A Non-Linear Schema Theorem for Genetic Algorithms

William A. Greene
Computer Science Department
University of New Orleans
New Orleans, LA 70148
bill@cs.uno.edu
504-280-6755

Abstract

We generalize Holland's Schema Theorem to the setting that genes are arranged, not necessarily in a linear sequence, but as the nodes in a connected graph. We have experimental results showing that the flourishing of building blocks can be expected for two distinct graphs we have investigated, one being a tree and the other being the lattice points in a cube in Euclidean 3-space.

1 INTRODUCTION

Holland developed the ground work for genetic algorithms in the 1970's (Holland, 1975). His Schema Theorem says that building blocks should flourish. Its proof relies on the fact that genes in a chromosome are arranged in a linear sequence. But why limit ourselves to this arrangement of genes? Ultimately genes should be arranged in geometries most appropriate to the problem at hand, which may not be sequential. For example, for Koza's work (Koza, 1992) in genetic programming, the natural arrangement is tree-like. Other arrangements which might be appropriate to some problem are (a) a high-dimensional cube of short edge size, such as $\{0, 1, 2\}^{20}$, or (b) a web of points on the surface of a sphere.

In this paper we generalize Holland's Schema Theorem to the case that genes are arranged, not necessarily in a linear sequence, but as the nodes in a connected graph.

In this paper, a gene will simply be a bit.

2 HOLLAND'S SCHEMA THEOREM

First we carefully review the setting in which the theorem is proved. We follow the development found in (Goldberg, 1989). There is a universe U of individuals, each of which is a bit string (a list of bits) of a fixed length, call that length L. In set-theoretics, $U = 2^L$. There is a fitness function $f : U \to \Re^+$ which assigns to each individual a fitness value which is a positive real number. Also present is a population P which is a subset of the universe U. There will be generations of the population.

A schema is to be a pattern which can match individual bit strings. We use the asterisk, *, to be a wildcard character, and then a *schema* is a string of length L each of whose components is either a bit value 0 or 1, or the wildcard character. The schema positions holding a 0 or 1 are called *fixed* positions; those holding the wildcard are *unfixed*. The schema *matches* an individual bit string if the latter agrees identically with the schema at the schema's fixed positions. If an individual bit pattern matches a schema, we say that bit pattern is a *representative* of the schema.

If $L = 12$, an example of a schema is s = (*, *, 0, *, *, 1, 0, *, *, *, *, *). This schema has 3 fixed positions, at indices 3, 6, and 7. The *order* of a schema, *order*(s), is the number of fixed positions in it, 3 for this example. The *defining length*, $\delta(s)$, of a schema is the distance between its first and last fixed positions. For the above example, the first and last fixed positions are at indices 3 and 7 respectively, so the defining length is 7 - 3 = 4.

Now we return to the population P which is a subset of the universe. The population undergoes generational changes by subjecting it to the evolutionary forces of survival of the fittest, mating with crossover, and mutation, soon to be detailed. Recall the presence of the fitness function f. Implicit in a generational change is the hope that fitter and fitter individuals begin to appear in the population as time t increases. Holland talks of *building blocks*, meaning schemas of low defining length (so, also low order) whose representatives tend to have higher fitnesses. The supposition is that the fixed values within a building block are ones that are beneficial to the individual. Holland's Schema Theorem asserts that we can expect an increase, as time advances, in the number of representatives of the good building blocks within P.

Let the size of the population be N. Let the individuals in P be denoted $a_1, a_2, ..., a_N$, and let their associated fitness values be $f_1, f_2, ..., f_N$. Create the next generation population as follows. We use the *weighted roulette wheel* approach for choosing individuals for parenting: an individual a_i is selected for parenting with a probability equal to its relative fitness within the population, that is, with probability $f_i / (\Sigma_j f_j)$. Having selected two parents in this way, we perform mating with one-point crossover:

at random choose one of the L-1 positions 2 through length L. Cut each parent in two, between the chosen position and the preceding one, next interchange parental fragments to form two children, then add the children to the next generation. Continue in this way until the next generation reaches the same size as the preceding one. Note that two parents produce two children. Finally, with some low probability p_m, subject each bit in each child to a mutational change.

