# Enhanced Gene Expression Programming for Signal-Background Discrimination in Particle Physics

**Liliana Teodorescu**
**Zhengwen Huang**

Brunel
UNIVERSITY
WEST LONDON

# Outline



- ***Gene Expression Programming***
- ***New developments on Gene Expression Programming***
  - ✓ *alternative solution representation*
  - ✓ *controlled evolution*
  - ✓ *dynamic classification threshold*
- ***Comparative studies***
  - ✓ *experiments*
  - ✓ *results*
- ***Conclusions***

# GEP - Evolutionary Algorithm



***Gene Expression Programming (GEP) – a new Evolutionary Algorithm (EA)***

***Multi-purpose algorithms inspired by natural evolution theories***

***String based***

- ***Genetic Algorithms (GA)*** *(J. H. Holland, 1975)*
- ***Evolutionary Strategies (ES)*** *(I. Rechenberg, H-P. Schwefel, 1975)*

***Tree based***

- ***Genetic Programming (GP)*** *(J. R. Koza, 1992)*

***Hybrid representation***

- ***Gene Expression Programming (GEP)*** *(C. Ferreira, 2001)*

# Terminology



- ***Individual* – candidate solution to a problem**

***decoding*** ↑ ↓ ***encoding***

- ***Chromosome* – representation of the candidate solution**

- ***Gene* – constituent entity of the chromosome**

- ***Population* – set of individuals/chromosomes**

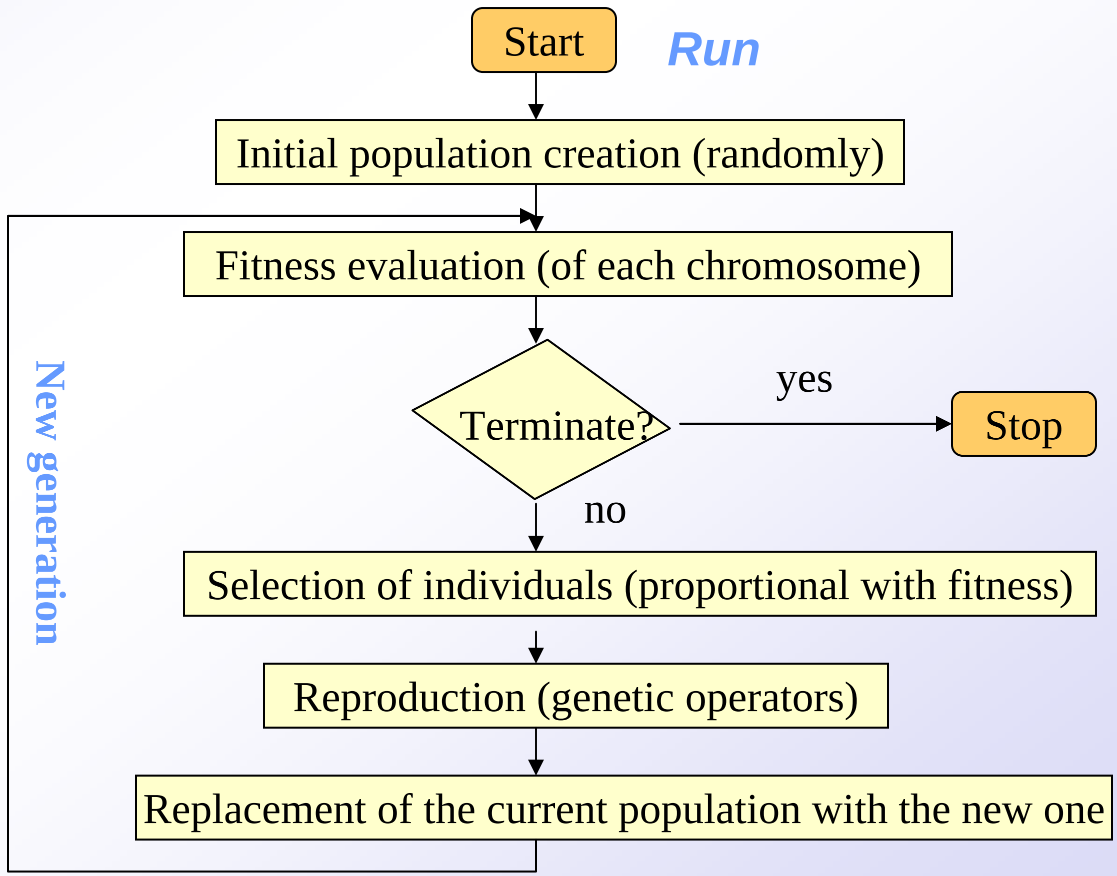
- ***Fitness function* – representation of how good a candidate solution is**

- ***Genetic operators* – operators applied on chromosomes in order to create *genetic variation* (other chromosomes)**

# Evolutionary Algorithms



*EA - iteratively improve the quality of the solution until an optimal/feasible solution is found*




- ***Problem definition***
- ***Solution representation (encoding the candidate solution)***
- ***Fitness definition***
- ***Run***
- ***Decoding the best fitted chromosome = solution***

# Gene Expression Programming



*Chromosome - sequence of symbols (functions and terminals)*

Head (h) Tail (t)

