Introduction to Systems Biology II

Amin Emad

NIH BD2K KnowEnG Center of Excellence in Big Data Computing
Carl R. Woese Institute for Genomic Biology
Department of Computer Science
University of Illinois at Urbana-Champaign

June, 2018
Systems Biology

- Systems biology is the computational and mathematical modeling of complex biological systems (wikipedia).

![System biology view](Image)

- Studies the interactions between the components of biological systems such as genes, proteins, metabolites, etc. (i.e. biological networks), and how these interactions give rise to the function and behavior of that system (phenotype)
A graphical representation of the interactions of the components of a biological systems

- **Cell signaling networks**
- **Gene regulatory networks**
- **Protein-protein interaction networks**
- **Gene co-expression networks**
- **Metabolic networks**
Biological Networks in Computational Biology

Analyzing network properties

Analyzing ‘omic’ data in light of networks

Reconstructing biological networks

Graph Theory

Machine Learning

Statistics
1) Analyzing network properties
What is a network/graph?

- **Graph**: A representation of relationship among objects
- A graph $G(V, E)$ is a set of vertices (nodes) $V$ and edges (links) $E$

Directed vs. Undirected:

**Undirected graph**
- Protein-protein interactions
- Co-expression network

**Directed graph**
- Gene regulatory network
- Signaling pathways
Weighted vs. Unweighted:

- Weights represent affinity in PPI, correlation coefficient in a co-expression network, confidence in a GRN, etc.
Degree and degree distribution:

- **Degree**: Number of connections of a node to other nodes

- **Indegree** (outdegree) of a node in a directed graph is the number of edges entering (leaving) that node

- **Degree distribution** of a network is the probability distribution of these degrees over the network:
Graph Properties

Adjacency matrix:

- A matrix representation of the graph

Graph Properties

Path and connectivity:

- **Path**: A sequence of distinct edges connecting a sequence of vertices: GFAB, EAC, etc.

- **Connectivity**: A graph that in which a path exists between any two nodes.
Important classes of graphs:

- **Tree**: Any two vertices are connected by exactly one path (e.g. dendogram in hierarchical clustering)

- **Complete graph**: Each pair of vertices are connected by an edge
2) Analyzing ‘omic’ data in light of biological networks
Analyzing ‘omic’ data in light of networks

How to analyze large ‘omic’ datasets?

Statistics

Machine Learning
Analyzing ‘omic’ data in light of networks

How to analyze large ‘omic’ datasets?

Machine learning is concerned with utilizing statistical techniques to give computers the ability to “learn”.
Analyzing ‘omic’ data in light of networks

How to analyze large ‘omic’ datasets?

Machine learning is concerned with utilizing statistical techniques to give computers the ability to “learn”.

However, it can do much more!
Some examples:

- Predicting whether a patient is sensitive or resistant to a drug
- Predicting the survival probability of a cancer patient
- Identifying the subtypes of a disease
- Identifying genes associated with a disease
- etc.
Training examples are provided with desired inputs and outputs to help learning the desired rule.

No training example exists and the goal is to learn structure in the data.
Machine Learning

- Supervised Learning
  - Classification
  - Regression
  - Supervised Feature Selection
- Unsupervised Learning
  - Clustering
  - Dimensionality Reduction
Unsupervised Machine Learning (Clustering)

• We have a set of samples characterized using several features (e.g. expression of thousands of genes for tumor samples)

• **Goal:** Group the sample such that those in the same group are more similar to each other than to those in other groups

• Many methods exist such as K-means, hierarchical clustering, matrix factorization, etc.

• **Example:** Identifying subtypes of breast cancer using transcriptomic data
Unsupervised ML (Dimensionality Reduction)

• We have a set of samples characterized using several features

• **Goal:** Reduce the number of features while preserving characteristics of the data

• Many methods exist such as principal component analysis, linear discriminative analysis, etc.

• **Example:** PCA identifies a few principal components, orthogonal to each other, such that they account for most of the variance in the data
Classification:

- We have a set of samples characterized using several **features** (e.g. expression of thousands of genes for tumor samples)
- The samples belong to set of known **categories**
- **Goal**: Given a new sample, to which category does it belong?
- Many methods exist such as KNN, SVM, logistic regression, decision trees, random forests, etc.
Example:

- We have ‘omic’ profiles and clinical information of breast cancer patients
- We also know which patients were resistant to a drug and which ones were not

Given the ‘omic’ profiles and clinical information of a new patient, will they be resistant to the drug or not?
Supervised Machine Learning (Regression)

• We have a set of samples characterized using several features (e.g. expression of thousands of genes for tumor samples)

• For each sample, we know a continuous-valued response (dependent variable) (e.g. number of years between diagnosis and occurrence of metastasis)

• Goal: Estimate the relationship between the response and features and predict the value of response for a new sample

• Many methods exist such as linear regression, LASSO, Elastic Net, Support vector regression, etc.
Supervised Machine Learning (Regression)

Example:

• We have transcriptomic profiles of breast cancer patients
• We also know number of months between diagnosis and occurrence of metastasis

What is the relationship between gene expression and time of metastasis?

