Grammatical Evolution Neural Networks for Genetic Epidemiology

Alison Motsinger-Reif, PhD

Bioinformatics Research Center
Department of Statistics
North Carolina State University
Overview

• Epistasis and its implications for genetic analysis
• GENN Method
  – Optimization and dissection of the evolutionary process
  – Comparison to other NN applications
  – Comparison the other methods used in genetic epidemiology
  – Power studies
  – Application to an HIV Immunogenetics dataset
• Future directions
Genetics of Human Disease

Single Gene $\rightarrow$ Single Disease

Multiple Genes $\rightarrow$ Complex Disease
Epistasis

gene-gene or gene-environment interactions;

two or more genes interacting in a non-additive manner to confer a phenotype
Epistasis

- Biologists believe bio-molecular interactions are common
- Single locus studies do not replicate
- Identifying “the gene” associated with common disease has not been successful like it has for Mendelian disease
- Mendelian single-gene disorders are now being considered complex traits with gene-gene interactions (modifier genes)
“gene-gene interactions are commonly found when properly investigated”

[Moore (2003)]
Traditional Statistical Approaches

• Typically one marker or SNP at a time to detect loci exhibiting main effects

• Follow-up with an analysis to detect interactions between the main effect loci

• Some studies attempt to detect pair-wise interactions even without main effects

• Higher dimensions are usually not possible with traditional methods
Traditional Statistical Approaches

• Logistic Regression
  – Small sample size can result in biased estimates of regression coefficients and can result in spurious associations (Concato et al. 1993)
  – Need at least 10 cases or controls per independent variable to have enough statistical power (Peduzzi et al 1996)
  – Curse of dimensionality is the problem (Bellman 1961)
Curse of Dimensionality

N = 100
50 Cases, 50 Controls

SNP 1

AA  Aa  aa
Curse of Dimensionality

N = 100  
50 Cases,  
50 Controls

<table>
<thead>
<tr>
<th>SNP 1</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNP 2</td>
<td>Bb</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bb</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Curse of Dimensionality**

N = 100

50 Cases, 50 Controls

<table>
<thead>
<tr>
<th>SNP 4</th>
<th>SNP 2</th>
<th>SNPs 1</th>
<th>SNP 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD</td>
<td>BB</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
<tr>
<td></td>
<td>Bb</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
<tr>
<td></td>
<td>bb</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
<tr>
<td>Dd</td>
<td>BB</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
<tr>
<td></td>
<td>Bb</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
<tr>
<td></td>
<td>bb</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
<tr>
<td>dd</td>
<td>BB</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
<tr>
<td></td>
<td>Bb</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
<tr>
<td></td>
<td>bb</td>
<td>AA Aa aa</td>
<td>BB Aa aa</td>
</tr>
</tbody>
</table>
Traditional Statistical Approaches

• Advantages
  – Easily computed
  – Easily interpreted
  – Well documented and accepted

• Disadvantages
  – Susceptibility loci must have significant main effect
  – Difficult to detect purely interactive effects
  – Need a very large sample size to explore interactions between more than two variables
Objectives for Novel Methods

- Variable Selection
  - Choose a subset of variables from an effectively infinite number of combinations
- Statistical Modeling
- Generate Testable Hypotheses
Objectives for Novel Methods

- **Variable Selection**
  - Choose a subset of variables from an effectively infinite number of combinations
- **Statistical Modeling**
- **Generate Testable Hypotheses**

**GOAL**: Detect genetic/environmental factors associated with disease risk in the presence or absence of main effects from a large pool of potential factors
Methods to Detect Epistasis

• Multifactor Dimensionality Reduction (MDR)
• Random Forests™
• Restricted Partition Method (RPM)
• Classification and Regression Trees (CART)
• Symbolic Discriminant Analysis (SDA)
• Focused Interaction Testing Framework (FITF)
• Set Association
• Combinatorial Partitioning Method (CPM)
• Patterning and Recursive Partitioning (PRP)
• ............
Methods to Detect Epistasis

- Multifactor Dimensionality Reduction (MDR)
- Random Forests™
- Restricted Partition Method (RPM)
- Classification and Regression Trees (CART)
- Symbolic Discriminant Analysis (SDA)
- Focused Interaction Testing Framework (FITF)
- Set Association
- Combinatorial Partitioning Method (CPM)
- Patterning and Recursive Partitioning (PRP)

There are theoretical and/or practical concerns with each!
How Many Combinations are There?

