HOMEWORK MODULE: SEARCHING FOR *SURPRISING SEQUENCES* WITH A GENETIC ALGORITHM

SIGVE SEBASTIAN FARSTAD

1. Introduction

This report presents a solution to Homework Module: Searching for *Surprising Sequences* with a Genetic Algorithm, IT3708, spring 2014, NTNU. The assignment is to use a previously developed genetic algorithm framework to find long Surprising Sequences of at most 20 symbols.

2. Surprising Sequences

A surprising sequence, as defined by the well-known puzzle writer Dennis Shasha, is a sequence of symbols free of repeating patterns. Formally:

"A sequence is surprising if and only if, for every pair of symbols, A and B, and any distance d, there is at most ONE instance in the sequence of AX_dB, where X_d is any subsequence of length d."[1]

This report considers two types of surprising sequences: locally surprising sequences and globally surprising sequences. The latter are sequences as defined above, while the former are sequences in which there are no repeat occurrences of AX_0B.

2.1. Upper Bounds.

Since the task is to find the longest surprising sequence given a symbol set, it is useful to calculate upper bounds for the length of the longest surprising sequence.

2.1.1. *Globally Surprising Sequences.*

It turns out that, given a symbol set of size s, the longest surprising sequence can be at most $3s - 1$ long.

Consider a longest surprising sequence for a given symbol set size s, Z. Let x_i be the number of symbols that occur i times in the sequence. Then, the length of of the longest surprising sequence is:

$$l = \sum_{i=1}^{s} i x_i \tag{1}$$

Since every symbol in the symbol set is trivially a part of a longest surprising sequence for that symbol set, the symbol set size can be expressed as:

$$s = \sum_{i=1}^{s} x_i \tag{2}$$

In order to show that $l < 3s$, it must be shown that:

$$\sum_{i=1}^{s} i x_i < \sum_{i=1}^{s} 3 x_i \tag{3}$$

, or simplified, that:

$$\sum_{i=4}^{s} (i-3) x_i < 2x_1 + x_2 \tag{4}$$

Since the sequence is of a finite length l, each symbol that occurs i times uses $\binom{i}{2}$ of the $l-1$ different distances between the symbols of the sequence. This means that:

$$\sum_{i=2}^{s} \binom{i}{2} x_i < l \tag{5}$$

, or simplified, that:

$$\sum_{i=4}^{s} (\binom{i}{2} - i) x_i < a + b \tag{6}$$

Since $(\binom{i}{2} - i) < 2(k-3)$, substituting gives:

$$\sum_{i=4}^{s} 2(i-3) x_i < x_1 + x_2 \tag{7}$$

, which, when simplified, shows that:

$$\sum_{i=4}^{s} (i-3) x_i < \frac{x_1 + x_2}{2} \tag{8}$$

Because $\frac{x_1+x_2}{2} < 2x_1 + x_2$, it follows that:

$$\sum_{i=4}^{s} (i-3) x_i < 2x_1 - x_2 \tag{9}$$

, which is what was to be shown.

[1] K. Downing, Searching for *Surprising Sequences* with a Genetic Algorithm.

2.2. **Genetic Coding.** Initially, this problem was attempted solved using a bit vector genotype. Two different coding strategies were used.

The first involved giving sequences of length l an implicit ordering, and interpreting the bit vector as a single number indicating which of the sequences it represents. This coding allowed for a continuous usage of the genotype space, but had two flaws. The first involved cases where the total number of possible sequences of length l was not a power of 2. In such cases, up to half of all possible genotypes would represent invalid sequences, which is counter-productive for a genetic solver. The other problem is that there was little correlation between mutations and fitness scores, which created a scenario in which exploration was too heavily prioritized over exploitation, yielding poor results.

The second bit vector coding approach involved grouping n bits into an integer representing one of the symbols from the set. This solved the problem of correlation between mutation and fitness scores, but still would, in the worst case, generate a lot of invalid sequences.

