Schema Disruption in Tree-Structured Chromosomes

William A. Greene
Computer Science Department
University of New Orleans
New Orleans, LA 70118 USA
504-280-6755
bill@cs.uno.edu

ABSTRACT

We study if and when the inequality $dp(H) \le rel\Delta(H)$ holds for schemas H in chromosomes that are structured as trees. The disruption probability $dp(H)$ is the probability that a random cut of a tree limb will separate two fixed nodes of H. The relative diameter $rel\Delta(H)$ is the ratio (max distance between two fixed nodes in H) / (max distance between two tree nodes), and measures how close together are the fixed nodes of H. Inequality $dp(H) \le rel\Delta(H)$ is of significance in proving Schema Theorems for non-linear chromosomes, and so bears upon the success we can expect from genetic algorithms. For linear chromosomes, $dp(H) = rel\Delta(H)$. Our results include the following. There is no constant c such that $dp(H) \le c \cdot rel\Delta(H)$ holds for arbitrary schemas and trees. This is illustrated in trees with eccentric, stringy shapes. Matters improve for dense, ball-like trees, explained herein. Inequality $dp(H) \le rel\Delta(H)$ always holds in such trees, except for certain atypically large schemas. Thus, the more compact are our tree-structured chromosomes, the better we can expect our genetic algorithms to work.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search – *heuristic methods*; F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems – *computations on discrete structures, geometrical problems and computations.*

General Terms

Algorithms, Theory

Keywords

Genetic algorithms, schema theory, schema disruption probability, alternative chromosomes, trees.

1. INTRODUCTION

In breve, this paper investigates the probability of disrupting a schema, when chromosomes are structured as a tree of bits.

The area of Genetic Algorithms concerns a heuristic problem-solving paradigm that takes Darwinian evolution as its metaphor. There is some problem at hand. The problem has numerous solutions, some better (fitter) than others. The number of solutions is typically enormous, too large to search exhaustively. A small-ish population of solutions is maintained, and subjected to such evolutionary forces as survival of the fittest, mating with crossover, and mutation. The hope is that better and better solutions will surface as the population evolves.

In classical genetic algorithms (GAs), as found in Holland [4] or Goldberg [1], individual solutions in the population mimic haploidal chromosomes from nature. Individual solutions get represented as bit strings; that is, as 0's and 1's that are arranged in a linear sequence, one after the other, like beads strung along a strand. All individuals have the same length, N. A don't-care symbol, *, is introduced, then a *schema* is defined to be a string, of length N, of symbols chosen from {0, 1, *}. A schema denotes a subspace of the space of all individuals, namely, an individual is in that subspace (the individual is a *representative* of the schema) if the individual agrees with the schema at all the positions where the schema is 0 or 1 (the *fixed* positions in the schema). (The positions with * are the *free* positions of the schema.) The letter H (for hyperplane) is often used to name a schema. Schemas are studied because they focus on the issue of when GAs work well. If the problem at hand is reasonably well-behaved, and if the representation of individual solutions is a natural one, then bits should group together into building blocks. A *building block* is a collection of related bits, which can take on values (0's versus 1's) that contribute significantly to the fitness of the individual. A schema simply expresses the characterizing properties of a building block.

Mating via one-point crossover is accomplished by clipping two parent chromosomes at some same *cutpoint* along their sequence of bits, and exchanging parental fragments to form two children, which then have the same length as their parents. If the cutpoint lies between the outermost fixed positions of a schema H, the cutpoint is said to *disrupt H*. The terminology acknowledges the fact that if one parent is a representative of H and if the cutpoint lies between the outermost fixed positions of H, then it is possible that neither child is again a representative of H. The *disruption probability* $dp(H)$ of a schema H is the probability that a uniformly randomly chosen cutpoint will disrupt H. For a chromosome structured as a linear sequence, the disruption probability is easily calculated: $dp(H) = \delta / (N-1)$, where δ is the distance between the outermost fixed positions of H. Probability $dp(H)$ plays an important role in Holland's Schema Theorem (confer [1] or [4]).

The lesson is clear. If a building block is to persist in the population over generations, it is better if its bits (that is, the fixed positions of the corresponding schema) are located close together. But

GECCO'05, June 25-29, 2005, Washington, DC, USA.

this points up a weakness with arranging the bits of an individual as a linear sequence. A bit has two nearest neighbors (the bits to either side of it, of course), but no more than two. What if it is in the nature of the problem at hand that a bit should be equally close to more than two other bits?

In this paper we allow a different structure for chromosomes. Namely, we consider the case that the 0's and 1's in a chromosome are arranged like the nodes of a tree, which we denote as *T*.