[Image of genes and samples]

https://www.cancer.gov/types/metastatic-cancer
Supervised Machine Learning (Feature Selection)

- We have a set of samples characterized using several features.
- We know a continuous-valued or categorical response for samples.
- **Goal**: What are the features most predictive of the response?

**Examples:**
- Differentially expressed genes (case vs. control)
- Correlation analysis (GWAS)
- etc.
How can biological networks help?

- When features correspond to genes or proteins (e.g. gene expression, mutation, etc.), these networks can provide information regarding the interactions and relationships of these features.
Network-guided gene prioritization using ProGENI
Background

- Phenotypic properties of a cell are determined (partially) by expression of its genes and proteins.

- Gene expression profiling measures the activity of thousands of genes to create a global picture of cellular function.
Background

• **Goal:**
  • Identifying genes whose basal mRNA expression determines the drug sensitivity in different samples (supervised feature selection)

• **Motivations:**
  • Overcoming drug resistance
  • Revealing drug mechanism of action
  • Identifying novel drug targets
  • Predicting drug sensitivity of individuals
Gene prioritization

Examples of current methods:

• Score each gene based on the correlation of its expression with drug response
Gene prioritization

Examples of current methods:

• Score each gene based on the correlation of its expression with drug response

• Use multivariable regression algorithms such as Elastic Net to relate multiple genes’ expression values to drug response

\[ \sum w_i x_i \]
Gene prioritization

Examples of current methods:

• Score each gene based on the correlation of its expression with drug response

• Use multivariable regression algorithms such as Elastic Net to relate multiple genes’ expression values to drug response

Shortcoming:

• These methods do not incorporate prior information about the interaction of the genes
Hypothesis:

- Since genes and proteins involved in drug MoA are functionally related, prior knowledge in the form of gene interaction network (e.g. PPI) can improve accuracy of the prioritization task.
ProGENI: Network-guided gene prioritization

An algorithm that incorporates gene network information to improve prioritization accuracy

DOI 10.1186/s13059-017-1282-3

Knowledge-guided gene prioritization reveals new insights into the mechanisms of chemoresistance

Amin Emad1, Junmei Cairns2, Krishna R. Kalari3, Liewei Wang2* and Saurabh Sinha4*

© The Author(s). 2017
Open Access
This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance.
Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance.
ProGENI

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance.

Step 2: Find genes most correlated with drug response (RCG set).
ProGENI

**Step 1:** Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance

**Step 2:** Find genes most correlated with drug response (RCG set)

**Step 3:** Score genes based on their relevance to the RCG set
ProGENI

Step 1: Generate new features representing expression of each gene and the activity level of their neighbors weighted proportional to their relevance

Step 2: Find genes most correlated with drug response (RCG set)

Step 3: Score genes based on their relevance to the RCG set

Step 4: Remove network bias by normalizing scores w.r.t. scores corresponding to global network topology
Datasets

- **Human lymphoblastoid cell lines (LCL)**
  - Gene expression (~17K genes of ~300 cell lines)
  - Drug response of 24 cytotoxic treatments

- **Publicly available dataset from GDSC**
  - Gene expression (~13K genes of ~600 cell lines from 13 tissues)
  - Drug response of 139 cytotoxic treatments

- **Publicly available prior knowledge**
  - Network of gene interactions (PPI and genetic interactions) from STRING (~1.5M edges, ~15.5K nodes)
Validation using drug response prediction

- Genes ranked highly using a good prioritization method are good predictors of drug sensitivity

---

1. **Cell lines**
2. **Gene expressions**
3. **Genes**
4. **Test set**
5. **Training set**
6. **Repeat $N$, times**

- Divide samples into training and test sets
- Rank all genes
- Train a SVR on training set using expression of highly ranked genes
- Predict drug sensitivity of the test set
Validation using drug response prediction