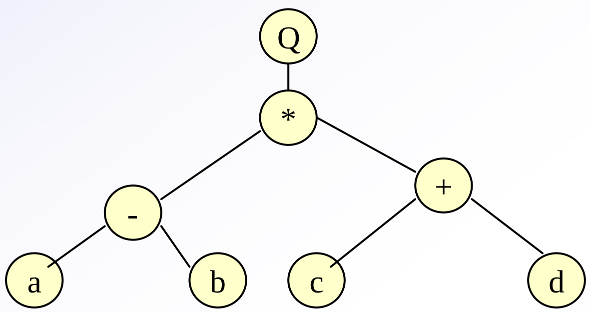
*Q*-+abcdaaabbb*

*t=h(n-1)+1*

*n – higest arity*

***mapping***

***Expression tree (ET)***



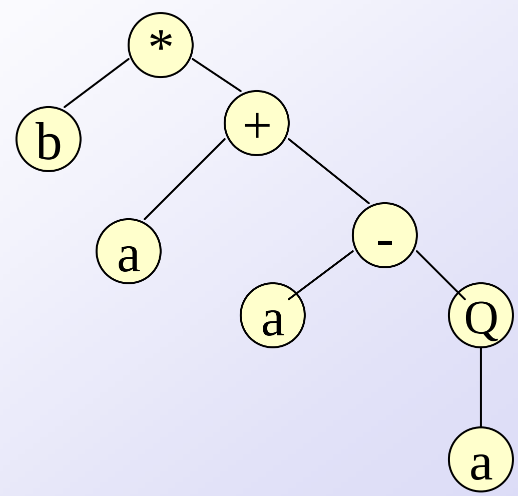

***Translation (as in GP)***

***Mathematical expression***

$$\sqrt{(a-b)\cdot(c+d)}$$

***ET ends before the end of the gene!***

***b+a-aQab+//+b+babbabbbababbaaa**



# GEP (cont.)



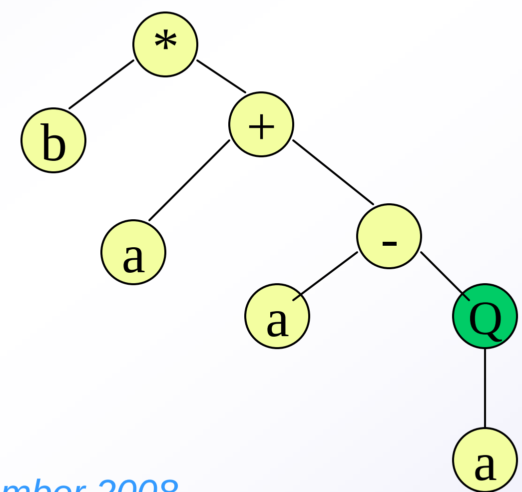
## Reproduction

**Genetic operators applied on chromosomes not on ET =>**
**always produce sintactically correct structures!**

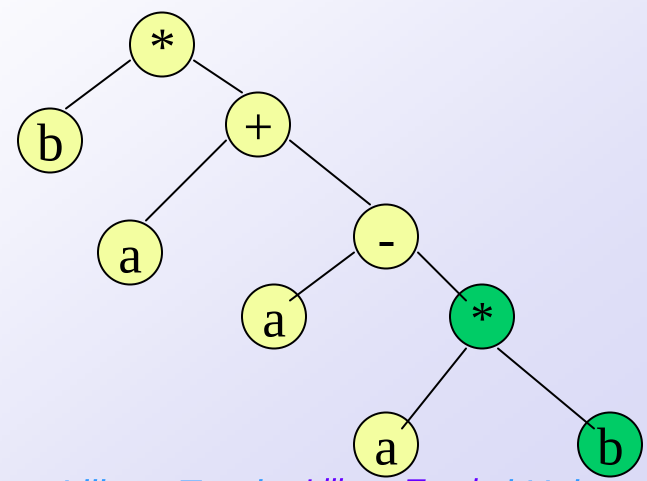
- ***Cross-over* – exchanges parts of two chromosomes**
- ***Mutation* – changes the value of a node**
- ***Transposition* – moves a part of a chromosome to another location in the same chromosome**

*e.g. Mutation: Q replaced with **

***b+a-aQab+//+b+babbabbbababbaaa**




***b+a-a*ab+//+b+babbabbbababbaaa**





Liliana Teodorescu

# GEP in PP - event selection



*L. Teodorescu, IEEE Trans. Nucl. Phys., vol. 53, no.4, p. 2221 (2006)*
*L. Teodorescu, D. Sherwood, Comp Phys. Comm. 178, p 409 (2008)*
*also talks at IEEE NSS 06, CHEP06 and ACAT 2007*
*CERN Yellow Report CERN-2008-02*

***cuts/selection criteria finding*** *for signal/background classification (statistical learning approach)*

- ***fitness function*** *- number of events correctly classified as signal or background (maximise classification accuracy)*
- ***input functions***
  - *logical functions => cut type rules*
  - *all common mathematical functions => continuous function*
- ***input data*** *- Monte-Carlo simulation from BaBar experiment for Ks production in $e^+e^-$ (~10 GeV),* $K_S \rightarrow \pi^+\pi^-$

*8 variables (used in cut-based analysis)*

- *doca (distance of closest approach)*
- *$|cos(\theta_{hel})|$ ( $K_S$ helicity angle)*
- *Fsig (Flight Significance)*
- *Mass ($K_S$ reconstructed mass)*
- *RXY, |RZ| (region around interaction point)*
- *SFL (Signed Flight Length)*
- *Pchi ($\chi 2$ probability of the vertex)*

