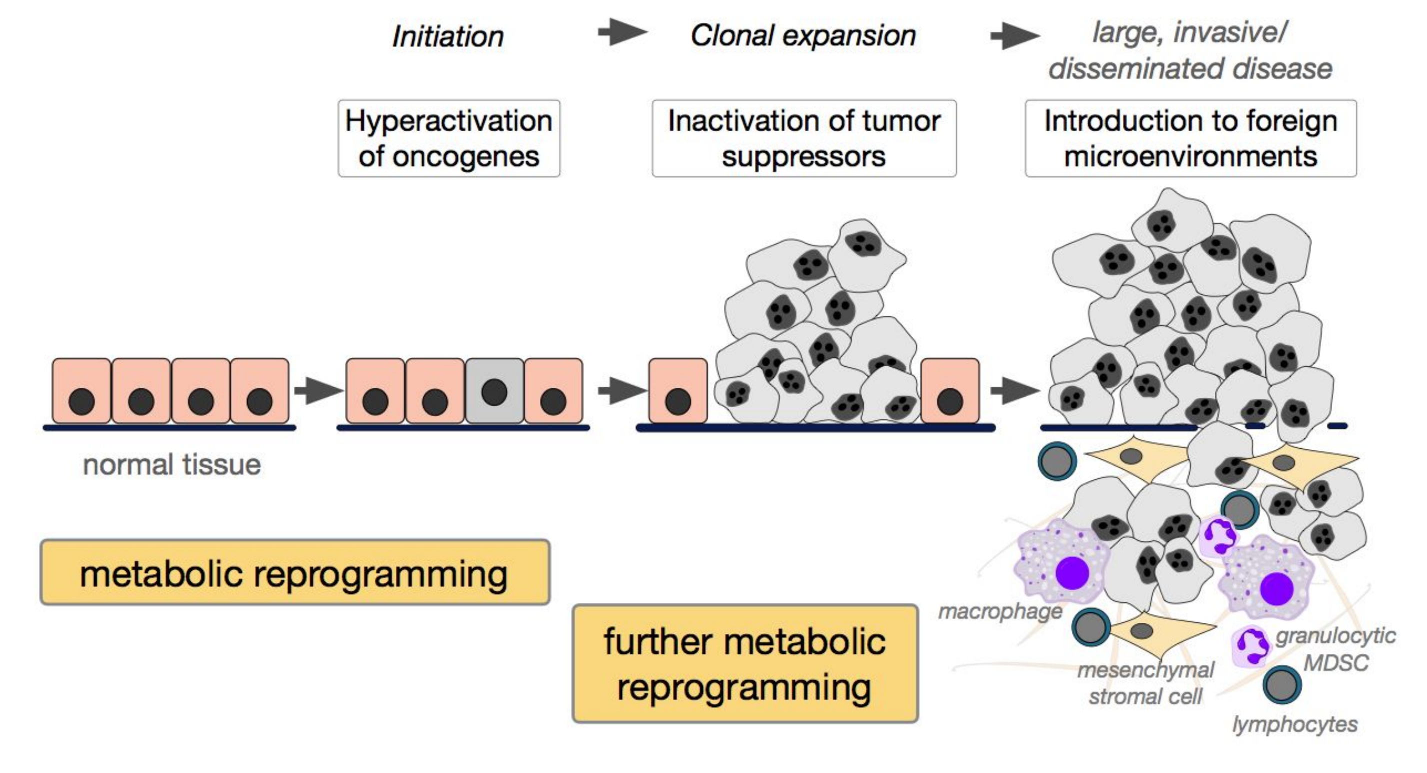
Treatment Prediction For Breast Cancer Patients

