

Genetic Programming for Association Studies

Robin Nunkesser¹ Thorsten Bernholt¹ Holger Schwender²
Katja Ickstadt² Ingo Wegener¹

¹Department of Computer Science, University of Dortmund

²Department of Statistics, University of Dortmund

Seminarvortrag IMBIE

1 Introduction

2 Application

- FrEAK
- Genetic Association Study

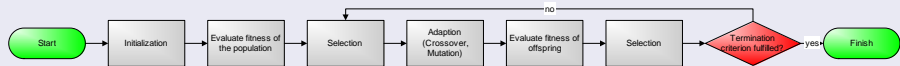
- General term for the usage of search heuristics inspired by natural evolution
 - possible solutions are represented by *individuals*
 - a set of individuals (a *population*) undergoes variation (*crossover* and *mutation*)
 - the *fitness* of the individuals is evaluated
 - a new generation is derived after a *selection* process
- Black-box optimization
- Often used for problems that are not easy to solve by conventional methods
 - Combinatorial optimization
 - Learning
 - ... (Banzhaf et al., 1998 alone lists 178 application examples of GP up to 1997)



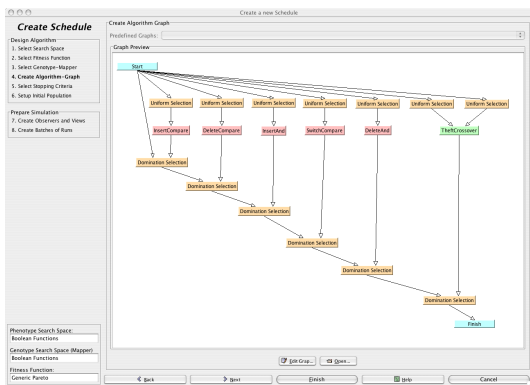
Layout of Algorithms for Evolutionary Computation

- 1 Create an initial random population.
- 2 Evaluate the fitness values of the population.
- 3 Perform the following steps on the current generation:
 - 1 Select individuals in the population based on a selection scheme.
 - 2 Adapt the selected individuals.
 - 3 Evaluate the fitness value of the adapted individuals.
 - 4 Select adapted individuals for the next generation according to a selection scheme.
- 4 If the termination criterion is fulfilled, then output the final population. Otherwise, set the next generation as current and go to step 3.

Graphical Representation



- Base of implementation: Free Evolutionary Algorithm Kit (FrEAK)
<http://sourceforge.net/projects/freak427/>
- Adaptions to easily access FrEAK from R with rJava



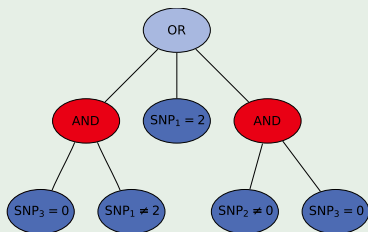
Genetic Programming for Genetic Association Studies

- Goal: Identification of SNPs and SNP interactions leading to a higher disease risk
- Search for logic expressions as predictors

Structure of the Individuals

- We grow trees representing logic expressions in disjunctive normal form
- Multi-valued logic expressions, because our input can take three states
 - homozygous reference (0)
 - heterozygous (1)
 - homozygous variant (2)