Our population at time t, $P(t)$, consists of individuals a_1, a_2, ..., a_N, with corresponding fitnesses $f_1, f_2, ..., f_N$. Let H be some given schema, and suppose within $P(t)$ there are m representatives of H, denote them $a_{i_1}, a_{i_2}, \ldots, a_{i_m}$.

Denote

$$\mu(H) = (\Sigma_k f_{i_k})/m$$

which is the average fitness of the H-representatives in $P(t)$, and denote

$$\mu(P) = (\Sigma_j f_j)/N$$

which is the average fitness in the entire population $P(t)$.

The probability that a particular H-representative, a_{i_k}, will be chosen for parenting is $f_{i_k}/(\Sigma_j f_j)$, therefore, the probability that some H-representative is chosen for parenting is $(\Sigma_k f_{i_k})/(\Sigma_j f_j)$. There are altogether N parents chosen, so the number of parents that are H-representatives has expected value $N \cdot (\Sigma_k f_{i_k})/(\Sigma_j f_j)$.

The latter number rewrites to $m \cdot \mu(H)/\mu(P)$.

Assuming a representative of H is chosen for mating, and recalling that a cutpoint for crossover is chosen with uniform randomness to cut just in front of the positions 2..L, then the bit values making the parent a representative of H will not be separated from one another under crossover if the cutpoint is chosen outside the span of those bit values, and the latter happens with probability $1 - (\delta(H)/(L-1))$. Thus, after mating with crossover, the number of children that are representatives of H should be at least as big as $m \cdot \left(\frac{\mu(H)}{\mu(P)}\right) \cdot \left(1 - \frac{\delta(H)}{L-1}\right)$.

Recall p_m is the probability that a single bit is mutated to the complementary value, and *order*(H) is the number of fixed positions in H. An H-representative remains such, despite mutation, with probability $(1-p_m)^{order(H)}$.

Now, in the next generation $P(t+1)$ of the population, the number of representatives of H should be at least as big as

(E1) $$m \cdot \left(\frac{\mu(H)}{\mu(P)}\right) \cdot \left(1 - \frac{\delta(H)}{L-1}\right) \cdot (1-p_m)^{order(H)}.$$

If the factor complementary to m,

(E2) $$\left(\frac{\mu(H)}{\mu(P)}\right) \cdot \left(1 - \frac{\delta(H)}{L-1}\right) \cdot (1-p_m)^{order(H)}$$

is greater than 1, we conclude that the number of H-representatives can be expected to increase from $P(t)$ to $P(t+1)$. The smaller $\delta(H)$ is, the closer factor $1 - (\delta(H)/(L-1))$ is to 1. The smaller p_m and *order*(H) are, the closer factor $(1-p_m)^{order(H)}$ is to 1. If $\mu(H)$, the average fitness of those H-representatives in $P(t)$, is somewhat greater than $\mu(P)$, the average fitness of all of $P(t)$, then the quantity (E2) can exceed 1. Whence:

Holland's Schema Theorem: Assume schema H has short defining length $\delta(H)$ (therefore, also low value for *order*(H)). Also assume the mutation rate p_m is low. If the representatives of H have somewhat above average fitness within the population $P(t)$ at time t, then the number of representatives of H can be expected to increase in the next generation $P(t+1)$.

3 A NON-LINEAR SCHEMA THEOREM

Now we want to consider scenarios where individuals in a universe are described by some set of bit values but where the bits are arranged in a configuration which is not necessarily a linear sequence. For example, the bits might be arranged as the nodes of a tree. Or they might be arranged as the lattice points (the points with integer coordinates) in a parallelepiped (a "box") in multi-dimensional real space $\Re^n$.

