Grammatical Evolution Neural Networks for Genetic Epidemiology

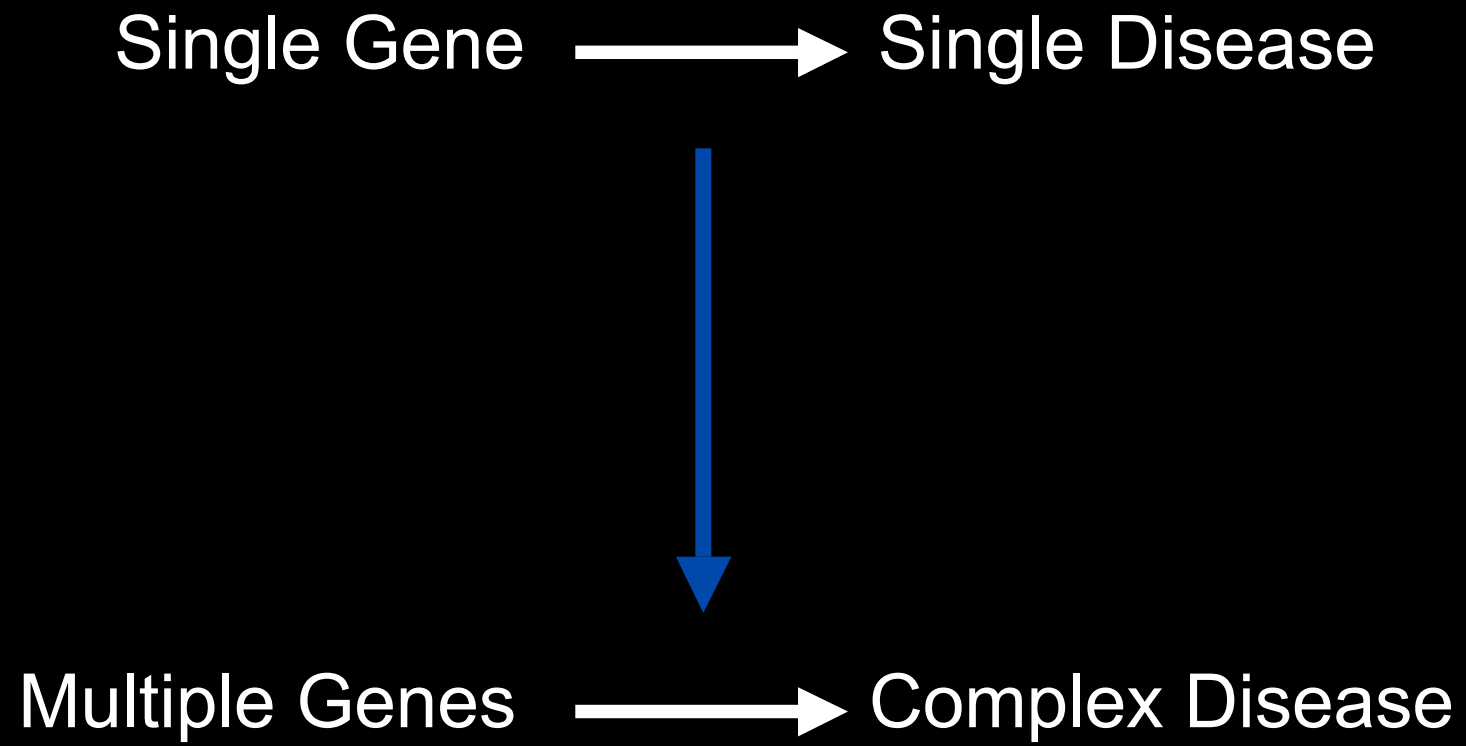
Alison Motsinger-Reif, PhD

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

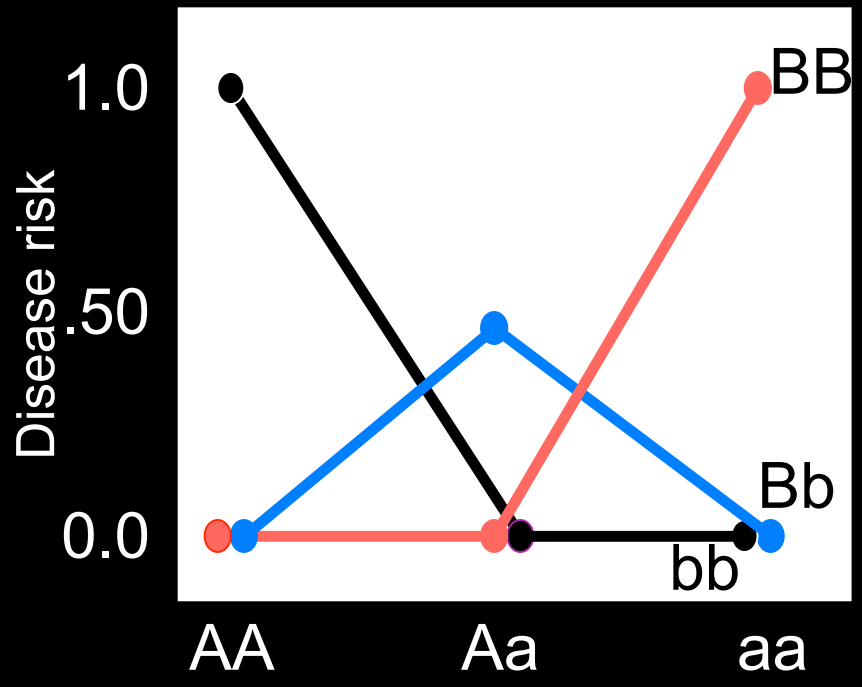


Epistasis

gene-gene or gene-environment interactions;

