Pan-Cancer Dysregulated Pathways and PolyTherapy AI

Margaret Liñán, MPH MS
Bioinformatics Consultant
Margaret.Linan@asu.edu

Valentin Dinu, PhD
Associate Professor of Biomedical Informatics
Translational Genomics Lab PI
College of Health Solutions
Arizona State University
Valentin.Dinu@asu.edu

CGC Virtual Webinar
Wednesday August 25th
This presentation contains information related to three experimental research studies by the authors and are not intended to guide clinical applications. The information presented here are the intellectual property of the authors and only represent experimental methods and results. External parties are prohibited from using any of the study related tables, graphics, results in this presentation without the permission of the authors. The third study's methods and results are already part of a manuscript in-preparation.
Overview of the TCGA Research Studies
Overview of the Translational Research Studies
The main workflows

Filter and collect all mRNA samples (solid normal and tumor)

Apply Prep Tools and PoTRA Pipeline (Docker/Rabix)

Collect Invasive Breast Cancer Pathway Ranks

Dispersion Analysis of Pathway Ranks

Build Graph Network with Cypher and Analyze with Neo4J

Determine if ncRNAs have predictive power

Build Rshiny platform mockup and test ability to host interactive graph networks

Use Unsupervised Learning to build and test multiple models

Detect most influential mRNA mediated dysregulated pathways in the pan-cancer graph network

Identify combination of solid normal and tumor that produces least deviation in the ranks

As our project progresses, we will be able to predict combination therapies

PoTRA Pipeline in the CGC Public Apps section soon
Browse 591 publicly available Common Workflow Language workflows and tools to enable reproducible bioinformatics.

Featured Apps

GRAF Germline Variant Detection Workflow

The GRAF Germline Variant Detection Workflow enables accurate alignment and variant calling by utilizing a genome graph reference that can address the bias and other limitations inherent in linear genome references. Seven bridges has constructed a comprehensive pan-genome graph that incorporates the diverse genetic composition of all populations around the world. By using this pan-genome graph, the GRAF Germline Variant Detection Workflow makes graph technology applicable at the whole genome level, enabling highly accurate and fast read alignment and variant calling. The current version of the workflow is 1.0 and supports both GRCh37 and GRCh38 versions of the Pan-Genome graphs.
Pan-Cancer Exploration of mRNA Mediated Dysregulated Pathways in the Cancer Genomics Cloud
Pathways of topological rank analysis (PoTRA)

- PoTRA measures the relative topological ranks of genes in each biological pathway, then selects hub genes for each pathway, and uses the Fishers Exact or Kolmogorov-Smirnov test to determine if the number of hub genes in each pathway is altered from normal to cancer (Li, Liu & Dinu, 2018).

Value of Detecting Dysregulated Pathways

- Recent studies have shown that both dysregulated or frequently mutated pathways should be used for the characterization of cancers instead of driver mutations (Zhang, Chien, Yong and Kuang, 2017). Network biology is an important approach that can detect frequently dysregulated pathways in distinct cancer types (Zhang, Chien, Yong and Kuang, 2017).
Graph Network Pan-Cancer Research with PoTRA and Neo4J

- The PoTRA algorithm was applied to 17 open-access TCGA project datasets of FPKM normalized mRNA genes from primary tumor and solid tissue normal samples in the CGC Cloud (Linan, Wang and Dinu, 2018).
- Docker and Rabix were utilized for wrapping libraries, computational tools and pipeline building.
- Neo4J’s cypher language was used to visualize and analyze the ranked dysregulated pathways.

https://www.biorxiv.org/content/10.1101/599225v6
Rabix Pipelines for PoTRA analysis
The mRNA mediated dysregulated pathways were detected and ranked by PoTRA and then visualized and analyzed with Neo4J. The Page Rank analysis determines the most important dysregulated pathway by traversing the graph network and checking which pathway node is connected to the most highly ranked pathways.

https://www.biorxiv.org/content/10.1101/599225v6
CGC App Preparation

Lessons Learned

1) Worthwhile to study other CGC Public Apps, to find helpful practices
2) CGC Apps require a development, pre-build and commit project
3) The workflows that will become CGC Apps can be tested in the CGC Platforms workflow environment – this is better than Rabix desktop because it will give you access to data sets in Project.
Results
Conclusion/Discussion

- Overall, the PoTRA tool can accurately identify dysregulated pathways even in tumor samples with intra-tumoral heterogeneity by using a majority-rules approach.

