

Applying Genetic Programming to Bytecode and Assembly

Eric C. Collom
University of Minnesota, Morris
coll0474@morris.umn.edu

ABSTRACT

Traditional genetic programming has not been able to evolve entire programs at the source code level. Instead only small sections within the programs are usually evolved. A proven method that solves this problem is by evolving programs in either bytecode or assembly language. This paper provides an overview of applying genetic programming to Java Bytecode and x86 to solve problems. We will discuss solving different problems that these methods have been proven to solve.

Keywords

evolutionary computation, x86 assembly code, Java bytecode, FINCH, automated bug repair

1. INTRODUCTION

Genetic Programming (GP) is a set of techniques used to automate computer problem solving. This is done by evolving programs with an evolutionary algorithm (EA) that imitates natural selection, to find a solution. Traditional GP has been used mostly to evolve only specific parts of programs and not full-fledged programs themselves. This is because traditionally we have to understand the structure of the program being evolved. A EA is usually designed for a specific program with knowledge about how the program works. Being able to evolve an entire program without knowing its structure would allow more flexibility in GP. A method that allows for this flexibility is by evolving bytecode and assembly instead of source code. This is possible because bytecode and assembly languages are less restrictive syntactically than source code. This will be further explained in Section 3.

Orlov et al., [5] propose a method of applying GP to full-fledged programs which only requires a program to be compiled to Java bytecode. Once in Java bytecode a EA is applied to the program to solve the desired problem. Eric et al., [7] also apply a similar method with both Java bytecode and the x86 assembly family.

These methods are important because they show that evolving entire programs is possible, which has yet to be done with traditional GP in source code. This is also useful since once the EA has been created there is little modification that has to be done to it to evolve different programs and the structure of the program being evolved does not have to be known.

This paper will examine how both Orlov et al., [6] and Eric et al., [7] apply GP to solve programs in Java bytecode and x86 assembly. We will also show that this methodology is feasible by solving an array of problems in Section 6. Orlov et al., [6] focused on automating solving simple programs while Eric et al., [7] focused on automated bug repair in programs.

This paper is organized as follows. Section 2 covers the background needed to understand this paper. It contains information on Evolutionary Computation (EC), x86 assembly, and Java bytecode. Section 3 describes the benefits of evolving assembly and bytecode. Section 4 demonstrates how Orlov, et al., [6] evolve Java bytecode. In Section 5 we will discuss how Eric, et al., evolve both x86 assembly and Java bytecode. In section 6 we summarize the results from both Orlov et al., [6] and Eric et al., [7] experiments. In section 7 we will discuss possible future work and ideas.

2. BACKGROUND

In this section we will explain what evolutionary computation (EC) and GP is, their components, and how they are used. We will also discuss what x86 and Java bytecode are. Additionally, we will also go over basic details of the JVM.

2.1 Evolutionary Computation

EC is a field of computer science and artificial intelligence that is loosely based off of evolutionary biology. EC imitates evolution through continuous optimization to solve problems. Optimization in EC is the selection of the best individual within a population. For example Figure 1 shows the process of evolution in EC. An initial population of individuals is taken and a selection process is done to choose the most fit individuals. The most fit individuals are then taken and modified with genetic operators that imitate procreation between two individuals and mutation. Then a check is done to see if any individuals from the population solve the desired problem. If not then the process of evolution is repeated. If the problem is solved then the individual with the best solution is returned.

GP is a tool that uses the EC process to evolve programs. A GP evolves an initial population of programs either until the

This work is licensed under the Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/us/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

UMM CSci Senior Seminar Conference, April 2014 Morris, MN.