*20 variables – previous and*

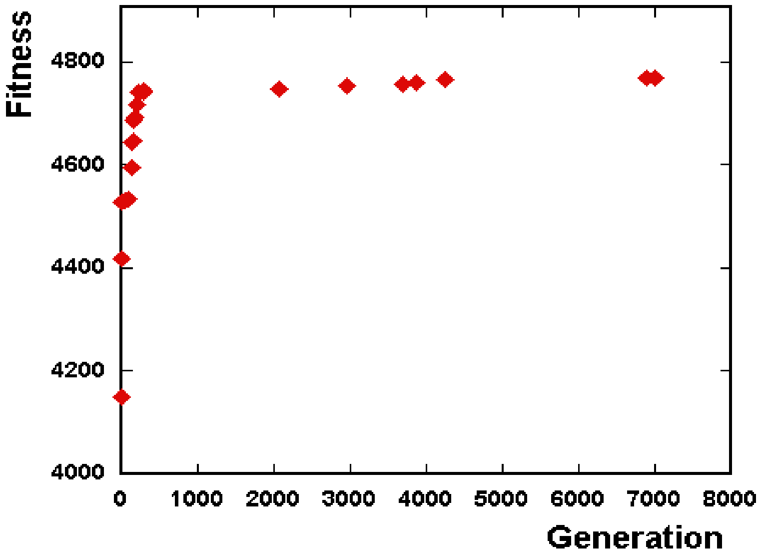
- *cartesian coordinates of $K_S$ vertex*
- *polar coordinate $K_S$ momentum*
- *polar coordinates of $\pi$ daughter particles*

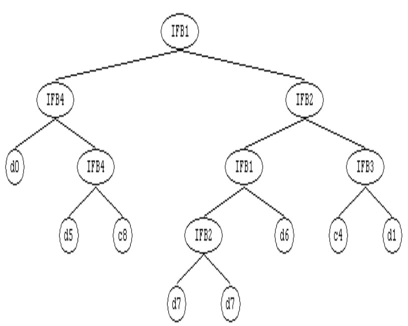
# Previous results



## GEP analysis

*optimises classification*



***Fsig ≥ 4.1***

***Rxy < 0.2cm***

***SFL>0.2cm***

***Pchi>0***

## Cut-based analysis

*optimises signal significance*

***Fsig ≥ 4.0***
***Rxy ≤ 0.2cm***
***SFL ≥ 0cm***
***Pchi > 0.001***

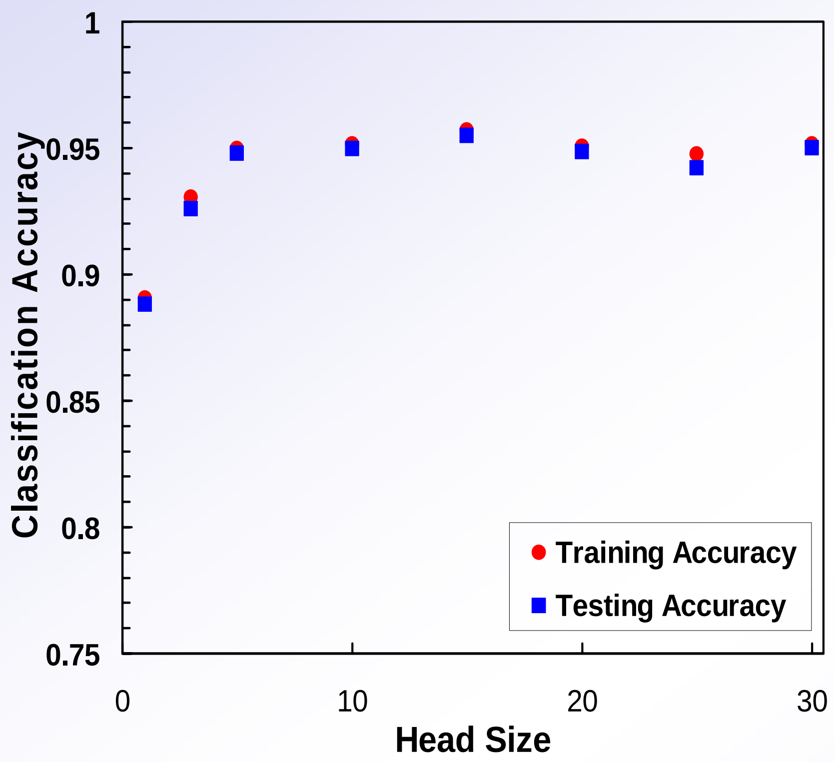
***Reduction***
***S: 15%***
***B: 98%***

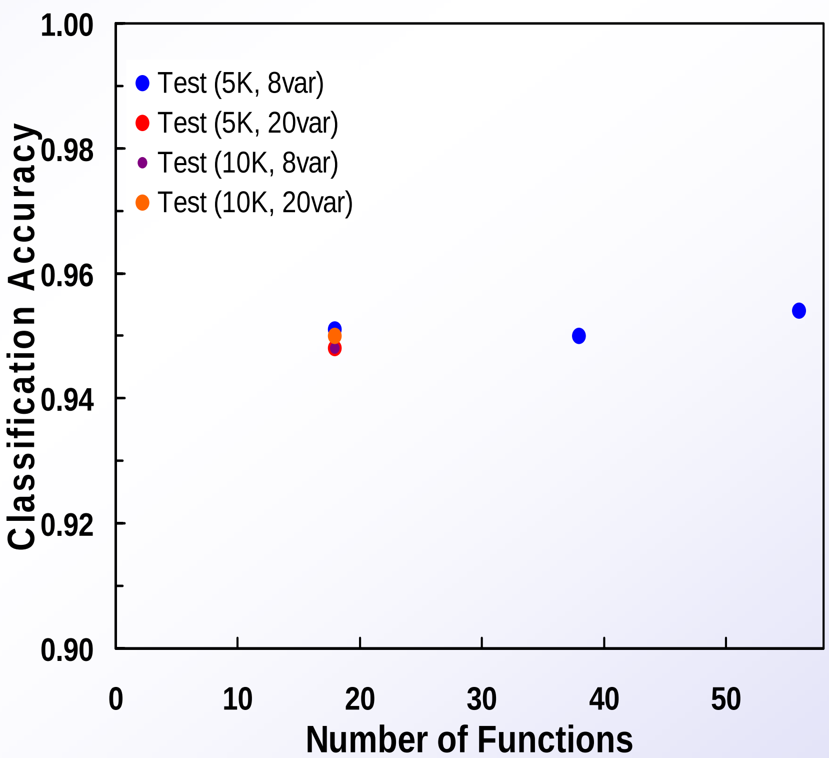
***doca ≤ 0.4cm***
***|Rz| ≤ 2.8cm***

***Reduction***
***S: 16%***
***B: 98.3%***

# Previous results (cont.)






- ***Solutions with good generalisation power***
- ***No overtraining observed***