Example of an individual

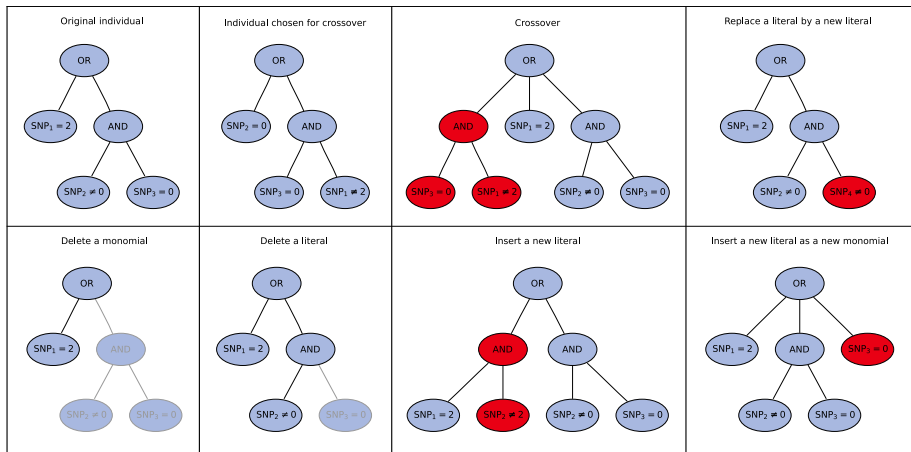


Genetic Programming Algorithm

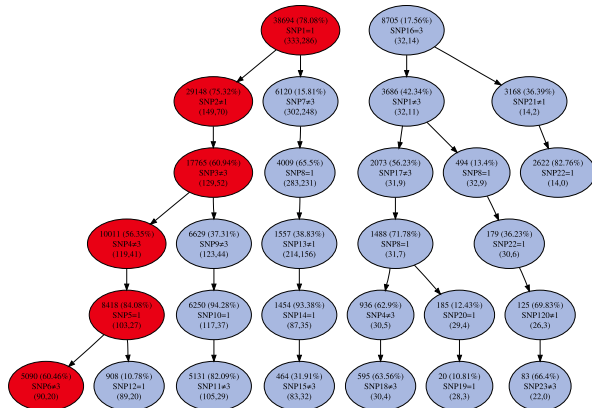
- 1 Create an initial random population consisting of two individuals.
- 2 Perform the following steps on the current generation:
 - 1 Select all individuals in the population, and reproduce them.
 - 2 Conduct mutations and crossovers to uniformly at random selected individuals in the population.
 - 3 Evaluate the fitness value of the adapted and reproduced individuals with fitness functions evaluating
 - 1 Number of predicted controls
 - 2 Number of predicted cases
 - 3 Size of the logic expression
 - 4 Select all adapted and reproduced individuals that are not pareto dominated for the next generation.
- 3 If the termination criterion is fulfilled, then output the final population. Otherwise, set the next generation as current and go to step 2.



Crossover and Mutations

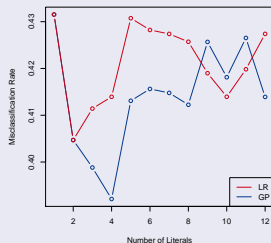


New Visualization to Detect Interesting Interactions



Results on GENICA

MCR of LR and the GP algorithm for a restricted number of variables



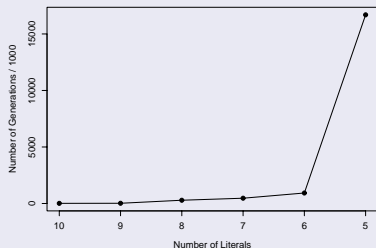
MCR and running times of discrimination

	GP Algorithm	Logic Regression	CART	Bagging	Random Forests
MCR	0.392	0.405	0.429	0.457	0.450
Runtime	6.31	11.75	1.37	21.77	9.03

MCR and running times of discrimination on significant genes

	GP Algorithm	Logic Regression	CART	Bagging	Random Forests
MCR	0.011	0.144	0.356	0.022	0.011
Runtime	1.1 (89.3)	1.15	0.83	5.01	0.3

Search on BRLMM genotypes



Running Time

- About 8 minutes for 10,000 generations.
- Other approaches do not work on a standard PC.

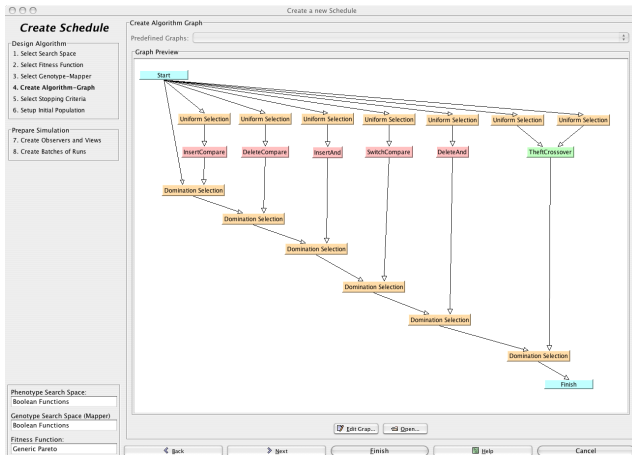


Genetic Programming for Association Studies



- provides easy to interpret models,
- is the only of the considered models working on more than 100,000 SNPs,
- provides the best results in the considered applications,
- is very flexible and adaptable,
- provides results on the fly.

- Interface to R
- Application to other categorical data
- Multi-class case

Short Demo



Thank you!

-  Banzhaf, W., Francone, F. D., Keller, R. E., Nordin, P., 1998. Genetic programming: an introduction: on the automatic evolution of computer programs and its applications. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
-  Nunkesser, R., Bernholt, T., Schwender, H., Ickstadt, K., Wegener, I., 2007. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. Tech. Rep. 24/2007, SFB 475, University of Dortmund.