- Genome-wide association studies
- ~500,000 SNPs to span the genome

Number of Possible Combinations

SNPs in each subset

- 1 SNP: $5 \times 10^5$
- 2 SNPs: $1 \times 10^{11}$
- 3 SNPs: $2 \times 10^{16}$
- 4 SNPs: $3 \times 10^{21}$
- 5 SNPs: $2 \times 10^{26}$
How Many Combinations are There?

- Genome-wide association studies
- ~500,000 SNPs to span the genome

Number of Possible Combinations vs. SNPs in each subset

- $2 \times 10^{26}$ combinations
- * 1 combination per second
- * 86400 seconds per day

---------

3.0 x $10^{21}$ days to complete (8.2 x $10^{18}$ years)
How Many Combinations are There?

- Genome-wide association studies
- ~500,000 SNPs to span the genome

We need methods to detect epistatic interactions without examining all possible combinations!!!
Novel Approaches

- Pattern Recognition
  - Considers full dimensionality of the data
  - Aims to classify data based on information extracted from the patterns
    - Neural Networks (NN)
    - Clustering Algorithms
    - Self-Organizing Maps (SOM)
    - Cellular Automata (CA)
Neural Networks

- Developed 60 years ago
- Originally developed to model/mimic the human brain
- More recently, uses theory of neurons to do computation
- Applications
  - Association, classification, categorization
Neural Networks

- NNs multiply each input node (i.e. variable, genotype, etc.) by a weight (\(a\)), the result of which is processed by a function (\(\Sigma\)), and then compared to a threshold to yield an output (0 or 1).
- Weights are applied to each connection and optimized to minimize the error in the data.
Neural Networks

• Advantages
  – Can handle large quantities of data
  – Universal function approximators
  – Model-free

• Limitations
  – Must fix architecture prior to analysis
  – Only the weights are optimized
  – Weights are optimized using hill-climbing algorithms
Neural Networks

• Advantages
  – Can handle large quantities of data
  – Universal function approximators
  – Model-free

• Limitations
  – Must fix architecture prior to analysis
  – Only the weights are optimized
  – Weights are optimized using hill-climbing algorithms

• Solution: Evolutionary computation algorithms can be used for the optimization of the *inputs, architecture*, and *weights* of a NN to improve the power to identify gene-gene interactions.
Grammatical Evolution

• Evolutionary computation algorithm inspired by the biological process of transcription and translation.

• Uses linear genomes and a grammar (set of rules) to generate computer programs.

• GE separates the genotype from the phenotype in the evolutionary process and allows greater genetic diversity within the population than other evolutionary algorithms.
DNA: The heritable material in GE is the binary string chromosome. The GE chromosome is divided into codons, undergoes crossover and mutation, and can contain non-coding sequence just as biological DNA.

RNA: In GE, the binary chromosome string in transcribed into an integer string. This integer string is a linear copy message of the original heritable material that can then be processed further.

Polypeptide String: The integer string is translated using the grammar provided into the code for a functional NN.

Protein Folding: The grammar encoding is then interpreted as a multi-dimensional NN. This NN produces a classification error, just as a protein produces a phenotype within an organism.

Function: In GE a lower classification error indicates higher fitness. Natural selection will work at the level of reproductive fitness, forcing changes in the heritable material of both biological organisms or GE individuals.
Step 1: A population of individuals is randomly generated, where each individual is a binary string chromosome (genetic material). The number of individuals is user-specified.

Step 2: Individuals are randomly chosen for tournaments – where they compete with other individuals for the highest fitness, and the tournament winners get to pass on their genetic material.

Step 3: Of the winners, user-specified proportions participate in crossover, mutation, or duplication of their genomes to produce offspring.

Step 4: When pooled together, these offspring will become the initial population for the next generation of evolution.

Steps 1-4 are repeated for a user-specified number of generations, to produce offspring with the highest possible fitness.
GE Neural Networks

**STEP 1**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>population_size</td>
<td>200</td>
</tr>
<tr>
<td>max_generations</td>
<td>50</td>
</tr>
<tr>
<td>pvm_exchange_generations</td>
<td>25</td>
</tr>
<tr>
<td>random_seed</td>
<td>7</td>
</tr>
<tr>
<td>crossover_rate</td>
<td>0.9</td>
</tr>
<tr>
<td>mutation_rate</td>
<td>0.01</td>
</tr>
<tr>
<td>codon_size</td>
<td>8</td>
</tr>
<tr>
<td>wrapper_count</td>
<td>2</td>
</tr>
<tr>
<td>min_chrom_size</td>
<td>50</td>
</tr>
<tr>
<td>max_chrom_size</td>
<td>1000</td>
</tr>
</tbody>
</table>