***No dependence on***

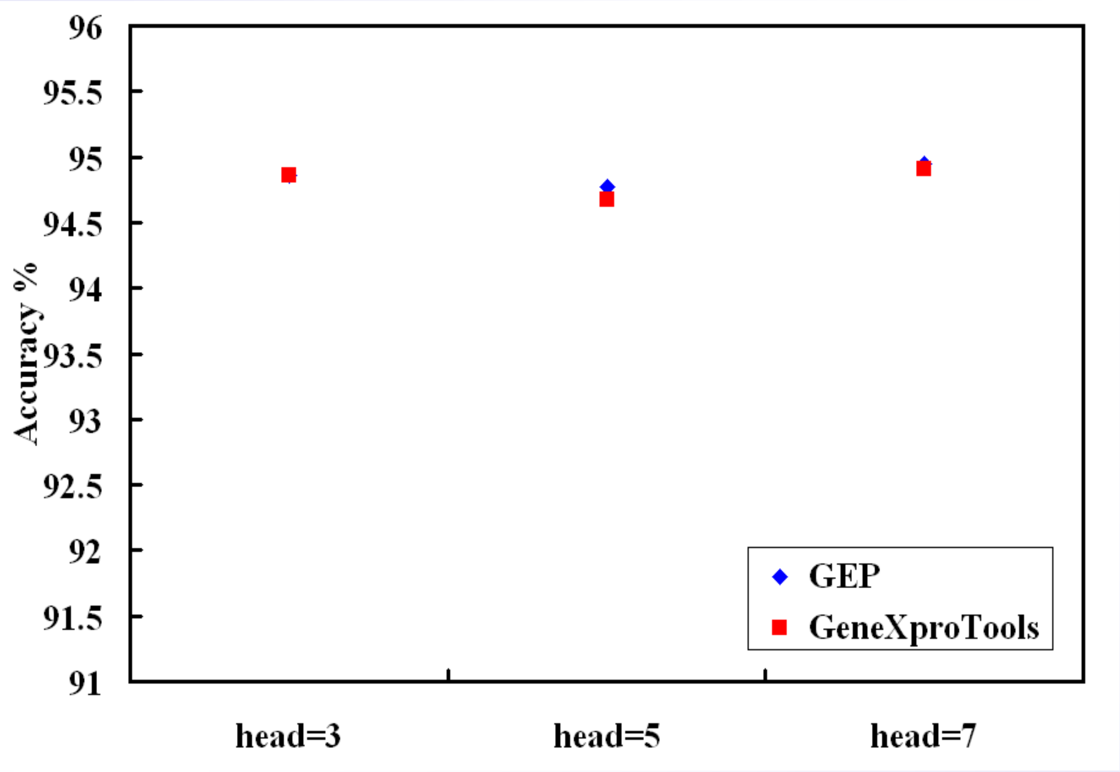
- *event variables (automatic selection of relevant variables)*
- *number of input functions,*
- *number of training events*

# New developments

# Software implementation



- ***Previous studies with GeneXproTools (commercial software package developed by the GEP developer)***
- ***Current studies with a private implementation***




Head=3 (5000 generations)

Head=5 (7000 generations)

Head=7 (15000 generations)