Finally, the bit vector genotype was replaced with a symbol vector genotype. This genotype is like the bit vector, but allows for s values per component, rather than just 2. This symbol vector genotype did not have any of the problems encountered with the initial bit vector genotypes.

2.3. **Phenotype.** The phenotype developed from the genotype discussed in the section above is an ordered list of integers representing the different symbols of the symbol set. The translation process is quite straight-forward, mapping the symbol vector of the genotype one-to-one to the ordered integer list that is the phenotype.

2.4. **Fitness.** For both the locally and globally surprising sequences' fitness calculation, the basic algorithm is the same. Every single AX_dB subsequence is considered for a given phenotype. For each occurrence of a AX_dB sequence beyond the first occurrence, a penalty is counted. The final fitness score is then given as $fitness = 1/(1 + total_penalties)$, which has the convenient property of being normalized.

2.5. **Mutation.** As described in the report presenting the algorithm framework, mutation can be quite problem-specific.

The mutation scheme for locally surprising sequences is quite simple: a symbol in the vector is selected at random, and substituted with a random symbol from the symbol set.

The mutation scheme for globally surprising sequences is a little more elaborate. It considers the number of occurrences of the symbols already present in the sequence. The replacement of a symbol is still random, but it is weighted so that symbols that are underrepresented in the string are prioritized. This is based on the intuition that a long surprising sequence is more likely to have a relatively even distribution of symbols, rather than a lot of occurrences of a single symbol.

3. Comparison of problems

The task also specifies that a comparison of the relative difficulties of three problems should be made. The three problems are One-Max, as elaborated upon in the report presenting the EA framework, finding the longest globally surprising sequence for a symbol set of size 9, and finding the longest locally surprising sequence for a symbol set of size 9.

Using the average number of generations before an acceptable solution as an indicator of difficulty, it is easy to see that One-Max is by far the easiest for the EA implementation to solve. That is because unlike most problems that are typically solved by EAs, it is easy to see when the optimal answer is reached.

That is not the case with finding the longest surprising sequence. Both for globally and locally surprising sequences, there is no simple way of determining if a found surprising sequence is in fact the longest one. As such, there is no good way of knowing when to stop the search. Because of this, finding locally surprising and globally surprising sequences can be considered to be equally hard.

4. Results

Table 2 and table 1 show the longest surprising sequences for symbol sets of size between 3 and 20, including. All sequences were found with the solver using full generational adult replacement, tournament selection with $k = 8$, $P_{crossover} = 1$ and per-genome $P_{mutation} = 0.001$. The population size was 200 for every run.

