# A Practical Search Index and Population Size Analysis Based on the Building Block Hypothesis

Zhenhua Li
School of Computer Science
China Univ. of Geosciences
Wuhan 430074, China
zhli@cug.edu.cn

Erik D. Goodman
Dept. of Electrical and Computer Engineering
2120 Engineering Building
Michigan State Univ., E. Lansing, MI 48824
goodman@egr.msu.edu

## ABSTRACT

Use of the Building Block Hypothesis to illuminate GA search behavior, as pursued by J. H. Holland and D. E. Goldberg, invites additional investigation. This paper re-examines the space actually searched by a GA, in light of the Building Block Hypothesis, GA sampling and population size, in an effort to develop more quantitative measures of GA difficulty for problems where building block sizes can be estimated. A Practical Search Index (PSI) is defined, related to the size of the space actively searched by the GA, in terms of sizes and numbers of building blocks. When BBs are hierarchical, the PSI can be used at various stages of BB assembly. Difficulty depends strongly on the sizes of the largest building blocks, rather than on the size of the entire search space, for GAs dominated by crossover. Premature convergence prevails when population size is not adequate to allow sampling and assembly of building blocks. Appropriate sizing depends on balancing the BB sampling and mixing costs. A set of simple GA experiments on classical test functions with clear building block structures (One-Max, RR1, RR2, RRJH, HIFF, etc.) at various population sizes, illustrates the relationship between the PSI, population size, and efficiency of search.

## Categories and Subject Descriptors

I.2.m.c [**ARTIFICIAL INTELLIGENCE**]: Miscellaneous: evolutionary computing and genetic algorithms

## General Terms

Algorithms, Performance

## Keywords

genetic algorithm, building blocks, search space, practical search index,building block sampling, population size, GA hardness

## 1. INTRODUCTION

John H. Holland declared that building blocks(BBs) are a ubiquitous feature at all levels of human understanding, from perception through science and innovation; and genetic algorithms are designed to exploit this prevalence[2].


*GECCO'08,* July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-130-9/08/07.

The Building Block Hypothesis, claimed to be supported by Holland's schema theorem, was formulated by David E. Goldberg in 1989, describing the abstract adaptive mechanism of the GA[1].

The past work has concentrated mainly on how to find and maintain the BBs, in order to speed GA search; however, apart from Holland's schema theorem and Building Block Hypothesis, few papers discuss BB theory or explain why we should track them — in other words, how to explain the GA adaptive mechanism from a BB viewpoint. In this paper, we will define a new measure or index on the GA search space, review GA sampling, and discuss sizing of populations based on the Building Block hypothesis.

## 2. SAMPLE SPACE, PRACTICAL SEARCH SPACE, AND PRACTICAL SEARCH INDEX

When we refer to the search space or sample space of a GA, we refer to the space of all possible samples. But we argue that any GA will not normally search the whole space. And indeed, for any biased search algorithm, only a small fraction of the sample space is normally searched (were that not true, enumeration would be superior to GA search for such problems). We will call that subset the Practical Search Space (PSS).

We cannot easily quantify the relationship of the PSS to the whole search space, given the GA parameters and problem characteristics. As a first step, we shall define a Practical Search Index (PSI), in order to seek insight into the fraction of the sample space actually visited in a typical GA search.

For those problems exhibiting a strong building block structure, or so-called decomposable problems, a Practical Search Index can be defined at the initial stage of GA search:

*Definition 1.* The *P*ractical Search Index (PSI) of a problem $P$ to be searched by a GA, $PSI_P$ is

$$\sum_{i=1}^{k} n^{l_i}$$

where k is the number of BBs, $l_i$ is the length of the i-th BB, and n is the number of alleles at each position (2, for example, for binary chromosomes).

Assessing the PSI for a Hierarchical Building Block Structure (HBBS) will be more complex than for the flat structure problems discussed in the last subsection. For HBBS prob-

lems, however, taking a BB as an element, we can calculate the PSI of some HBBS problems.

*Definition 2.* After the lower BBs are formed,the GA search space $PSI_{BB}$ is

$$\sum_{i=1}^{k} A_i$$

where $k$ is the number of BBs, and $A_i$ is the number of alleles of the i-th BB.