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

p(D) | Genotype			
	BB	Bb	bb
AA	0.0	0.0	1.0
Aa	0.0	0.5	0.0
aa	1.0	0.0	0.0

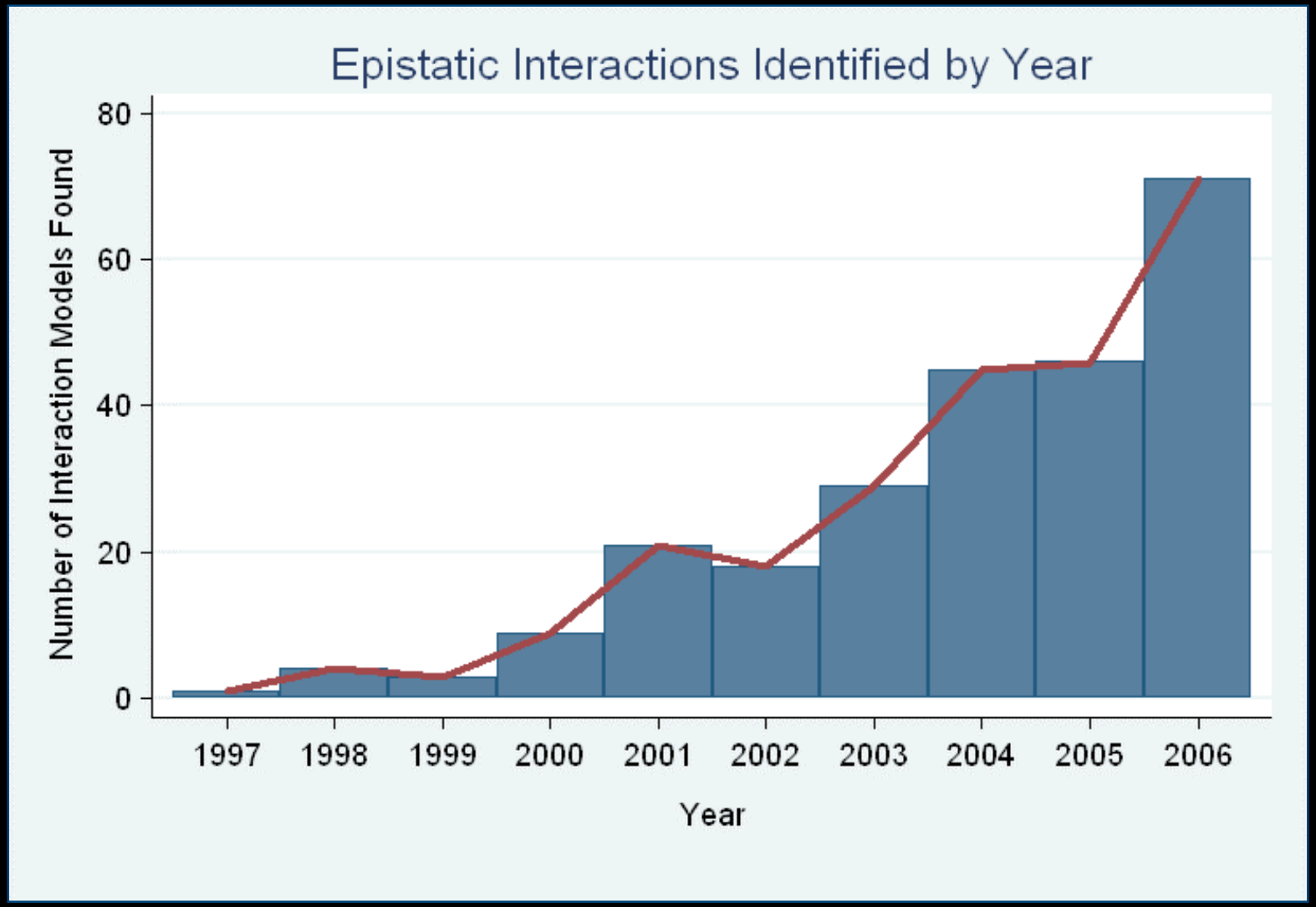


Epistasis

- 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)

