Homework Module:

Searching for *Surprising Sequences* with a Genetic Algorithm

Purpose: Apply genetic algorithms to problems, such as surprising sequences, in which the common genotypic representations exhibit epistasis: interactions among genes.

1 Background

The well-known puzzle writer, Dennis Shasha, defines *surprising sequences* as those that are completely free of repeating patterns. Defined 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.

Note that **order matters** in assessing patterns, so AX_dB is not equal to BX_dA. Also, the criteria include the case where A=B.

As a simple example, the sequence ABCCBA is surprising, but AABCC is not, since AX_2C occurs twice. Similarly, ABBACCA is not surprising due to 2 occurrences of AX_2A.

For this assignment, we will consider two types of surprising sequences:

1. **Globally** surprising sequences are those defined above: all values of distance d must be considered.
2. **Locally** surprising sequences are those in which there are no repeat occurrences of AB (i.e. AX_0B) for any symbols A and B. That is, the only distance of relevance is d = 0.

Thus, AABCC is not globally surprising, but it is locally surprising. The same goes for ABBACCA. A sequence that is not locally surprising is ABCBC due to the repeat of BC.

2 Task

You will use the EA designed for the earlier half of this assignment to search for the longest possible surprising sequences that can be constructed from symbol sets of different lengths. For instance, when their are only 2 symbols, then the largest globally surprising sequence is ABBA (or BAAB or AABA, etc.), while the largest locally-surprising sequence is ABBAA (or BAABB or AABBA, etc.).

You will create tables containing information about the sequences that your EA finds, along with the key EA parameters. They must be of the following format:

Size of symbol set	Population size	Generations	Sequence Length	Sequence
3	50	2	7	(1 2 3 2 2 1 3)
4	50	10	9	(4 3 1 1 2 1 3 4 2)
:	:	:	:	:
20	200	:	:	:

Note that here we use numbers instead of letters in the sequences. Either option is fine for this assignment.

For each size (S) of symbol set, do your best to find what you think is the absolute longest sequence (though this is hard to prove). In short, try several runs of the EA for each S and various lengths (L) in the attempt to find long sequences. Hint: Don't expect to find extremely long sequences, especially for the global-surprise cases.

Just to give you an idea of the size range: for S = 10, a globally-surprising sequence of L = 26 can be formed. However, there is no guarantee that your EA can find it. Just do the best you can and clearly document each of the longest sequences found by your EA in your 2 tables (one for locally- and the other for globally-surprising sequences). For a given S and L, it is wise to run your EA 5 or 10 times to see if it can find a solution of length L.

If you don't feel like testing a lot of different L values (for a given value of S), you can let your EA explore different-length sequences as well. You will have to figure out the details yourself, but here is one hint:

> You can include *dummy* alleles in a chromosome of length L. Any gene with a dummy value/allele is ignored in the phenotype, so the chromosome actually encodes sequences of any length, from 0 to L. For example, if 0 is a dummy allele, then the sequence of alleles 1202103 actually represents the sequence 12213.

Also, for each table, include general parameter settings for your EA, including:

- mutation rate
- crossover rate
- adult selection strategy
- parent selection mechanism

If these vary for different runs, then include the different parameter lists and the cases on which they were used.

3 Deliverables

1. Draw a diagram of the genetic encoding used for this exercise: genotypes, phenotypes, translation process. Include a very brief description. (**1 point**)
2. Describe (with text and a little mathematics) the fitness functions used to evaluate globally and locally-surprising sequences. (**1 point**)
3. Create a single table showing the best locally-surprising sequences that your EA finds for 18 different sizes: S=3,4,5...20.(**3 points**)
4. Create a single table showing the best globally-surprising sequences that your EA finds for 18 different sizes: S=3,4,5...20.(**3 points**)
5. Given a 20-gene chromosome and these 3 tasks (1-max, global surprising sequence (S=9), local surprising sequence (S=9)), rank them in terms of their degree of difficulty (for an EA). Explain your decision (**2 points**)

Your complete report for this homework module should be NO LONGER than TWO pages. Longer reports can incur point loss. It is important to write clearly and concisely.

4 The Harsh Reality of EAs

In general, the programming of EA solutions to hard problems involves two significant phases:

1. Coding the genome representation, genetic operators, genotype-phenotype mapping, fitness function, etc.
2. Tuning the parameters of the system to actually solve the problem.

The second phase, tuning, often requires **much** more time than the first phase. Be aware of this when scheduling time for this (and other) EA assignments. Just getting the system to run, bug-free, is normally half (or less) of the total work.

5 Warning

Depending upon the actual course and semester - your instructor uses this module in various courses - you may or may not be required to do any or all of the following:

1. Demonstrate this module to the instructor or a teaching assistant.
2. Upload the report and/or code for this module to a particular site such as *It's Learning.*

Consult your course web pages for the requirements that apply.