

# Optinformatics for Schema Analysis of Binary Genetic Algorithms

Minh Nghia Le  
School of Computer  
Engineering  
Nanyang Technological  
University  
Singapore  
lemi0005@ntu.edu.sg

Yew Soon Ong  
School of Computer  
Engineering  
Nanyang Technological  
University  
Singapore  
asysong@ntu.edu.sg

Quang Huy Nguyen  
School of Computer  
Engineering  
Nanyang Technological  
University  
Singapore  
nguy0046@ntu.edu.sg

## ABSTRACT

Given the importance of *optimization* and *informatics* which are the two broad fields of research, we present an instance of *Optinformatics* which denotes *the specialization of informatics for the processing of data generated in optimization so as to extract possibly implicit and potentially useful information and knowledge*. In particular, evolutionary computation does not have to be entirely a black-box approach that generates only the global optimal or good quality solutions. How the solutions are obtained in evolutionary search may be brought to light through Optinformatics. In this paper, we present a Frequent Schemas Analysis (FSA) technique for extracting knowledge from the search process by using the historical optimization data, which are otherwise often discarded. FSA bring about greater understanding of GA dynamics through mining for frequent schemas that exists implicitly within the optimization data via the design of frequent pattern techniques (LoFIA) in informatics. To illustrate the principle of optinformatics, a case study using the Royal Road problem is used to explain the search performance of Genetic Algorithm (GA) in action.

**Categories and Subject Descriptors:** I.2.m [Artificial Intelligence]: Miscellaneous

**General Terms:** Algorithms, Performance, Experimentation

**Keywords:** Genetic Algorithms, Frequent Pattern Mining, Schema Theory, Royal Road problem

## 1. DEFINITION OF FREQUENT SCHEMA

Let function  $Freq(s, t)$  define the frequency of the schema  $s$  in the population at generation  $t$  and  $Freq(s, [m, n])$  denote the frequency of schema  $s$  in the populations over generations  $m$  to  $n$ .

$$Freq(s, [m, n]) = \frac{\sum_{t=m}^n Freq(s, t)}{(n - m)} \quad (1)$$

We define a schema  $s$  as **frequent schema** with a level  $\theta$  in the period  $[m, n]$  if and only if  $Freq(s, [m, n]) \geq \theta$ . One possible interpretation of a frequent schema  $s$  is that GA has spent at least  $\theta$  percentage of its sampling budget on the hyperplane defined by  $s$ ; or  $\theta$  is a lower bound of the probability that a point in the hyperplane  $s$  is sampled by GA during the period  $[m, n]$ .

As stated in Holland's book [1], "..., if some schema begins to occupy a large fraction of the population (through consistent above-average performance), its rate of increase will come very close to

$[\mu_\xi(t)/\mu(t)] - 1$ ", it is expected that the frequencies  $Freq(s, t)$  of a schema with consistent above-average performance in a period will form a non-decreasing sequence and the set of consistently above average schemas would likely contribute to the set of frequent schemas.

## 2. FREQUENT SCHEMAS ANALYSIS

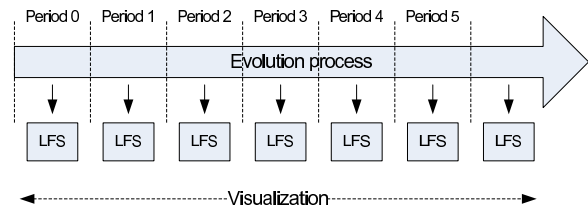


Figure 1: Frequent Schemas Analysis

In our technique of frequent schemas analysis (FSA) as shown in Figure 1, data which is collected from the evolution process is divided into consecutive and non-overlapping periods. The sampling of GA in each period of the search space is analyzed by investigating on the set of frequent schemas ( $Freq(s, P) \geq \theta$ ) found in that period. Alternatively, frequent schemas can also be compared across periods to understand the change in GA dynamics. Large value of  $\theta$  gives more confidence on the located convergence regions but the frequent schemas are generally less specific (lower order schemas), thus, interesting information may be not captured.

Each chromosome (binary string) in the data generated by GA in the period is first transformed to a set of items, so as to allow a two-way transformation from a chromosome or schema to an item-set and vice versa. From the possibly numerous frequent schemas, it is up to the analyzer to select *interesting* schemas from the pool to investigate. In this paper, the *interestingness* metric is defined as the longest frequent schema (LFS) which provides a sketch on how GA progressively reduces the number of dimensions of its search space or biases its search towards promising regions. Most specific frequent schemas are then found using our LoFIA algorithm which employs bottom-up and depth-first approach to quickly identify the *longest* frequent schemas from the optimization data. A visualization method is also introduced to capture the change of the schemas across the periods of evolution. Scalar vector  $x$  of length  $L$  represents the set of  $M$  most specific frequent schemas. The value of element  $x_i$  for loci  $i$  is then calculated by  $x_i = \frac{N_1 - N_0}{M}$  where  $N_1$  and  $N_0$  are the number of schemas in the set has value 1 and 0, respectively, at loci  $i$ . Vectors  $x$  of consecutive periods are plotted