*(less than 0.1% difference)*

# Chromosom - ET mapping

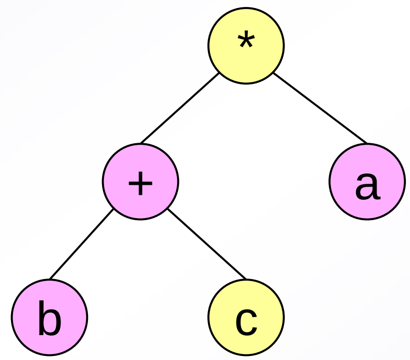


## Postfix order - original GEP (Ferreira, 2001)

**Chromosome**

| | | | | | | |
|---|---|---|---|---|---|---|
| * | + | a | b | c | d | b |

**ET**




**Mathematical expression**

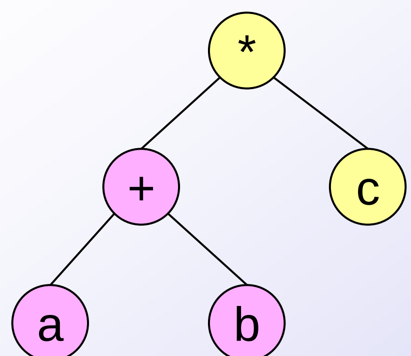
$$(b+c)*a$$

## Prefix order - pGEP (X. Li et.al. ,GECCO2005)

**Chromosome**

| | | | | | | |
|---|---|---|---|---|---|---|
| * | + | a | b | c | d | b |

**ET**



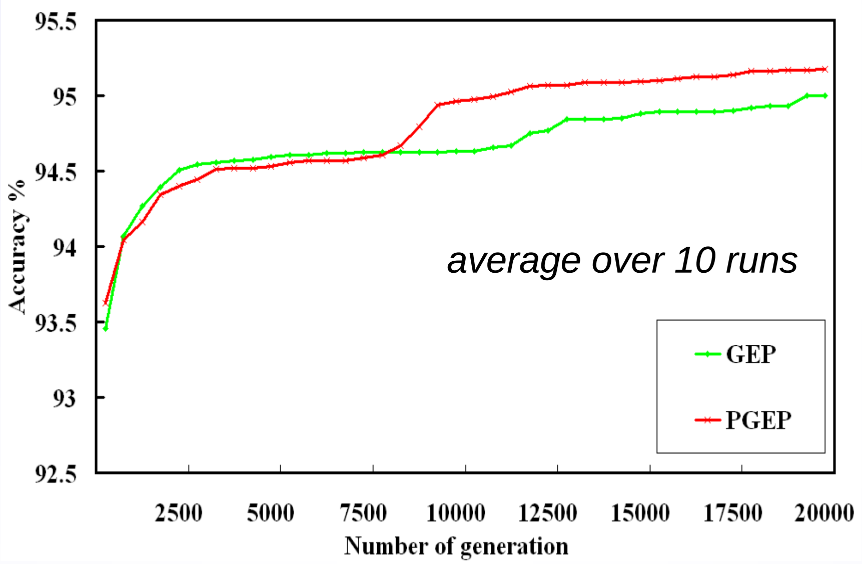

**Mathematical expression**

$$(a+b)*c$$

# GEP vs. pGEP



pGEP keeps the proximity of the genetic material during the translation process → expected lower destructive effect of the genetic operators




*pGEP- earlier convergence*
*- slightly higher accuracy*

*student t-test significance = 35%*

***Proximity of the related genetic material - not controlled during the evolution process***

***Further developments - enforce keeping the related genetic material together might help the evolution***

# Controled evolution



- ***Eliminate the weak individuals ( individuals with fitness lower than a threshold) from the evolution process***
- ***Setting the value of Fitness Threshold (FT)***

  ***Population Diversity vs. Convergence***

  - ✓ ***Static FT - fixed value for all individuals/generations***
  - ✓ ***Online FT – guided by the average fitness per generation***

    *FT = average fitness per generation * scaling factor*

    *Scaling factor should be optimised (typical values between 0.5 to 1.5 )*

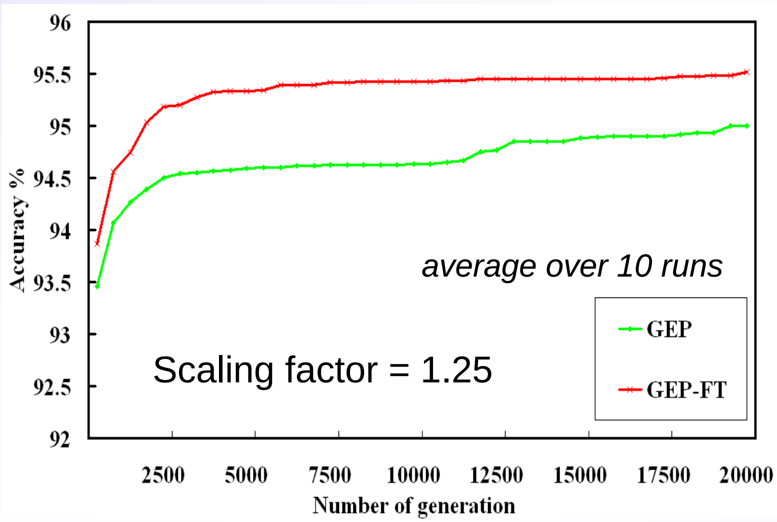
***Versions developed: GEP-FT, pGEP-FT***

