Breaking the curse of dimensionality

Explainable-AI and Evidence Mining as Applied to Systems Biology

Dan Jacobson

This research is supported by an INCITE award and uses resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE AC05-00OR22725.
Experimental Data Types

• Natural Variation
  – Genome Wide Association Studies
  – 28 Million of SNPs
  – ~140,000 Primary Phenotypes
    • Morphology/Phenology
    • Molecular

• Microbiomes & Metagenomes

• Omics & Meta-omics
  – Genomics, Transcriptomics, Proteomics, Metabolomics

• All publically available Genomes

• Differential/Time Series Expression Studies

• Systems Biology Approach
  – Combining datasets across omics layers, sample sets, and species
<table>
<thead>
<tr>
<th>Gene</th>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pavir.Aa00004</td>
<td>23.03260874</td>
<td>-0.772176419</td>
<td>0.235718754</td>
<td>-3.275837864</td>
<td>0.00105349</td>
<td>0.036650136</td>
</tr>
<tr>
<td>Pavir.Aa00007</td>
<td>3.617339133</td>
<td>-3.277187207</td>
<td>0.925328577</td>
<td>-3.541647029</td>
<td>0.000397637</td>
<td>0.016344905</td>
</tr>
<tr>
<td>Pavir.Aa00318</td>
<td>11.69495376</td>
<td>-1.375554763</td>
<td>0.421360908</td>
<td>-3.264552399</td>
<td>0.001096372</td>
<td>0.037862673</td>
</tr>
<tr>
<td>Pavir.Aa01140</td>
<td>432.2298561</td>
<td>-0.920355344</td>
<td>0.087912301</td>
<td>-10.46901667</td>
<td>1.20E-25</td>
<td>1.26E-22</td>
</tr>
<tr>
<td>Pavir.Aa01336</td>
<td>14.76644122</td>
<td>-7.964343955</td>
<td>1.643802037</td>
<td>-4.84507488</td>
<td>1.27E-06</td>
<td>0.000109099</td>
</tr>
<tr>
<td>Pavir.Aa01612</td>
<td>63.51089454</td>
<td>1.524126268</td>
<td>0.377624869</td>
<td>4.03608552</td>
<td>5.44E-05</td>
<td>0.029659159</td>
</tr>
<tr>
<td>Pavir.Aa01614</td>
<td>86.61299946</td>
<td>1.970704034</td>
<td>0.235133135</td>
<td>8.381226395</td>
<td>5.24E-17</td>
<td>2.16E-14</td>
</tr>
<tr>
<td>Pavir.Aa01686</td>
<td>45.57577197</td>
<td>-2.776318341</td>
<td>0.3350917</td>
<td>-8.285249514</td>
<td>1.18E-16</td>
<td>4.66E-14</td>
</tr>
<tr>
<td>Pavir.Aa01805</td>
<td>7.78468493</td>
<td>1.72469978</td>
<td>0.269249957</td>
<td>6.405571227</td>
<td>1.50E-10</td>
<td>2.64E-08</td>
</tr>
<tr>
<td>Pavir.Aa01856</td>
<td>15.77390176</td>
<td>-3.03656463</td>
<td>0.739522148</td>
<td>-4.106117228</td>
<td>4.02E-05</td>
<td>0.00228249</td>
</tr>
<tr>
<td>Pavir.Aa01950</td>
<td>246.4158349</td>
<td>0.749398201</td>
<td>0.130879565</td>
<td>5.725861023</td>
<td>1.03E-08</td>
<td>1.35E-06</td>
</tr>
<tr>
<td>Pavir.Aa02015</td>
<td>194.2868719</td>
<td>0.55688662</td>
<td>0.146658117</td>
<td>3.797209232</td>
<td>0.000146334</td>
<td>0.007032352</td>
</tr>
<tr>
<td>Pavir.Aa02010</td>
<td>71.8661413</td>
<td>-0.894576165</td>
<td>0.223959112</td>
<td>-4.222539364</td>
<td>2.42E-05</td>
<td>0.001450415</td>
</tr>
<tr>
<td>Pavir.Aa02130</td>
<td>45.08826603</td>
<td>-2.821545181</td>
<td>0.381707372</td>
<td>-7.391906442</td>