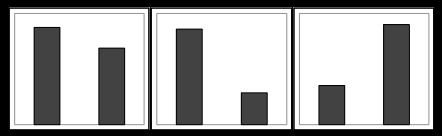
Curse of Dimensionality

N = 100

50 Cases,
50 Controls

SNP 1

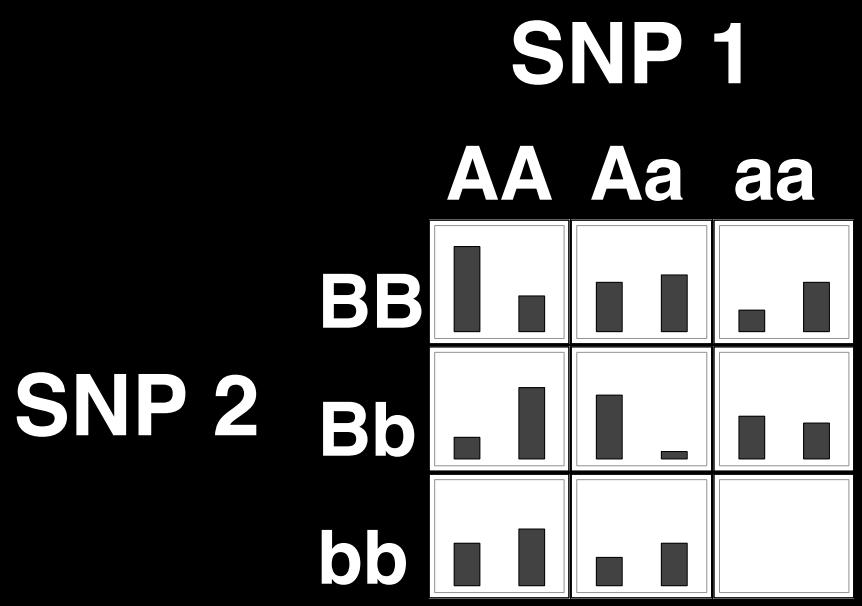
AA Aa aa



Curse of Dimensionality

N = 100

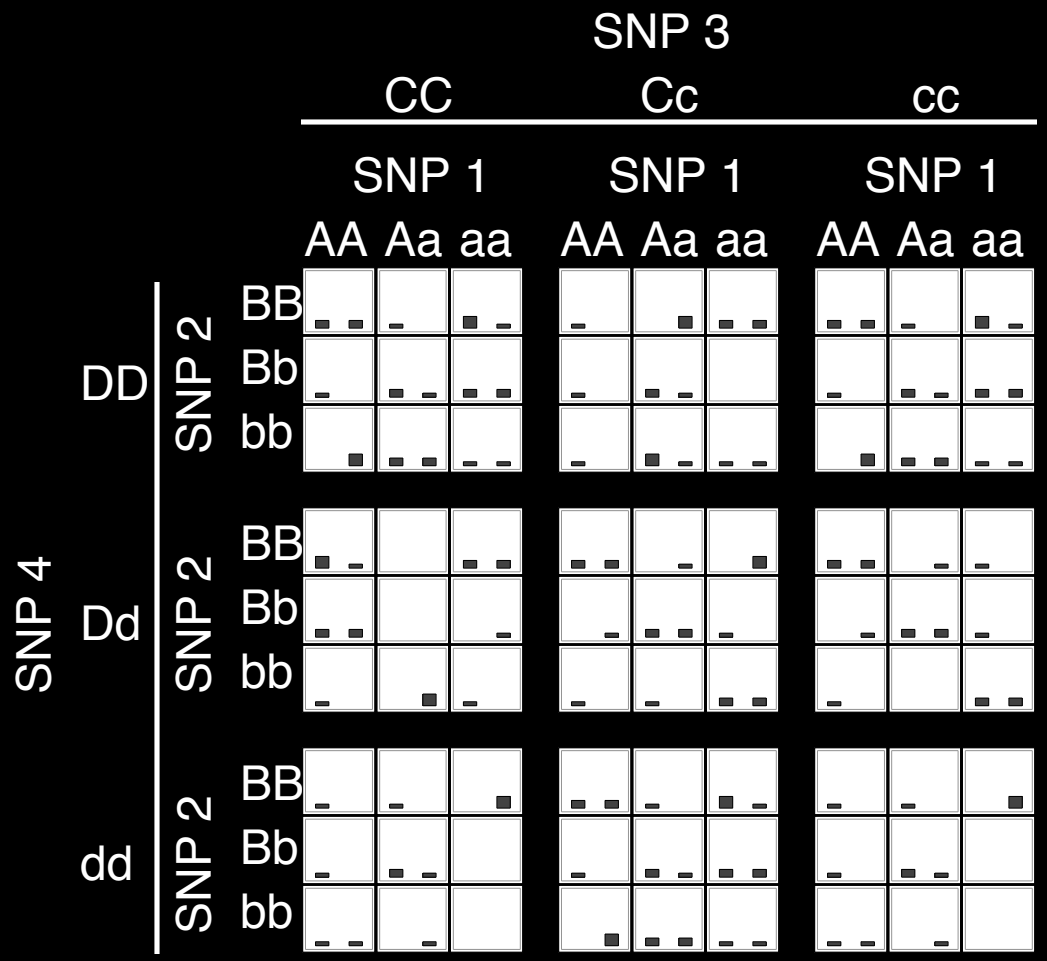
50 Cases,
50 Controls



Curse of Dimensionality

N = 100

50 Cases,
50 Controls



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)
- ……

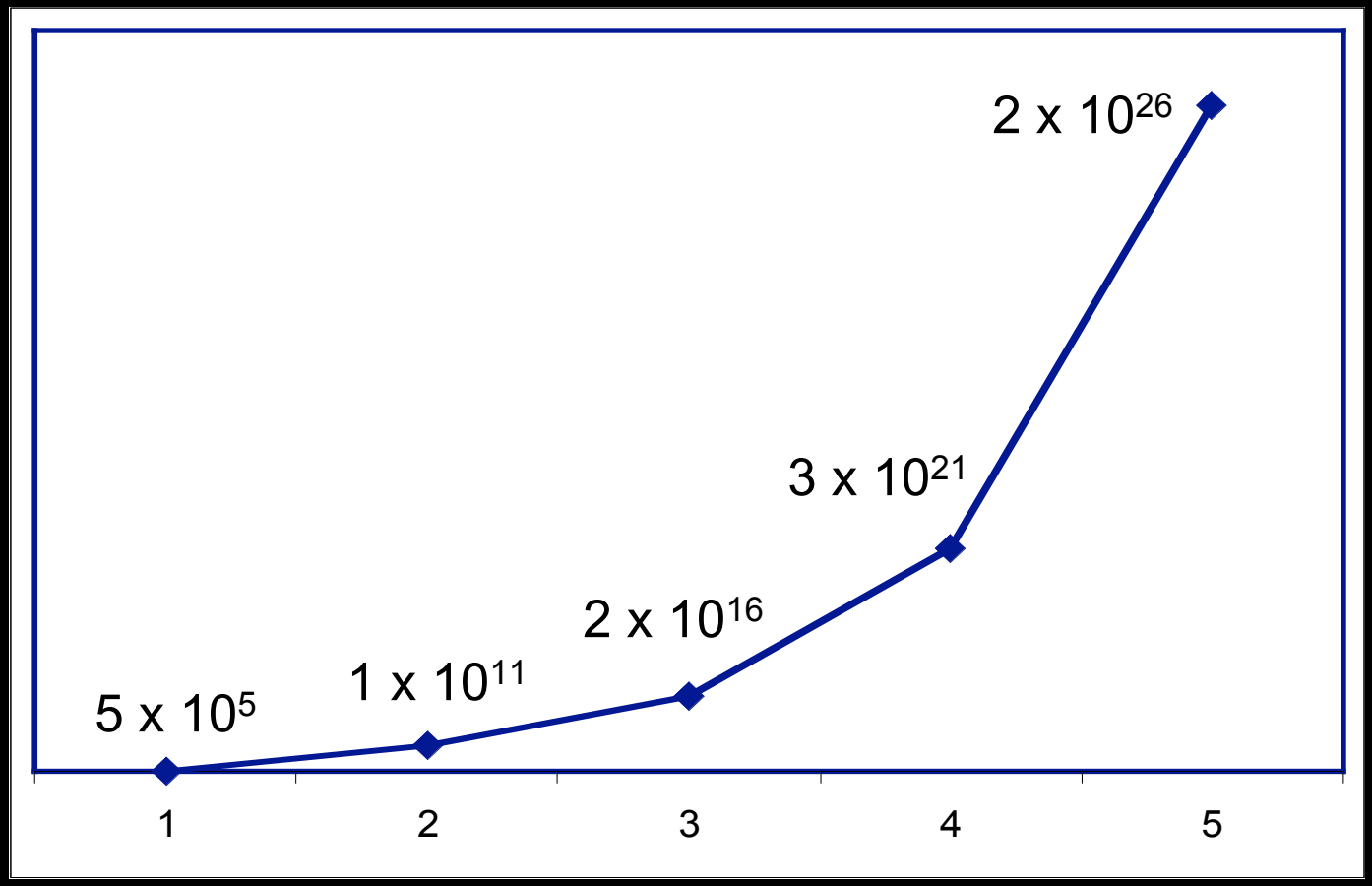
Methods to Detect Epistasis

- 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!