# GEP vs. GEP-FT & pGEP-FT

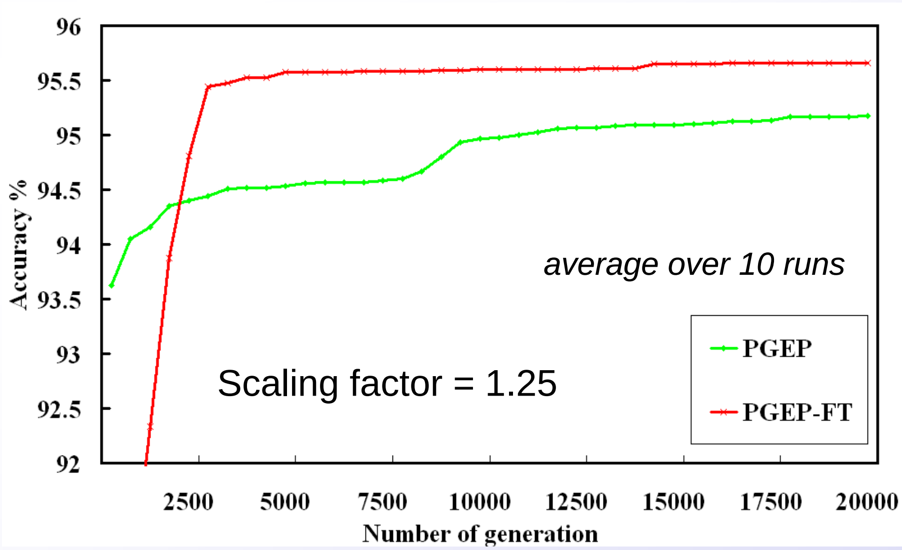


***Static FT** – creates uniformity in the population => convergence problems*

***Online FT** – better pressure on the evolution if FT properly chosen (FT too high => convergence problems)*




*student t-test significance = 0.6%*




*student t-test significance = 0.4%*

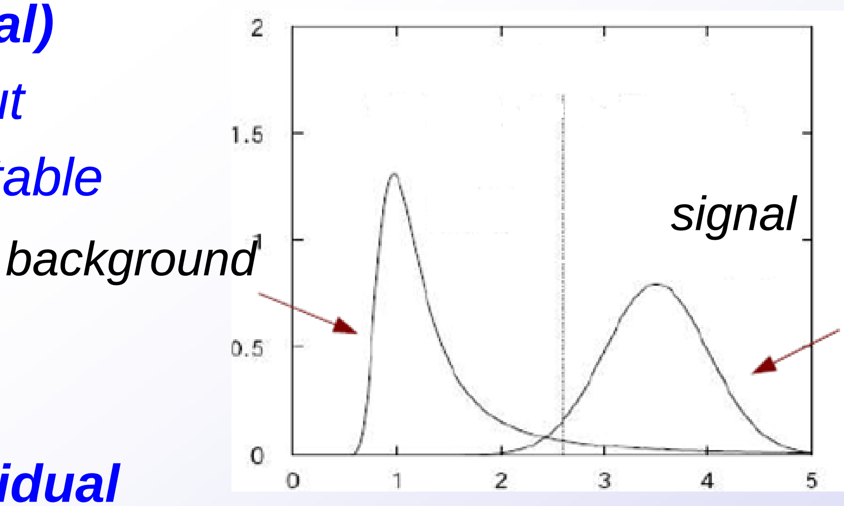
***GEP-FT and pGEP-FT** - earlier convergence*
*- slightly higher accuracy*

# Dynamic classification threshold



**Fixed classification threshold**

- ***for other methods - chosen at the end of the process (on the final output)***
- ***not suitable for GEP (and EA, in general)***
  - ✓ *each individual provides its own output*
  - ✓ *threshold for one individual is not suitable for another*



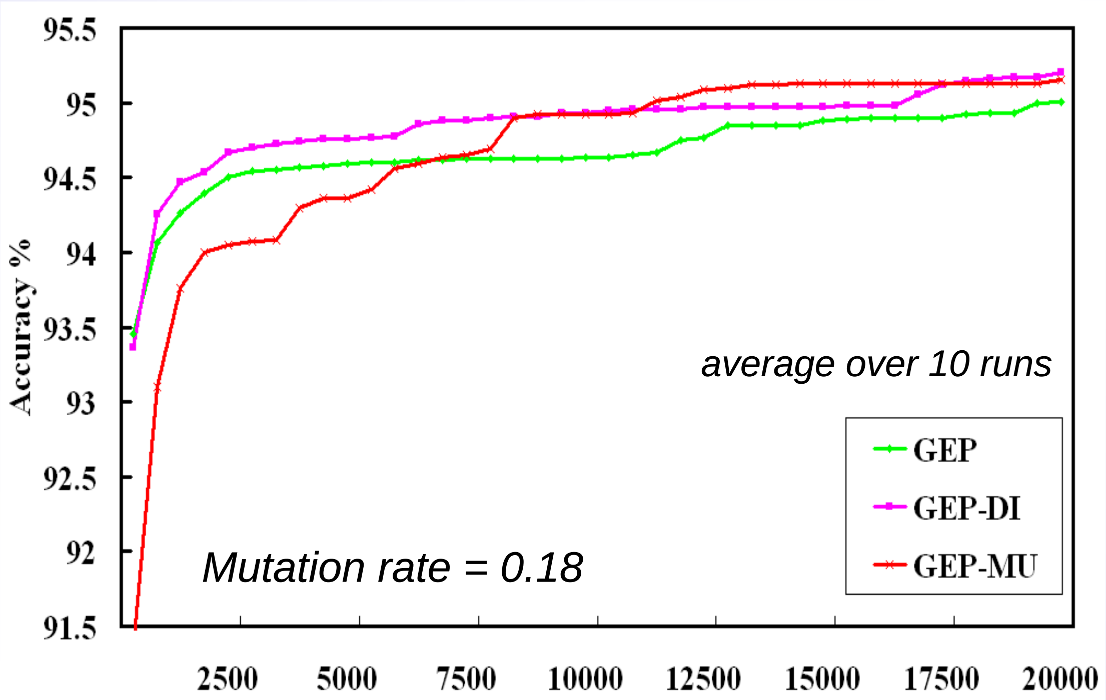

**Dynamic classification threshold**

- ***threshold value adapted to each individual***
- ***two implementations** (GEP-DI and GEP-MU)*

# GEP vs. GEP-DI & GEP-MU



- ***For each individual the optimal threshold is determined by scanning the full range of the output function (GEP-DI)***
- ***Each chromosome has an additional element which contains the potential threshold value which is evolved with a mutation operator (GEP-MU)***