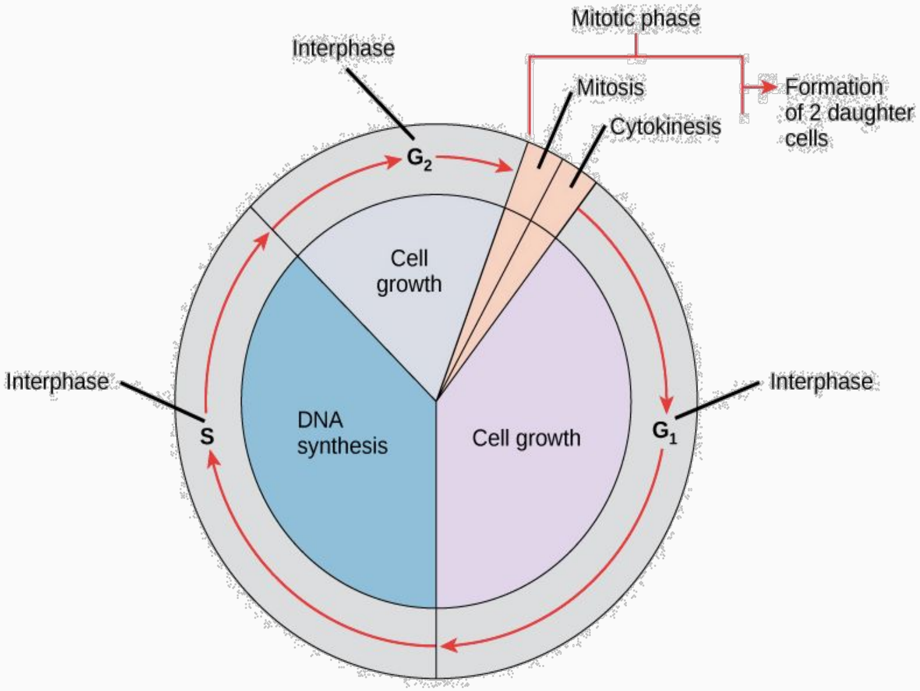
Prashant Rajput & Ruchi Jain

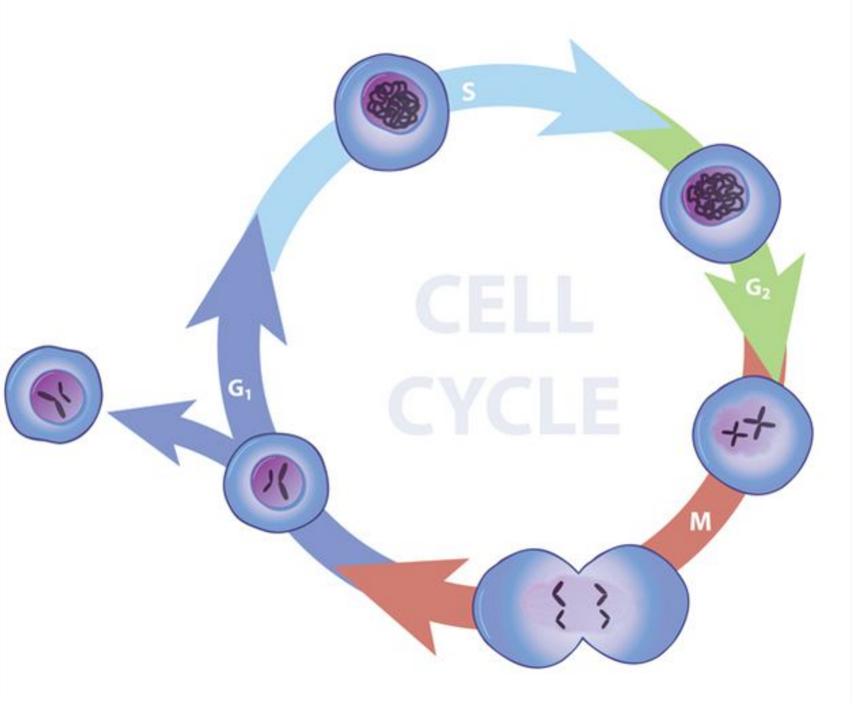
Cancer at Microscopic Level


Initiation
Clonal expansion
large, invasive/ disseminated disease
Hyperactivation of oncogenes
Inactivation of tumor suppressors
Introduction to foreign microenvironments
normal tissue
metabolic reprogramming
further metabolic reprogramming
macrophage
mesenchymal stromal cell
granulocytic MDSC
lymphocytes