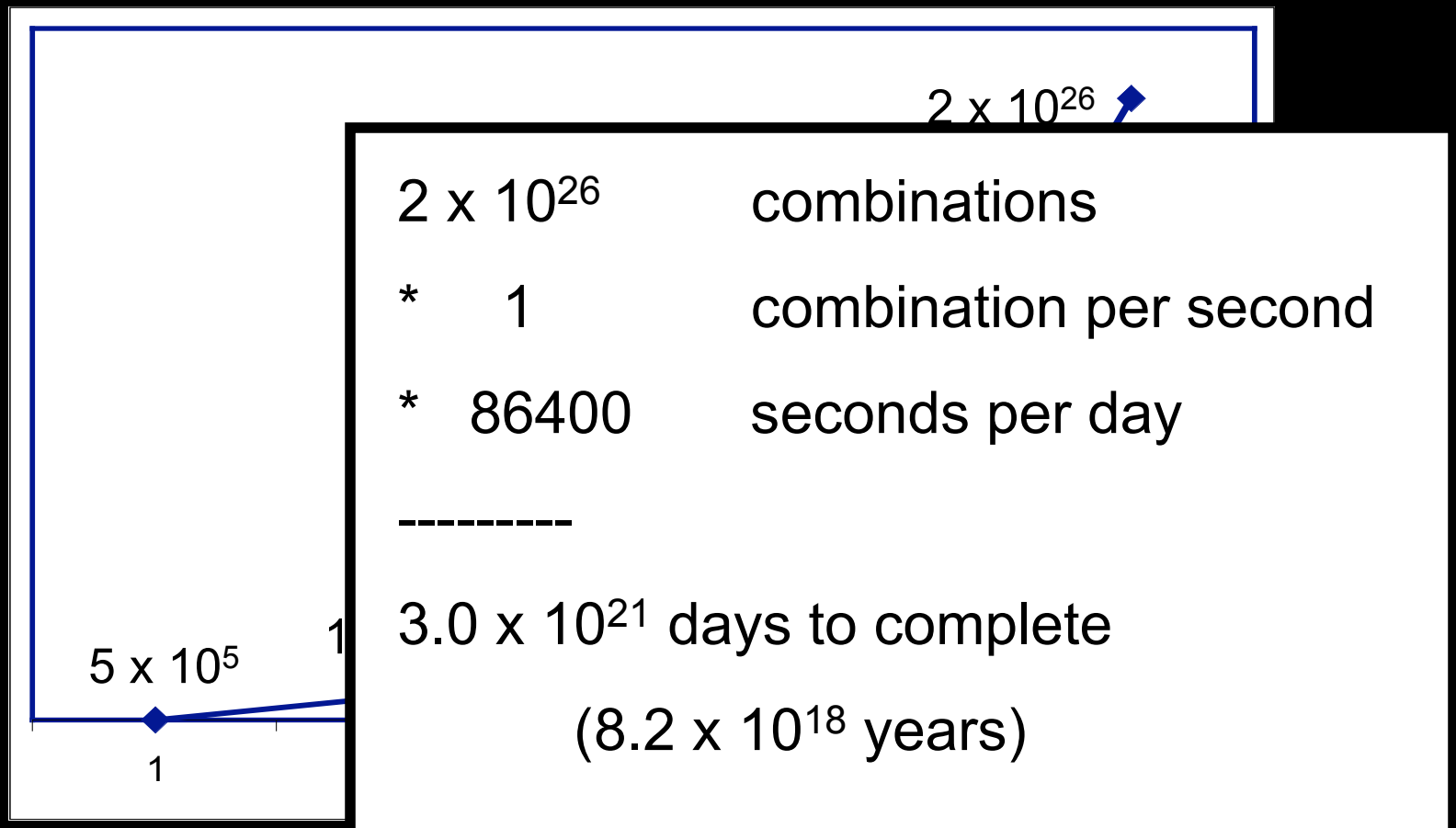
How Many Combinations are There?

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



How Many Combinations are There?