<td>1.45E-13</td>
<td>3.90E-11</td>
</tr>
<tr>
<td>Pavir.Aa02199</td>
<td>82.09354863</td>
<td>2.652283666</td>
<td>0.48092843</td>
<td>5.514923839</td>
<td>3.49E-08</td>
<td>4.08E-06</td>
</tr>
<tr>
<td>Pavir.Aa02377</td>
<td>48.01170214</td>
<td>1.765138681</td>
<td>0.318940668</td>
<td>5.534379463</td>
<td>3.12E-08</td>
<td>3.70E-06</td>
</tr>
<tr>
<td>Pavir.Aa02382</td>
<td>4.900020424</td>
<td>-6.641133505</td>
<td>1.55203693</td>
<td>-4.278971603</td>
<td>1.88E-05</td>
<td>0.001166295</td>
</tr>
<tr>
<td>Pavir.Aa02400</td>
<td>3.536708979</td>
<td>-2.288869563</td>
<td>0.396004267</td>
<td>-5.77991361</td>
<td>7.47E-09</td>
<td>1.01E-06</td>
</tr>
<tr>
<td>Pavir.Aa02455</td>
<td>100.2653536</td>
<td>0.851939179</td>
<td>0.154407276</td>
<td>5.517480799</td>
<td>3.44E-08</td>
<td>4.03E-06</td>
</tr>
<tr>
<td>Pavir.Aa02456</td>
<td>74.76890191</td>
<td>0.900755926</td>
<td>0.267107154</td>
<td>3.37226319</td>
<td>0.000745529</td>
<td>0.027702451</td>
</tr>
<tr>
<td>Pavir.Aa02462</td>
<td>129.7507791</td>
<td>1.878568856</td>
<td>0.195429139</td>
<td>9.612532015</td>
<td>7.08E-22</td>
<td>5.19E-19</td>
</tr>
<tr>
<td>Pavir.Aa02463</td>
<td>0.855875118</td>
<td>-3.952874961</td>
<td>1.177355482</td>
<td>-3.357418402</td>
<td>0.00078674</td>
<td>0.028956754</td>
</tr>
<tr>
<td>Pavir.Aa02517</td>
<td>239.8175815</td>
<td>3.424148863</td>
<td>0.634311687</td>
<td>5.398211843</td>
<td>6.73E-08</td>
<td>7.46E-06</td>
</tr>
<tr>
<td>Pavir.Aa02526</td>
<td>20.12897762</td>
<td>-1.829988585</td>
<td>0.513742501</td>
<td>-3.56207357</td>
<td>0.000367937</td>
<td>0.015318345</td>
</tr>
<tr>
<td>Pavir.Aa02574</td>
<td>1.957536218</td>
<td>-5.978272647</td>
<td>1.222823914</td>
<td>-4.888907208</td>
<td>1.01E-06</td>
<td>8.89E-05</td>
</tr>
<tr>
<td>Pavir.Aa02621</td>
<td>0.909365395</td>
<td>-6.53529993</td>
<td>1.672432432</td>
<td>-3.907661562</td>
<td>9.32E-05</td>
<td>0.004726253</td>
</tr>
<tr>
<td>Pavir.Aa02666</td>
<td>26.2769212</td>
<td>0.691682664</td>
<td>0.195671446</td>
<td>3.534918755</td>
<td>0.000407901</td>
<td>0.01668753</td>
</tr>
<tr>
<td>Pavir.Aa02688</td>
<td>20.64051337</td>
<td>1.419916888</td>
<td>0.311120505</td>
<td>4.563880767</td>
<td>5.02E-06</td>
<td>0.000367199</td>
</tr>
<tr>
<td>Pavir.Aa02777</td>
<td>32.70837314</td>
<td>0.824566433</td>
<td>0.256714392</td>
<td>3.211999724</td>
<td>0.001318147</td>
<td>0.044226251</td>
</tr>
<tr>
<td>Pavir.Aa02799</td>
<td>5.953157198</td>
<td>1.635139531</td>
<td>0.489562315</td>
<td>3.340002856</td>
<td>0.000837775</td>
<td>0.030512025</td>
</tr>
<tr>
<td>Pavir.Aa02841</td>
<td>4.061306867</td>
<td>-1.69398357</td>
<td>0.345840001</td>
<td>-4.898171305</td>
<td>9.67E-07</td>
<td>8.51E-05</td>
</tr>
<tr>
<td>Pavir.Aa03067</td>
<td>7.20334301</td>
<td>-6.09679446</td>
<td>1.535018046</td>
<td>-3.971806374</td>
<td>7.13E-05</td>
<td>0.003773958</td>
</tr>
</tbody>
</table>
Integrated Vision: From Systems Biology to 3D Structural Interactions