The analogue of *dp(H)* for tree-structured chromosomes is addressed in the next section. For now, suffice it to say that the calculation of the analogue of *dp(H)* is not so easily done. This paper is devoted to finding an easily calculated upper bound for *dp(H)*.

We close this section by commenting on related research. Non-linear bit arrangements and in particular tree-structured arrangements, and a schema theory for such, have been studied by others. Principally this has come from those in the Genetic Programming (GP) community, although Greene in [2] and [3] has investigated non-linear bit arrangements in the abstract. In GP approaches, individuals are programs, specifically functions, realized as expression trees. Mating with crossover consists of clipping out and exchanging subtrees between the two parents. The individuals in a population can have quite different shapes, which fact complicates a number of issues, such as, what will be the definition of a schema, and what relation will hold between the locations of the cutpoints in the two parents? For Koza [5], O'Reilly [6], [7], and Whigham [10], schemas are expression fragments which incorporate don't care symbols, and which are further characterized by not being anchored to some fixed position within the expression tree and moreover can be instantiated multiple times within the same individual. In Rosca [9] the innovation is that a schema is an expression fragment which is anchored at the root of the expression tree.

Our own interest in non-linear bit arrangements did not originate from a prior interest in genetic programming. Rather, our intuition has been that strictly linear bit arrangements are simply too confining and inflexible. From within the GP community, the work that comes closest to our own efforts is that of Poli and Langdon [8]. Their definitions of schema, mutation, and crossover are the closest carryover to GP of the allied notions from the standard GA approach with its linear bit arrangements. For Poli and Langdon, a schema is a rooted tree of symbols, where the root is to correspond to the root of an expression tree that is an individual in the population. The symbols in the schema are of three kinds: function symbols, terminal symbols (variables and constants), and a don't-care symbol. Don't-care symbols can appear at interior nodes or leaf nodes of a schema, so the schema dictates a minimum size of any individual expression tree which can instantiate the schema. To perform crossover, Poli and Langdon search the structures of two parent individuals, starting at their roots and working downwards. They identify the largest rooted subtrees which are isomorphic between the two parents, and at random they choose one of the branches in the isomorph for a one-point cut, then parental fragments are exchanged. (Finally, for example, mutation of an interior node amounts to substituting one function by another of the same function type, meaning the same type of return value and the same number and types of parameters.) We will remark on similarities between our present research and the work of Poli and Langdon [8] in passing.

2. SCHEMA DISRUPTION IN TREES

We begin with some definitions. Many are standard, but we give them for the sake of clarity and completeness.

A graph *G* consists of points called *nodes*, certain pairs of which have a *edge* between them, in which case the pair are termed *adjacent*. A *path* in *G* from node *x* to node *y* is a sequence of distinct nodes $x = v_0, v_1, v_2, ..., v_n = y$ in *G*, such that v_{i-1} is adjacent to v_i for $1 \le i \le n$. The *length* of this path is *n*. Note that insisting the nodes in the path are distinct means that a path does not cross or retrace itself. A *cycle* is similar to a path, and is a sequence $v_0, v_1, v_2, ..., v_n$ of adjacent nodes, all distinct except that $v_0 = v_n$.

A graph *G* is *connected* if for each pair of nodes *x* and *y*, there exists a path between them. A graph is *acyclic* if it contains no cycles.The *degree* of a node in a graph is the number of nodes to which it is adjacent.

A(n *ordinary*) *tree* is a finite acyclic graph. The number of edges in a tree is always one less than the number of nodes. A *tree of degree k* is a tree for which the degree of every node is at most *k;* such a tree is also called one of *bounded degree* with bound *k*.

We also need the next definitions. A *rooted tree* is a tree which has one node distinguished as the *root* of the tree. The nodes adjacent to the tree are its *children* and the root is their *parent*. The children are themselves the parents of the non-parental nodes to which they are adjacent, etc. A *rooted k–ary tree* is one in which every node has at most *k* children. A rooted 2–ary tree is the familiar binary tree. A *leaf* is a node with no children.

In a rooted *k*-ary tree, a node is adjacent to its parent, so a rooted *k*-ary tree is also a(n ordinary) tree of degree *k*+1. But a tree of degree *k*+1 might not meet the definition of a rooted *k*-ary tree, since the root is adjacent to at most *k* (not *k*+1) nodes, namely, its children.

In a tree, there is a unique path between any two nodes *x* and *y*; the *distance* $dist(x, y)$ between *x* and *y* is the length of (i.e., number of edges in) that path. Function *dist* satisfies the triangle inequality: $dist(x, z) \le dist(x, y) + dist(y, z)$.

