New computational approaches for cancer therapeutics

Jonathan Allen, Ya Ju Fan, Stewart He (LLNL)
The dose response prediction problem

Problem motivated by NCI’s Molecular Profiling-Based Assignment of Cancer Therapy (MPACT) clinical trial study

Menden et al. (2013) PLoS ONE

Presents a challenging data driven modeling problem with large input feature dimensions
Focusing on two prediction problems

1. **Query** New cell (or tumor)
   - **Cell lines**
     - Genomic features
       - Feature matrix
     - Drugs
       - Response variable
   - Predict growth rate for new tumor

2. **Query** New drug
   - **Drug**
     - Drug properties
     - Feature matrix
     - Drugs
     - Response variable
   - Predict growth rate for new drug
Investigating the use of unsupervised feature learning

**Big Data Hypothesis:**
larger unlabeled collections of observations: tumors, normal tissue, chemical compounds can be used to learn descriptive features

*GDC:* includes gene expression for 11,574 tumors (and related normal tissue)

*ChEMBL:* public repository of 1.6 million compounds

**Transfer Learning Hypothesis:**
models (or encodings) of smaller subsets of features can be transferred to smaller pre-clinical trial data to improve accuracy of dose response predictions

*cell line models applied to patient derived tumor models*

**Mechanisms of Action Hypothesis:**
feature learning can help inform and be informed by biological models
Experimental data for model evaluation

- **NCI-60: 60 cell lines**
  - Each cell line has microarray gene expression, protein, miRNA, SNP, (RNAseq in progress).
  - 5-dose response measurements for 52,672 compounds.
    - Size of individual cell culture measured after an initial 24 hour incubation period ("Time Zero"). ~5K – 40K cells depending on cell/tumor type.
    - Size of cell culture growth measured after 48 hours with initial compound treatment and without compound treatment (control)
    - Ratio reports change in tumor size relative to control.
      - 0 -> drug has no impact on tumor growth
      - -100 -> tumor exhibits large reduction in size
      - 100 -> tumor exhibits large increase in size

- **Other labeled datasets being examined:**
  - CCLE, GDSC: Additional cell line repositories, with more (~1K) cell lines but future drugs tested.
  - Ultimate goal is application to new patient derived tumors being established at Frederick National Laboratory (NCI).
Scalable deep learning tools used to explore feature representation space

- Learn 500 dimension feature representation on 50K chemical compounds from 5000 dimension input feature: \( \sim 6 \) hours x 24 cores x 16 nodes x 104 runs = \( \sim 239,616 \) CPU hours

Experiments run using Livermore Big Artificial Neural Network (LBANN)

Only select models converge on low reconstruction error

Initial use of autoencoder indicated potential to find low reconstruction error but results were sensitive to the choice of parameters
Autoencoders show potential to retain information for downstream prediction

1. Unlabeled training: 24,789 (samples) x 3,809 features
2. Build logistic regression model on single concentration dose response using a single cell line.
3. Test response prediction on ~2K held out compounds for 6 representative cell lines. Labeled training: ~14k-16k examples.
Next steps

- Evaluate feature learning with larger unlabeled dataset - ChEMBL’s 1.6 million compound library
- Evaluate response prediction on novel compound classes, not just near neighbors.
- Incorporate new molecular features, such as gene expression and SNPs
- Explore methods to guide model complexity reduction and integrate models with biological knowledge (no time to discuss today)
Framework for developing predictive models

Compile and maintain modeling data

- Primarily public cell line
- Single or multiple "-omics" data types
- Training dataset
- Public data resources
- Clinical cohorts

Evaluate prediction confidence

- NCI MPACT clinical trials
- Further independent validations

Build data driven predictive models

- Statistical or machine learning techniques
- Classification and regression
- Prediction models
- Model cross-validation and selection
- Selected candidate model
- Model interpretation and reporting

Integrate biological mechanisms

Improve drug treatment selection for patient
Increase understanding of drug efficacy mechanisms
Cancer pilot is a multi-laboratory effort

- Rick Stevens (Lead), Fangfang Xia, Maulik Shukla (ANL)
- Yvonne Evrand, Susan Holbeck (NCI / Frederick)
- Jason Gans, Judith Cohn, John Hodge (LANL)
- Adam Zemla, Marisa Torres (LLNL)
Deeper networks show potential to learn novel feature encoding

- Gene expression feature encoding (5-fold cross validation)
  Experiments run using Livermore Big Artificial Neural Network (LBANN)

Data size of 60,483 features x 11,574 examples presents a computational challenge
Non-linear feature encoding improves performance in some conditions

Indicates potential to find more robust molecular features with autoencoder