RADU Lab, UCLA

Background: Cell Cycle



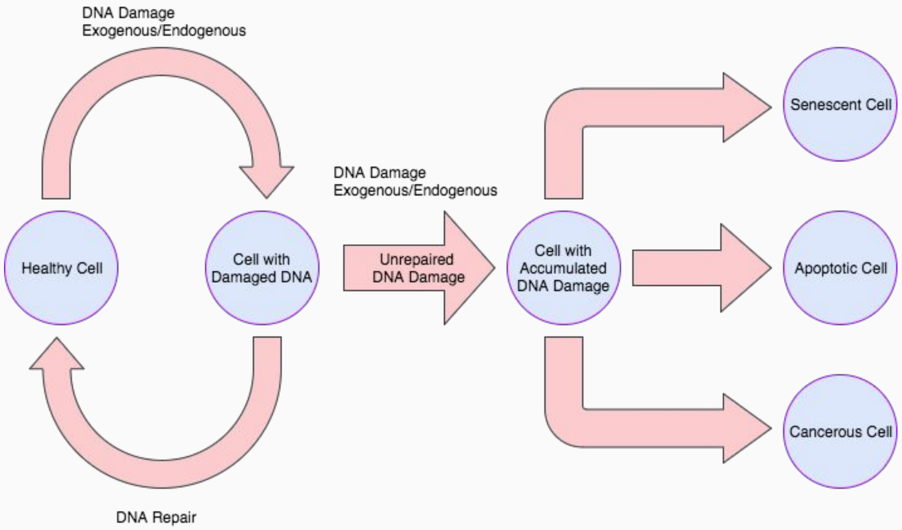


DNA Repair

Senescent Cell: These cells cease to divide.

Apoptotic Cell: Cells killed by Apoptosis (programmed cell death).

Cancerous Cell: Cells that divide relentlessly.



Basis of all Treatment

- In cancer cells, the molecules that decide whether a cell should repair itself are faulty. For instance, a protein called p53 normally checks to see if the genes can be repaired or if the cell should die.
- But many cancers have a faulty version of p53, so they don’t repair themselves properly.
- Once a cell’s DNA is damaged beyond repair, the cell goes into apoptosis or programmed death.