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



2 x 10^{26} combinations

* 1 combination per second

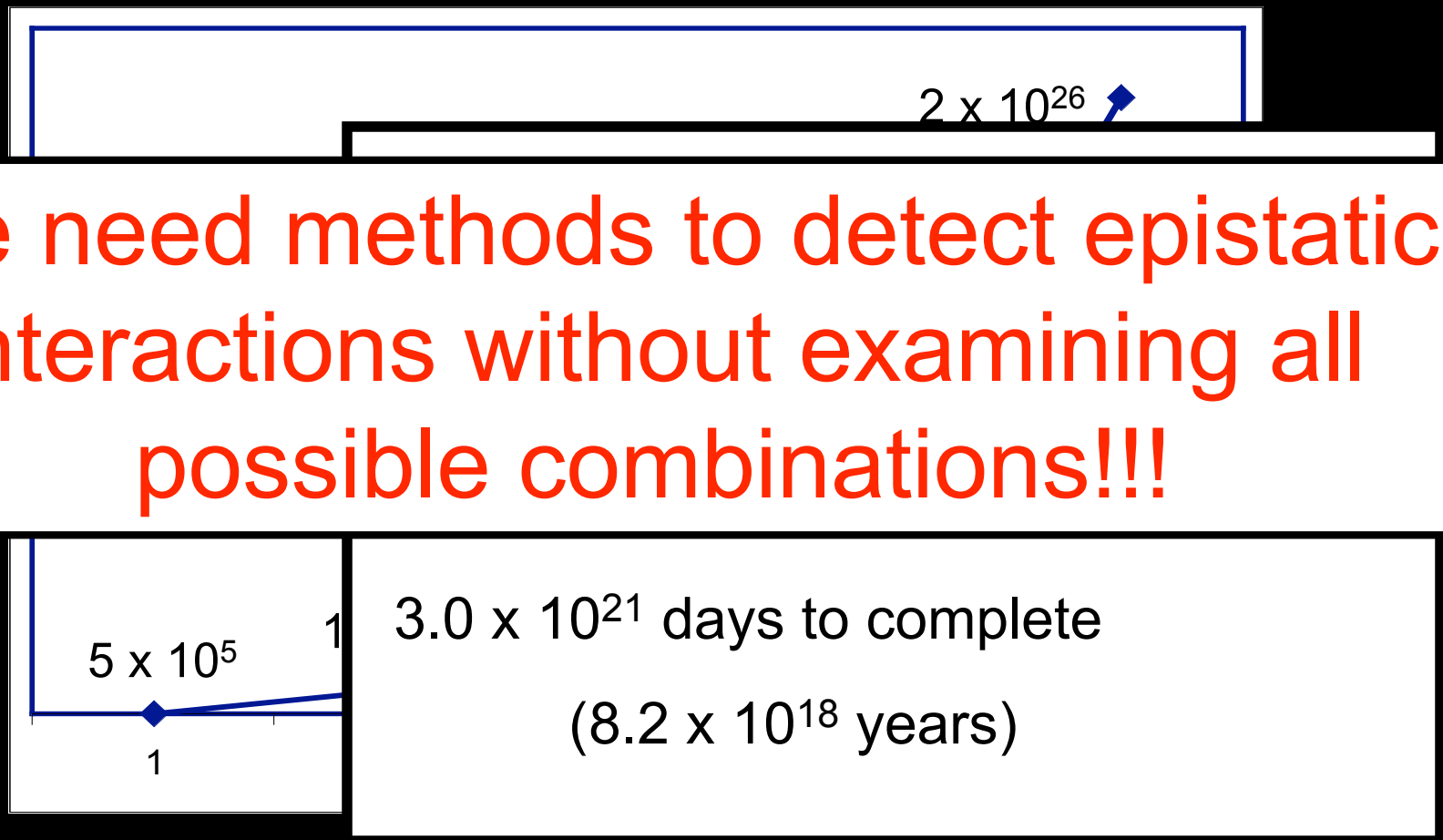
* 86400 seconds per day

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



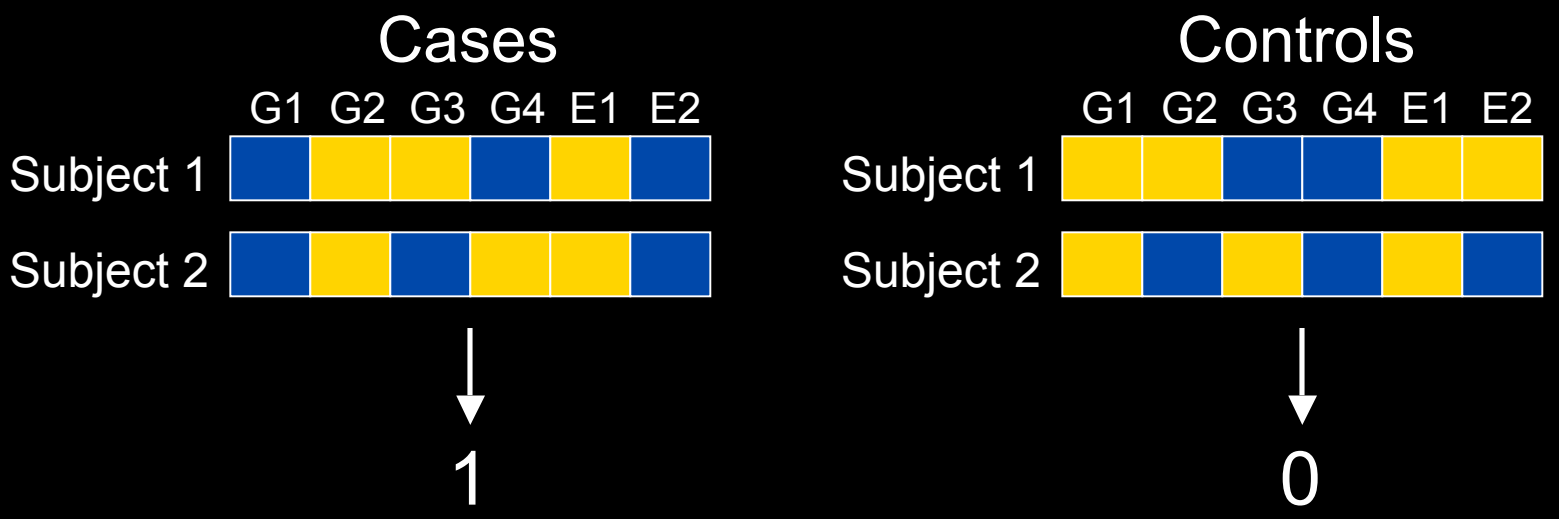
We need methods to detect epistatic interactions without examining all possible combinations!!!

3.0×10^{21} days to complete

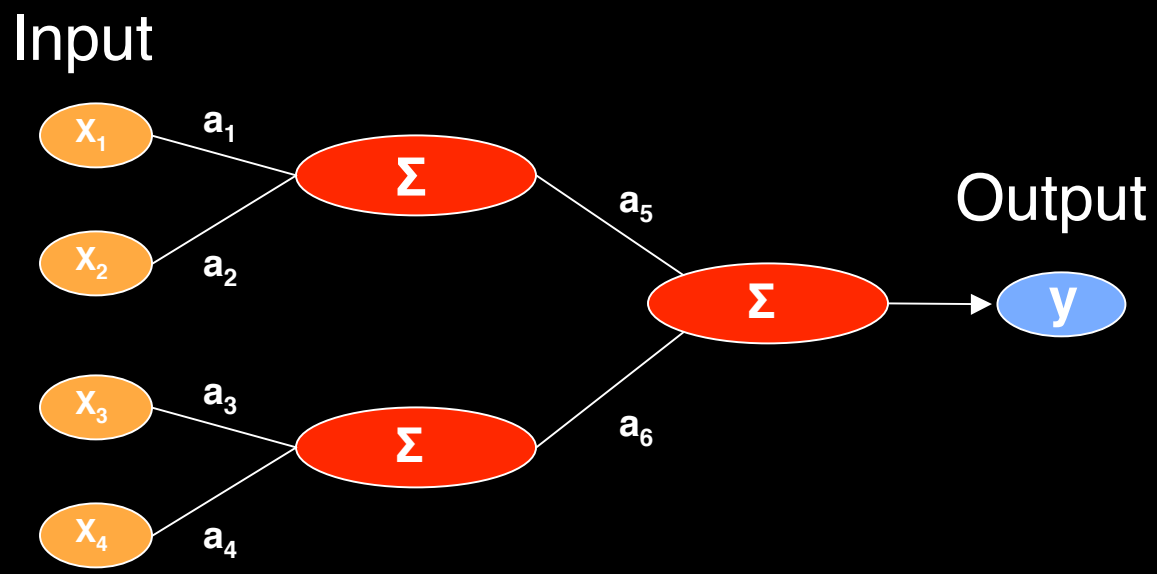
(8.2×10^{18} years)

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)

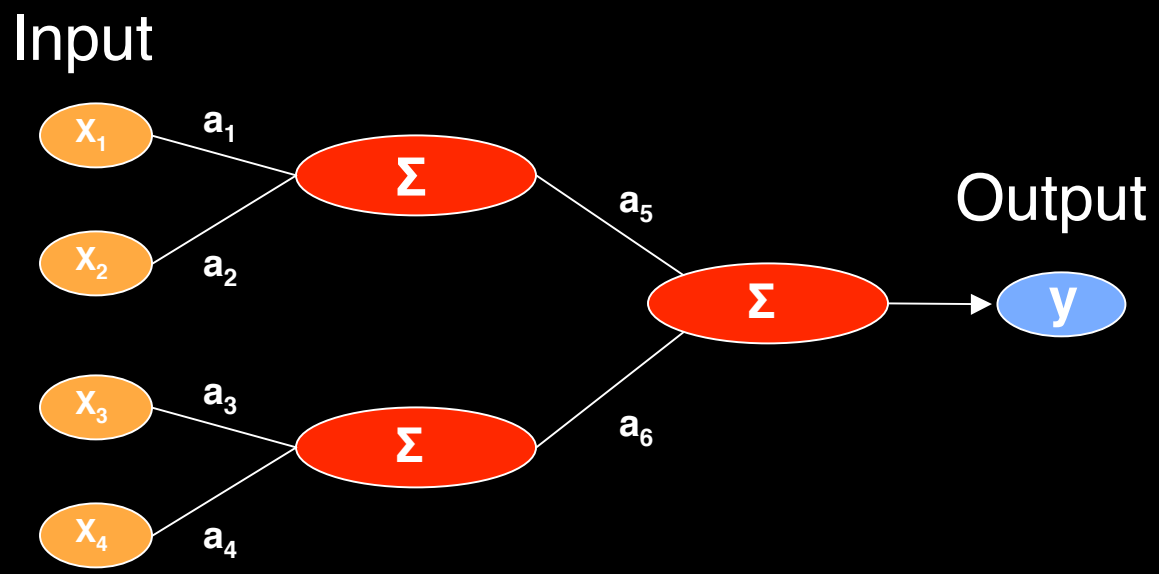


Neural Networks



- 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 (Σ), 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