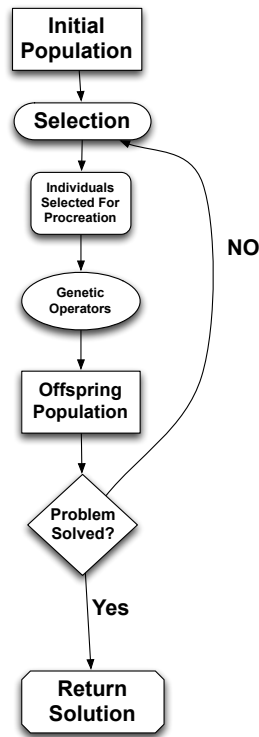


Figure 1: The process of of Evolutionary Computation

desired solution to the problem is found, or a specified number of generations is reached. The fitness of each individual is the deciding factor on how likely it is chosen for evolution. The fitness is a value which indicates how well we think they solve the specific problem. The fitness is usually represented as a numerical value. For the research discussed in this paper a higher fitness indicates a more fit individual. One way that selection of individuals for procreation is done is through tournament selection. This is where a certain number of individuals are chosen to compete and the individual with the highest fitness wins and then is selected for procreation. One way that procreating is implemented in GP is through the genetic operator called crossover. Crossover is the process of taking two individuals and extracting a section of code from one and replacing it with a section from the other program to form an offspring. Another genetic operator that is used to produce offspring is mutation. Mutation takes a program and then randomly changes a section of it. Mutation can be used along with crossover to produce offspring. Also, in some cases, the most fit individuals are passed on to the next generation unchanged, which is called elitism.

2.2 x86 Assembly & Java Bytecode

Through out the rest of the paper I will use the term instruction-level code when referring to both Java bytecode and x86 assembly together. Instruction level code consist of operation codes (opcodes) which are each one byte in length. Hover, some opcodes that take parameters are multiple bytes long [3, 8].

<pre>float x; int y=7; if(y>=0){ x=y; }else{ x=-y; }System.out.println(x);</pre> <p>(a)</p>	<pre>int x=7; float y; if(y>=0){ y=x; x=y; }System.out.println(z);</pre> <p>(b)</p>
--	--

Figure 2: Both (a) and (b) are valid code syntactically however only (a) is valid semantically.

2.2.1 X86

x86 is a family of backward compatible low-level programming languages [9] created by Intel and designed for specific computer architectures. This means that it is designed to only run on certain physical machines. This is one of the main differences between x86 assembly and Java bytecode.

2.2.2 Bytecode and the JVM

Java bytecode “is the intermediate, platform-independent representation” [5] originally designed for the Java compiler. However, other high-level languages now also have compilers for the Java Virtual Machine (JVM). Some of those languages are Scala, Groovy, Jython, Kawa, JavaFx Script, Clojure [6], Erjang, and JRuby.

The JVM executes bytecode through a stack-based architecture. Each JVM thread has a private stack which can only be popped and pushed to. The private stack stores frames. Each frame contains an array of local variables, its own operand stack, and a reference to the class of the current method¹. A new frame is created when a method is invoked. When a method is done executing the frame is destroyed [1]. Only the frame of the current executing method can be active at a time.

Figure 3 is a simple example of what Java bytecode looks like and how the stack works. In this example we are assuming that the frame already contains the local variables 1 and 3 to retain simplicity. After `iload_1` is executed it takes the element from the frame at index 1 and pushes it onto the stack. `iload_2` does the same thing but with index 2. `ladd` pops two elements from the stack, which both must be of type `int`, and then adds them and pushes the result on the stack. `ireturn` simply pops the stack and returns that element.

3. INSTRUCTION-LEVEL CODE BENEFITS

There are many problems we run into when trying to evolve source code. One problem is that it is extremely difficult to evolve an entire program due to source code syntactical constraints. Another problem is that we cannot just take a program and evolve it, we have to design a EA for that specific problem. At times we have to modify the program so that it can be evolved or use templating. Evolving instruction-level code can bypasses both of these problems.

¹It is a reference to the run time constant pool of the class of the current method. More details can be found in chapter 2.5.5 of [2]

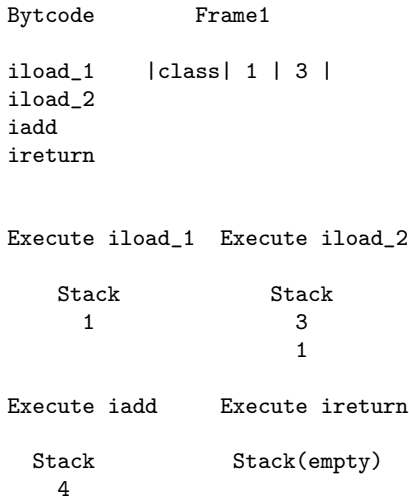


Figure 3: Index zero of the frame is a reference to the methods current class. Istore_n pushes an element from index n from frame onto stack. Iadd pops two elements from operand stack add them together and then pushes to the result to the stack. Ireturn pops an element from the stack and returns it