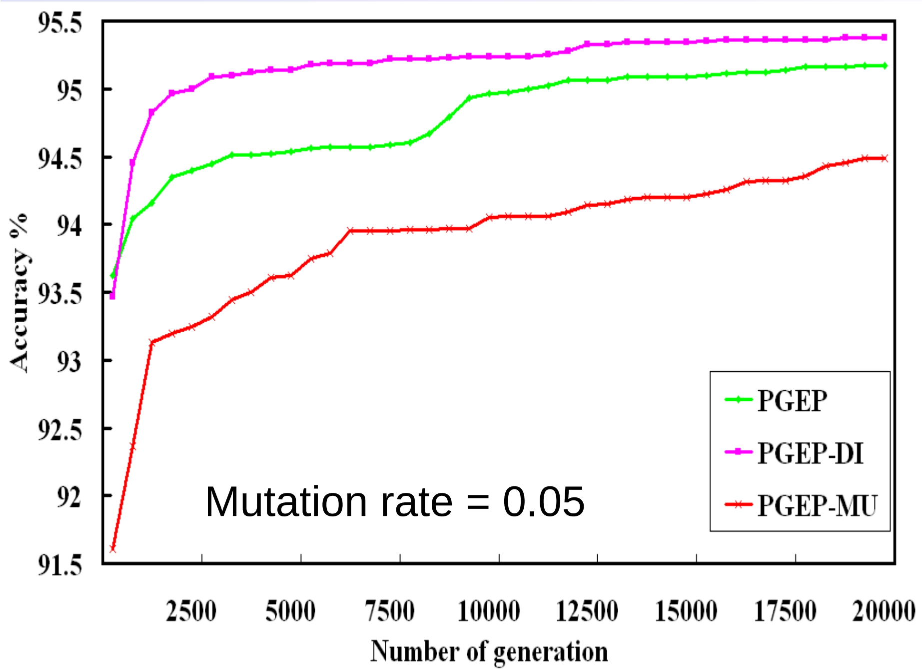

*GEP-DI & GEP-MU - similar accuracy, slightly higher that GEP*

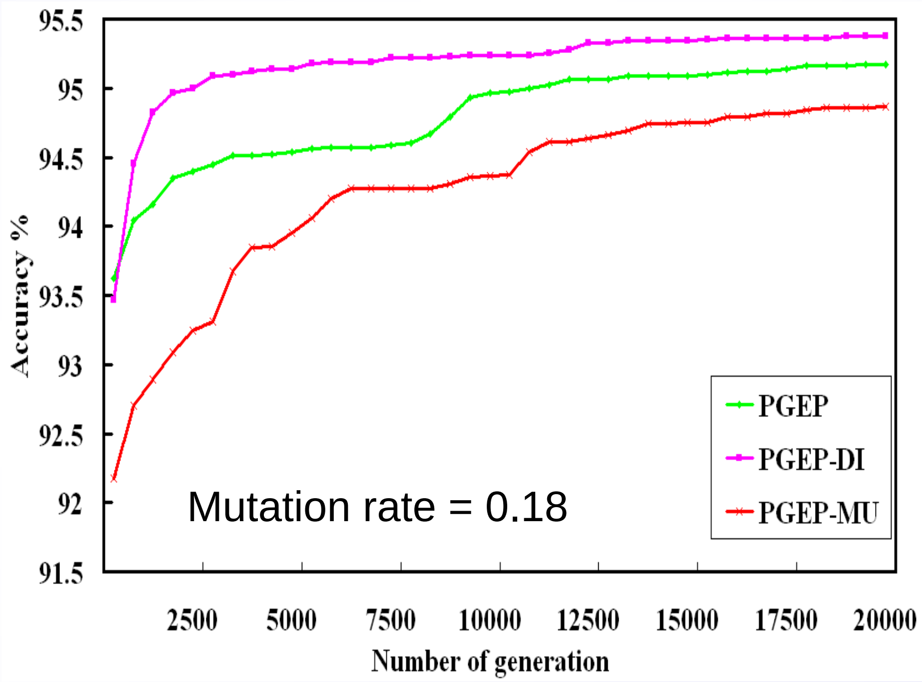
*GEP-MU – slower early evolution but earlier convergence that GEP-DI*