More generally, we will assume the bits are arranged as the nodes in some connected finite graph G. We shall assume there is some notion of the distance between two nodes (perhaps path length in G). Also we will assume there is some natural notion of ways to cut G into two non-empty subsets of its nodes (possibly by clipping one or several edges in G).

A schema will be as before: at G's nodes we assign values of either 0 or 1 or the wildcard, *. Fixed positions in the schema, versus unfixed positions, means the same as before. The order of a schema means (as before) the number of fixed positions in it.

We need a notion analogous to a schema's defining length. We have assumed there is a notion of distance between two nodes in G. If B is some subset of G-nodes, define the *diameter* $\Delta(B)$ of B to be the maximum distance between any two nodes in B. Since graph G is finite, the diameter of any G-subset, including G itself, is a well-defined (finite) positive real number. Given a schema H, define its *relative diameter*

$$rel\Delta(H) = \Delta(fixed(H))/\Delta(G)$$

where *fixed*(H) is the set of fixed nodes in H. Figure 1 suggestively illustrates the diameters $\Delta(G)$ and $\Delta(fixed(H))$. The relative diameter of a schema is a real number in the unit interval [0, 1]. Note that in the original Holland scenario, when L bits are arranged in a linear

sequence, the relative diameter of a schema H is $\delta(H)/(L-1)$.

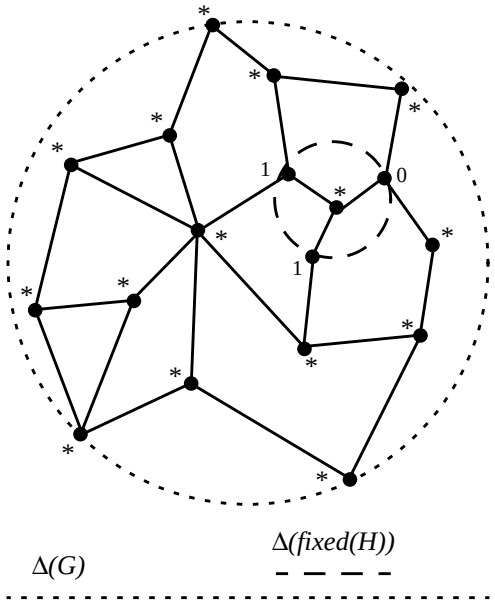

Figure 1: Diameters of Graph G and *fixed*(H) within it.

When graph G is cut in two (in some natural but as yet unspecified way), it is possible for two fixed positions in a schema to become separated (wind up in separate subsets of G-nodes); term this a *disruption* of the schema. Define the *disruption probability*, $dp(H)$, of a schema H to be the probability that a random cut (whatever that may mean) disrupts the schema. Note that in the original Holland scenario, when L bits are arranged in a linear sequence, the disruption probability of a schema H is $\delta(H)/(L-1)$, so, is equal to the relative diameter $rel\Delta(H)$.

Now we prove the non-linear analogue of Holland's Schema Theorem. Let schema H be given. Suppose there are m representatives of H in the population $P(t)$. As before, $\mu(H)$ denotes the average fitness of the H-representatives in $P(t)$, and $\mu(P)$ the average fitness of the entire population $P(t)$. Again, when creating a generational change, we select individuals for parenting by a weighted roulette wheel. Again the number of parents which are H-representatives has expected value $m \cdot \mu(H)/\mu(P)$. Two children are formed from two parents by making a random cut of G, then interchanging parental fragments. A child of an H-representative continues to be an H-representative if the random cut did not disrupt H, which happens with probability 1 - $dp(H)$. At this point we make the critical assumption: Assume $rel\Delta(H) \geq dp(H)$. Then the number of children which are H-representatives should be at least as big as $m \cdot \left(\frac{\mu(H)}{\mu(P)}\right) \cdot (1 - rel\Delta(H))$, and after mutation, the number of children which are H-representatives should be at least as big as

(E3) $$m \cdot \left(\frac{\mu(H)}{\mu(P)}\right) \cdot (1 - rel\Delta(H)) \cdot (1 - p_m)^{order(H)}.$$

Of course, this is the analogue to the expression (E1) in section 2 above.

