

Artificial Intelligence based Analysis of Postprandial Triglyceride Response using Genetic and Clinical Data

Ioannis Valavanis¹, Stavroula Mouggiakakou¹,
Keith Grimaldi², and Konstantina Nikita¹

¹ Biomedical Simulations and Imaging Laboratory,
School of Electrical and Computer Engineering,
National Technical University of Athens, Greece

² SCIONA USA, R&D Department

Outline

- Introduction
- Current Study
- Novel Algorithm
- Results
- Conclusions

Multi-Factorial Diseases

- Many different influences acting together to cause the appearance of the disease in a person:
 - Combination of genetic factors
 - Environmental factors
 - Weight
 - Diet / Drinking Habits
 - Exercise
 - Smoking habits
 - Industrial pollution, toxic substances, excessive exposure to sunlight
- **Disease examples:** Cardiovascular Disease (CVD), diabetes, asthma

Target: Select a number of factors that contribute to a disease phenotype and find a corresponding multi-factorial complex pattern

Introduction ◀

Current Study

Novel Algorithm

Results

Conclusions

Literature Review

- Soft computing methods
 - Artificial Neural Networks (NNs) and Hybrid Methods
 - Use of Genetic Algorithm (GA) and Genetic Programming as module in NN-based methods
 - GA as variable selection method
- Combinatorial methods
 - Combinatorial partition method
 - Restricted partition method
 - Multidimensional reduction method
- Set association method
- Tree-based methods (e.g. random forest)
- Logistic regression methods and hybrids

Introduction ◀

Current Study

Novel Algorithm

Results

Conclusions

Cardiovascular Diseases

Introduction ◀

Current Study

Novel Algorithm

Results

Conclusions

- CVD refers to the class of diseases that involve the heart and/or blood vessels (arteries and veins) and can affect the cardiovascular system
- CVD is the main cause of early death in America and Europe
- At least 20 million people survive heart attacks and strokes every year, many require continuing costly clinical care
- Many different influences acting together cause the appearance of CVD:
 - Genetic factors
 - Environmental factors
 - Weight, diet / drinking habits, exercise, smoking habits etc
 - Many environmental factors are indicated by clinical measurements (e.g. diet: body mass index, cholesterol, non-esterified folic acids etc)

Current Study

Introduction

Current Study

Novel Algorithm

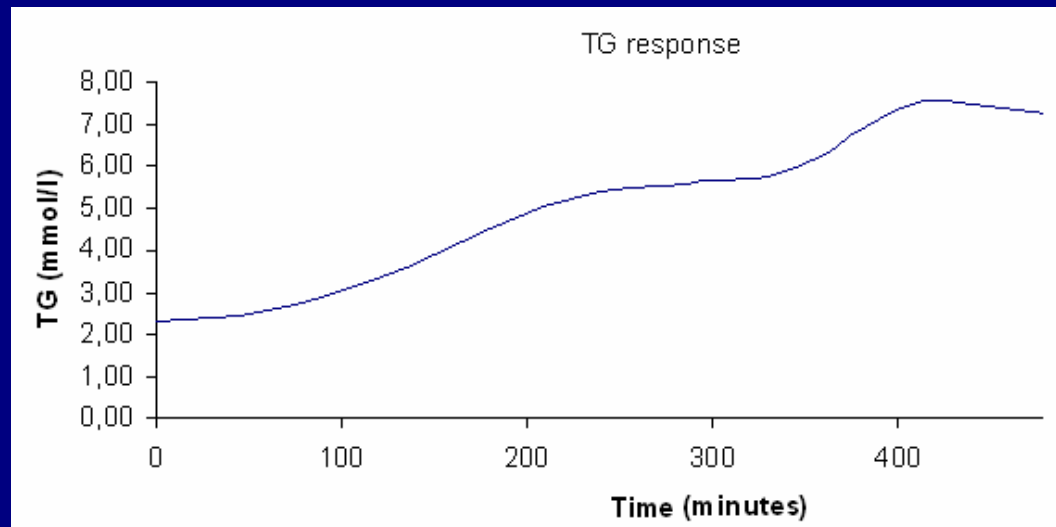
Results

Conclusions

- **Scope:** Associate post-prandial metabolism of triglyceride (TG) with genetic profile and other measurements (sex, age, clinical measurements)
- Elevated post-prandial TG response is recognized as a predictor of CVD
- Dataset
 - 213 subjects who had fat breakfast and meal on time 0 and 330 min
- Methods used: Hybrid GA-NN algorithm

