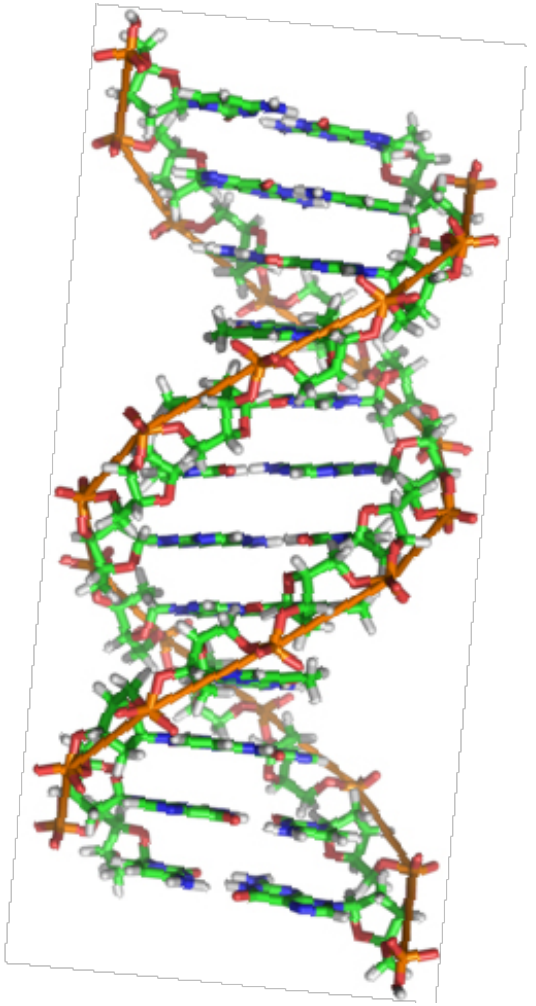
Typewritten symbols recognition using Genetic Programming

I.L. Bratchikov, A.A. Popov

Saint-Petersburg State University

Petrozavodsk, 2010

Table of contents



Purposes and goals

- The main purpose:
  - To estimate the application of GP for the problem of typewritten symbols recognition

- The goals:
  - To determine the superiorities of GP in comparison with the other approaches
  - To develop specific terminals, functions, fitness measure, certain parameters for controlling the run, the termination criterion and method for designating the result of the run.

Description of a problem

- The main problem is to recognize the typewritten Cyrillic and Latin symbols.

- It means the electronic or mechanical translation of scanned images of printed or typewritten symbols into machine-encoded text.

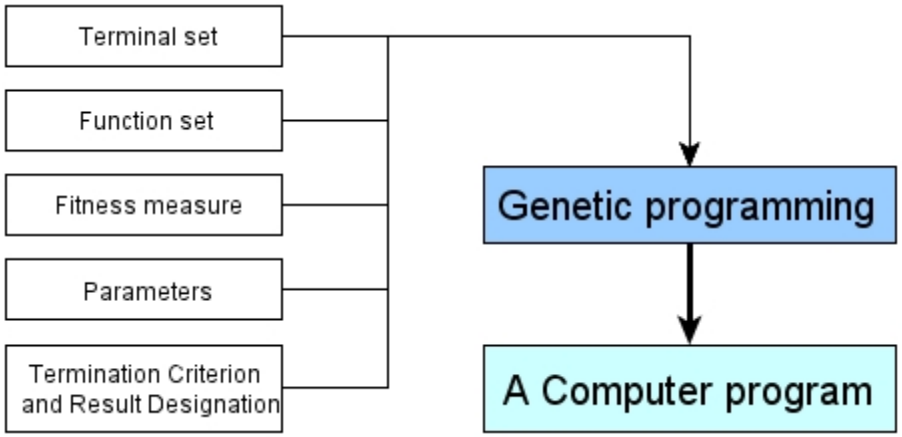
What is GP?

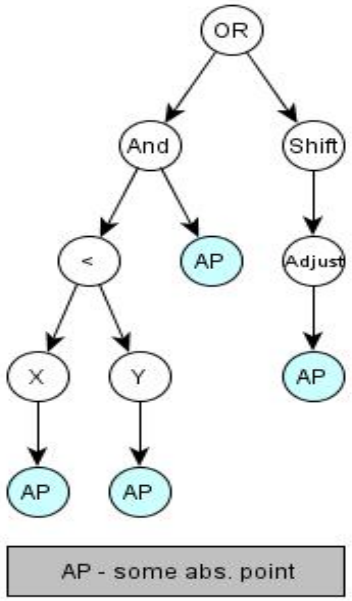
- GP is an evolutionary algorithm-based methodology inspired by biological evolution to find computer programs that perform a user-defined task.

- It is a specialization of genetic algorithms where each individual is a computer program.