Symbol set size	Generations	Length	Sequence
3	0	7	(3 1 2 1 1 3 2)
4	7	10	(1 2 3 4 3 1 4 2 2 1)
5	0	12	(4 5 3 3 1 2 5 1 4 2 4 3)
6	70	15	(5 3 2 6 4 1 1 2 4 5 4 3 5 1 6)
7	2608	18	(3 2 1 5 4 4 3 1 6 7 6 2 4 7 2 5 1 3)
8	670	20	(7 3 4 7 5 6 4 1 3 1 2 8 8 7 4 5 2 3 7 6)
9	2621	23	(8 4 7 2 6 8 3 6 5 4 3 1 9 9 2 7 6 7 1 4 2 8 5)
10	4537	26	(9 10 4 3 8 7 10 5 1 5 2 9 6 6 7 9 2 8 1 3 4 5 7 4 8 10)
11	111	28	(5 9 8 1 6 10 7 6 4 2 11 8 3 3 4 6 1 3 9 5 11 5 7 1 2 8 4 10)
12	11182	31	(1 12 3 9 9 8 6 2 10 3 2 5 6 4 11 7 12 1 2 11 12 5 10 8 7 8 4 9 6 1 3)
13	10316	33	(3 4 2 8 12 1 11 11 6 13 4 5 10 4 9 6 9 7 10 8 2 3 1 13 12 3 5 8 13 11 7 2 1)
14	14916	36	(12 14 8 7 2 1 3 9 6 11 4 12 1 13 8 10 3 5 13 4 3 6 9 7 9 5 7 6 2 2 8 14 12 10 1 11)
15	9579	39	(3 10 4 13 7 9 5 1 11 2 14 9 15 6 1 12 7 4 8 12 2 8 6 10 14 3 3 11 15 5 15 13 12 10 1 9 13 5 4)
16	12049	41	(9 12 15 3 1 7 2 5 16 4 6 8 1 10 13 11 13 8 15 5 8 3 4 14 14 2 3 10 16 15 11 9 6 7 10 12 5 1 16 12 13)
17	2600	43	(15 7 2 14 1 2 15 6 4 8 12 9 6 5 7 3 16 16 4 9 10 13 17 8 14 12 11 1 13 2 11 10 3 5 3 8 1 15 17 7 6 9 14)
18	96	45	(14 15 16 1 4 12 10 9 10 17 13 11 7 18 9 16 15 3 14 16 17 5 8 8 6 11 4 1 7 4 13 5 17 6 2 1 3 10 12 14 2 18 15 12 9)
19	21602	49	(5 13 17 2 10 7 18 7 15 12 2 4 9 3 4 16 9 18 6 11 14 13 8 8 17 14 1 12 19 3 13 9 5 11 19 6 1 10 16 12 5 18 4 7 2 17 11 15 8)
20	70948	52	(8 16 9 19 18 6 12 14 4 14 20 1 5 3 13 10 15 8 4 2 5 19 1 11 20 2 7 17 9 16 7 10 11 5 17 15 13 6 6 18 15 16 18 12 8 1 10 4 9 3 14 19)

TABLE 1. Longest globally surprising sequences for symbol sets of sizes between 3 and 20, inclusive, as found by the EA solver. The population size in every case was 200.

