

Hybrid Artificial Neural Network Methods for the Analysis of Complex Disease Traits

Ioannis Valavanis

Electrical and Computer Engineer, MSc, PhD

Post Doctoral Research Associate (NHRF, Athens, Greece)

Visiting Lecturer (University of Peloponnese, Tripolis, Greece)

Athens Information Technology

March 17, 2011

Introduction

- Oxford English Dictionary

"bio – informatics: bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical chemistry) and applying "informatics techniques" (derived from disciplines such as applied maths, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications"

- Traditional Bioinformatics

- DNA, RNA, Proteins Sequences Alignment
- Genes Identification and Mapping to Chromosomes
- Proteins Structures Prediction
- Protein Function Classification and Prediction
- Protein-Protein Interaction and Binding

ds

s

ome Project (2003)

ing Biology and Medicine: Personalized Medicine

Bioinformatics

Multifactorial
Diseases

Multifactorial
Analysis Methods

Introduction

- Multifactorial patterns
 - Genes and/or environment
 - n -way interactions, $n > 1$
 - Main effects, 1-way interactionsGalton (1875) «*nature vs nurture*»
«*blending characters*»

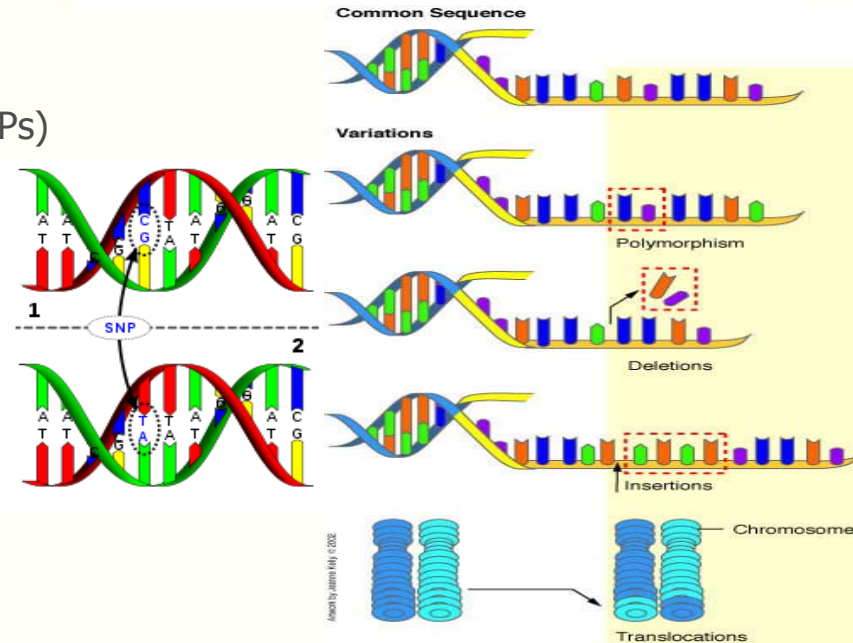


Bioinformatics

Multifactorial Diseases

Multifactorial Analysis Methods

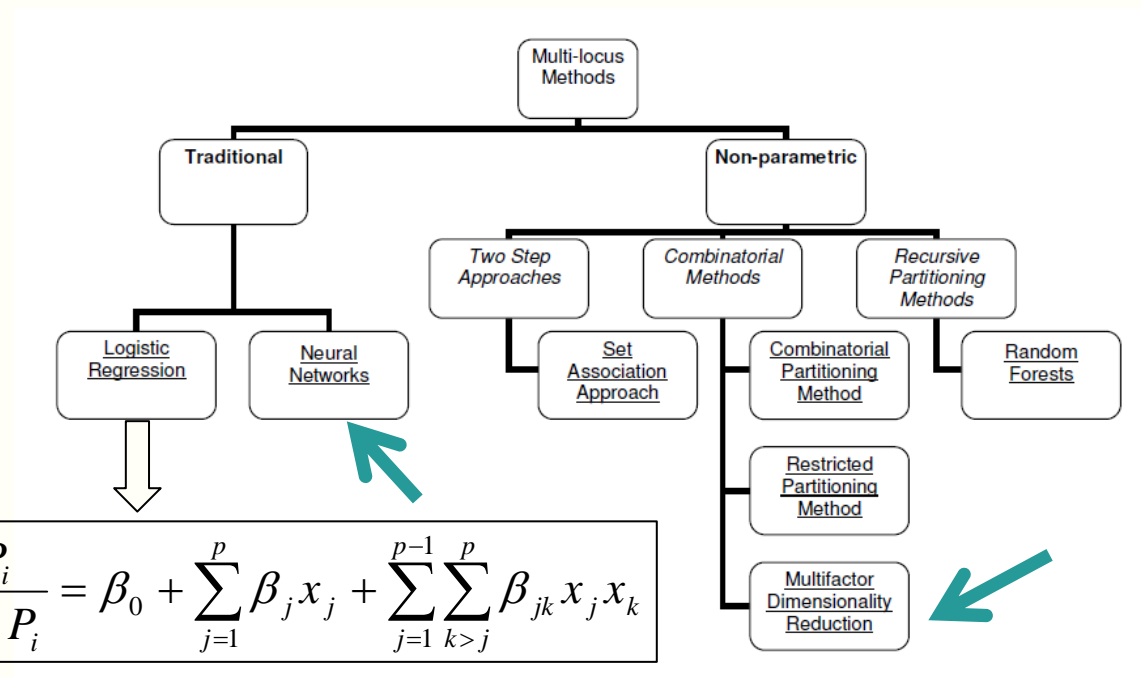
- Genes: Genetic Polymorphisms
 - Single Nucleotide Polymorphisms(SNPs)
 - Insertions/Deletions (indels)
 - Haplotypes



- Environment
 - Nutrition
 - Exercise
 - Smoking
 - Pollution ...
- Multifactorial Diseases
 - Cardiovascular Diseases (CVDs), some cancers, Type 2 diabetes, Alzheimer, Obesity,..

Introduction

- Scopes of Analysis
 - Find factors that affect the disease onset
 - Find patterns of interactions among the selected factors
- Methods (Heidema et al., 2006)




Multifactorial Analysis of Postprandial Lipemia

Postprandial Lipemia

- Postprandial increase of lipids in blood
- Atherogenesis: Happens postprandially (Zilverman, 1979)

Introduction

Data

- Exaggerated postprandial lipemia (Triglycerides, TG) 
Independent CVD risk factor

Methods

- Multifactorial trait

Results

- Genes
- Nutrition
- Sex, Age
- Weight
- Exercise

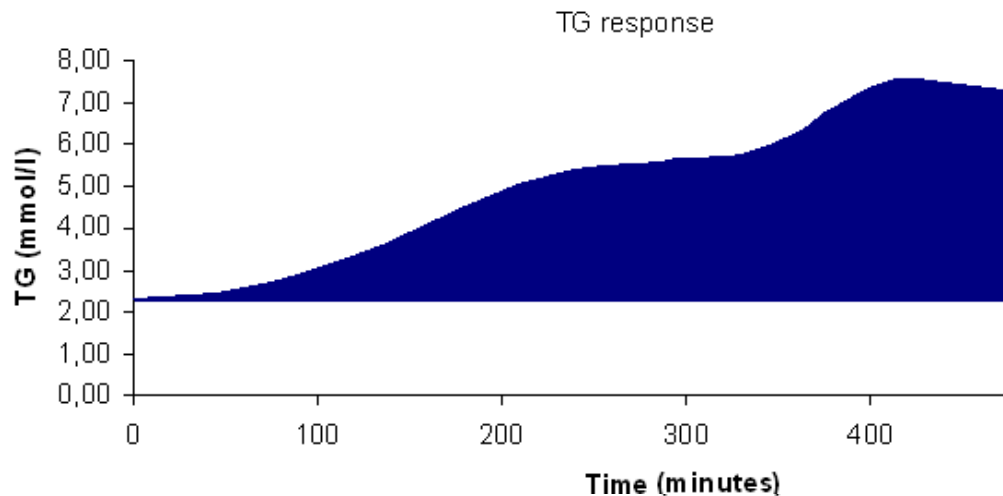
Conclusions

Multifactorial Analysis of Postprandial Lipemia

- 213 subjects – High Fat Meal (HFM)