3.1 Source Code Constraints

One problem with trying to evolve entire programs at the source code level is that there is a very high risk of producing a non-compilable program. This is due to the fact that high-level programming languages are designed to make it easy for humans to read and write programs. Most high-level programming languages are defined using grammars which are used to represent the syntax of the programming language [1, 7]. This means that the grammar does not represent the semantic constraints of a program. The grammar does not capture the languages type system, variable visibility and accessibility, and other constraints [6]. For example, in Figure 2 both 1.(a) and 1.(b) comply to the syntactical rules of Java. However, when taking into account semantics 1.(b) is illegal code. In 1.(b) variable *y* is uninitialized before the test in the if statement, and assigning *y* to *x* violates a type constraint. Plus, variable *z* is not defined anywhere. In order to write a program to evolve source code we would have to deal with all these constraints. While this task is possible it would require creating a full-scale compiler to check for these semantic constraints. Thus, it is easier to evolve instruction-level code.

Instruction-level code consist of a smaller set of instructions [7]. They are simpler syntactically and there are less semantic constraints to violate. This means that there is a lower risk of producing a noncompilable program during evolution.

3.2 Flexibility

In tradition EC when creating a EA to evolve a program there is a lot of scaffolding that has to be done. Usually we have to design the GP for that specific program. When a EA is designed it is common to just evolve expressions

or formulae [6] within the program. Most of the program as a whole would be already written and remain the same after evolution. The EA has to know what it can evolve and where it is located. Also, usually the EA isn't useful in evolving other programs that contain different problems. Instead an entirely new EA would have to be designed and the program along with it.

When evolving instruction-level code the program does not have to be "intentionally written for the purpose of serving as a EA representation". [5] The only requirement is that it can be compiled to instruction-level code [5, 7]. The EA doesn't have to focus on a specific part of the program in order to evolve it. Instead the program that is being evolved stays as is and the EA goes through minor modifications to fit the problem trying to be solved. The EA has to be given a fitness algorithm in order to attempt to find a desirable individual. It also needs to know what genetic operator and the desired number of generations is needed. Also, depending on the program being evolved one might also have to put a growth limit on the code. For example, Orlov et al., [6] enforced a growth limit when performing crossover on code so that the programs would not become too large.

A huge benefit that comes from most of the scaffolding happening in the EA end is that the initial program being evolved does not have to contain a whole lot of code. The EA can evolve a minimal amount of code into a full-scale working program. However, evolving instruction-level code is also capable of being very focused like traditional EA is on source code. For example, Eric et al., [7] in their research focused on only evolving small areas of code and would only make minor changes to the code as a whole. The resulting code would usually only have one line of code changed. However, since they were evolving instruction-level code they were able to fix bugs such as incorrect type declarations, which was not possible in their earlier work done in source code.

As previously stated, in traditional GP the program being evolved is mostly all written before hand and only a section of it is evolved. For evolving instruction-level code this is not the case. Infact, a program can contain the bare minimum elements needed for evolution. This will be further shown in Section 6.1.1.

4. FINCH

FINCH is a program developed by Orlov et al., [5, 6] that evolves programs that have already been compiled into Java bytecode. FINCH also makes sure that all offspring produced are executable Java bytecode through a set of strict rules.

4.1 Selecting Offspring

Evolving Java Byte code only reduces the chance of evolving non-compilable bytecode to a certain extent. Even though Java byte code has a simpler syntax than source code, it still has syntactical constraints. Orlov et al.,[5] address this issue by checking if a good offspring has been created before letting it join the evolved population, thus ensuring offspring produced through crossover contain valid bytecode. If an illegal offspring is produced then another offspring with the same parents is made. This process is repeated until a good offspring is produced or a predetermined number of attempts have been made.

These checks makes sure that stack depth, variable type,

1.	iconst_1	iconst_4	iconst_4	iconst_4
2.	iconst_3	iconst_2	iconst_2	istore_1
3.	istore_1	istore_1	istore_1	istore_2
4.	istore_2	istore_2	istore_2	istore_2
5.	iload_1	iload_1	iload_2	iload_2
6.	iload_2	iload_2	iload_2	iload_1
7.	iadd	isub	isub	isub
8.	ireturn	ireturn	ireturn	ireturn
	(A)	(B)	(C)	(D)

Figure 4: This is an example of good and bad crossover. `iconst_n` pushes onto the stack an int of value `n`. `istore_n` pops the stack and saves the value on the frame at index `n`. Look at figure 3 for a description of `iload`, `iadd`, `isub`, and `ireturn`.

and control flow is respected.