Symbol set size	Generations	Length	Sequence
3	1	10	(3 3 2 2 1 1 2 3 1 3)
4	14	17	(2 2 3 4 4 1 4 2 1 3 3 2 4 3 1 1 2)
5	24	26	(4 1 2 1 1 5 4 5 2 4 3 5 3 1 3 4 4 2 2 3 3 2 5 5 1 4)
6	491	37	(3 1 4 6 2 5 2 6 6 4 4 1 6 3 6 1 3 5 5 4 5 1 1 5 6 5 3 2 1 2 4 3 3 4 2 2 3)
7	33	50	(1 3 4 3 3 5 4 1 4 7 5 5 2 1 5 6 6 3 6 2 7 4 6 4 4 2 4 5 1 6 5 3 7 7 6 1 2 6 7 3 2 5 7 2 2 3 1 7 1 1)
8	755	64	(3 1 5 2 1 7 6 4 5 3 6 5 8 1 4 2 7 5 7 1 6 6 2 4 8 5 5 1 2 8 2 5 6 1 1 3 7 8 3 5 4 7 3 3 8 7 7 4 3 4 1 8 4 6 3 2 6 8 8 6 7 2 2 3)
9	160	82	(5 8 5 4 8 2 9 4 3 8 3 7 4 1 7 7 9 7 5 9 2 4 2 1 2 5 7 8 6 4 7 2 6 6 9 5 3 1 4 4 5 2 3 6 5 6 3 9 6 7 6 8 9 9 8 1 8 7 1 1 3 2 2 7 3 3 4 6 2 8 8 4 9 1 5 5 1 6 1 9 3 5)
10	198	100	(10 4 5 5 6 7 8 8 1 3 2 7 9 2 8 9 5 2 3 6 5 4 9 6 4 6 2 5 1 6 10 9 1 2 10 2 2 9 9 8 4 3 5 7 7 1 7 3 10 1 10 5 9 3 8 5 3 9 10 6 1 8 7 6 8 2 4 2 6 6 3 1 5 8 6 9 4 8 10 3 3 4 7 10 10 7 2 1 1 9 7 4 4 1 4 10 8 3 7 5)
11	169	110	(9 11 1 4 2 10 10 11 11 7 6 5 8 8 7 9 5 1 5 6 8 6 10 8 5 4 3 9 4 1 3 8 11 10 9 3 3 1 8 10 1 6 1 11 4 4 6 7 3 7 10 2 2 7 4 10 7 5 11 3 5 2 1 9 8 9 6 4 5 9 10 6 9 2 9 9 7 11 8 3 11 2 3 10 5 10 3 6 2 5 7 7 2 11 9 1 10 4 8 4 11 6 3 4 7 8 2 8 1 1)
12	45	117	(3 10 3 8 10 7 2 12 2 11 4 9 3 7 12 11 1 1 10 9 12 5 4 12 9 2 6 9 7 3 4 6 8 1 7 10 8 8 3 3 9 9 6 1 2 5 2 2 1 4 4 11 8 12 10 5 5 1 12 12 1 9 4 2 10 4 1 6 11 9 1 5 8 7 8 6 12 8 9 8 4 8 11 7 4 5 11 11 12 6 7 5 6 5 7 7 11 2 4 10 2 9 5 10 11 10 12 3 5 3 11 5 9 10 10 6 10)
13	517	165	(8 11 2 2 9 11 5 6 13 10 3 9 9 1 1 5 12 3 7 5 13 13 1 8 2 5 8 8 12 6 6 3 4 9 2 3 2 10 7 12 11 3 8 3 3 11 12 5 7 13 7 10 13 9 6 4 10 12 9 7 1 7 7 4 11 13 11 8 4 1 3 6 8 6 11 11 7 11 10 10 8 10 9 3 1 2 1 12 7 8 9 10 4 8 1 6 2 13 6 12 13 12 10 11 6 1 4 4 13 5 3 5 4 7 6 10 1 11 1 10 6 5 11 9 5 10 5 1 9 13 4 5 2 11 4 2 4 3 12 4 6 9 4 12 8 13 2 6 7 3 10 2 7 2 8 5 5 9 8 7 9 12 12 2 12)
14	4776	194	(3 12 1 11 11 4 13 6 9 1 7 9 11 10 8 12 3 3 4 12 11 8 6 13 13 3 1 5 1 13 12 12 8 2 1 12 2 6 2 7 10 2 2 9 14 7 13 4 6 5 2 11 2 5 3 8 13 14 5 8 8 3 10 7 6 3 14 6 7 2 4 14 4 10 11 12 7 3 7 11 3 6 14 14 1 1 3 9 5 14 9 10 5 5 9 6 6 10 6 11 5 12 9 13 2 13 9 3 13 1 8 1 9 4 8 14 10 9 2 3 11 7 1 10 1 4 1 14 11 13 5 6 1 6 12 5 13 7 12 10 12 4 9 8 5 10 3 2 12 13 10 14 13 8 4 3 5 4 7 7 14 8 9 9 12 6 8 7 4 2 14 12 14 2 10 10 13 11 6 4 11 1 2 8 10 4 4 5 11 9 7 8 11 14)