30 Factors – Input variables

- 21 Polymorphisms (ApoE, ApoA5, LPL, FABP2, CETP, MTP ..)
- Sex, Age
- 7 clinical measurements: BMI, Fasting levels of TG, Glucose, total cholesterol (TC), HDL-C, LDL-C, Non-esterified fatty acids (NEFA)
- Postprandial TG Response (0-480 min)
Incremental Area under Curve (TG_iAUC) – Output variable (bottom 50%, top 50%)



Introduction

Data

Methods

Results


Conclusions

Multifactorial Analysis of Postprandial Lipemia

- Statistical Methods

- χ^2 Independence test (p -value)
 - Input – Input (redundancy?)
 - Input – Output (main effect?)
- ANOVA (p -value)
 - n -way interactions ($n \geq 1$)

- Hybrid ANN methods

- PDM-ANN (Tomita et al., 2004)
 - Multilayer feed forward artificial neural network (ANN)
 - Parameter Decreasing Method(PDM)
- GA-ANN
 - ANN
 - Genetic Algorithm (GA)  Select factors, optimize ANN structure and ANN training

- Multifactor Dimensionality Reduction(MDR) (Hahn et al., 2003)

- Combinatorial method
- Finds the optimal subset of factors of dimensionality M
- Output rules that describe interactions

- Introduction
- Data
- Methods
- Results
- Conclusions

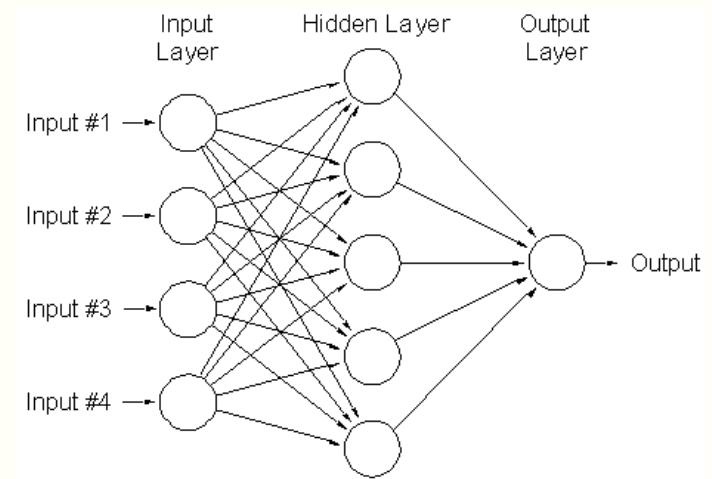
Multifactorial Analysis of Postprandial Lipemia

- Artificial Neural Networks (ANNs)

- Artificial Neuron: Information Processing Unit
- ANN: Parallel connected artificial neuron organized in layers
- Training and Adjustment
- Pattern Recognition and Classification

Proper Encoding of Input and Output Variables
Recursive Training and Evaluation
(n -fold cross validation)

Curse of Dimensionality -> Feature Selection



- Genetic Algorithms (GAs)

- “Biological” (non mathematical) Optimization
- Survival of the fittest (Darwin)
- A chromosome encodes a candidate solution
- Genetic Operators act on chromosomes for a number of generations

Introduction

Data

Methods

Results

Conclusions

Multifactorial Analysis of Postprandial Lipemia

- PDM-ANN

- ANN

- One hidden layer with variable number of neurons (2,4,6,8)
 - One output neuron
 - Training: back propagation (fixed initial learning rate and momentum)
 - 3-fold cross-validation (training and testing sets)

- PDM

- Start from the initial set of 30 input variables (factors)
 - Sequentially subtract input variables that reduce average accuracy in training and testing sets (fitness function: F_N , $1 \leq N \leq 30$)