Non-Linear Schema Theorem: Let H be a schema. Assume (i) the mutation rate p_m is low, (ii) H has low order, (iii) H has low relative diameter $rel\Delta(H)$, and (iv) $rel\Delta(H) \geq dp(H)$. If representatives of H have somewhat above average fitness within the population $P(t)$, then the number of representatives of H can be expected to increase in the next generation $P(t+1)$.

Of course, for this theorem it would be enough to replace (iii) and (iv) with (iii´) $dp(H)$ is low. But we anticipate that, in general, $dp(H)$ is considerably harder to calculate than $rel\Delta(H)$.

With regard to proving theorems about the burgeoning of building blocks, possible relations between $rel\Delta(H)$ and $dp(H)$ are:

1. $rel\Delta(H) \geq dp(H)$ for all schema H;
2. $rel\Delta(H) \geq dp(H)$ for all schema H which have low relative diameter $rel\Delta(H)$;
3. $rel\Delta(H) \geq dp(H)$ for the majority of schema H which have low relative diameter.

4 BITS ARRANGED IN A TREE

Now we investigate connecting bits as the nodes in a tree. In particular, consider the full binary tree T on 63 nodes. In this tree, levels 0 through 5 are full; level 5 contains 32 leaf nodes. The edges in this graph are those that connect children to parents. Only the root node has no parent, so there are 62 edges altogether. For our notion of distance, we will use path length between two nodes (we mean the shortest path that goes through their nearest common ancestor). The longest path in this tree has length 10; this is the length of path that connects any leaf in the root's right subtree to any leaf in the root's left subtree.

Manufacturing a schema H amounts to making some of the 63 nodes have fixed values of 0 or 1, and assigning an asterisk to the other nodes. We can write a function that returns the length of the longest path between two fixed nodes in a schema. Thus $rel\Delta(H)$ is easily calculated directly.

Given two parent bit trees, we perform mating with crossover in the style of (Koza, 1992): at random select

Table 1: Tree Experiments

Schema order	Number trials	$dp(H)$			dp / $rel\Delta$			# times $dp > rel\Delta$
		least	most	avg	least	most	avg	
2	100	0.0161	0.1613	0.1089	0.1613	0.1613	0.1613	0
3	800	0.0323	0.2258	0.1606	0.1613	0.2419	0.1937	0
4	12K	0.0484	0.2903	0.2007	0.1613	0.3024	0.2237	0
5	140K	0.0806	0.3387	0.2368	0.1613	0.3629	0.2532	0
6	1400K	0.1129	0.3871	0.2691	0.1613	0.4032	0.2810	0
8	1000	0.1935	0.4355	0.3265	0.2097	0.4480	0.3330	0
16	1000	0.3548	0.6290	0.4948	0.3548	0.6290	0.4954	0
32	1000	0.5806	0.8387	0.7165	0.5806	0.8387	0.7165	0
48	1000	0.7742	0.9677	0.8737	0.7742	0.9677	0.8737	0

Column 2 is the number of random schema generated for the given order. Columns 3-5 and 6-8 give values of the disruption probability $dp(H)$ and of the ratio $dp(H)/rel\Delta(H))$. Column 9 shows that $dp(H)$ never exceeded $rel\Delta(H))$ in these experiments.

one of the 62 edges in the tree and interchange the parental subtrees beneath that edge.

Now we address the calculation of $dp(H)$, the probability that a random cut will disrupt a schema H. Of course, here by disrupt we mean that at least one fixed node of H is in the subtree underneath the edge being cut, and at least one fixed node of H lies in the rest of the tree. Since there are only 62 edges which can be cut, we write a function that explicitly calculates $dp(H)$ by examining the consequences of the 62 possible cuts.