15	1855	219	(13 11 3 10 3 2 4 12 9 1 11 15 2 10 8 9 13 5 5 2 7 15 4 5 9 5 13 10 10 12 1 15 12 14 3 8 3 14 12 8 15 6 1 13 12 7 4 15 1 2 3 3 13 6 12 12 5 15 13 15 10 1 14 9 9 11 8 6 5 1 9 2 9 3 9 14 8 2 5 14 6 14 14 13 1 5 3 1 1 6 11 5 10 5 6 10 6 9 10 11 14 10 4 6 7 9 15 8 14 11 2 13 4 10 2 2 6 8 5 11 4 9 4 14 7 7 2 1 8 1 10 7 8 7 11 7 5 12 4 13 7 10 14 2 12 2 11 6 2 8 11 9 6 4 1 12 15 5 4 3 5 8 10 13 14 5 7 6 13 3 15 3 4 2 15 9 8 12 10 15 15 7 13 9 7 3 7 12 3 6 6 3 11 11 10 9 12 6 15 11 1 4 4 7 14 1 3 12 11 12 13 8 8 13 13 2 14 4 8)
16	1952	245	(9 8 1 11 7 14 10 4 10 16 7 15 1 13 3 14 6 10 14 14 8 6 1 2 1 14 7 5 15 14 2 12 15 10 7 3 8 7 4 7 13 2 4 8 4 14 9 15 12 8 8 10 1 15 6 16 13 13 14 4 13 1 5 8 9 10 11 3 4 11 4 16 16 4 5 1 12 3 6 5 14 11 13 12 5 13 7 12 10 12 4 6 9 16 6 4 3 15 13 5 12 14 5 3 3 9 4 2 16 5 16 1 1 7 1 8 16 8 2 3 10 15 4 15 5 4 1 4 12 16 2 10 6 8 13 16 14 12 13 6 12 12 11 8 11 14 15 8 15 11 11 9 14 13 11 10 13 10 9 9 7 11 6 3 2 15 9 6 6 11 2 8 12 9 13 8 3 12 7 16 10 8 14 1 3 13 15 16 12 2 14 16 3 7 8 5 7 9 2 7 7 2 9 12 1 9 5 10 5 9 3 11 12 6 2 2 11 5 5 11 16 11 1 16 15 3 5 2 5 6 7 6 15 15 7 10 2 6 13 9 1 10 10 3 16)
17	241	261	(7 15 8 1 2 10 11 5 17 5 7 14 16 3 13 9 16 2 3 5 14 8 12 12 7 13 1 16 4 11 10 13 17 10 14 1 4 4 14 12 3 15 14 11 17 16 6 2 6 10 15 2 1 8 9 5 1 7 1 15 1 11 7 4 6 7 17 17 15 10 16 9 10 10 5 12 5 11 11 1 3 2 5 8 7 6 14 7 2 2 12 14 13 5 13 11 8 13 16 11 15 7 9 6 3 8 10 2 8 17 9 13 8 8 14 5 6 6 11 13 7 5 3 10 9 14 15 16 10 8 3 16 12 9 7 11 2 11 9 1 13 2 17 2 15 12 8 4 8 15 15 17 3 11 4 16 7 10 6 17 1 1 6 13 14 6 15 5 15 6 4 7 12 2 9 11 12 11 14 2 13 6 1 17 12 1 14 17 6 9 2 14 3 3 14 9 12 15 11 6 16 5 4 10 12 13 15 9 9 15 13 10 4 17 14 10 17 13 13 4 5 5 2 4 1 10 3 17 4 9 4 13 3 12 16 14 4 3 1 5 16 17 11 3 6 12 10 7 8 5 9 8 2 16 1 9 3 7 16 8 11)
18	311	290	(12 9 2 4 2 8 3 17 18 13 4 13 17 7 9 7 14 3 11 4 9 3 6 3 9 4 4 3 8 15 7 3 14 13 10 4 7 2 11 3 4 14 7 17 14 16 3 16 18 4 16 13 1 6 2 13 6 8 11 10 11 6 17 9 12 2 15 16 10 12 14 6 14 10 14 11 16 7 4 1 12 10 9 16 15 15 17 10 3 10 5 9 11 5 17 15 3 18 6 4 17 13 9 5 8 6 12 13 5 11 14 12 8 5 1 17 12 15 2 7 1 14 8 9 10 16 14 2 9 13 2 18 5 13 8 7 16 8 17 4 15 5 5 18 9 15 10 10 15 1 16 4 18 18 10 2 16 1 2 2 12 6 7 8 2 6 18 15 8 12 7 13 14 14 5 15 6 9 17 1 4 10 17 8 1 1 13 15 9 9 6 11 13 13 16 9 1 7 7 5 6 5 16 12 1 8 10 1 10 18 16 5 7 11 12 3 3 7 15 4 11 18 1 18 3 1 3 5 10 8 16 17 16 2 1 15 12 16 11 9 8 8 18 17 11 2 5 14 4 12 5 12 11 1 9 18 7 10 13 3 2 3 12 4 8 13 12 17 6 16 16 6 13 11 7 6 15 14 18 8 4 5 3 13 7 18 2 10 6 10)