F_N is kept while redundant or non-affecting factors are affected

Introduction

Data

Methods

Results

Conclusions

Multifactorial Analysis of Postprandial Lipemia

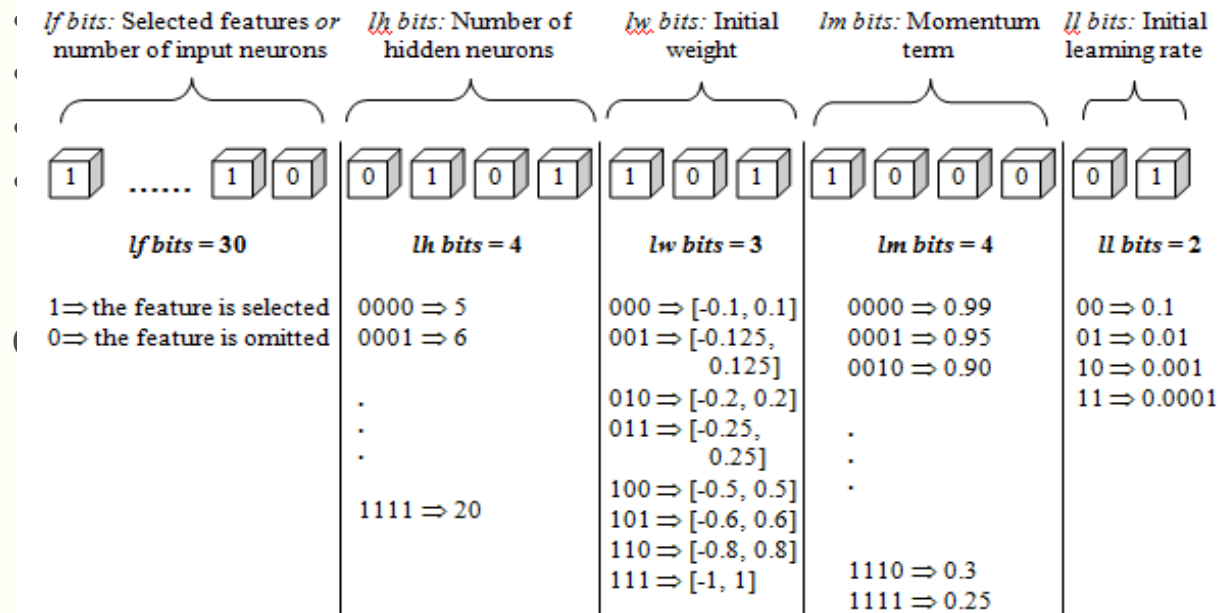
- GA-ANN

- ANN

- One hidden layer with variable number of neurons
 - One output neuron
 - Training: back propagation
 - Training set (60%), Validation Set (20%), Testing Set (20%)