The first check makes sure that the stack and stackframe are both type compatible. Stack underflow is also checked for, where stack underflow is an attempt to pop from an empty stack. This done by assuring that stack pops of β have identical or narrower types and depth as α , and Stack pushes of α must have identical or narrower types as stack pushes of β .

There are also checks done within the bytecode that make sure that by inserting α into β that variables written before and after those sections are compatible with the change. First, variables written by α must have identical or narrower types that are read after β . It is also made sure that all variables read after β and not written by α must be written before β . Finally, all variables read by α must be written before β .

The last checks are for control flow. These checks are done to make sure that when jumps within the bytecode is done that it doesn't cause the program to break. First, a check is done to make sure that there are no jumps in β and jumps out of α . Secondly, a check is done to make sure that the code before β transitions into β and that α transitions into code post α .

These checks, simply put, make sure that all the variables will be written before read, type-compatible, and will not cause stack underflow[5]. In Figure 4, for example, FINCH would not accept 3(D) as a good offspring since it would fail the checks for stack and frame compatibility. These checks are important so that good offspring is selected but also because they provides good variability in the population. Good variability is where there are many individuals that vary in code and fitness. Noncompilable code would result in a fitness score of zero. Since noncompilable code would occur frequently enough it would cause a large portion of the population to have a zero fitness score. This would possibly result in having individuals with a zero fitness being selected for procreation and thus consistently resulting in bad crossover and bad offspring.

4.2 Crossover

The FINCH uses two-point crossover to evolve Java bytecode. It takes two programs A and B and extracts sections α and β respectively. It then takes section α and inserts it into where section β used to be. It only selects an α that will

compile after being inserted into B. For example in Figure 4 let the Java bytecode snippets 4(A) and 4(B) be A and B. Let α be line 6 from A and β be line 5 from B. A(C) is the product of replacing β with α in B. This is considered good crossover because the value at index 2 on the frame can be called twice. It is also an int type which keep isub from breaking. Now let α be lines 3-4 in A and β be lines 2-3 in B. A(D) is the result of replacing β with α in B. However in this case A(D) is incorrect crossover since only one integer is saved on the frame and at line 3 `istore_2` tries to pop an empty stack causing stack underflow.

4.3 Non-Halting Offspring

Another issue that arises from evolving unrestricted bytecode is that the resulting program might enter a non-halting state. These problems do not arise when the check to see if an offspring is good bytecode. Instead, this is a runtime issue. This, is especially true when evolving programs that contain loops and recursion. The way that Orlov et al., [6] deal with this, before running the program, is count how many calls are made to each function. If too many calls are made to a function than an exception is thrown. The lowest possible fitness is assigned to an individual who fails this test.

Orlov et al., chose to count the calls to each function before running the code to prevent from having to either run it on a separate thread or setting a run time limit. They decided against running each program on a separate thread because killing a thread can be unreliable and be unsafe for the GP as a whole. They also decided against setting a time limit due to the difficult of defining how long a program should run. How long a program runs could vary greatly depending on the program being run and the CPU load. Also, this would limit the search space since an individual that is a desired solution could run longer than the time limit.

5. EVOLVING ASSEMBLY CODE

Eric, et al., [7] focus evolving x86 assembly and Java bytecode for the purpose of program repair and debugging. In their tests they took medium to large sized programs in Java, C, and Haskell that contained a bug. The types of bugs they used were common human errors such as having a for loop index off by one. Most of their experiments focused on evolving a small section of the program.

5.1 Selecting Offspring

In selecting offspring a Eric et al., [7] chose a different path than Olov, et al., [6]. They decided to not make sure produced offspring were compilable. Instead they decided to let all produced offspring into the next generation. This produced a considerable amount of individuals with fitness zero due to being noncompilable.

Each program to be evolved was provided with a set of tests that passed and one test that failed. The failed test was used to check if an offspring fixed the bug. The tests that passed were used to make sure the program retained functionality. For each offspring they compiled them into either an executable binary(x86) or class file (Java). If the program failed to compile then it would obtain a fitness of zero. If the program did compile then it was ran against the tests. The fitness score was calculated as the weighted sum of tests passed, the test that initially failed being worth

more. A greater weight was placed on the non passing testing since that was the main goal.

5.2 Genetic Operators

Eric et al., [7] used mutation on 90% of each population and crossover on the rest to produce the offspring population. Multiple tournaments consisting of three individuals was performed to select fit individuals for reproduction. They used mutation over crossover 90% of the time because it produced better results for the type of problems they were solving. Since, each bug was a minor change, such as change a zero to a one, using a large amount of crossover or more complex operators generally lengthened their search time.

