Knowledge-Guided Analysis with KnowEnG Lab

KnowEnG Center

Powerpoint by Charles Blatti
Exercise

In this exercise we will be doing the following:

1. Deploying the **KnowEnG Analytics Suite** to cluster somatic mutation data while integrating **STRING network** of protein interactions.

2. Using **KnowEnG Analytics Suite** to characterize gene sets of subtypes while integrating **Protein Similarity** network of evolutionary conservation.
**Genomic Data Analysis Using Prior Knowledge**

### User Interface

- **Genes**
- **Samples**
- **User Spreadsheet**

### Analysis Pipelines

- RNA-seq, Somatic Mutations, etc.
- Subtype Stratification
  - Network Smoothing
  - NMF Clustering
  - Consensus Clustering

### Knowledge Network

- Physical interactions, co-expression, pathways, biological processes, text mining, etc.

---

**Using Prior Knowledge**
Genomic Data Analysis Using Prior Knowledge in a Scalable Cloud

**Genes**
RNA-seq, Somatic Mutations, etc..

**Samples**

**User Spreadsheet**

**User Interface**

**Analysis Pipelines**
- Network Smoothing
- NMF Clustering
- Consensus Clustering

**Subtype Stratification**

**Knowledge Network**
Physical interactions, co-expression, pathways, biological processes, text mining, etc.

amazon web services
General Inputs to KnowEnG

- Collection of Gene Signatures from multiple samples / experiments called the “Genomic Spreadsheet”
- Types of Gene Signatures
  - Gene Sets
  - Gene Importance (e.g. p-value)
  - Gene Scores
- Examples
  - \( z \)-scores of transcriptomic profile
  - Differential expression gene sets/pvalues
    - after drug treatment
    - after gene knockdown
    - after environmental stimuli
  - Genes with non-silent somatic mutations
  - Genes near regions of open chromatin
  - Genes with promoter binding of TFs
  - Etc.
Value of KnowEnG Tools

• Integration of analysis with many and varied public data sets
  • Called the “Knowledge Network”
• Gene Interactions
  • STRING, BioGrid, Pathway Commons,
• Gene Annotations
  • Gene Ontology, KEGG, MSigDB, LINCcs, GEO
Three Pipelines Available Today

• Sample Clustering
  • Given genomic data from many samples find clusters (subtypes) that relate to relevant phenotype

• Gene Set Characterization
  • Find biological annotations related to your gene set(s)

• Gene Prioritization
  • Find top ranking genes that relate to a phenotype
    • ProGENI applied to drug response of LCLs
Step 0A – Form Groups of Two

For this laboratory session we will be using the KnowEnG Center’s interface for their Analytics Suite.

Because the interfaces are not yet deployed at full scale, we will be working through these laboratory exercises as pairs.
Step 0B: Download and Extract Data Files

For viewing and manipulating the files needed for this laboratory exercise, download the following archive:

http://veda.cs.uiuc.edu/CompGen2017/labs/08_Network_Analysis.zip

Right Click and Extract the contents of the archive to your course directory. We will use the files found in:

[course_directory]/08_Network_Analysis/data/
Robust Methods for Clustering Somatic Mutation Data

In this exercise, we will use the analysis workflows implemented by the KnowEnG Center to find genomic subtypes of uterine corpus endometrial carcinoma patients from somatic mutation data. The approach we demonstrate here is called Network Based Stratification and was inspired by Hofree, et al.

We will explore the ability of Network Based Stratification to identify clusters that related to compelling clinical features.
## Dataset Characteristics

<table>
<thead>
<tr>
<th>Name</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>uterine_clinical_data.tsv</td>
<td>A matrix of phenotype labels for 248 uterine cancer TCGA samples with phenotypes including survival, surgical outcome, tumor stage and grade, etc.</td>
</tr>
<tr>
<td>uterine_somatic_mutations.tsv</td>
<td>A matrix that indicates non-silent somatic mutations that are identified in the protein coding region of a gene for all 248 uterine cancer samples across more than 16,000 genes</td>
</tr>
</tbody>
</table>
Dataset Characteristics II

- Phenotypic data:
  - Table with
    - Samples as the rows
    - Phenotypic categories as the columns
    - Categorical or Quantitative values
  - Enables searching for clusterings that are consistent with particular phenotypes
    - survival, surgical success, treatment outcome, metastasis, etc.
Dataset Characteristics III

Here is an analogous visualization of our non-silent somatic mutation cancer data (**uterine_somatic_mutations.tsv**) from the [UCSC Cancer Genome Browser](http://genome.ucsc.edu).