- GA

- Initial Population: $M=100$ binary chromosomes of 43 digits



Introduction

Data

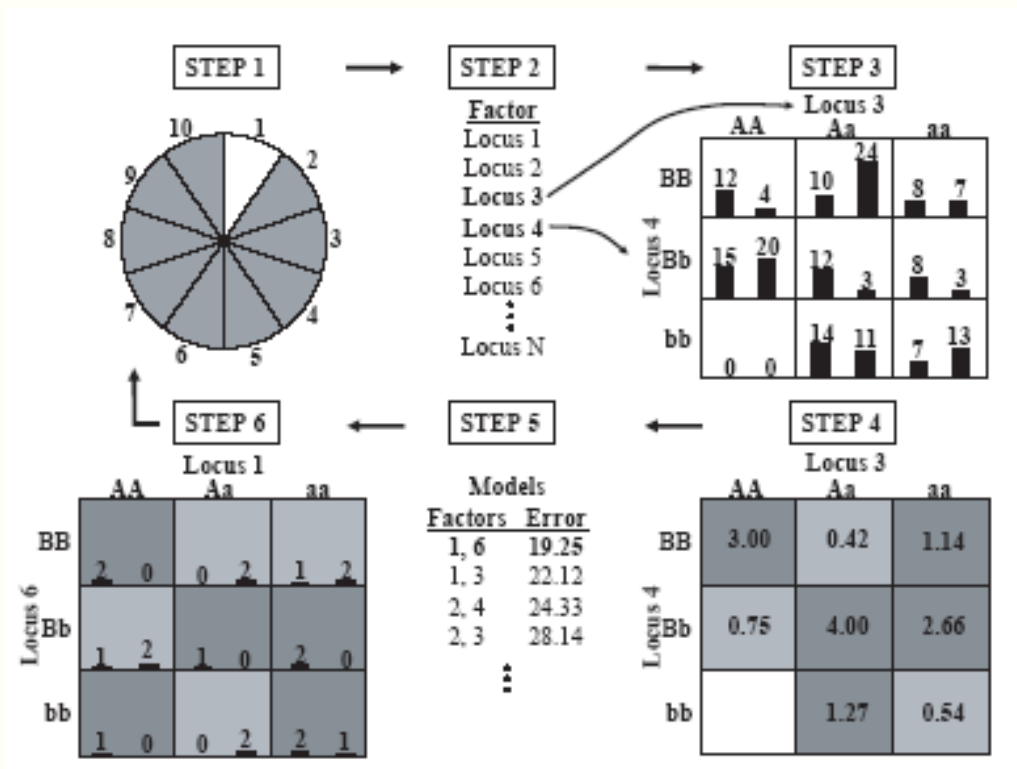
Methods

Results

Conclusions

Multifactorial Analysis of Postprandial Lipemia

- MDR (Hahn *et al.*, 2003)
 - Combinatorial Method
 - Selects the optimal subset of M factors
 - 10-fold cross validation (training set, test set)
 - Variable Subset Consistency (VSC)



- Introduction
- Data
- Methods
- Results
- Conclusions

Multifactorial Analysis of Postprandial Lipemia

- χ^2 Test

Introduction

Data

Methods

Results

Conclusions

	ApoE haplotype	FABP2	ApoB	CETP	INS	LPL Hind III	MTP	LPL S447	TNF	ESR1 XXbaI	ESR1 PPvuII	ApoC3 C3238G	LEPR Gln233Arg	ApoA4 T347S	ApoA5 1131	ApoA5 SGG	ApoA5 haplotype	PPAR α	ApoA4 Q360H	IRS1	ApoE Promoter	Gender	Age	BMI	TC	TG	HDL-C	LDL-C	NEFA	Glucose		
ApoE haplotype																																
FABP2																																
ApoB																																
CETP		0.011																														
INS			0.042																													
LPL Hind III					0.016																											
MTP						0.075																										
LPL S447							0.000																									
TNF																																
ESR1 XXbaI																																
ESR1 PPvuII										0.000																						
ApoC3 C3238G											0.028																					
LEPR Gln233Arg													0.1																			
ApoA4 T347S												0.012																				
ApoA5 1131																																
ApoA5 SGG		0.06			0.058					0.033		0.000																				
ApoA5 haplotype																																
PPAR α						0.083			0.027		0.000		0.064	0.000	0.000																	
ApoA4 Q360H			0.026			0.051		0.088				0.085																				
IRS1			0.07							0.075	0.054										0.012											
ApoE Promoter	0.000								0.01																							
Gender										0.017						0.04																
Age			0.091							0.009												0.065										
BMI			0.067							0.000												0.073	0.000	0.011								
TC			0.001							0.069				0.02								0.067		0.000	0.017							
TG	0.006				0.008	0.054		0.000									0.094						0.000	0.000	0.000	0.000						
HDL-C									0.074	0.031												0.000	0.000	0.000	0.000							
LDL-C	0.083		0.014										0.09			0.012	0.074					0.007	0.000	0.016	0.000	0.000	0.000					
NEFA		0.088						0.009														0.027	0.012		0.032	0.077	0.000					
Glucose												0.09				0.041	0.057					0.053	0.018		0.018							
TG Δ AUC									0.033						0.044	0.084						0.091	0.000	0.042	0.001	0.001	0.002	0.000	0.002			

- Redundant information in the input space
- Main effects to the output

Multifactorial Analysis of Postprandial Lipemia

- ANOVA
 - Main effects

Μεταβλητή Εισόδου	<i>p</i> -value
ApoB	0.008
MTP	0.072
LPL S447	0.029
ApoC C3238G	0.015
ApoA4 T347S	0.030
PPARα	0.293
ApoE Promoter	0.205
Age	0.174
HDL-C	0.055
Glucose	0.229

- 2-way interactions

Μεταβλητές Εισόδου	<i>p</i> -value
ApoB* ApoE Promoter	0.071
MTP*HDL-C	0.202
LPL S447 * ApoE Promoter	0.082
LPL S447* HDL-C	0.013
ApoA4 T347S *Age	0.053
ApoA4 T347S*HDL-C	0.085
PPARα*HDL-C	0.169