<table>
<thead>
<tr>
<th></th>
<th>LCL Dataset</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson</td>
<td>Elastic Net</td>
<td></td>
</tr>
<tr>
<td>Num. Drugs (out of 24)</td>
<td>14</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>ProGENI &gt; Baseline</td>
<td>6.5 E-3</td>
<td>9.6 E-5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GDSC Dataset</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson</td>
<td>Elastic Net</td>
<td></td>
</tr>
<tr>
<td>Num. Drugs (out of 139)</td>
<td>66</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>ProGENI &gt; Baseline</td>
<td>9.1 E-4</td>
<td>4.0 E-21</td>
<td></td>
</tr>
</tbody>
</table>

SPCI (ProGENI-SVR): 17-AAG FDR = 3.2E-29 PIF = 88.00 %
SPCI (ProGENI-SVR): 681640 FDR = 1.2E-31 PIF = 87.20 %
SPCI (ProGENI-SVR): A-443654 FDR = 1.2E-33 PIF = 87.20 %
SPCI (ProGENI-SVR): A-770941 FDR = 1.2E-31 PIF = 87.20 %
SPCI (ProGENI-SVR): ABT-263 FDR = 2.9E-34 PIF = 86.80 %

SPCI (PCC-SVR): 17-AAG FDR = 3.2E-29 PIF = 88.00 %
SPCI (PCC-SVR): 681640 FDR = 1.2E-31 PIF = 87.20 %
SPCI (PCC-SVR): A-443654 FDR = 1.2E-33 PIF = 87.20 %
SPCI (PCC-SVR): A-770941 FDR = 1.2E-31 PIF = 87.20 %
SPCI (PCC-SVR): ABT-263 FDR = 2.9E-34 PIF = 86.80 %

SPCI (EN-SVR): 17-AAG FDR = 3.2E-29 PIF = 88.00 %
SPCI (EN-SVR): 681640 FDR = 1.2E-31 PIF = 87.20 %
SPCI (EN-SVR): A-443654 FDR = 1.2E-33 PIF = 87.20 %
SPCI (EN-SVR): A-770941 FDR = 1.2E-31 PIF = 87.20 %
SPCI (EN-SVR): ABT-263 FDR = 2.9E-34 PIF = 86.80 %
We validated role of 33 (out of 45) genes (73%) for three drugs.

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Rank (ProGENI)</th>
<th>Rank (Pearson)</th>
<th>Absolute value of Pearson correlation coefficient</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATF1</td>
<td>1</td>
<td>1</td>
<td>0.2000</td>
<td>Direct (this study)</td>
</tr>
<tr>
<td>MIS12</td>
<td>2</td>
<td>4</td>
<td>0.1887</td>
<td>Direct (this study)</td>
</tr>
<tr>
<td>OSBPL2</td>
<td>5</td>
<td>6</td>
<td>0.1865</td>
<td>Direct (this study)</td>
</tr>
<tr>
<td>CSNK2A1</td>
<td>7</td>
<td>1587</td>
<td>0.0752</td>
<td>Direct (literature)</td>
</tr>
<tr>
<td>PSIP1 (LEDGF)</td>
<td>8</td>
<td>46</td>
<td>0.1537</td>
<td>Direct (literature)</td>
</tr>
<tr>
<td>CAMK2A</td>
<td>9</td>
<td>6991</td>
<td>0.0157</td>
<td>Direct (literature)</td>
</tr>
<tr>
<td>CSNK2A2</td>
<td>10</td>
<td>4870</td>
<td>0.0347</td>
<td>Direct (literature)</td>
</tr>
<tr>
<td>GOSR1</td>
<td>11</td>
<td>6867</td>
<td>0.0167</td>
<td>Direct (this study)</td>
</tr>
<tr>
<td>MAPK8</td>
<td>13</td>
<td>7574</td>
<td>0.0112</td>
<td>Direct (literature)</td>
</tr>
<tr>
<td>SPI1</td>
<td>14</td>
<td>6287</td>
<td>0.0217</td>
<td>Direct (literature)</td>
</tr>
<tr>
<td>CREB1</td>
<td>15</td>
<td>665</td>
<td>0.1000</td>
<td>Direct (literature)</td>
</tr>
<tr>
<td>NOC3L</td>
<td>3</td>
<td>3</td>
<td>0.1893</td>
<td>Not found</td>
</tr>
<tr>
<td>IL27RA</td>
<td>4</td>
<td>2</td>
<td>0.1911</td>
<td>Not found</td>
</tr>
<tr>
<td>MGEA5</td>
<td>6</td>
<td>7</td>
<td>0.1814</td>
<td>Not found</td>
</tr>
<tr>
<td>WAPAL</td>
<td>12</td>
<td>8</td>
<td>0.1805</td>
<td>Not found</td>
</tr>
</tbody>
</table>

**Functional validation**

**BT549**

- p-value < 0.0001
- p-value < 0.0001
How about other ML tasks?