19	630	324	(10 4 16 4 8 19 2 2 4 5 10 6 12 18 14 15 6 11 14 8 7 13 18 4 3 12 14 5 8 6 14 9 13 16 7 3 4 2 19 11 6 3 17 14 13 15 1 15 19 5 11 2 1 3 14 3 8 5 9 8 13 6 15 12 1 12 3 19 6 16 10 14 18 18 15 14 14 7 18 1 5 2 13 4 11 4 14 2 3 9 2 8 8 15 4 7 8 18 10 13 1 6 2 16 8 14 16 17 18 17 15 15 3 5 5 7 4 19 18 8 3 15 11 8 12 5 6 4 10 15 5 14 12 12 7 6 5 3 11 13 12 16 11 9 7 14 17 1 2 11 10 11 7 9 5 1 7 17 17 12 10 10 18 6 10 19 9 12 9 9 18 5 4 4 17 3 18 13 8 16 19 15 8 11 12 2 15 17 11 15 7 10 5 19 7 12 8 1 1 18 12 6 7 1 11 1 14 10 16 5 17 13 19 19 1 9 11 3 10 12 11 16 14 6 1 16 12 15 16 1 17 5 15 18 11 18 16 16 2 5 18 2 9 16 18 19 17 19 8 10 3 3 2 18 7 2 10 7 5 13 13 11 19 4 13 7 11 5 12 17 10 17 4 1 19 14 4 12 13 9 10 9 3 7 15 9 14 1 8 2 12 19 16 6 6 13 3 16 9 17 8 9 1 13 17 6 17 7 16 13 10 8 4 6 8 17 16 3 6 18 3 13 2 17)
20	14809	401	(9 2 17 7 6 20 16 3 12 7 12 20 9 12 13 8 9 1 7 19 17 18 9 5 3 16 5 1 2 12 11 4 18 11 19 12 1 16 1 12 5 20 20 11 5 19 16 8 1 6 1 8 17 17 14 4 12 10 3 3 9 10 16 2 7 3 10 10 8 11 18 12 6 18 7 2 16 6 9 13 13 19 7 7 16 18 3 2 18 16 16 9 8 7 9 20 10 1 15 19 19 10 9 4 5 7 4 8 20 18 5 18 18 15 10 19 2 19 3 6 12 9 6 16 15 5 8 12 2 4 10 7 8 13 15 4 13 6 4 20 1 11 13 18 6 14 20 12 16 19 14 9 9 17 5 6 17 11 16 7 18 1 20 6 6 2 3 11 17 9 16 12 19 6 19 5 16 11 1 18 10 2 13 5 5 2 1 5 11 11 10 11 6 13 16 13 14 15 11 2 2 15 7 20 4 14 6 11 8 10 18 14 19 13 12 18 4 17 6 5 14 17 3 5 4 7 10 20 5 13 1 10 5 9 11 14 7 11 12 3 4 6 3 13 7 15 15 9 18 8 18 17 4 2 14 2 11 7 14 14 1 9 19 11 3 20 17 1 4 16 10 4 15 14 8 14 18 19 15 8 3 19 20 2 8 19 9 15 6 8 15 18 2 6 7 13 2 5 12 15 3 1 13 9 14 12 12 17 10 15 1 1 19 18 13 3 7 1 14 5 15 16 14 3 18 20 3 8 2 10 6 15 17 20 7 17 12 4 19 1 3 14 11 20 8 16 4 3 15 20 14 16 20 19 4 11 15 13 20 13 4 1 17 15 2 20 15 12 14 13 11 9 7 5 17 8 4 4 9 3 17 16 17 13 10 12 8 6 10 13 17 19 8 8 5 10 14 10 17 2 9)

TABLE 2. Longest locally surprising sequences for symbol sets of sizes between 3 and 20, inclusive, as found by the EA solver. The population size in every case was 200.