Introduction

Data

Methods

Results

Conclusions

Multifactorial Analysis of Postprandial Lipemia

- PDM-ANN
 - F_N is kept 82%-85% up to dimensionality D=10

	Input Variables (N)	Accuracy in training sets	Accuracy in testing sets	F_N
Introduction	HDL-C (1)	66.67	63.38	65.02
	HDL-C, TC (2)	79.58	62.91	71.24
Data	LPL S447, HDL-C, TC (3)	79.11	66.67	72.89
	ApoE haplotype, LPL S447, HDL-C, TC (4)	88.73	70.90	79.81
Methods	ApoB, ApoE haplotype, LPL S447, HDL-C, TC (5)	94.37	68.08	81.22
	ApoB, ApoE haplotype, LPL S447, Sex, HDL-C, TC (6)	94.13	69.01	81.57
Results	ISR1, ApoB, ApoE haplotype, LPL S447, Sex, HDL-C, TC (7)	93.66	70.42	82.04
	ISR1, ApoB, ApoE haplotype, LPL S447, Sex, HDL-C, TC, NEFA (8)	97.18	70.42	83.80
Conclusions	ISR1, ApoB, ApoE haplotype, ApoA5 1131, LPL S447, Sex, HDL-C, TC, NEFA, (9)	96.48	68.08	82.28
	ApoE haplotype, ApoB, LPL S447, ESR1 XXba1, ApoA5 1131, ISR1, Sex, TC, HDL-C, NEFA (10)	99.53	70.89	85.21
(11)	99.30	69.01	0.8415
(12)	97.18	69.95	0.8357
(13)	98.83	69.48	0.8415

Multifactorial Analysis of Postprandial Lipemia

- χ^2 and ANOVA: In a pre-processing step helped observe the input space and reveal up to 2-way interactions
- PDM-ANN selected optimal factor subsets and constructed predictive models
- Hybrid ANN method performed better than MDR

Introduction

Data

Methods

Results

Conclusions

Multifactorial Analysis of Obesity

Obesity

- Independent CVD risk factor (Framingham Study; Wilson et al., 2002)