# 3. GA ADAPTIVE MECHANISM

Here, we argue (after Goldberg) that although the population size is much smaller than the whole search space, we will proceed toward optimality not by repeatedly sampling for the best individual, but by ferreting out BBs, for combining into the solution in later stages.

## 3.1 Population Size: BB Sampling vs. BB Mixing

For many crossover-dominated GAs, their search process could divided into two stages: the initial stage and the evolutionary stage. In the former phase, the BBs are sampled; in the latter, the BBs are mixed. BB sampling is easier when more individuals are available, so the BB can appear and have more instances present (which means more chances to persist to the later stages); while the BB mixing often works well with fewer individuals, since too many individuals, especially in later stages when the population is almost homogenous, result in many wasted mixing operations among similar individuals.

So, BB sampling and mixing are somewhat competitive in population sizing. When the population size is small (but not smaller than the BB sample space size), although the cost of mixing is less, the availability of BB instances is also less. The GA still needs a long time to collect them. When the population is larger, the number of BB instances goes up, however the evaluations in each generation also jumps, therefore making the simple GA more costly to run. An appropriate population size should balance the BB sampling and mixing costs.

This also suggests the well-known fact that a small population size and long evolution time is not equivalent to a large population size and a short evolution time. This tradeoff can be made only on a small scale near the balance point, where neither the sampling nor mixing are overwhelming each other and these two operations can compensate for each other to some extent.

# 4. EXPERIMENTS

### *HBBS and Evaluations.*

The result of RR1, RR2, and RRJH64, whose BB structures are the same at the bottom level but differ on upper levels (RR1 is flat, RR2 is fully hierarchical, and RRJH64 has a stronger hierarchical structure and weaker deception within a single BB),suggests that the HBBS have weak effects on the GA process when the population size is small, and gradually lose their impact as the population size increases. The lack of a striking difference in performance on RR1 and RR2 indicates that, once again, the Royal Road is "not taken", i.e., the additional reward for higher-level building blocks in RR2 does not speed its search at any of the population sizes tested.

### *Population Size.*

Five functions (One-Max, RR1, RR2, RRJH64, HIFF64)with the same string length are tested under different population sizes. The results suggests: 1) Smaller-BB-size problems demand smaller population sizes. 2)Population size should be at least big enough to sample the BBs. And we regard that in order to avoid the premature convergence, the population size should not be smaller than the size for sampling BBs, for crossover-dominated GAs. Introduction of deception in the BBs requires still larger sampling in the initial population for efficient solution.

The experiments indicate that the population size should be at least somewhat larger than the minimal size needed to contain all values for potential BBs, but not an order of magnitude larger, and not of a size that would probabilistically guarantee that all bit combinations within a BB would be present in an initial randomly generated population. The value of an initialization method that systematically generates all bit combinations below a certain size may allow a tightening of that bound, but was not tested in these experiments.

# 5. CONCLUSIONS

A Practical Search Index measure of problem difficulty for problems with strong building block structure has been defined. It can be calculated by summing the search space of each building block. According to this measure, large BB sizes, rather than long strings per se, account for the exponential increases in the space that must actually be sampled by a GA in solving a problem, and therefore make the problem hard.

Hierarchical Building Block structures contribute less to the PSI since once formed, lower-level Building Blocks can, under appropriate operators and population sizes, be treated as 1-bit units in the PSI measure. That may help to explain in part why the introduction of second-level fitness bonuses in hierarchically structured BB problems do not greatly speed problem solution (i.e., the "Royal Road" is not taken).

It is argued that a GA features Building Block sampling, rather than individual sampling, so the population size should not be smaller than what is needed to sample the space of the largest Building Block, for crossover-dominated GAs. Premature convergence is regarded as evidence that the population size was not large enough to discover all the BBs. At the same time, too large a population results in too many wasted evaluations in each generation, resulting in a high mixing cost. Therefore, the appropriate population size should balance the BB sampling and mixing costs.

# 6. REFERENCES


[1] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1989.

[2] J. H. Holland. Building blocks, cohort genetic algorithms, and hyperplane-defined functions. *Evolutionary Computation*, 8(4):373–391, 2000.