**Mutated Genes**

[Image of a genome browser visualization showing mutated genes in cancer patients.]
Dataset Characteristics IV

**Issue 1**

- **Difficulty** - Basic clustering methods will be susceptible to **noisy data** (passenger mutations)
- **Hypothesis** - **Bootstrapping** provides more robust clustering results by running method on many samples of the data and returning consensus results

**Issue 2**

- **Difficulty** - Typical clustering methods have difficulty because of the **sparseness of the data**
- **Hypothesis** - Although two tumors may not share the same somatic mutations, they may affect the **same pathways and interaction networks**

**Network Based Stratification (NBS)** approach presented here will attempt to address both issues.
Issue 2: Value of Network-Guided Analysis

• Sparsity of Data Problem
  • Distance between all samples is the same
Issue 2: Value of Network-Guided Analysis

- Sparsity of Data Problem
  - Distance between all samples is the same
  - Knowledge of pathways helps produce clusters
Issue 2: Value of Network-Guided Analysis

• Sparsity of Data Problem
  • Distance between all samples is the same
  • Knowledge of pathways helps produce clusters

• Network Smoothing
Step 1A – Sign into KnowEnG Scientific Collaboration Portal

Follow the link to the appropriate HubZero login screen
  • https://hub.knoweng.org/

Enter your user name and password for the course into the credentials boxes

Click “Sign In”
Step 2A – Upload Cancer Data Files

Click on “Data” among the options along the top.

Click on “Upload New Data”.

Click on “browse” link.

Find and highlight the data files in 08_Network_Analysis/data/sample_clustering:
- **uterine_clinical_data.tsv**
- **uterine_somatic_mutations.tsv**

Click “Open” to load files.
Step 3A – Configure NBS Analysis

Click on “Analysis Pipelines” along the top

In the “Select A Pipeline” section:
Hover over “Sample Clustering” and click on the “Start Pipeline” button
Step 3B – 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 **Knowledge Network** to all of the genes with somatic mutations in the patient sample.

Repeats clustering analysis using **bootstrap sampling** to build a **consensus patient similarity matrix**.
Step 3C – Select User Data Files

Leave the default species “Human”

To choose the genomics spreadsheet:
• Click on “Select genomic feature file”
• Then click on the checkbox next to “uterine_somatic_mutations.tsv “

To choose the clinical spreadsheet:
• Click on “Select response file”
• Then click on the checkbox next to “uterine_clinical_data.tsv “

Click “Next” in the bottom right corner
Step 3D – Configure Algorithm Parameters

We will use ‘K-Means’ clustering as the “final clustering algorithm” to run on our patient similarity consensus matrix.

We will set the number of clusters (genomic subtypes) to find to be “4”
Click “Next”
Step 3E – Configure KnowNet Parameters

Click “Yes” for question about using the Knowledge Network

The Knowledge Network we will use is the high confidence protein interactions from the STRING database (“STRING Experimental PPI”)

A amount of network smoothing controls how much importance is put on network connections instead of the somatic mutations. We will use the default of 50%

Click “Next”
Step 3F – Configure Bootstrap Parameters

Click “Yes” for question about using the bootstrapping

For the purposes of the demo, we will sample the data 20 times by setting the number of bootstraps to 20

We will keep the default bootstrap sample percentage of 80%

Click “Next”
Step 3G – Launch NBS Workflow

Review your selections

Change your “Job Name” if you wish

Click “Submit Job”
Step 3H: Wait for Results to Complete

Once submitted, should see Running icon. Click on “Go to Results Page”.