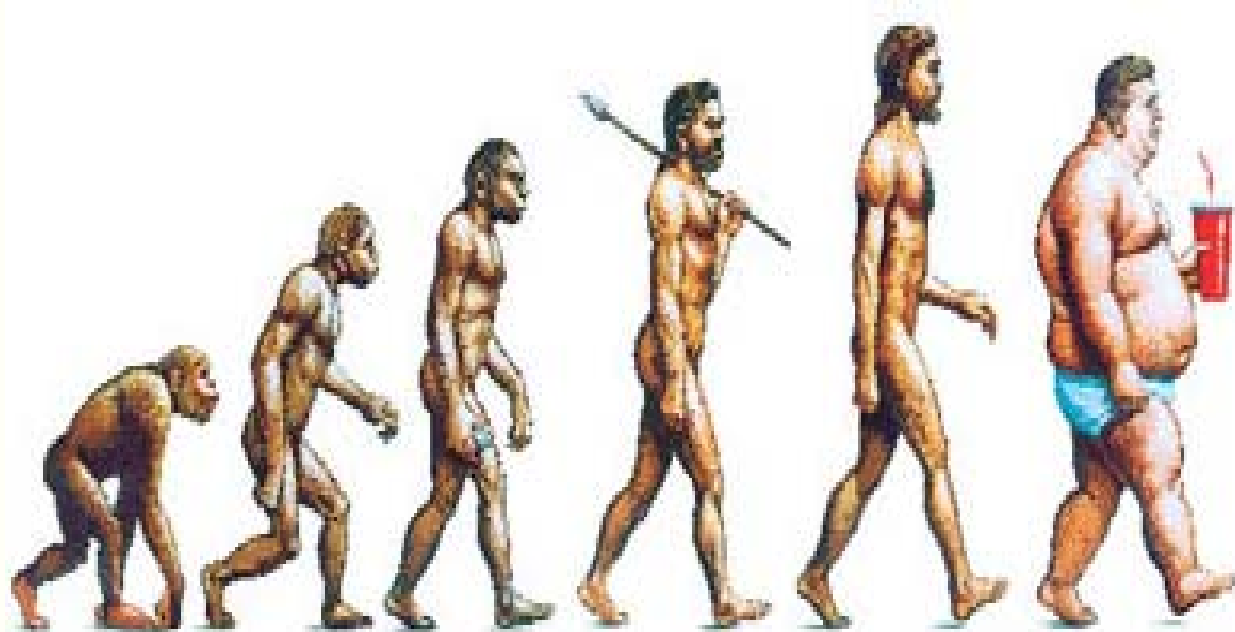
Introduction

Data

Methods

Results

Homo erectus → ***Homo sapiens*** → ***Homo fatuous***



The Economist, December 2003. *The shape of things to come*

Multifactorial Analysis of Obesity

Obesity

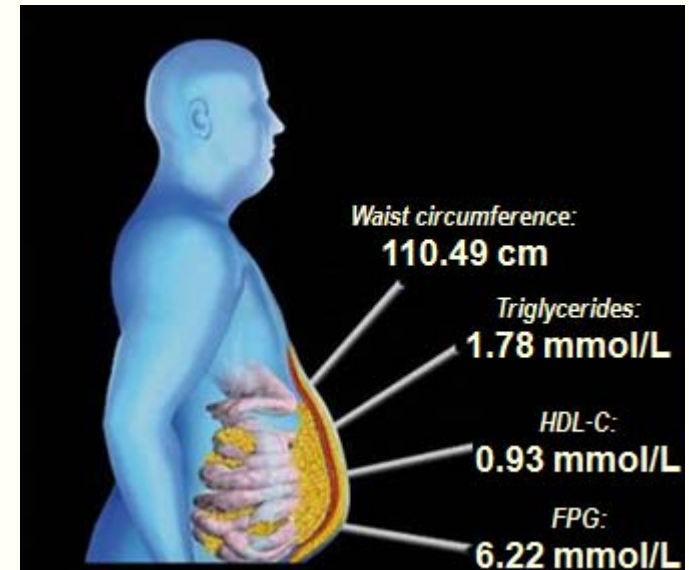
- Independent CVD risk factor (Framingham Study; Wilson et al., 2002)
 - Positive correlation between Mortality and Body Mass Index (BMI)

Introduction

Data

Methods

Results



Multifactorial Analysis of Obesity

Obesity

- Independent CVD risk factor (Framingham Study; Wilson et al., 2002)
 - Positive correlation between Mortality and Body Mass Index (BMI)
- Multifactorial trait
 - Genes
 - Nutrition
 - Exercise
- PDM-ANN, GA-ANN
 - Select factors (genes, sex, nutrition) that affect BMI
 - Construct predictive models for BMI status



Nutrigenetics/Nutrigenomics

Introduction

Data

Methods

Results

Multifactorial Analysis of Obesity

- 2341 subjects (Age: 20-78)
 - Nutrigenetics Test (Sciona, U.S.A.)
- 63 Input Variables
 - Sex
 - 24 Polymorphisms (SNPs, Insertion/Deletion)
 - ✓ 19 genes (CVD or nutrition related)
APOC3, LPL, TNFa, PPAR gamma,....
 - Average daily nutrition measurements (38)
 - ✓ Calories
 - ✓ Several substances (Cholesterol, fatty acids, Vitamins,..) (taken as supplements or food)
- Output Variable: BMI (Kgr/m²)
 - BMI > 25 (Overweight, Obese)
 - BMI ≤ 25 (Normal Weight)

Introduction

Data

Methods

Results

Multifactorial Analysis of Obesity

- PDM-ANN
 - 3-fold Cross Validation
 - F_N : Mean accuracy in training and testing sets
- GA-ANN
 - 3-fold Cross Validation
 - F_N
 - Penalty ($T=30$) or not ($T=Inf$)
- Validation indices
 - Accuracy
 - Sensitivity (BMI>25)
 - Specificity (BMI≤25)

Introduction

Data

Methods

Results

Multifactorial Analysis of Obesity

- PDM-ANN
 - F_N is kept (77.7%-79.6%) up to D=32

	Input Variables (N)	F_N	Accuracy in training sets (%)	Accuracy in testing sets (%)
Introduction	Cholesterol-Intake in Food (1)	62.23	63.16	61.29
	(N=1) + Gender (2)	64.24	64.17	64.32
Data	(N=2) + Vitamin A-Total Intake (3)	65.52	65.71	65.34
	(N=3) + Omega 3-Intake in Supplement (4)	66.12	67.37	64.87
Methods	(N=4) + VDR Fok1 (5)	65.75	68.85	62.66