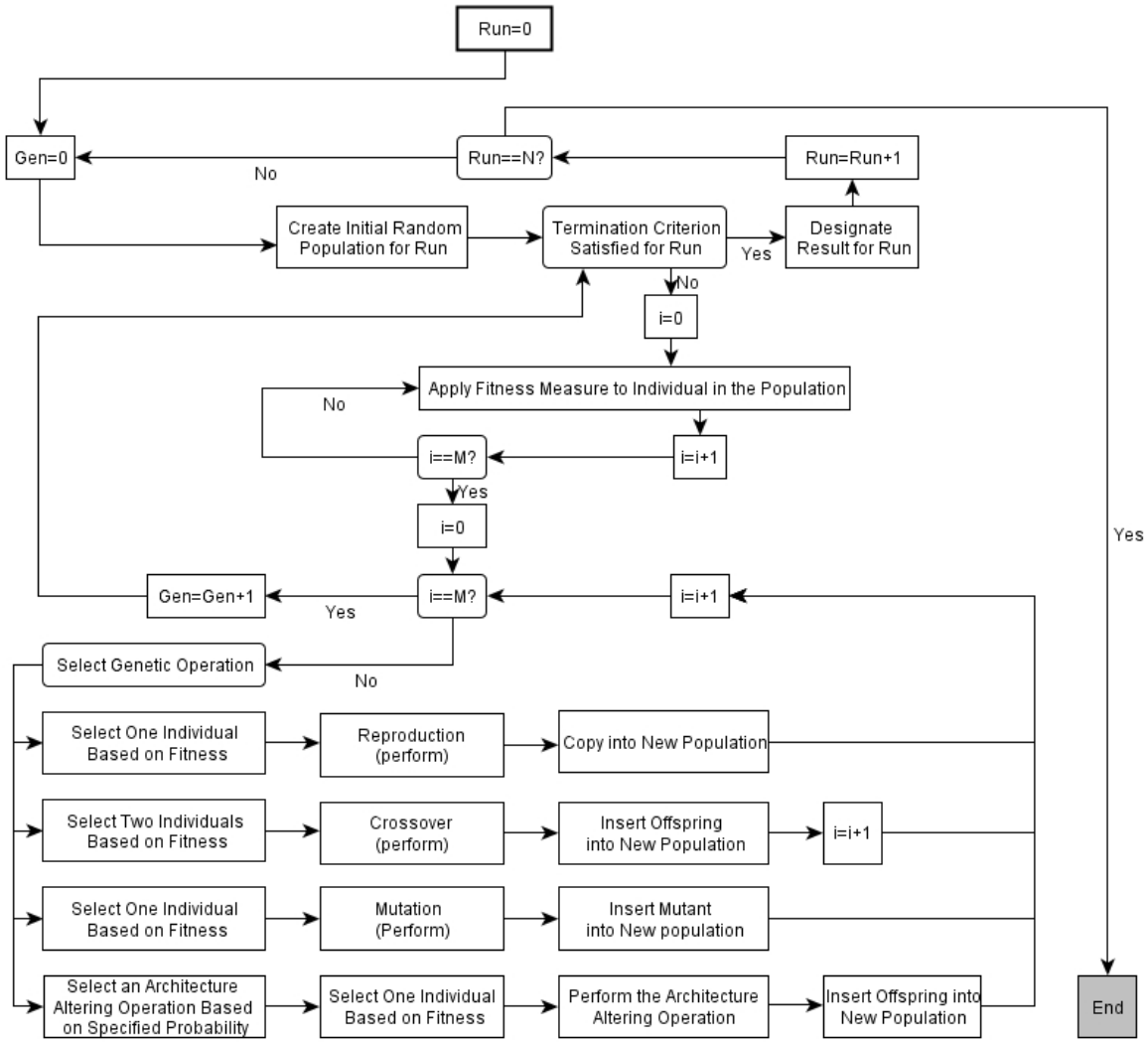
How does it work?

- In so few words GP is a method of solving problems using computers through an analogue of natural selection.
- GP evolves computer programs, traditionally represented in memory as tree





Typical scheme



Adding GP to the problem

- Evaluation of a certain solution is based on a **set** of entities and collects the behavior of the solution on individual elements of this set.

- It's a characteristic for machine learning, where **solutions** are hypotheses, the set contains training cases, and the **evaluation function** is the accuracy of such classification.

Adding GP to the problem

- For some hypothesis the evaluation function returns its accuracy of classification on the training set.
- Incomparability involves a partial order in the solution space and the possibility of existence of many best solutions at the same time.
- We can prevent the algorithm from losing good solutions by replacing the scalar evaluation function with a pairwise comparison of solutions

Outranking relation

- Let’s define formally the **outranking relation** between two solutions (hypotheses), given the sets of examples correctly classified by these hypotheses.
- **Outranking** means that first hypothesis is at least as good as a second one.
- This condition has to hold separately and simultaneously for examples representing some decision classes.

How to select the best solutions?

- Tournament selection scheme cannot work properly in solving this problem due to the fact, that the incomparability decreases the selection pressure, so some tournaments might remain undecided.

- Therefore we have to select some non outranked solutions (hypotheses).

Fitness cases, symbol representation

- The solutions (programs-candidates) performing image analysis and recognition are evaluated on a set of training cases (pictures), called **fitness cases**.
- The data source should be the database of typewritten symbols. It might consist of two subsets, testing and training.
- The symbols could be easily represented by matrix of gray level pixels.
- Let’s assume that the symbols are scaled and centered.

Estimated values

- **population size**: 2000;
- **probability of mutation**: 0.05;
- **maximal depth** of a randomly generated tree (initialization): 3 or 4;
- **maximal number of generations**: 100 (stopping condition);
- **training set size**: 200 cases (100 images per each class);
- **tournament selection**.

Results

- Though GP has some evident superiorities in comparison with the other approaches such as statistics, neural networks and the other techniques, it is not an ideal approach to solve the problem.

- But it could be used **simultaneously** with the other methods in some disputable issues.

In perspective

1. Font normalization (deskewing);
2. Development of recognition system (programming complex or toolbox);
3. Transition from typewritten to handwritten symbols;
4. Integration with the other systems.

Thanks for your attention!