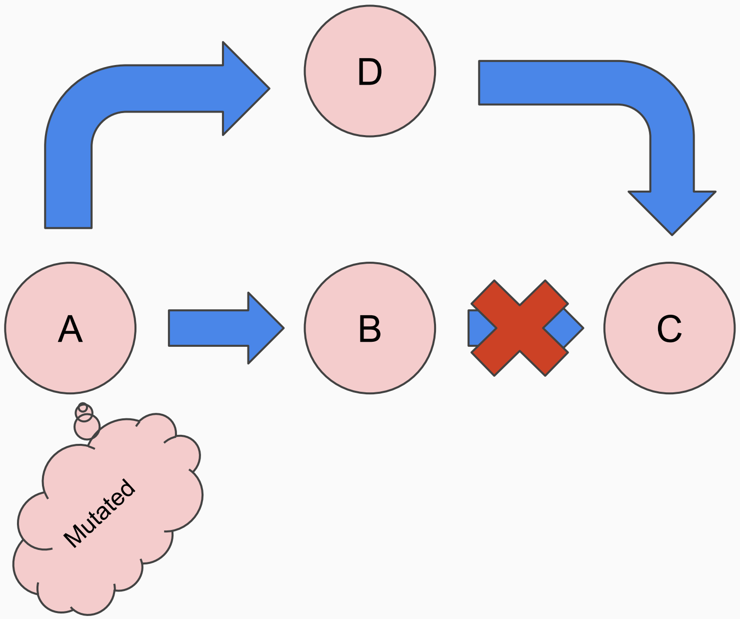
Chemotherapy

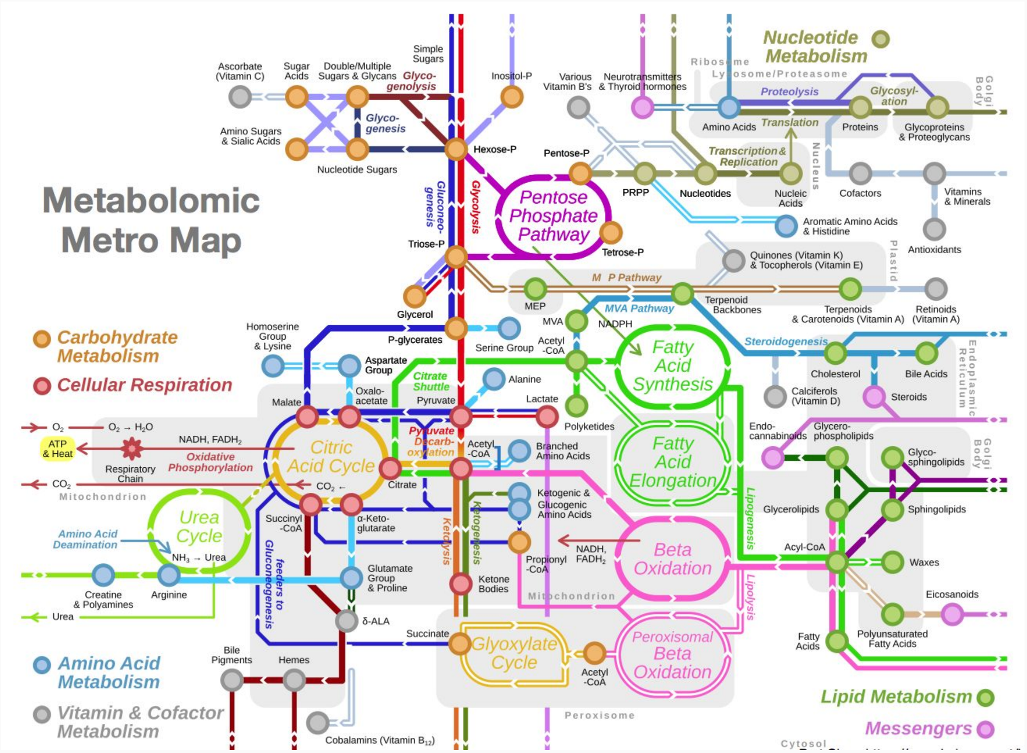
- Damage cell’s DNA beyond repair so that it enters Apoptosis.
- Given through blood stream.
- Harms regular cells as well, but they stop themselves in stage G1 and go under repair.
- Radiotherapy does the same thing but the effect is localized.