- Also, the PoTRA algorithm can identify dysregulated pathways by using only mRNA while other algorithms arrive at similar conclusions by using a multi-omics approach.

- The most consistent ranks were found for the following primary sites: Breast, Colorectal, Kidney, Liver, Lung, Prostate and Thyroid.

- The PoTRA algorithm accurately identified the dysregulated pathways associated with each TCGA cancer project. It was also easily used in the cancer genomics cloud and therefore is suited for large scale analyses.

- The PoTRA algorithm will continue to be developed by the Dinu Lab at ASU’s College of Health Solutions.
Lessons Learned

- The CGC’s TCGA data browser is an excellent tool for properly stratifying the existing datasets and helps researchers accurately select only the most important datasets and data types.
- The CGC Task environment is critical for tracking down previous analyses and previous results files, as well as the methodologies that were utilized to perform the analysis.
- Rabix is a powerful tool that can help speed up bioinformatics analyses by enabling researchers to quickly select large amounts of data as input for a single algorithm or complex pipeline.
- The CGC bioinformatics staff are effective partners in troubleshooting unexpected errors.
- The CGC platform makes it easy to run very large-scale bioinformatics analyses.
Dispersion analysis of PoTRA ranked mRNA mediated dysregulated pathways in Breast Invasive Cancer from a TCGA Pan-Cancer study
• In data mining and machine learning class imbalance impacts the accuracy and error rate of classifiers.

• Similarly in bioinformatics, computational tools that are applied to unbalanced data sets will have more variation in its results.
Background
To identify the optimal solid tissue normal:primary tumor sample ratios, the SD analysis used random combinations of 1:N unbalanced solid tissue normal:primary tumor data sets:
- 1:N with N = 1-9

To identify the minimum sample size, random resampling of solid tissue normal and primary tumor samples of various sizes are used:
- 3 vs 3
- 5 vs 5
- 10 vs 10
- 25 vs 25
- 50 vs 50
- 75 vs 75
- 100 vs 100
- 113 vs 113
Rabix Pipelines for PoTRA dispersion analysis
Results

These results suggest that the 1:1 ratio (solid tissue normal and primary tumor) achieves the lowest average rank variation and that the minimum sample size of 50 normal and 50 tumor samples reaches a steady state in the average rank variation.
Conclusion/Discussion

• In conclusion, future applications of the PoTRA algorithm to analyze gene expression data sets such as TCGA should use balanced data sets as well as a minimum sample size of 50 for both solid tissue normal and primary tumor to ensure the most robust performance.

• The present work also demonstrates how pathway ranks are changed by data set size. Interestingly, the MAPK pathway had the least variation in the different ratios of solid tissue normal: primary tumor, perhaps because this pathway is very active in breast invasive carcinoma.

• In contrast, the cAMP 199 signaling pathway had the greatest variability, perhaps because this pathway is associated with tumor progression and therefore targeted by chemotherapies that are prescribed to the BRCA 201 patients.