Results	(N=31) -Vitamin B12-Intake in Food (30)	77.29	94.74	59.83
	(N=32) - TNF alpha G308A (31)	77.30	94.21	60.39
	Gender, Calories, Calcium-Intake in Food, Calcium-Supplement Only, Allium-Intake in Food, Folic Acid-Supplement, Cholesterol-Intake in Food, Cholesterol-Intake in Supplement, Omega 3-Intake in Food, Omega 3-Intake in Supplement, Saturated Fat-Intake in Supplement, Vitamin A-Total Intake, Vitamin A-Intake in Food, Vitamin A-Intake in Supplement, Vitamin B6-Total Intake, Vitamin B6-Intake in Food, Vitamin B6-Intake in Supplement, Vitamin B12-Total Intake, Vitamin B12-Intake in Food, Vitamin C-Total Intake, CBS C699T, CETP G279A, COL1A1 G Sp1 T, GSTM1 deletion, GSTP1 A313G, GSTT1 deletion, IL 6 G634C, MTHFR C677T, SOD2 C28T, TNF alpha G308A, VDR Fok1, VDR Bsm1 (32)	77.89	95.56	60.22

Multifactorial Analysis of Obesity

- GA-ANN
 - Increase of F_N during GA evolution

	<i>T=Inf</i>	<i>T=30</i>
Introduction	<p>F_N</p> <p>80 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0</p> <p>Generations</p>	<p>Gender, Calcium- Total Intake, Calcium- Intake in Food, Allium- Intake in Food, Cruciferous-Intake in Food, Folic Acid- Intake in Food, Folic Acid- Intake in Supplement, Cholesterol-Intake in Food, Cholesterol-Intake in Supplement, Omega 3-Total Intake, Omega 3- Intake in Food, Omega 3-Intake in Supplement, Vitamin A-Total Intake, Vitamin A-Intake in Food, Vitamin B₆- Total Intake, Vitamin B₆-Intake in Food, Vitamin B₁₂-Total Intake, Vitamin B₁₂-Intake in Supplement, Vitamin C-Intake in Supplement, Vitamin D-Total Intake, Vitamin D- Intake in Food, Vitamin D-Intake in Supplement, Vitamin E-Total Intake, Vitamin E-Intake in Supplement, CBS C699T, COL1A1 G Sp1 T, GSTP1 C341T, LPL 1595G, MTHFR C677T, MTHFR A1298C, MTR A2756G, NOS3 G894T (32 factors in total)</p>
Data		
Methods		
Results		

Multifactorial Analysis of Obesity

- ANN architectures' evaluation

Index	ANN architecture(N)	Mean value in training sets (%)	Mean value in testing sets (%)
Accuracy	PDM-ANN (32)	95.56	60.22
	GA-ANN, $T=Inf$ (32)	97.67	60.69
	GA-ANN, $T=30$ (25)	97.10	61.46
Sensitivity	PDM-ANN (32)	98.14	69.15
	GA-ANN, $T=Inf$ (32)	99.39	70.79
	GA-ANN, $T=30$ (25)	98.90	69.80
Specificity	PDM-ANN (32)	91.15	46.08
	GA-ANN, $T=Inf$ (32)	94.73	44.62
	GA-ANN, $T=30$ (25)	94.54	48.63

Valavanis et al. *BMC Bioinformatics* 2010, 11:453
<http://www.biomedcentral.com/1471-2105/11/453>



RESEARCH ARTICLE

Open Access

A multifactorial analysis of obesity as CVD risk factor: Use of neural network based methods in a nutrigenetics context

Ioannis K Valavanis^{1*}, Stavroula G Mouggiakakou^{1,2,3}, Keith A Grimaldi⁴, Konstantina S Nikita¹



Thanks to

Prof. Konstantina Nikita (NTUA, Greece)

Dr. Stavroula Mougiakakou (NTUA, Greece)

Dr. Keith Grimaldi (Sciona, USA, Italy)

Dr. Rosalyn Gill (Sciona, USA)

Prof. Ann Minihane (Reading University, UK)

Stathis Marinos (NTUA, Greece)

George Karkalis (NTUA, Greece)

(FP6-IST-4-027333-STP "Micro2DNA: Integrated polymer-base microfluidic micro system for DNA extraction, amplification, and silicon-based detection")

Questions ?