**STEP 2**

**STEP 3**

**STEP 4**

**STEP 5**

**STEP 6**

<table>
<thead>
<tr>
<th>Classification Error</th>
<th>Prediction Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.25</td>
<td>21.55</td>
</tr>
</tbody>
</table>

**GENN Model**

**GENN Models**

<table>
<thead>
<tr>
<th>Classification Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.25</td>
</tr>
<tr>
<td>22.12</td>
</tr>
<tr>
<td>24.33</td>
</tr>
<tr>
<td>28.14</td>
</tr>
</tbody>
</table>

Tournament
<table>
<thead>
<tr>
<th>CV</th>
<th>Factors in Model</th>
<th>CE</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SNP_1 SNP_200</td>
<td>0.26</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>SNP_1 SNP_200 SNP_630 SNP_755</td>
<td>0.38</td>
<td>0.22</td>
</tr>
<tr>
<td>3</td>
<td>SNP_1 SNP_200 SNP_512</td>
<td>0.32</td>
<td>0.29</td>
</tr>
<tr>
<td>4</td>
<td>SNP_1 SNP_200 SNP_333 SNP_467 SNP_987</td>
<td>0.19</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>SNP_1 SNP_200 SNP_814 SNP_900</td>
<td>0.12</td>
<td>0.32</td>
</tr>
<tr>
<td>6</td>
<td>SNP_1 SNP_200 SNP_665</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>7</td>
<td>SNP_1 SNP_200 SNP_742 SNP_801</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td>8</td>
<td>SNP_1 SNP_200 SNP_245 SNP_294</td>
<td>0.19</td>
<td>0.28</td>
</tr>
<tr>
<td>9</td>
<td>SNP_1 SNP_200 SNP_410 SNP_502 SNP_873</td>
<td>0.18</td>
<td>0.28</td>
</tr>
<tr>
<td>10</td>
<td>SNP_1 SNP_200 SNP_311</td>
<td>0.26</td>
<td>0.18</td>
</tr>
</tbody>
</table>
# Example Results

<table>
<thead>
<tr>
<th>CV</th>
<th>Factors in Model</th>
<th>CE</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SNP_1 SNP_200</td>
<td>0.26</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>SNP_1 SNP_200 SNP_630 SNP_755</td>
<td>0.38</td>
<td>0.22</td>
</tr>
<tr>
<td>3</td>
<td>SNP_1 SNP_200 SNP_512</td>
<td>0.32</td>
<td>0.29</td>
</tr>
<tr>
<td>4</td>
<td>SNP_1 SNP_200 SNP_333 SNP_467 SNP_987</td>
<td>0.19</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>SNP_1 SNP_200 SNP_814 SNP_900</td>
<td>0.12</td>
<td>0.32</td>
</tr>
<tr>
<td>6</td>
<td>SNP_1 SNP_200 SNP_665</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>7</td>
<td>SNP_1 SNP_200 SNP_742 SNP_801</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td>8</td>
<td>SNP_1 SNP_200 SNP_245 SNP_294</td>
<td>0.19</td>
<td>0.28</td>
</tr>
<tr>
<td>9</td>
<td>SNP_1 SNP_200 SNP_410 SNP_502 SNP_873</td>
<td>0.18</td>
<td>0.28</td>
</tr>
<tr>
<td>10</td>
<td>SNP_1 SNP_200 SNP_311</td>
<td>0.26</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Significance Testing

- Final Model is forced
- Average PE is calculated
- Permutation testing is used to ascribe statistical significance to the model

\[
\begin{align*}
\text{SNP}_1 & \rightarrow -2.38 \\
\text{SNP}_200 & \rightarrow 12.58 \\
\end{align*}
\]

Prediction Error: 15.4%
p<0.01
Successes of GENN

• High power to detect a wide range of main effect and interactive models

• Robust to changes in the evolutionary process

• Higher power than traditional BPNN, GPNN, or random search NN
Successes of GENN

• Robust to class imbalance
  – Hardison NE, Fanelli TJ, Dudek SM, Ritchie MD, Reif DM, Motsinger-Reif AA. Balanced accuracy as a fitness function in Grammatical Evolution Neural Networks is robust to imbalanced data. Genetic and Evolutionary Algorithm Conference. In Press.