Structures: 28 million compounds
Integrated Vision: From Human Systems Biology to 3D Structural Interactions - Pharmacogenomics and Personalized Medicine

- Co-evolution
  - 1000 Genomes Project
  - Protein cross species
- Human RNA-seq
- Crystal Structures
- Protein-protein interaction
- Human interactome
- ENCODE
- Explainable-AI
  - iRF/TiRF
  - DNNs
  - MIPs
- Public GWAS data
- As available from collaboration with VA
  - Genetic data
  - Clinical phenotypes
  - Polypharmacy data

Structures:
- 28 million compounds

Sequences
- Existing Structure
- Protein Complexes
- Co-evolution Network
- 1000+ Genomes
- Phenotypes
- GWAS Networks

Experimentation

summit
Integrated Vision: From Human Systems Biology to 3D Structural Interactions - Pharmacogenomics and Personalized Medicine

Structures: 28 million compounds

Experimentation
Single QTL mapping: 28 million tests per phenotype

- SNP Vectors
- Phenotype Vectors

- SNP Matrix expansion from interpolation (ANL)
- Control for effects of population structure
140,000 Manhattan Plots???
Network Theory

• Networks can be used to represent biological systems
  – Nodes
    • Represent any object (genes, SNPs, proteins, metabolites, species, microbiomes, etc.)
  – Edges
    • Represent a relationship between two nodes (correlation, co-occurrence, physical contact, etc.)
    • Relationships can be quantitative (represented by the thickness of the line)

• Integration and Visualization of Systems Biology Models

• Mathematical Structure
  – Allows to be computed upon
  – Millions of nodes
  – Trillions of edges
Deeper Discoveries in Systems Biology: The Balance Between Type 1 and Type 2 Error

Our ability to reconstruct the entirety of a complex biological system improves as the number of population-scale endo-, meso- and exo-phenotypes are measured and combined with deep layers of experimental data collected on individual genotypes.

GWAS: Single QTL Mapping

• Very Powerful

• Frequently does not capture a significant portion (often the majority) of the genetic signal

• Often does not find complete genetic architectures for complex phenotypes (Dementia, Alzheimer’s, Schizophrenia, Cardiovascular disease, PTSD, Suicide, Addiction, etc.)
Epistatic Example: Transcription Initiation Complex
The Need for Speed

Pairwise Epistatic Compute Time per Phenotype

4-way combinations = $2.4 \times 10^{20}$ CPU hours per phenotype
Breaking the curse of dimensionality

10M Genetic Variants in >40k genes

Genes do not work in isolation: $10^{170}$ potential interactions among variants

Linking genetic variants to phenotypes requires the exploration of an enormous space

To obtain accuracy and insight, we are developing procedures to detect interactions of any form or order at the same computational cost as main effects

Explainable-AI
Machine and Deep Learning Algorithms

• Great at classification
• Essentially black boxes
  • Don’t reveal the interactions between variables that lead to the classification
• Need Explainable AI
Finding Higher Order Combinatorial Interactions in Complex Systems

- X matrix and Y vector
- Iterative Random Forests
iRF Workflow

Training Data → Test Data → Prediction Accuracy

Tree Ensemble

Feature Interactions

Weighted Random Forest

Feature Importance

Interactions (refined)

Random Intersection Trees

Bootstrap Aggregation (Bagging)

Node Filtering

Interactions (stabilized)

DOE Collaboration: Ben Brown - LBNL
iRF – X Matrix and 1 Y Vector

SNP Vectors

Phenotype Vectors
iRF – X Matrix and 1 Y Vector

SNP Vectors

Phenotype Vectors

4-way combination = 1000 CPU hours per phenotype (140,000 phenotypes)
Tensor iterative Random Forests (TiRFs)

- Effectively build forests that can be mined for interactions within a multi-dimensional X, a multi-dimensional Y and interactions between multiple dimensions in X and Y, all at the same time.
SNP Vectors

Phenotype Vectors
TiRF – X Matrix and Y Matrix Simultaneously

SNP Vectors

Phenotype Vectors
Clinical Genomics and Human Systems Biology: DOE & VA – MVP Champion

- ORNL
  - Clinical records 23+ million patients, 20 years
  - 358,000 Genotypes
  - => 4 million genotypes
VA Use Case: Polypharmacy

• Simultaneous use of multiple medication
• Of concern if 5 or more medications are used
Why Worry About Polypharmacy?