It becomes too expensive to explicitly determine by exhaustion whether $rel\Delta(H) \geq dp(H)$ for all schemas H (nor has a mathematical proof occurred to us). For example, the number of schemas of order 4 is a multiple of C(63, 4), where C(63, 4) = the number of ways of combining 63 items 4 at a time = 595,665. Instead we make the following statistical analysis. For some orders 2, 3, 4, etc., we manufacture many random schema, and for those schema we explicitly calculate and compare $rel\Delta(H)$ and $dp(H)$. The results are given in Table 1. As the last column in Table 1 shows, in our experiments, $dp(H)$ never exceeded $rel\Delta(H)$. Moreover, as columns 6-8 show, for low order schemas, $dp(H)$ was on average quite a bit smaller than $rel\Delta(H)$.

We thank our colleague Dr. Terry Watkins of the University of New Orleans Mathematics Department for the statistical assertions of this paragraph. Suppose we test the hypothesis that $rel\Delta(H) \geq dp(H)$ by taking n = 1000 randomly chosen schema H and for those seeing if the hypothesis holds. These are Bernoulli trials where we will assume the probability of success (that is, $rel\Delta(H) \geq dp(H)$) is constant. If every one of the 1000 trials results in success, then with confidence level 0.9999 we can assert that the proportion of schemas for which the hypothesis $rel\Delta(H) \geq dp(H)$ holds must be at least 0.9999999.

Thus, the experiments reported in Table 1 suggest with extraordinary statistical persuasion that $rel\Delta(H) \geq dp(H)$ for all the schema in our setting.

5 BITS ARRANGED IN A CUBE

Now we consider bits arranged as the lattice points within a parallelepiped in 3-dimensional Euclidean space $\Re^3$. A *lattice point* is one whose coordinates are integers. Our parallelepiped will actually be a cube. We will consider several cubes. The first cube has 8 points on each edge and consists of the vectors (l, m, n) in $\Re^3$ such that each of the coordinates l, m, n is an integer in the range 0..7. Such vectors are the nodes in our graph G. We intend there to be an edge between two vectors (= nodes) if the two vectors agree in two of their coordinates and differ by 1 in the remaining coordinate. But, in truth, since the notion of distance we will use will be Euclidean distance, the edges are less significant than the physical proximity of nodes to one another. The other two cubes are similar but with, respectively, 10 and 12 points on each edge. The number of bits in our three cubes are, respectively, 8^3 = 512, 10^3 = 1000, and 12^3 = 1728.

We will cut such a cube with a hyperplane. In 3-dimensional Euclidean space $\Re^3$, a *hyperplane* is determined by four real numbers a, b, c, d, and consists of those vectors (x, y, z) such that $a \cdot x + b \cdot y + c \cdot z = d$. The half-spaces into which the hyperplane cuts the world consists of those (x, y, z) which satisfy

Table 2: Cube Experiments

Edge size	Ball size	Avg order	Avg $rel\Delta$	% $rel\Delta \geq dp$	Avg ratio $dp/rel\Delta$	
					all 1000	when $dp > rel\Delta$
8	19	7.63	0.211	73.5	0.915	1.439
10	27	11.20	0.195	64.0	1.015	1.486
12	57	22.92	0.229	66.5	1.034	1.526

For all three edge sizes, 1000 random schema were manufactured. Column 2 is the maximum number of points in a ball with $rel\Delta = 0.25$. Column 5 is the percentage of the 1000 schema which satisfy $rel\Delta(H) \geq dp(H)$. Column 6 has the average ratio of $dp(H)$ to $rel\Delta(H)$, taken over all 1000 schema, and column 7 has that average ratio over just the minority of schema for which $dp(H) > rel\Delta(H)$.