When your NBS workflow status turns from “Running” (gray) to “Completed” (green), click on the name of the run

Then click on the blue “View Results” button

Results should be ready after several minutes
Step 4A: Consensus Patient Similarity Matrix

Patients by Patients matrix

The color of each cell indicates how often a pair of patients fell within the same cluster across all samplings.

Dark blocks show genomic subtypes of patients that may relate to a phenotype.

To the right we see the number of patients in each cluster.
Step 4B: Exploring Genomic Subtypes

Click on the “Add Phenotypes to Compare”

One at a time, check the boxes next to the various clinical features.

Click on the color scale to see the phenotype legend.

Are there genomic subtypes that enrich with various phenotypes?
Step 5A: Genes Specific to a Subtype

Download the detailed results. Click on the “Results” link at the top of the page.

Click on the job name of the desired run

Click on the download icon

A zip file with the same job name should download to your machine. Right Click and Extract the contents of the archive to your course directory.
Step 5A: Genes Specific to a Subtype

The file “top_genes_by_cluster.tsv” contains genomic spreadsheet with four gene sets, one for each genomic subtype we found.

Each of the gene sets contain the **100 most important genes** per subtype based on the network-smoothed clustering. Many of these will have frequent somatic mutations in the subtype and some of these will be the network neighbors of frequently mutated genes.

We will use Gene Set Characterization to find out what these genes are related to.
Network Based Methods for Gene Set Characterization

In this exercise, we will use the analysis workflows implemented by the KnowEnG Center to find standard and network-based characterizations of the gene sets that relate to each genomic subtypes of uterine cancer. The network-guided approach we demonstrate here is based on the DRaWR algorithm published by Blatti, et al.
Step 6: Gene Set Characterization

Purpose:
Find relationships between novel gene sets of the user to known annotations in the Knowledge Network …

… in order to anchor understanding of findings to previous literature and generate consistent hypotheses/mechanisms

Related webtools
DAVID - https://david.ncifcrf.gov/
Enrichr – http://amp.pharm.mssm.edu/Enrichr/
GSEA - http://software.broadinstitute.org/gsea/index.jsp
Step 6A - Upload Gene Set Spreadsheet

Click on “Data” among the options along the top.

Click on “Upload New Data”

Click on “browse” link

Find and highlight the data file
• top_genes_by_cluster.tsv

Click “Open” to load files
Step 6B – Configure Enrichment Analysis

Click on “Analysis Pipelines” along the top

In the “Select A Pipeline” section:
Hover over “Gene Set Characterization” and click on the “Start Pipeline” button
Step 6C – Select User Data Files

Leave the default species “Human”

To choose the genomics spreadsheet:
- Click on “Select a spreadsheet”
- Then click on the checkbox next to “top_genes_by_cluster.tsv”

Click “Next” in the bottom right corner
Step 6D – Configure Algorithm Parameters

We will choose to use a subset of 3 possible gene set collections available in the knowledge network:

- **Ontologies**: Gene Ontology (default)
- **Pathways**: KEGG Pathway (default)
- **Disease/Drugs**: Enrichr Phenotype Signature (needs to be added)
- (unclick Protein Domains: PFam Protein Domains)

Click “Next”
Step 6E – Configure KnowNet Parameters

Click “No” for question about using the Knowledge Network

Do you want to use the Knowledge Network?

Yes  No

This will run the one-sided Fisher exact test that is used by the DAVID webserver which you used in the lab earlier

Click “Next”
Step 6F – Launch GSC Workflow

Review your selections

Change your “Job Name” if you wish

Click “Submit Job”
Step 6G: Wait for Results to Complete

Once submitted, should see Running icon. Click on “Go to Results Page”.

When your NBS workflow status turns from “Running” (gray) to “Completed” (green), click on the name of the run

Then click on the blue “View Results” button

Results should be ready after several minutes
Step 7: Pathways that Overlap with Important Genes

Move the “Filter Matches by Score” slider all of the way to the right to see more results.

These pathways overlap with our 100 genes that are important for each genomic subtype.

Right click “Cluster_0” to sort columns to see the relationship of this type to other biological processes.