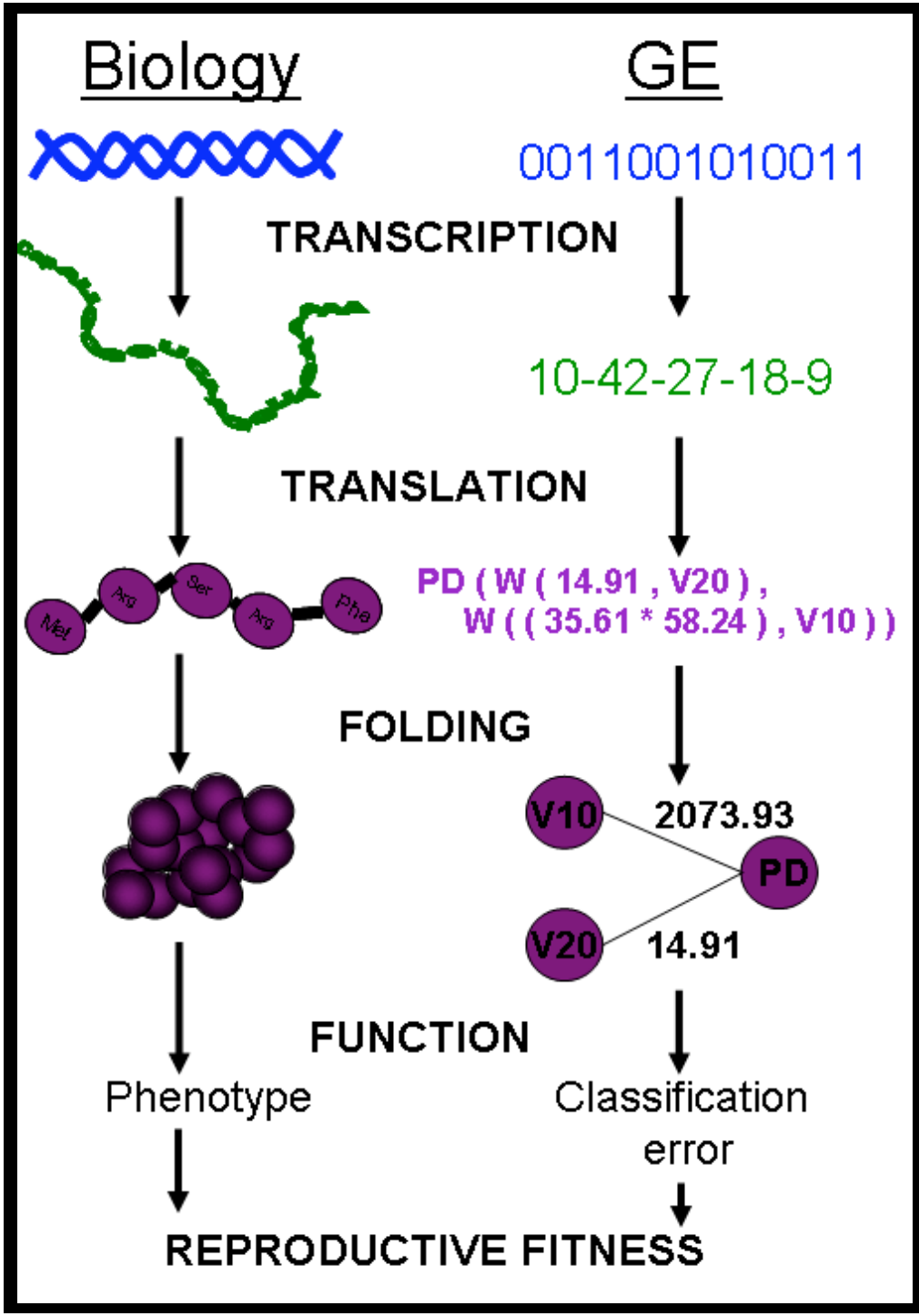
Neural Networks

- 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.



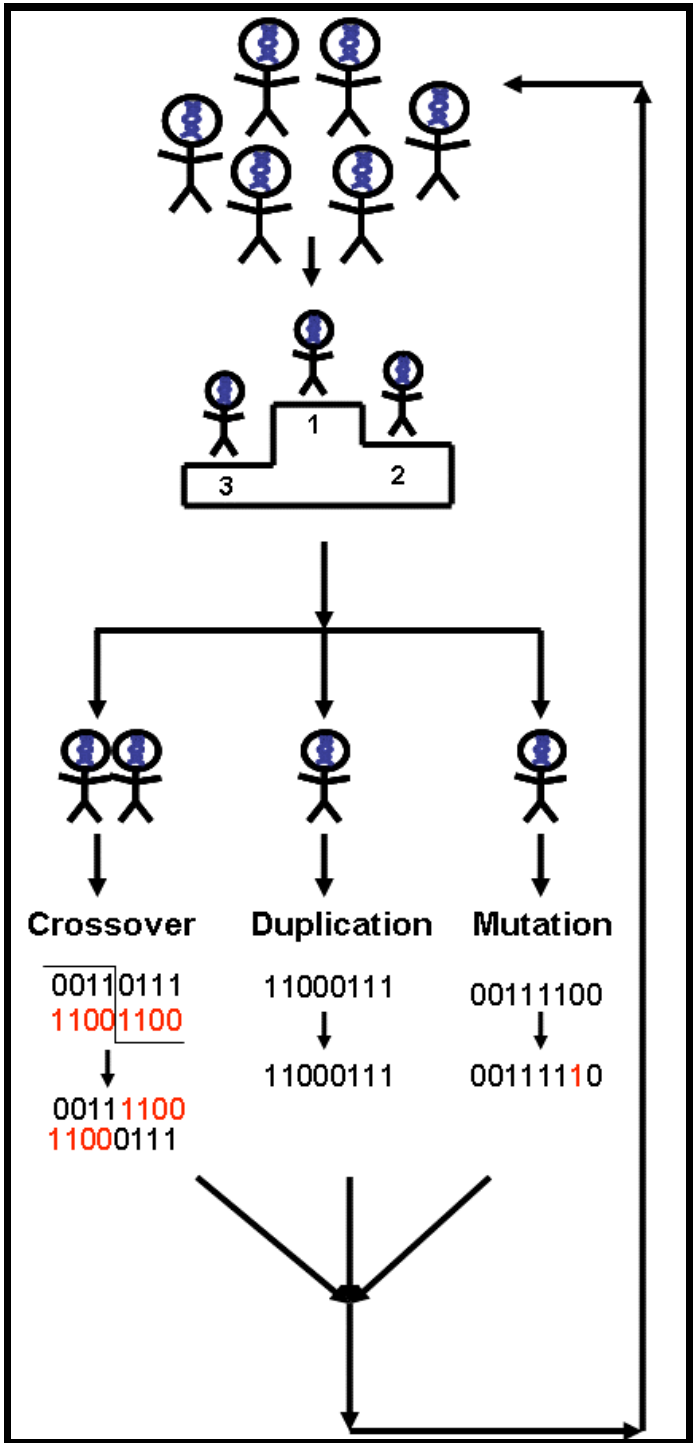
DNA: The heritable material in GE is the binary string chromosome. The GE chromosome is divided into codons, undergoes crossover and mutation, and can contain non-coding sequence just as biological DNA.

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.

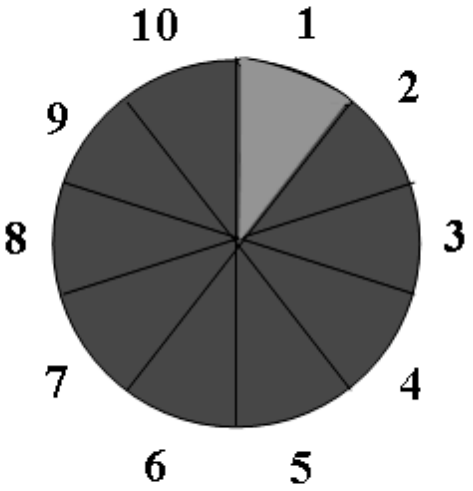
Steps 1-4 are repeated for a user-specified number of generations, to produce offspring with the highest possible fitness.