Small Molecule Inhibitors and Other Targeted Therapy

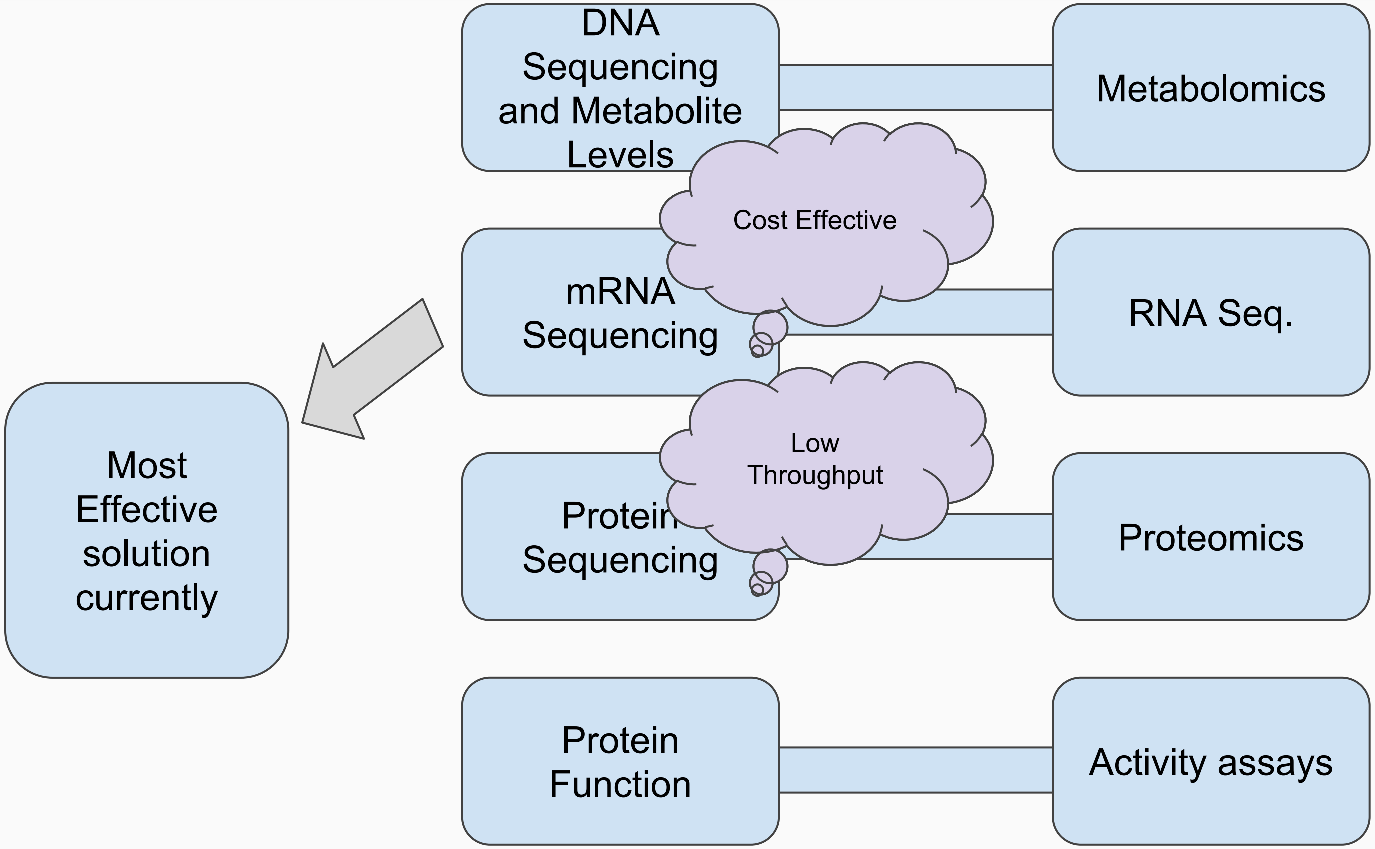
- All cells need proteins to function.
- Stop the DNA synthesis so as to stop the synthesis of proteins. For example, Hydroxyurea is a FDA approved Ribonucleotide reductase inhibitors which inhibits DNA synthesis.
- Once cancer cells are deprived of proteins, they cannot function and die.
- Less harmful than Chemotherapy.

What is Drug Resistant Cancer





Sequencing



DNA Sequencing and Metabolite Levels

Metabolomics

Cost Effective

mRNA Sequencing

RNA Seq.

Most Effective solution currently

Low Throughput

Protein Sequencing

Proteomics

Protein Function

Activity assays

Limitation of Current Solutions

Treatment is limited to the localization and biomarkers (tissue biopsy) of the cancer and is not focused on the genetic signature of the patient.