• In fact, the pathways with greatest variability (cAMP signaling, Human Papillomavirus infection, PI3K-Akt signaling pathway, Proteoglycans in cancer) also had no detectable differences (FE Test P-value > 0.05) in hub genes between solid tissue normal and primary tumor mRNA networks in ratios 1:7 and 1:9. This may indicate that PoTRA can be used to measure the efficacy of a chemotherapy that targets genes in particular pathways.
Lessons Learned

- In the CGC project environment it is easy reuse
  - Data files that were generated in previous projects.
  - Previous tools that were created in Rabix.
- The CGC task environment makes it easy to generate and collect images that require a lot of RAM.
- The Tag feature makes it easy to quickly label and re-label results files.
- The CGC safely stores all the algorithms and pipelines that were built in Rabix for every project.
- The CGC bioinformatics staff are effective partners in troubleshooting unexpected errors.
PolyTherapy AI

Principal Investigators

- Dr. Valentin Dinu, Associate Professor of Biomedical Informatics at the College of Health Solutions at Arizona State University
- Dr. Marinka Zitnik, Assistant Professor of Biomedical Informatics at the Harvard Medical School
Petabytes of data, complex bioinformatics algorithms and the international crisis known as the Covid19 pandemic highlighted the need for a gold standard platform capable of accelerating massive scale analyses of heterogenous datasets (multi-omics, clinical, etc).

The PolyTherapy AI platform project was created by me and my team so that the Microsoft Quantum Azure cloud and quantum machine learning could make it possible to accelerate the prediction of multi-omics based combination therapies using new and trusted bioinformatics algorithms, all by the speed of light.
Initial Stage of Development

1. Pan-Cancer Druggable Mutations Graph Networks
2. Build Rshiny platform mockup and test ability to host interactive graph networks
3. Determine ncRNAs that have predictive power
4. Use Unsupervised Learning to build and test multiple models
1. Created TCGA GRCh38 Drug Gene Interactions, Interactive and Searchable Environment for each open access TCGA project using Rshiny.
2. Drafted PolyTherapy AI Dashboard Mockup in Rshiny.
3. Stress tested Rshiny for Large Graph Networks with CGC staff.
1. Tested hypothesis that ncRNAs can be predicative biomarkers for cancers.
2. Only a subset of these ncRNAs can be used as biomarkers for each cancer type.
3. Validated our results with a biomedical literature search.
4. Trained multiple deep learning models.
Pre-Training using Unsupervised Learning

Classification result for balanced data
Worst performing model: Epoch 1 – Accuracy 97.5%
Best performing model: Epoch 5 – Accuracy 99.5%
Impact of Imbalanced Datasets in Unsupervised Learning
1. Severely skewed datasets will impact the accuracy of model
2. Permutations are often used to pad the imbalanced groups
3. Variety of solutions for balancing these types of data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Machine Learning Metrics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-balanced</td>
<td>Validation Loss</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>F1 Score (weighted)</td>
<td>0.98</td>
</tr>
<tr>
<td>Imbalanced</td>
<td>Validation Loss</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td>F1 Score (weighted)</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Lessons Learned

• The CGC’s TCGA data browser is an excellent tool for properly stratifying the existing datasets and helps researchers accurately select only the most important datasets and data types.

• The CGC Task environment is critical for tracking down previous analyses and previous results files, as well as the methodologies that were utilized to perform the analysis.

• Rabix is a powerful tool that can help speed up bioinformatics analyses by enabling researchers to quickly select large amounts of data as input for a single algorithm or complex pipeline.

• The CGC bioinformatics staff are effective partners in troubleshooting unexpected errors and benchmarking.

• The CGC platform makes it easy to run very large-scale bioinformatics analyses.
Important Points
Important Points

• Reduce variation in PoTRA’s results by using
  • Data sets with a minimum of 50 cases and controls
  • Applying non-replacement random resampling to create data sets with a 1:1 ratio of cases and controls

• ncRNAs are good predictors of cancer.
Future Work

• Expand the PoTRA algorithm (available on Bioconductor) to include other genomic data types and advanced visualizations.
Thank You

**Acknowledgments**
Cancer Genomics Cloud Team
National Cancer Institute
The Cancer Genome Atlas
PolyTherapy AI Team

Margaret Linan, MPH MS
Bioinformatics Consultant
Margaret.Linan@asu.edu

Valentin Dinu, PhD
Associate Professor of Biomedical Informatics
Translational Genomics Lab PI
College of Health Solutions
Arizona State University
Valentin.Dinu@asu.edu