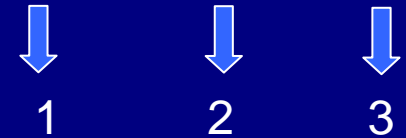
Data Sets

- 21 genes associated with CVD (ApoE, ApoA5, LPL, FABP2, etc)
- Sex, age
- Seven (7) clinical measurements: BMI, fasting levels for TG, Glucose, total cholesterol (TC), HDL-C, LDL-C, and Non-esterified fatty acids (NEFA)
- Response of triglyceride through time 0 – 480 min



Input Data

- **21 genes:** 21 categorical variables transformed through proper numbering
 - e.g. ApoE-219 three possible variants 219GG, 219GT, 219TT



A maximum of six (6) variants were found within all genes (case of haplotypes for ApoE gene)

- **Sex:** Male \Rightarrow 1 Female \Rightarrow 2
- **Age:** continuous variable
- **Clinical measurements:** continuous variable

A total of 30 input variables is obtained

Introduction

Current Study

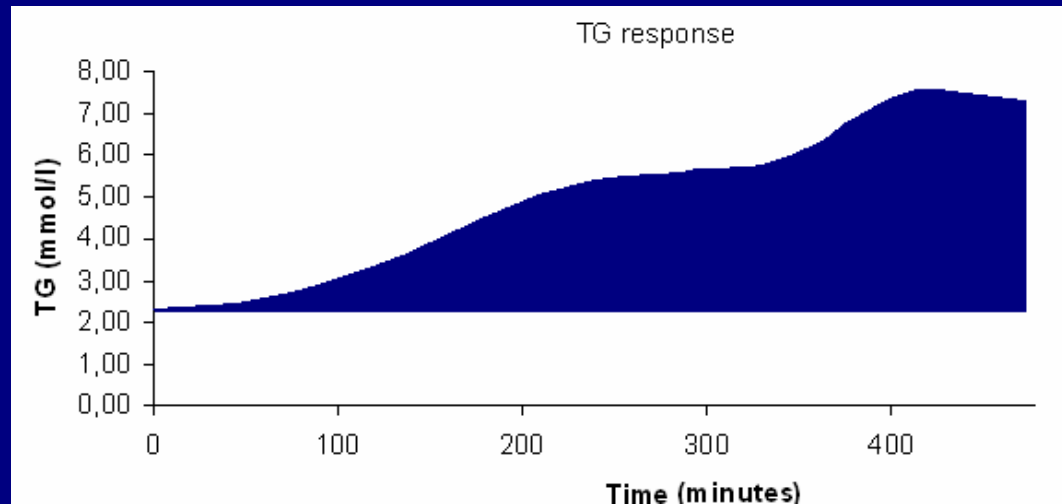
Novel Algorithm

Results

Conclusions

Output Data

- Incremental Area Under response Curve (IAUC) is calculated (baseline correction)



- All IAUC values are sorted and two (2) equal-sized output categories are obtained:
 - bottom 50% (0)
 - high 50% (1)