• Similar principles can be used for ML tasks other than feature selection/prioritization

• “Network-smoothing” of the features used in ProGENI can be used as a preprocessing step to regression and classification algorithms

• Network-smoothing can also be used for clustering and dimensionality reduction (e.g. Network-based stratification)


**Network-based stratification of tumor mutations.**

Hofree M\(^1\), Shen JP, Carter H, Gross A, Ideker T.
Goal:
• Stratification (clustering) of tumor samples based on somatic mutation profiles

Main Issue:
• The mutation data is very sparse and most conventional clustering techniques fail to identify reasonable patterns
• Although two tumors may not share the same somatic mutations, they may affect the same pathways and interaction networks
Data sparsity:

- Due to the sparsity of the data, all samples are at equal distance of each other
Value of network-guided analysis

Data sparsity:

• Due to the sparsity of the data, all samples are at equal distance of each other

• Pathway information clarifies the similarity among some samples
Value of network-guided analysis

Data sparsity:

- Due to the sparsity of the data, all samples are at equal distance of each other
- Pathway information clarifies the similarity among some samples
- Conventional clustering methods can identify clusters based on network-smoothed features
NBS (Algorithm Overview)

- Employs network smoothing to mitigate sparsity by transforming the binary gene-level somatic mutation vectors of patients into a continuous gene importance vector that captures the proximity of each gene in the network to all of the genes with somatic mutations in the patient sample.
- Bootstrap sampling enables robust clustering.
3) Reconstruction of Biological Networks
Gene Co-expression Networks

- Nodes represent genes
- An edge exists between two genes that are highly co-expressed across different samples
Gene Co-expression Networks

- Such networks provide a **global** view of co-expression patterns

**WGCNA: an R package for weighted correlation network analysis**

Reviewed by Peter Langfelder and Steve Horvath

**ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context**

Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano

- But do not provide information on how these networks relate to the **variation** in a phenotypic outcome
Gene Co-expression Networks

How can we relate these networks to the phenotypic variation?

gene-gene correlation
gene-phenotype association
Approach 1: In reconstructing the network, we can limit our samples to one manifestation of the phenotypic outcome

- For example, build a Basal-like co-expression network by looking at the gene correlations across Basal breast cancer samples

- Issues:
  1. Only works if we have categorical phenotype
  2. Does not relate the network to the variation in the phenotypic outcome
**Approach 2:** If the phenotype is binary, reconstruct two networks (one for each manifestation of the phenotype) and compare the two to build a **differential network**

- Shows changes in the co-expression pattern
Gene Co-expression Networks

- Issues:
  1. Becomes very cumbersome if phenotype is not binary
  2. Does not work for continuous-valued phenotypes
  3. By dividing the samples into two groups, we will have less statistical power in identifying co-expression patterns
  4. Fails in a case shown below
**Approach 3:** First, find genes associated with the phenotype and then reconstruct a **context-specific** network only using those genes.

- **Issue:** Ignores the strength (p-value) of gene-phenotype association.
**Approach 4:** Calculate p-values of gene-gene correlation and gene-phenotype associations separately and combine together using Fisher’s method or Stouffer’s method → **Simplified-InPheRNo**

- Specifically useful to identify transcription factor-gene-phenotype associations
Questions?
Regression algorithms

- **Lasso**: learns a linear model from the training data using only a few features (sparse linear model)

  \[
  \hat{\beta} = \arg\min_{\beta} \left( ||y - X\beta||^2 + \lambda_1 ||\beta||_1 \right)
  \]

- **Elastic Net**: learns a linear model from the training data by linearly combining ridge and Lasso regression regularization terms (a generalization of both Lasso and ridge regression)

  \[
  \hat{\beta} = \arg\min_{\beta} \left( ||y - X\beta||^2 + \lambda_2 ||\beta||_2 + \lambda_1 ||\beta||_1 \right)
  \]
Regression algorithms

- **Kernel-SVR:**
  - Linear SVR learns a linear model such that it has at most $\varepsilon$-deviation from the response values and is as flat as possible

(Smola and Schölkopf, 1998)

- Kernel-SVR generalizes the idea to nonlinear models by mapping the features to a high-dimensional kernel space