• Scales linearly in regards to computation with the number of variables

• Robust to genotyping error, missing data, and phenocopies
Successes of GENN

• Has higher power in the presence of heterogeneity than MDR

• The presence of LD increases the power of GENN

• Has been favorably compared to other methods in the field in a range of genetic models
Real Data Application: HIV Immunogenetics

- Applied GENN to the AIDS Clinical Trials Group #384 dataset to identify potential gene-gene interactions that predict EFV pharmacokinetics and long-term responses.
Real Data Application: HIV Immunogenetics

- Participants from ACTG 384, a multicenter trial that enrolled from 1998-99.
- Participants were randomized to 3- or 4-drug therapy with EFV, nelfinavir (NFV), or both EFV plus NFV, given with ddI+d4T or ZDV+3TC.
- 340 were randomized to receive EFV (± NFV) had genetic data available.
- 3 years follow up
- Baseline characteristics:
  - 83% male
  - 50% white, 32% black, 17% Hispanic, 1% other race/ethnicity
  - CD4 count 270 ± 220 cells/mm3
  - baseline HIV-1 RNA 5.0 ± 0.9 log10 copies/ml
Real Data Application: HIV Immunogenetics

- Polymorphisms identified in the immune system and drug metabolism gene

- Outcome of interest:
  - CD4 increases in HIV patients undergoing potent antiretroviral therapy
  - <200 CD4 cells/mm³ increase from baseline with 48 weeks of virologic control
Real Data Application: HIV Immunogenetics

<table>
<thead>
<tr>
<th>CV</th>
<th>Factors in GENN Model</th>
<th>CE</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CD132_9823 IL2RB_6844</td>
<td>0.4153</td>
<td>0.4000</td>
</tr>
<tr>
<td>2</td>
<td>CD132_9823 IL2RB_6844 IL2RB_6844 IL15RA_19029 IL15RA_19411</td>
<td>0.4268</td>
<td>0.4091</td>
</tr>
<tr>
<td>3</td>
<td>CD132_9823 IL2RB_6844 IL2RB_6844 IL2RB_6844 CD132_9823</td>
<td>0.4140</td>
<td>0.4227</td>
</tr>
<tr>
<td>4</td>
<td>IL2_9352 CD132_9823 IL2RB_6844 IL15RA_19371 IL15_87710</td>
<td>0.4173</td>
<td>0.4368</td>
</tr>
<tr>
<td>5</td>
<td>CD132_9823 IL2RB_6395 IL2RB_6844 IL15RA_18856 IL15RA_198866</td>
<td>0.4186</td>
<td>0.4253</td>
</tr>
<tr>
<td>6</td>
<td>CD132_9823 IL2RB_6844 IL15RA_18856 IL15_4526 IL2RB_6844</td>
<td>0.4122</td>
<td>0.5862</td>
</tr>
<tr>
<td>7</td>
<td>IL2_9511 CD132_9276 CD132_9823 IL2RB_6844 IL2RB_6844 IL2RB_29015 IL2RB_29015</td>
<td>0.4160</td>
<td>0.4483</td>
</tr>
<tr>
<td>8</td>
<td>CD132_9823 IL2RB_6844 IL2RB_6844 IL2RB_28628 IL15RA_19029 IL15RA_19009</td>
<td>0.4109</td>
<td>0.4828</td>
</tr>
<tr>
<td>9</td>
<td>CD132_9276 CD132_9823 IL2RB_6844 IL15_4526 IL15_87191</td>
<td>0.4262</td>
<td>0.4828</td>
</tr>
<tr>
<td>10</td>
<td>CD132_9823 IL2RB_6844 IL15RA_18856 IL15_87435</td>
<td>0.4198</td>
<td>0.5402</td>
</tr>
</tbody>
</table>

Avg PE = 32.3%
P<0.02
Real Data Application: HIV Immunogenetics

IL2 Receptor beta chain (IL2RB:16491)

- **CC**
  - **CC**
    - CD4 change >200 cells
  - **CG**
    - IL2 Receptor common gamma chain (CD132: 9823)
      - **CT**
        - CD4 change >200 cells
      - **CC/TT**
        - CD4 change <200 cells
- **GG**
  - IL2 Receptor common gamma chain (CD132: 9823)
    - **CC**
      - CD4 change >200 cells
    - **CT/TT**
      - CD4 change <200 cells
Future Directions

• Family data
• Both continuous and discrete input and output variables
  – Combine data types
• Empirical studies to aid in NN interpretation
• Improve computation time and evolutionary optimization
Acknowledgments

• Vanderbilt University
  – Center for Human Genetics Research
    • Scott Dudek
    • Lance Hahn, PhD
    • Marylyn Ritchie, PhD
  – CFAR
    • David Haas, MD
    • Todd Hulgan, MD MPH
    • Jeff Canter, MD MPH
    • Asha Kallianpur, MD
    • Tim Sterling, MD

• NCSU
  – Nicholas Hardison
  – Sandeep Oberoi

• EPA
  – David Reif, PhD

• Penn State
  – Theresa Fanelli
Questions?