$a \cdot x + b \cdot y + c \cdot z > d$ (which we term the *sky side*) versus those which satisfy $a \cdot x + b \cdot y + c \cdot z \leq d$ (the *sod side*). We manufacture a random hyperplane by choosing four random real numbers a, b, c, d out of the real interval [-1, 1]. A random hyperplane may not pass through the cube at all; for instance, the entire cube might lie in the sky side. If a hyperplane does pass through the cube, it disrupts a schema if at least one fixed position in the schema is on the sky side and at least one schema fixed position is on the sod side. Crossover between two parents consists of intersecting each parent with the same random hyperplane, then interchanging parental sky sides (or sod sides, of course) to form two children.

Our distance function will be Euclidean distance, the square root of the sum of the squares of the differences of corresponding vector coordinates.

For speaking of a Schema Theorem, we really want to make an assertion about schemas of low relative diameter. So, for our experiments, we manufactured numerous trial schemas of low relative diameter, done as follows. The relative diameter we here report on is $rel\Delta = 0.25$ (behavior for this value is typical). The schema's fixed positions are all chosen to lie within a geometric ball, of this relative diameter, about some center lattice point. Geometry necessitates that there is some maximum number (see column 2 of Table 2) of lattice points that can fit in this ball. (As can be deduced from Table 2, this maximum number amounts to just several percentage points of the entire cube.) To manufacture random schemas, we choose a center point at random, then, with 50-50 chance upon each addition, add each possible cube point that fits in this ball as another fixed position in the schema we are manufacturing. In Table 2, the average order (column 3) of the schemas we manufacture is a bit less that half the maximum number (column 2) because balls centered near a cube face do not receive points outside the cube.

Having manufactured a random schema H of low relative diameter as just described, we estimate the probability $dp(H)$ as follows. We choose 1000 hyperplanes at random, then use the ratio of hyperplanes that disrupt the schema, to hyperplanes that pass through the cube, as our approximation of $dp(H)$. (We would not be very interested at crossover time in hyperplanes that do not pass through the cube.)

Column 5 gives the percentage of the 1000 trial schema which satisfy $rel\Delta(H) \geq dp(H)$. This proportion ranges between 73.5% and 64.0%.

In the ranking of desirable relationships between $rel\Delta(H)$ and $dp(H)$ which is given at the end of section 3, the relationship shown in Table 2 is the third: $rel\Delta(H) \geq dp(H)$ for a majority of schema of low relative diameter. But the picture is actually a little rosier than that. Column 6 of Table 2 shows the average ratio $dp(H)/rel\Delta(H)$ over all 1000 schema. This evidence says values $rel\Delta(H)$ and $dp(H)$ are on average rather close to one another. Column 7 looks at the average ratio $dp(H)/rel\Delta(H)$ over just that minority of schemas for which $dp(H) > rel\Delta(H)$. This evidence says that when $dp(H)$ exceeds $rel\Delta(H)$, it does not do so excessively.

These experiments indicate that arranging bits at lattice points in boxes, which for purposes of crossover are cut by hyperplanes, can entail the desired behavior that building blocks have their representation increased. If a schema's representatives in the population have somewhat above average fitnesses, then their numbers should increase in the next generation provided the schema has low disruption probability $dp(H)$. Quantity $dp(H)$ is hard to calculate, but is usually less than relative diameter $rel\Delta(H)$ and on average is rather close to the value $rel\Delta(H)$.

6 CONCLUSIONS

We have generalized Holland's Schema Theorem, to the setting that bits are arranged, not necessarily in a linear sequence, but as the nodes of a connected graph G. To do

so we posited the existence of a distance function, and our theorem incorporates the premise that $rel\Delta(H) \geq dp(H)$.

We gave experimental results showing that Holland-like behavior, meaning the flourishing of building blocks, can reasonably be expected for two distinct graphs *G*, one being a tree and the other being a cube of lattice points in Euclidean 3-space.

We have given both theory and experimental evidence that bits (genes) need not be limited to alignment in a sequence. Ultimately genes should be arranged in ways that are natural to the problem at hand.

References

Goldberg, David (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley Publishing.

Holland, John (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Koza, John (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: The MIT Press.