Many of the programs being evolved were very large; consisting of thousands of lines of instruction-level code. Because of this Eric et al., used a "weighted path" to select what lines of instruction to apply mutation and crossover too. Each line of instruction-level code was given a weight based off of if the tests executed that instruction. This weight is supposed to indicate how relevant the statement is to the bug. A path weight of 1.0 is assigned if the instruction is only executed by the negative test case. A weight of .1 is given if the instruction is executed by both the negative test case and at least one positive test case. For all other cases a path weight of 0 is given. The path weight is the sum of the weight of all the instructions in that path.

They used three mutation operators. Mutate-insert selects an instruction based off its positive weight. Mutate-delete selects an instruction by proportionally by negative weight and deletes it. Mutate-swap operator selects an instruction proportionally by negative weight and swaps them. They also implement a simple version of crossover that swaps sections from two programs creating two new offspring.

5.3 Non-Halting Offspring

Eric, et al., also chose a different approach than that of Olav, et al., when dealing with non-halting offspring. They decided to not check for non-halting cases and instead run each individual on a virtual machine (VM) with an eight second timeout on the process. They did run into one of the problems that Orlov et al., were worried about. They had trouble having programs not respond to termination requests. They also ran into stack buffer overflow which occurs when there is an attempt to write to a memory address outside of data structures size. This can cause the program and system as a whole to crash. However, this was expected since one of the bugs they were trying to fix was buffer overflow. Also, since they ran each individual on a VM buffer overflow was not a significant problem.

6. RESULTS

Both Orlov et al., and Eric et al., both were able to evolve programs successfully by evolving instruction-level code. Many of the different designs of their EAs was due to the type of problems they were trying to solve. Orlov et al., focused on evolving simple programs that performed a specific task while Eric et al., focused on evolving programs to fix bugs within them.

6.1 FINCH

Orlov et al. [6] focused on evolving small programs to solve problems. The five problems they focused on were symbolic regression, artificial ant, intertwined spirals, array sum, and

```
class SymbolicRegression{
    Number symbolicRegression(Number num){
        double x = num.doubleValue();
        return Double.valueOf((x+x)*x);
    }
}
```

Figure 5: Example of a possible starting program to evolve and solve symbolic regression.

tic-tac-toe. We will discuss their results from the symbolic regression and array sum problems. Their program, FINCH, was able to evolve programs and solve each of these problems.

In each of tests FINCH was given each time a program that had a zero fitness. The elements included in the programs were the minimal components to successfully evolve and solve the problem.

6.1.1 Symbolic Regression

Symbolic regression is a method of finding a mathematical function that best fits a set of points between two intervals. Olov, et al., [6] chose to use 20 random points, between -1 and 1, from various polynomials. Fitness was calculated as the number of points hit by the function. The function set $\{+, -, *, \%, \sin, \cos, e, \ln\}$ ²³ was used for most of their experiments. This was done in order to mimic previous experimentation done [4] in order to compare the results to traditional GP. They were able to evolve up to a 9-degree polynomial with just the function set $\{+, *\}$. This is a good example of how little needs to be written of the original program to be able to evolve it and find a solution.

A more complex fitness was also tested. This symbolic regression consisted of a fitness that checked if all twenty points were within 10^{-8} degree of their polynomials output.

For each experiment they started off with an offspring of fitness zero and usually with a minimal amount of code, such as in Figure 5. Figure 5 only contains a simple function set of $+$, and a number object. However they were able to evolve programs like this to full-fledged programs that solved symbolic regression of up to 9-degree polynomials. They evolved programs with a 90% crossover. Using the simple fitness algorithm 99% of the time a maximum fitness individual was found. With the more complex fitness algorithm a maximum fitness individual was found 100% of the time.

6.1.2 Array Sum

The array sum problem consists of adding up all the values in an array. This problem is important because it would require evolving a loop or recursion to solve it. This would show that FINCH is capable of evolving more complex programs

FINCH was able to solve this problem quite easily though recursion and a for loop. This experiment also showed that FINCH is able to evolve different list abstractions such as

² e is not being referred as the constant but as the function in the java.lang.Math library.