*student t-test sig. = 20%*

# pGEP vs. pGEP-DI & pGEP-MU





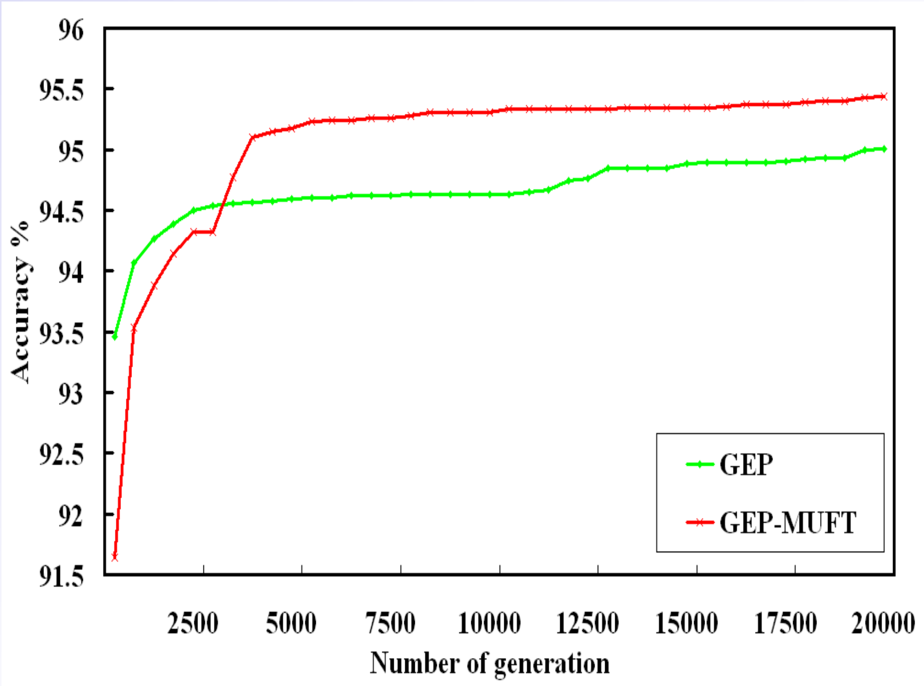



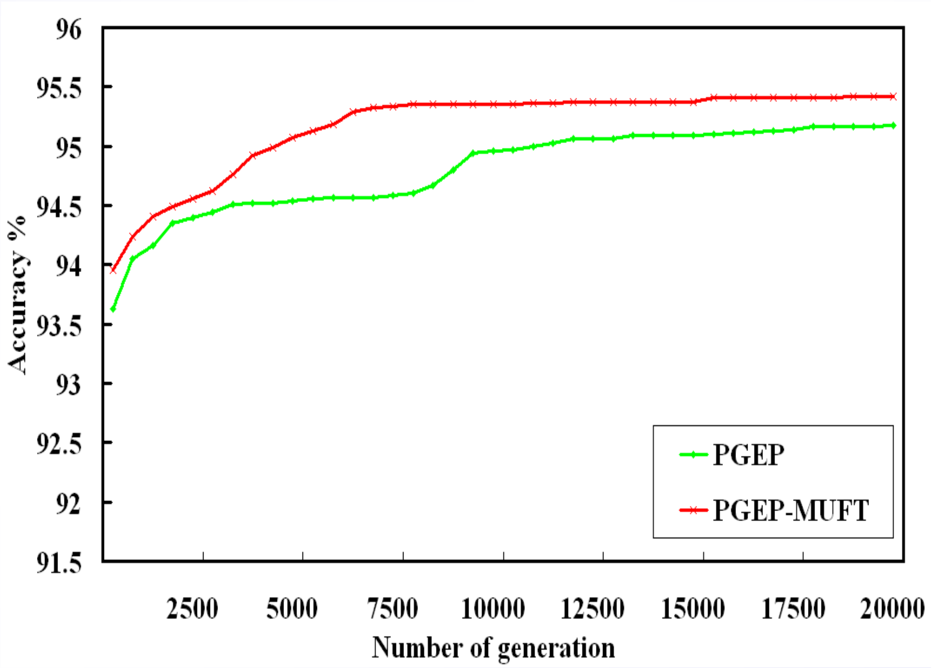

*student t-test sig. (pGEP vs. pGEP-DI) = 23%*

***Mutation rate – not optimised***

# Combined developments






*Student t-test sig. = 2%*

*Student t-test sig. = 15%*

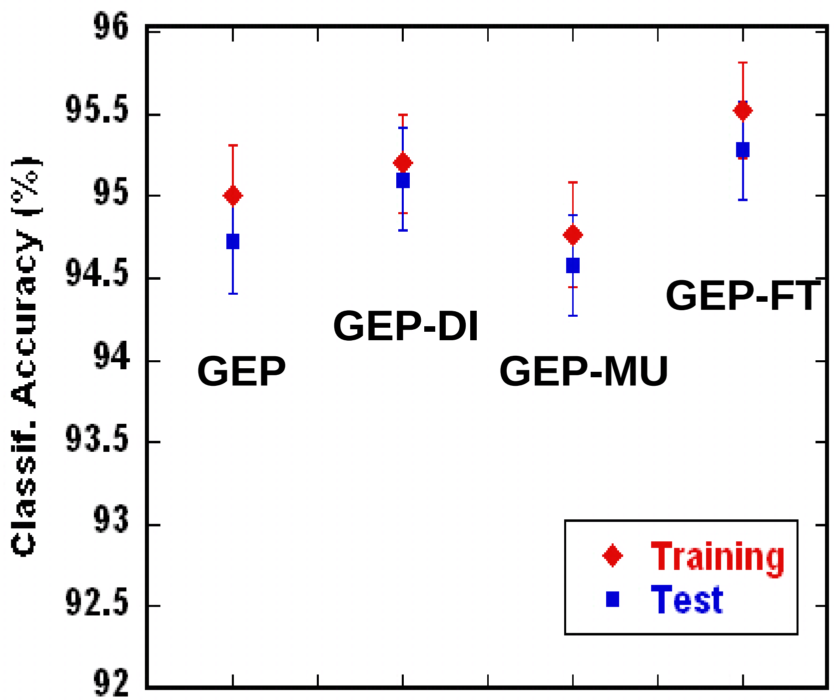
*Mutation rate – not fully optimised in this case*

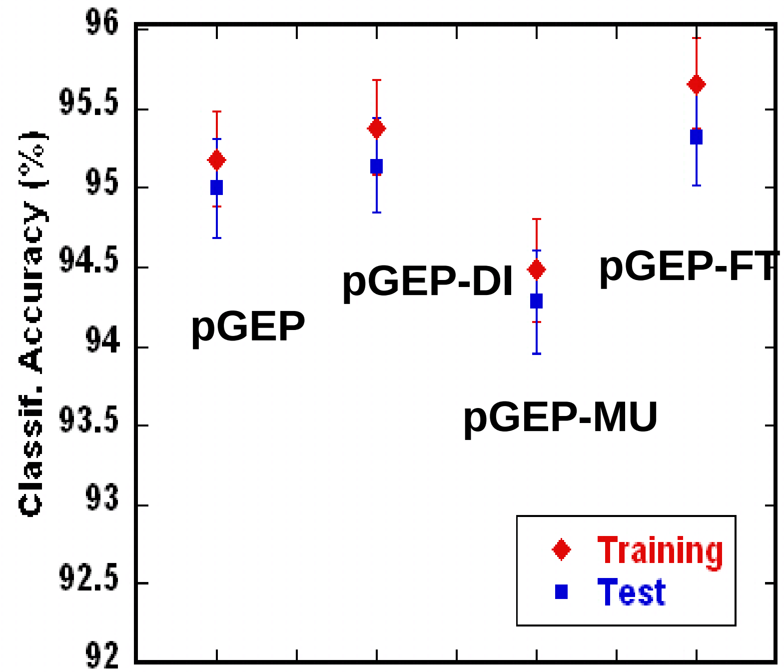
*Improvements - earlier convergence*

*- slightly higher accuracy*

# Training - test comparison






***All models – good generalisation power***

# Conclusions



## Current developments of GEP

- ❖ *software development* – ***allowed us flexibility***
- ❖ *algorithmic research*
  - ✓ ***prefix order mapping***
  - ✓ ***controlled evolution – online fitness threshold***
  - ✓ ***dynamic classification threshold (mutation based & range scanning)***

*New developments* – ***earlier convergence and higher accuracy at various levels (slightly higher accuracy – for this problem)***

## Further developments

- ❖ *algorithms research* – ***further control of the evolution***
- ❖ *software development* – ***extensions to more fitness functions, multi-objective optimisation***
- ❖ *other applications*