Genes dictate the fate:
Gene Signature based
Treatment

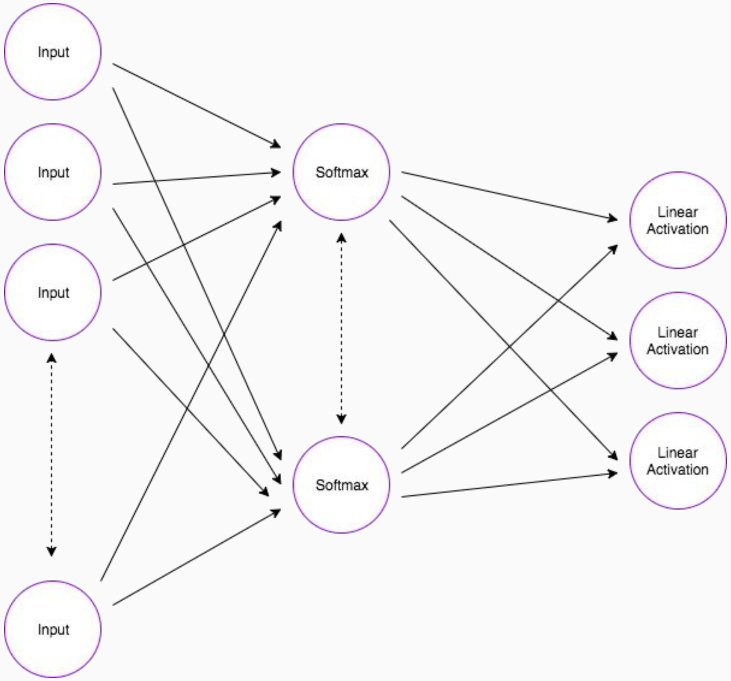
Problem Statement

- Build a model that combines genetic and clinical data from breast cancer patients to predict treatment

The Dataset

- Features:
  - Genomic: EFGR, ERBB2, BRCA1, CDH1, PTEN, FOXO3, BRCA2
  - Binary data: Inferred menopausal state (Pre, post), Breast surgery (Mastectomy, conserving)
  - One hot encoded: vital status, cellularity, her_snp6, claudin_subtype
- Labels: multi-class classification
  - Chemotherapy
  - Hormone Therapy
  - Radio Therapy

Artificial Neural Network



- One input layer, one hidden layer, one output layer
- Activation functions: softmax, linear activation
- Hidden layer size: 5 nodes

Accuracy: 63.1 (K: 3)

Multi-label Learning (MLL)

- Each sample can have more than one label in its output
- Models tested:
  - k-Nearest Neighbors
  - Decision Trees
  - Random Forest Classifier
- Have to use different evaluation metrics because traditional metrics are too harsh for MLL

Evaluation Metrics

Coverage Error

$$coverage(y, \hat{f}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \max_{j:y_{ij}=1} \text{rank}_{ij} \qquad \text{rank}_{ij} = \left|\left\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\right\}\right|$$

- “Average number of labels that have to be included in the final prediction such that all true labels are predicted”
- The best value of this metric is the average number of true labels. (In our case that was 2.055.

Label Ranking Average Precision

$$LRAP(y, \hat{f}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \frac{1}{|y_i|} \sum_{j:y_{ij}=1} \frac{|\mathcal{L}_{ij}|}{\text{rank}_{ij}} \qquad \mathcal{L}_{ij} = \left\{k : y_{ik} = 1, \hat{f}_{ik} \geq \hat{f}_{ij}\right\}$$

- “Average over each ground truth label assigned to each sample, of the ratio of true vs. total labels with lower score” **in other words “this metric will yield better scores if better rank is given to the labels associated with each sample”**
- Performance is best when LRAP is 1.

Ranking Loss

$$\text{ranking\_loss}(y, \hat{f}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \frac{1}{|y_i|(n_{\text{labels}} - |y_i|)} \left|\left\{(k, l) : \hat{f}_{ik} < \hat{f}_{il}, y_{ik} = 1, y_{il} = 0 \right\}\right|$$

- “Averages over the samples the number of label pairs that are incorrectly ordered, i.e. true labels have a lower score than false labels, weighted by the inverse number of false and true labels”
- Best performance is when ranking loss is zero.

Results

Model	Coverage Error	Label Ranking Average Precision Score	Ranking Loss
kNN (n=64)	2.575	0.838333	0.3175
Decision Tree (entropy, max_depth = 6)	2.655	0.809167	0.3825
Random Forest (n_estimators = 11, entropy, Max_depth = 6)	2.575	0.834167	0.325

Future Work

- Implement Multi-label Learning Neural Networks
- Obtain more (smooth) data
- Study the features set and use better feature extraction methods
- Incorporate information about treatment effectiveness in model
- Experiment with other MLL algorithms, such as AdaBoost or Kernel Methods
- Experiment with other error/loss functions, such as one-error, and hamming loss

References

- Zhang, M. And Zhou, Z. “Multi-Label Neural Networks with Applications to Functional Genomics and Text Categorization.” *IEEE Transactions On Knowledge and Data Engineering*