Click on individual heatmap cell to see the number of shared genes.
Step 8: Value of Network-Guided Analysis

Take advantage of gene neighbors

Integrating multiple data types
Step 8: DRaWR Method for GSC

DRaWR – using random walks on a network

- **Construct a network of interest**
- Find stationary distribution on network
- Find gene set specific distribution
- Return annotation nodes that are especially related to the query

---

Knowledge-Guided Analysis | KnowEnG Center | 2017
Step 8: DRaWR Method for GSC

DRaWR – using random walks on a network

- Construct a network of interest
- **Find stationary distribution on network**
- Find gene set specific distribution
- Return annotation nodes that are especially related to the query
Step 8: DRaWR Method for GSC

DRaWR – using random walks on a network

- Construct a network of interest
- Find stationary distribution on network
- Find gene set specific distribution
- Return annotation nodes that are especially related to the query
Step 8A – Configure Enrichment Analysis

Click on “Analysis Pipelines” along the top

In the “Select A Pipeline” section:

Hover over “Gene Set Characterization” and click on the “Start Pipeline” button

Leave the default species “Human”

To choose the genomics spreadsheet:
• Click on “Select a spreadsheet”
• Then click on the checkbox next to “top_genes_by_cluster.tsv”

Click “Next” in the bottom right corner
Step 8B – Configure Algorithm Parameters

We will choose to use a subset of 3 possible gene set collections available in the knowledge network

- Ontologies: **Gene Ontology** (default)
- Pathways: **KEGG Pathway** (default)
- Disease/Drugs: **Enrichr Phenotype Signature** (needs to be added)
- (unclick Protein Domains: PFam Protein Domains)

Click “Next”
Step 8C - Configure KnowNet Parameters

Click “Yes” for question about using the Knowledge Network

The Knowledge Network we will use is the evolutionary conservation network from protein homology (“BlastP Protein Sequence Similarity”)

A amount of network smoothing controls how much importance is put on network connections instead of the somatic mutations. We will use the default of 50%

Click “Next”
Step 8D – Launch GSC Workflow

Review your selections
Change your “Job Name” if you wish
Click “Submit Job”

Once submitted, should see Running icon.
Click on “Go to Results Page”.

When your NBS workflow status turns from “Running” (gray) to “Completed” (green), click on the name of the run

Then click on the blue “View Results” button
Step 9: Network-based Pathways that Overlap with Important Genes

Move the “Filter Matches by Score” slider all of the way to the right.

These pathways are especially well connected to the 100 genes that are important for each genomic subtype and/or their most similar proteins.

Sort by “Cluster_0” to see the additional heart and stress related terms that are found.

Click on the “Cluster_0” and “Heart Failure” individual heatmap cell to see that the number of overlap genes is low.
Logout and Wrapup:

**Feedback Survey:**
https://goo.gl/forms/L9L06tqoB1xPvjuM2

**KnowEnG Future Pipelines:**
https://hub.knoweng.org/features

**Knowledge Network Contents:**
https://hub.knoweng.org/app/site/knownet.html

**Other Test Examples (must be logged in):**
https://hub.knoweng.org/quickstart

**Contacts:**
Charles Blatti, Postdoc, blatti@illinois.edu
Subha Srinivasan, Program Manager, ssubha@illinois.edu
Sample Clustering
Clusters samples and scores clusters for their relation to provided phenotypes.

Gene Set Characterization
Finds systems-level properties, e.g., pathway, biological process, of gene set.

Gene Prioritization
Finds genes associated w/ phenotype, based on genomic profiles of a cohort and reports top ranking genes.

Literature Mining
Searches the corpus for occurrences of specified genes ranking the genes for relevance to each specified term.

Signature Analysis
Maps transcriptomic and other omics profiles of samples to best matching signatures.

Phenotype Prediction
Trains and uses stats model to predict numeric phenotype from genomic profile.

GRN Reconstruction
Predicts each gene’s expression as a function of expression levels of a few TFs.

Multi-Omics Analysis
Reports TFs and TF-gene relationships that can be linked to the phenotype.

KnowEnG Pipelines