GE Neural Networks

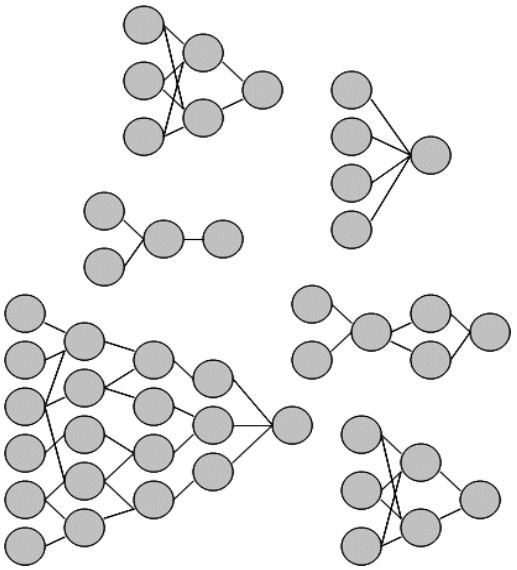
STEP 1

population_size	200
max_generations	50
pvm_exchange_generations	25
random_seed	7
crossover_rate	0.9
mutation_rate	0.01
codon_size	8
wrapper_count	2
min_chrom_size	50
max_chrom_size	1000

STEP 2



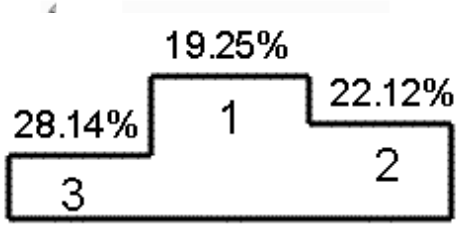
STEP 3



STEP 4

GENN Models
Classification Error
19.25
22.12
24.33
28.14
⋮

STEP 5



Tournament

STEP 6

GENN Model

Classification Error	Prediction Error
19.25	21.55

Example Results

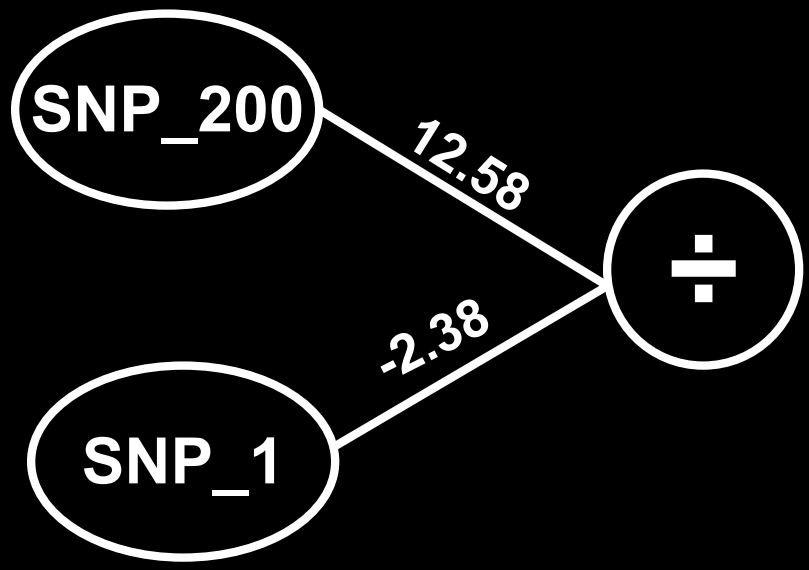
CV	Factors in Model					CE	PE
1	SNP_1	SNP_200				0.26	0.15
2	SNP_1	SNP_200	SNP_630	SNP_755		0.38	0.22
3	SNP_1	SNP_200	SNP_512			0.32	0.29
4	SNP_1	SNP_200	SNP_333	SNP_467	SNP_987	0.19	0.35
5	SNP_1	SNP_200	SNP_814	SNP_900		0.12	0.32
6	SNP_1	SNP_200	SNP_665			0.20	0.19
7	SNP_1	SNP_200	SNP_742	SNP_801		0.21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

Example Results

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

Significance Testing

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



Prediction Error: 15.4%
p<0.01

Successes of GENN

- High power to detect a wide range of main effect and interactive models
  - Motsinger-Reif AA, Dudek SM, Hahn LW, and Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genetic Epidemiology 2008 Feb 8 [Epub ahead of print]
- Robust to changes in the evolutionary process
  - Motsinger AA, Hahn LW, Dudek SM, Ryckman KK, Ritchie MD. Alternative cross-over strategies and selection techniques for Grammatical Evolution Optimized Neural Networks. In: Maarten Keijzer et al, eds. Proceeding of Genetic and Evolutionary Computation Conference 2006 Association for Computing Machinery Press, New York, pp. 947-949.
- Higher power than traditional BPNN, GPNN, or random search NN
  - Motsinger AA, Dudek SM, Hahn LW, and Ritchie MD. Comparison of neural network optimization approaches for studies of human genetics. Lecture Notes in Computer Science, 3907: 103-114 (2006).
  - Motsinger-Reif AA, Dudek SM, Hahn LW, and Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genetic Epidemiology 2008 Feb 8 [Epub ahead of print]

Successes of GENN

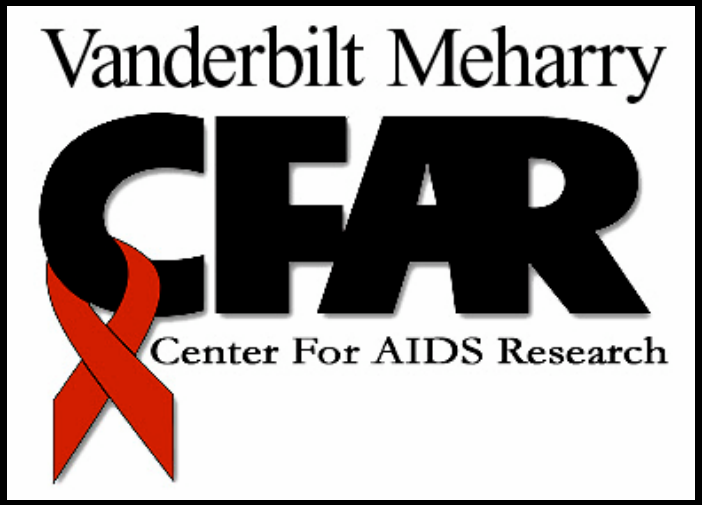
- Robust to class imbalance
  - Hardison NE, Fanelli TJ, Dudek SM, Ritchie MD, Reif DM, Motsinger-Reif AA. Balanced accuracy as a fitness function in Grammatical Evolution Neural Networks is robust to imbalanced data. Genetic and Evolutionary Algorithm Conference. *In Press.*
- Scales linearly in regards to computation with the number of variables
  - Motsinger AA, Reif DM, Dudek SM, and Ritchie MD. Dissecting the evolutionary process of Grammatical Evolution Optimized Neural Networks. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2006 pp. 1-8.
- Robust to genotyping error, missing data, and phenocopies
  - Motsinger AA, Fanelli TJ, Ritchie MD. Power of Grammatical Evolution Neural Networks to detect gene-gene interactions in the presence of error common to genetic epidemiological studies. BMC Research Notes *In Press.*