• Drugs interact with each other, the more you are on, the more interactions can occur

• Side effects add up and are more pronounced in older individuals

• Medications are approved by FDA based on short term trails that typically exclude:
  – Those with other diagnoses on other medications
  – 65+ year olds
Interaction Network

• Drug Set Simulations
  – For all set sizes 2 – 30
    • Create 20 million random sets of drugs for each set size
    • 58 million sets
  – Check for drug to drug edges amongst all possible pairs in each set for the shared target and shared pathway networks
    • 567 Billion interaction tests

• Clinical Data
  – Create drug sets from clinical records
  – Check for drug to drug edges amongst all possible pairs in each set for the shared target and shared pathway networks
Drug Interaction: Simulation vs Clinical Practice

Known Interactions vs Drug Set Size

- Non_HIV_Mean
- Simulation
Polypharmacy Morbidity & Mortality

![Graph showing the relationship between the number of drugs and interactions, with hazard ratios for different conditions and simulations.](image-url)
VA: Preliminary Results

- **Polypharmacy**
  - Clinically relevant patterns
  - iRF
    - Steps toward automated phenotyping
    - Interaction edges -> morbidity & mortality

- **Diseaseome**
  - iRF on diagnostic codes
  - 600,000 patients
  - Relationships between all known human conditions
    - Co-morbidity map
  - Discovered 9\textsuperscript{th}-order combinations
    - \(1.5 \times 10^{26}\) possible 9-way combinations
Life Science data: Multi-omics, multi-technology.
TiRF – Any Set of Matrices or Tensor Dimensions *Simultaneously*

- Spatial and temporal/longitudinal information
- Different Omics layers (genome, transcriptome, proteome, metabolome, microbiome…)
- Quantum chemical tensors
Tensors: Matrices → Cubes
Tensors: Matrices → Cubes → Polytopes
From data matrix to cube to polytopes.
Tensor iterative Random Forests (TiRFs)

- Effectively build forests that can be mined for interactions within a multi-dimension X, a multi-dimension Y and interactions between multiple dimensions in X and Y, all at the same time.

Applications in Systems Biology
- Plants
- Microbes
- Humans, Mice
- *Drosophila*

Applications in Text Mining
- Electronic Health Records
- Scientific Literature

Simulation Models
- Combinatorial parameter sweeps (X) model output (Y)

- Any domain with high a dimensional set of matrices

Iterative Deep Neural Networks (iDNNs)

- Unpacking the black box
- Discovering the interactions encoded in DNNs
High Order Interactions: Explainable AI: Machine and Deep Learning Integration

MIPs

DNNs

TiRFs

Experimentation

1000+ Genomes

GWAS Networks

Phenotypes

Co-evolution Network

Protein Complexes

3D Protein Structure

Existing Structure

Sequences

Small Molecule

Existing Structure

Sequences
High Order Interactions: Explainable AI: Machine and Deep Learning Integration
Exposome
High Order Interactions: Exposome – Adverse Outcome Networks
Explainable AI: Machine and Deep Learning Integration

TiRFs
DNNs
MIPs

Titan/Summit

Experimentation

1000+ Genomes
Phenotypes
GWAS Networks
Co-evolution Network
Protein Complexes
3D Protein Structure
Sequences
Existing Structure

Small Molecule

General external environment
Specific external environment

TiRFs

DNNs
MIPs
Acknowledgements

Collaborators

- Ben Brown
- Jerry Tuskan
- Steve DiFazio
- Wayne Joubert
- Amy Justice
- Edmon Begoli

Computational Infrastructure

- Oak Ridge Leadership Computing Facility (OLCF)
- Compute and Data Environment for Science (CADES)

Joint Genome Institute
Acknowledgements

CBI
PMI
LDRD
VA
Oak Ridge Leadership Computing Facility (OLCF) at ORNL
Compute and Data Environment for Science (CADES) at ORNL
INCITE
Joint Genome Institute (JGI)

Oak Ridge National Laboratory (ORNL)
Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville
Acknowledgements

• JAIL Team effort

– Debbie Weighill
– Piet Jones
– Carissa Bleker
– Armin Geiger
– Marek Piatek
– Ben Garcia
– Ashley Cliff
– Jonathon Romero
– David Kainer
– Annie Fouche
– Sandra Truong
– Ryan McCormick
– Priya Ranjan
– Manesh Shah
– Doug Hyatt
– Blake Wiley
– Jesse Marks
– Ian Hodge
– Annabel Large
– Chris Ellis
Questions?