³ $\%$ and \ln are protected division and logarithm. This means that if there is something like division by zero the EA avoids the computation and outputs some pre-selected value. Orlov et al. [6] do not specify how they deal them.

```

int sumlistrec(List<Integer> list) {
    int sum = 0;
    if (list.isEmpty())
        sum *= sumlistrec(list);
    else
        sum += list.get(0)/2 + sumlistrec(
            list.subList(1, list.size()));
    return sum;
}

```

Figure 6: Initial population function, for array sum, that enters into infinite recursion in the if statement

```

int sumlistrec(List list) {
    int sum = 0;
    if (list.isEmpty())
        sum = sum;
    else
        sum += ((Integer)list.get(0)).intValue() +
            sumlistrec(list.subList(1,
                list.size()));
    return sum;
}

```

Figure 7: FINCH’s solution to the initial program presented in Figure 6

List and ArrayList.

When evolving array sum with recursion the initial population consisted of an individual which entered infinite recursion as shown in figure 5. Due to the way FINCH deals with non-halting programs this wasn’t a problem and evolution was able to ensure. The resulting code from evolving Figure 5 is shown in Figure 6.

6.2 x86

looking into other research/papers by this group to get more to write about.

Eric, et al., were able to show that it is possible to fix human programming errors through bytecode and assembly. They were able to successfully debug various programs containing bugs such as infinite loops and remote buffer overflows. The interesting thing about these experiments was that it was done on real programs with real bugs and at times the programs consisted of thousands of lines of code. Another interesting thing that they discovered was that there were some bug repairs they were able to solve through x86 assembly that were not possible in the source code.

Even though they produced a considerable amount of offspring with fitness zero this did not seem to hurt their result very much. The average number of fitness evaluations required to produce an offspring that passed all the tests was 63.6 for C and 74.4 for assembly. This indicates that not much more computational work is needed to evolve repairs in assembly. Even programs that contained thousands of lines of code only required a few runs. This shows that evolving programs in assembly is feasible.

It was also discovered that most repairs only required a small number of operations. This was also due to optimization of mutation.

6.3 Future Work

Future work that could be done, through evolving instruction-level code is to evolve programs whose solutions are much longer and complicated. Even though Orlov et al., [6] proved that using bytecode evolution it is possible to solve many simple problems. Most of the problems only consisted of a few lines of code.

For the work done by Eric et al., [7] it would be nice to see debugging done on less focused areas. Possibly a program as a whole. This would be a good development since it would greatly reduce how much scaffolding is required to set up the evolution. However, this might not be extremely feasible in a real world situation where test coverage is usually minimal.

7. CONCLUSION

There are things that evolving at the instruction-level can be done that can not be done at the source code level. Benefits such as evolving the program as a whole and reduced scaffolding to the GP, and no scaffolding for the program being evolved. Eric et al., and Orlov et al., showed that evolving instruction-level code is just as good as source code and at times even better due to the fact there are some things that you can only do at the instruction-level. In conclusion, evolving instruction-level is feasible and exciting for the field of EC since it opens up many possibilities that were once unavailable.

8. ACKNOWLEDGMENTS

Nic McPhee, Elena Machkasova

9. REFERENCES

- [1] *The Java Language Specification*. Oracles America, Inc. And/or affiliates, 500 Oracle Parkway, Redwood City, California 94065, U.S.A., 2013.
- [2] *The Java Virtual Machine Specification*. Oracles America, Inc. And/or affiliates, 500 Oracle Parkway, Redwood City, California 94065, U.S.A., 2013.
- [3] Java bytecode. February 2014.
- [4] J. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. A Bradford book. Bradford, 1992.
- [5] M. Orlov and M. Sipper. Genetic programming in the wild: Evolving unrestricted bytecode. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, GECCO ’09*, pages 1043–1050, New York, NY, USA, 2009. ACM.
- [6] M. Orlov and M. Sipper. Flight of the finch through the java wilderness. *Evolutionary Computation, IEEE Transactions on*, 15(2):166–182, April 2011.
- [7] E. Schulte, S. Forrest, and W. Weimer. Automated program repair through the evolution of assembly code. In *Proceedings of the IEEE/ACM International Conference on Automated Software Engineering, ASE ’10*, pages 313–316, New York, NY, USA, 2010. ACM.
- [8] Wikibooks. X86 assembly/machine language conversion — wikibooks, the free textbook project, 2013. [Online; accessed 23-March-2014].
- [9] Wikipedia. X86 assembly language — wikipedia, the free encyclopedia, 2014. [Online; accessed 23-March-2014].