Successes of GENN

- Has higher power in the presence of heterogeneity than MDR
  - Motsinger AA, Fanelli TJ, Ritchie MD. Power of Grammatical Evolution Neural Networks to detect gene-gene interactions in the presence of error common to genetic epidemiological studies. BMC Research Notes *In Press.*
- The presence of LD increases the power of GENN
  - Motsinger AA, Reif DM, Fanelli TJ, Davis AC, Ritchie MD. Linkage disequilibrium in genetic association studies improves the power of Grammatical Evolution Neural Networks. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2007 pp. 1-8.
- Has been favorably compared to other methods in the field in a range of genetic models
  - Random Forests, Focused Interaction Testing Framework, Multifactor Dimensionality Reduction, Logistic Regression
  - Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD. Comparison of computational approaches for genetic association studies. Genetic Epidemiology *In Press.*

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

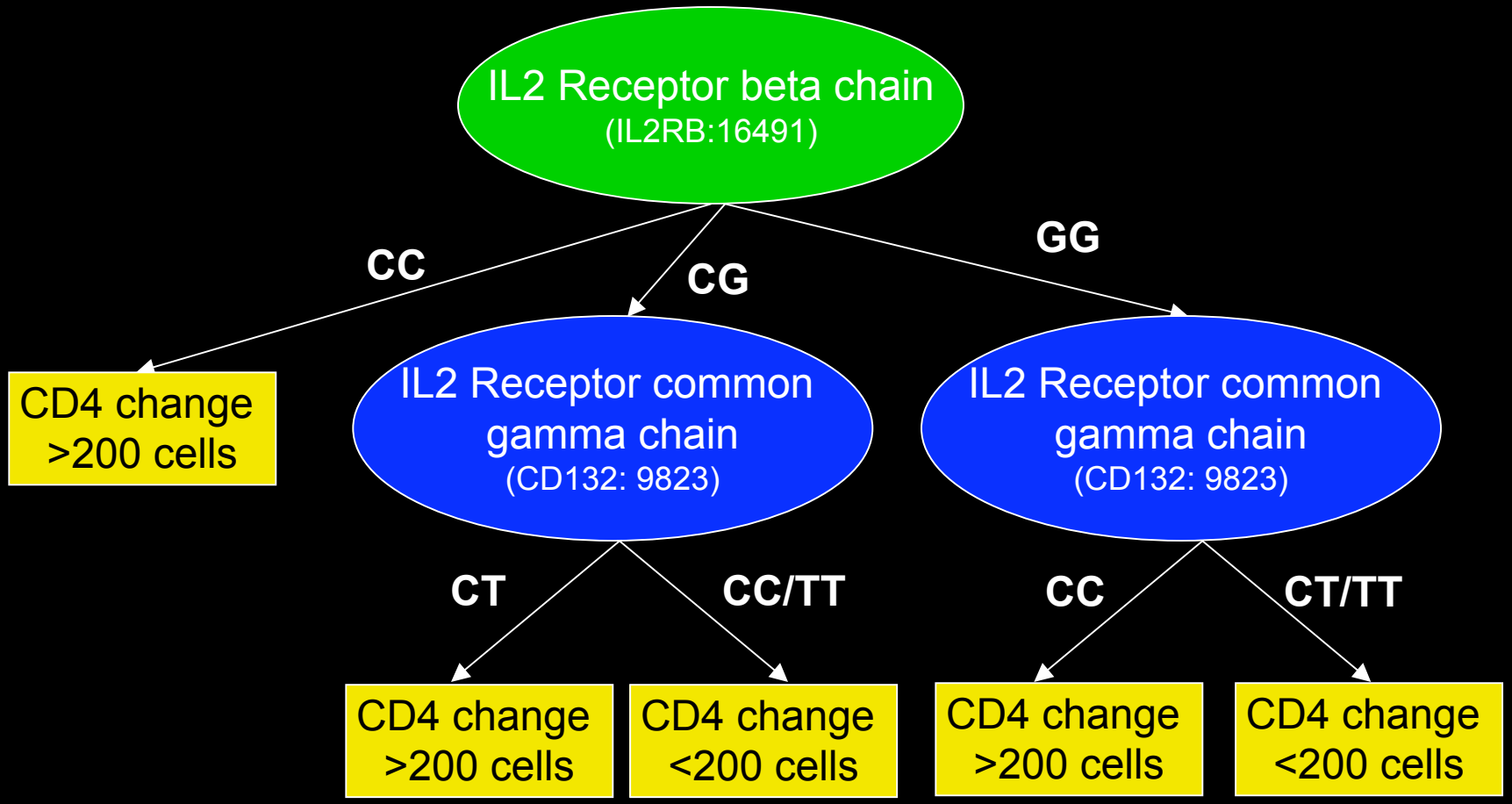
- Polymorphisms identified in the immune system and drug metabolism gene
- Outcome of interest:
  - CD4 increases in HIV patients undergoing potent antiretroviral therapy
  - <200 CD4 cells/mm3 increase from baseline with 48 weeks of virologic control

Real Data Application: HIV Immunogenetics

CV	Factors in GENN Model							CE	PE
1	***CD132_9823***	***IL2RB_6844***						0.4153	0.4000
2	***CD132_9823***	***IL2RB_6844***	***IL2RB_6844***	*IL15RA_19029*	*IL15RA_19411*			0.4268	0.4091
3	***CD132_9823***	***IL2RB_6844***						0.4140	0.4227
4	*IL2_9352*	***CD132_9823***	***IL2RB_6844***	*IL15RA_19371*	*IL15_87710*			0.4173	0.4368
5	***CD132_9823***	*IL2RB_6395*	***IL2RB_6844***	*IL15RA_18856*				0.4186	0.4253
6	***CD132_9823***	***IL2RB_6844***	*IL15RA_18856*	*IL15_4526*				0.4122	0.5862
7	*IL2_9511*	*CD132_9276*	***CD132_9823***	*IL2RB_6443*	***IL2RB_6844***	*IL2RB_29015*	*IL2RB_29015*	0.4160	0.4483
8	***CD132_9823***	***IL2RB_6844***	***IL2RB_6844***	*IL2RB_28628*	*IL15RA_19029*			0.4109	0.4828
9	*CD132_9276*	***CD132_9823***	***IL2RB_6844***	*IL15_4526*	*IL15_87191*			0.4262	0.4828
10	***CD132_9823***	***IL2RB_6844***	*IL15RA_18856*	*IL15_87435*				0.4198	0.5402

Avg PE = 32.3%
P<0.02

Real Data Application: HIV Immunogenetics



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?