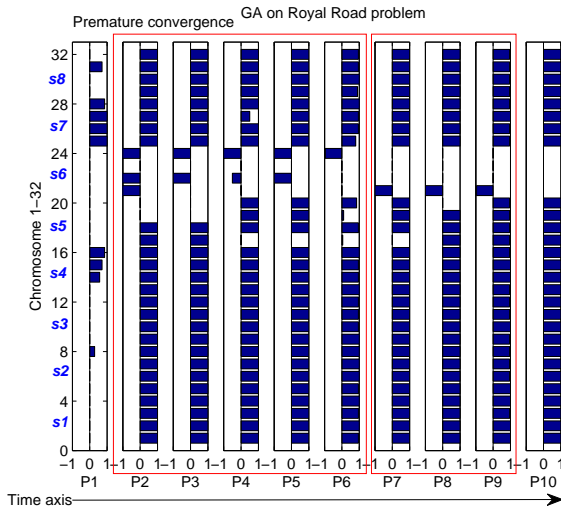
Methods	Royal Road (32K4)	Royal Road (64K8)
GA	7587.32 ± 7045.26	102880.96 ± 71723.45
RMHC	412.22 ± 206.61	5876.86 ± 2595.55

**Table 1: Hill-climbing outperforms GA on Royal Road problem**

against the time axis in the final visualization. Through this visualization, the plot of one period displays the current convergence regions of GA and the differences observed across periods serves to provide hints to the dynamics of GA.

### 3. FREQUENT SCHEMA ANALYSIS OF GA ON ROYAL ROAD PROBLEM

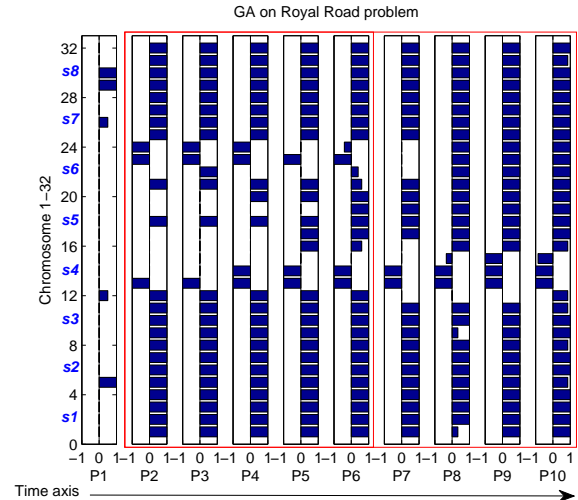
It is worth noting that Random Mutation Hill Climbing (RMHC) outperforms GA on the Royal Road problem. Table 1 shows the average number of evaluations incurred by each algorithm in reaching the optimal solution on the problem of 32 bits (block size  $K = 4$ ) and 64 bits (block size  $K = 8$ ) over 50 independent runs. Our configuration of GA is one-point crossover  $p_{cross} = 0.8$ , bit-flip mutation  $p_{mut} = 0.003$  and fitness-proportional selection with  $popsiz = 50$ .



**Figure 2: Frequent schemas analysis of GA at  $\theta = 0.8$**

To investigate what slowed down GA on the Royal Road problem, Frequent Schemas Analysis (FSA) was used to analyze the archived optimization data of a GA run on the problem (32 bits,  $K = 4$ ). Here, the evolutionary search is divided into 10 periods, with each period consisting of 15 GA search generations. The most specific frequent schemas of each period at  $\theta = 0.8$  were then obtained and the results are then plotted against time in Figure 2. Firstly, the plot well illustrates the Building Block Hypothesis where blocks are shown to have been discovered and combined in reducing the complexity of the problem. As expected, the length of longest frequent schemas increases as the search progresses. In Figure 2, note that block  $s_6$  was incorrectly identified in Period 2, containing three 0 bits ( $x_{21,22,24} = -1$ ). While other correct blocks of 1's were quickly found in early periods of the evolutionary process, GA took approximately 8 periods  $P_2$ - $P_9$  (120 generations) to correctly identify the good configuration of block  $s_6$  and were able to locate the global optimum afterward. Here, the block with incorrect alleles which hitchhikes in the previous gener-

ations took a long time to be corrected, thus highlighting a possible premature convergence of GA on the Royal Road problem. The supposition was confirmed when FSA was used to investigate the situation in which GA could not find the global optimum within the limited time (150 generations), as shown in Figure 3. In Figure 3, block  $s_6$  took 5 periods to be repaired while blocks  $s_4$  remained incorrect till the end of the search.



**Figure 3: Frequent schemas analysis of GA at  $\theta = 0.8$**

Here, the regions of many blocks of 1's with some 0's incorrectly identified at one block are considered as convergence regions of GA on the Royal Road landscape. When a population is in the region with high probability, mutating other bits of 1 will decrease the fitness of an individual significantly due to the loss of correct building blocks. Therefore, the mutated individual will cease to appear in the reproduction pool, thus also the next generation. Since one-point crossover working on the reproduction pool is also unlikely to be helpful in this case, an event of improvement which is defined as when an individual with *less* number of 0's appearing in the reproduction pool of the subsequent population can only be achieved but at a small probability by mutating bit  $0 \rightarrow 1$  while not changing other bits of 1.

### 4. CONCLUSIONS

In this paper, a Frequent Schemas Analysis (FSA) technique, which takes its roots from informatics, is introduced for analyzing GA dynamics through mining of frequent schemas that exists implicitly within archived optimization data. In particular, FSA is used to mine for interesting frequent schemas from Binary GA data that is often discarded and investigate the schemas of different search periods visually. FSA provides a comprehensive picture of how the search process evolves, hence bringing new insights into the properties of GA on different problem landscapes. Using the Royal Road problem, we demonstrated the ability of FSA in identifying the premature convergence of GA search which is also confirmed by previous studies. Note that FSA represents an instance of Optinformatics which aspires to make the evolutionary search more transparent instead of being an entirely black-box approach that serve only to provide good quality solutions.

### 5. REFERENCES

[1] J. Holland. *Adaptation in natural and artificial systems*. MIT Press Cambridge, MA, USA, 1992.