Hybrid GA-NN Algorithm

Introduction

Current Study

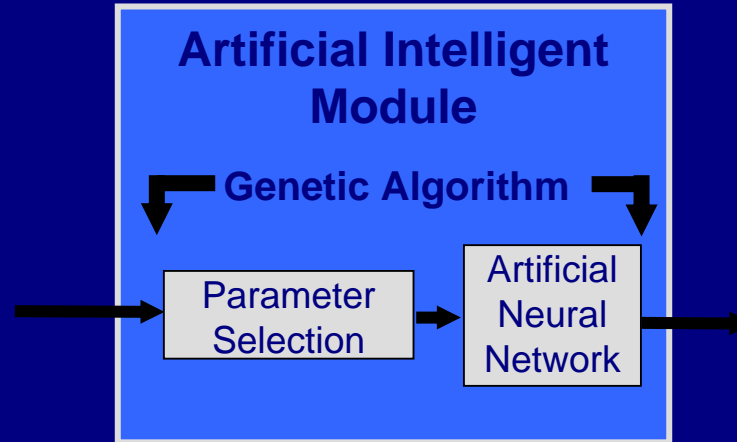
Novel Algorithm

Results

Conclusions

- NN features
 - Feed forward NN
 - One hidden layer
 - Back propagation training algorithm with variable learning rate and momentum
- Chromosome from binary digits represents a NN encoding:
 - Inputs to be considered
 - NN architecture parameters (inputs and hidden nodes)
 - NN training algorithm parameters (initial weight, momentum, learning rate)
- Data is split into training, testing and validation tests
- Resampling technique is used
- Fitness function: Classification accuracy in the validation test
- Use of GA operators to evolve the NN with maximum value in fitness function
- Evaluation of the generalization ability of the NN into the testing set for all resampling cases

Hybrid GA-NN Algorithm



In the randomly selected initial population of N chromosomes the following procedure was applied *while (termination condition is false)*

{
 refine the weight matrices with application of Back propagation with adaptive learning mate and momentum to each of N chromosomes using the training set
 calculate the fitness function of each chromosome in the validation set,
 select $N/2$ pairs from population using the elitist selection method
 mate selected chromosomes using two-point crossover
 switch value of chromosome bits
 update population, assign fitness values to new population and store best results }

The procedure is repeated for N_G generations

Hybrid GA-NN Algorithm

- All 30 initial inputs are normalized in the range [-1 1] → use of tansig function in input layer
- The 2-categories output is encoded in one output neuron: [0], [1] → use of logsig function in output layer
- Resampling yields to 10 random training, validation and testing sets (60%, 20%, 20% of dataset)
- Procedure followed for 10 times
- The hybrid algorithm will conclude to an optimal and parsimonious input feature set to be used in the discrimination of the two categories

Introduction

Current Study

Novel Algorithm

Results

Conclusions

Results

CV Consistency (genes)

Classification Accuracy

Training Accuracy	Validation Accuracy	Testing Accuracy
1,000	0,857	0,857
1,000	0,833	0,214
0,977	0,786	0,286
0,993	0,857	0,786
1,000	0,881	0,238
1,000	0,833	0,333
1,000	0,833	0,810
1,000	0,810	0,833
1,000	0,786	0,786
1,000	0,833	0,761

CV Consistency (clinical)

Feature	Consistency
Sex	7
Age	5
BMI	7
TC	8
TG	6
HDL-C	7
LDL-C	3
NEFA	4
Glucose	5

Feature	Consistency
ApoE haplotype	6
FABP2	7
ApoB	7
CETP	3
INS	3
LPL Hind3	3
MTP	7
LPL S447	7
TNF	4
ESR1 XXba1	4
ESR1 PPvu2	6
ApoC C3238G	3
LEPR Gln233Arg	4
ApoA4 T347S	3
ApoA5 1131	4
ApoA5 SGG	5
ApoA5 haplotype	5
PPARA	5
ApoA4 Q360H	5
ISR1	5
ApoE Promoter	5

Introduction

Current Study

Novel Algorithm

Results

Conclusions

Conclusions

- Features of high consistency are the most informative ones
- High classification accuracies were achieved
- Aggregation of optimal NNs can improve the results

Introduction

Current
Study

Novel
Algorithm

Results

Conclusions ◀

Future Work

- Comparative study with other methods (ANOVA, Multifactor Dimensionality Reduction - MDR, Parameter Decreasing Model Neural Network - PDM NN)
- Use the method to classify other multifactorial patterns
- Development of a system which will follow-up subjects and evaluate CVD risk

Introduction

Current Study

Novel Algorithm

Results

Conclusions

Acknowledgments

- This work was partially supported by the European Commission under the FP6-IST4-027333 project “Micro2DNA: Integrated polymer-based micro fluidic micro system for DNA extraction, amplification, and silicon-based detection”.

Introduction

Current
Study

Novel
Algorithm

Results

Conclusions