We look ahead to mating with crossover. Cutting a tree-structured chromosome will mean clipping a tree edge. That disconnects the individual into connected subtrees, which will serve as the fragments to be exchanged under crossover. All our individuals will be trees of the same shape (a shape dictated by the problem at hand which we are trying to solve), and when parents are cut for crossover, the cutpoints have the same locations in the parents. Children will thus have the same shape as their parents. It is easy to imagine tree analogues to one-point crossover, two-point crossover, etc., from standard GAs. This paper exclusively considers one-point crossover.

Given two nodes in a tree *T*, we say an edge *separates* them provided that cutting *T* at that edge results in *x* being in one of the fragments and *y* being in the other; this happens if the cut severs the unique path between *x* and *y*. We say an edge separates a subset *A* of nodes if there are two nodes in *A* which are separated by the edge.

For our exploration of tree-structured chromosomes, a *schema* will mean the expected analogue from standard genetic algorithms. Namely, a *schema* is the subspace of all possible individuals determined by fixing the values (0's and 1's) at some designated subset

of the bits (tree nodes) and letting the other bits range over their values. The schema can be denoted by labeling the nodes of the tree with characters 0, 1, or *. We will use the same name H for the schema and for its set of fixed nodes. The *disruption probability* of a schema H is the probability that a uniformly randomly chosen (one-point) cut will separate H. It equals the number of edges that separate H, divided by the total number of edges in our tree-structured individuals. (The number of edges that separate H comes closest to what is termed the defining length of a schema, in the GP work of Poli and Langdon [8].)

Define the *diameter*, $\Delta(S)$, of a set S of tree nodes to be the maximum distance between any two elements of S. The *relative diameter*, $rel\Delta(H)$, of a schema H means the ratio $\Delta(fixed(H)) / \Delta(G)$, where $fixed(H)$ is the set of fixed positions of H.

This paper concerns if and when inequalities of a form like $dp(H) \leq rel\Delta(H)$ hold. Amount $rel\Delta(H)$ is our candidate for an easily calculated upper bound on $dp(H)$ to which we earlier alluded.

3. THE GENERAL CASE

In this section we present some results that hold for arbitrary tree-structured chromosomes T. Let a schema H be given. Recall, we let the same name H also designate the set of fixed nodes of the schema. Now let T_H denote the smallest subtree of T that contains H. Tree T_H is the intersection of all the subtrees of T which contain H. Now we make some observations about T_H. (Subtree T_H is termed the minimum tree fragment for the schema H, in the GP work of Poli and Langdon [8].)

Each leaf of T_H is an element of H. For, if a leaf of T_H is not an element of H, then we can remove it and the edge to it and so obtain a smaller subtree of T which still contains H, in contradiction to the minimality of T_H.

The set of edges of host chromosomal tree T which separate H is the same as the set of edges which separate T_H. One set containment is obvious, since $H \subseteq T_H$. And if an edge separates T_H then in particular it must separate two leaves of T_H, but those are elements of H. Now we can also conclude that $dp(H) = dp(T_H)$.

Also $\Delta(H) = \Delta(T_H)$. Why? Since $H \subseteq T_H$, we see $\Delta(H) \leq \Delta(T_H)$. On the other hand, the two farthest apart elements of T_H must be two leaves of T_H, but those must be elements of H, and it follows that $\Delta(T_H) \leq \Delta(H)$. Conclude $\Delta(H) = \Delta(T_H)$. Hence, also $rel\Delta(H) = rel\Delta(T_H)$.

It follows that, when investigating whether relation $dp(H) \leq rel\Delta(H)$ holds, if need be we can assume our schema H of fixed nodes is in fact a subtree of T.

In general, the inequality $dp(H) \leq rel\Delta(H)$ does not hold, as our first example will show. The chromosomal tree T and schema H of Example 1 are pictured in Figure 1. Schema H consists of s spoke nodes arranged around a hub node, and the remainder of T consists of t tail nodes aligned in a row. There are altogether $s + t$ edges in this tree, and of those, s will separate nodes of H, so $dp(H) = s / (s + t)$. On the other hand, $rel\Delta(H) = 2 / (t + 2)$, so that the ratio $\frac{dp(H)}{rel\Delta(H)} = \frac{s}{2} \cdot \frac{t+2}{s+t}$. Denote the right-hand side of this equation by R. We can arrange that R is arbitrarily large. For instance, if s and t are a same very large number, then R is approximately $s / 4$

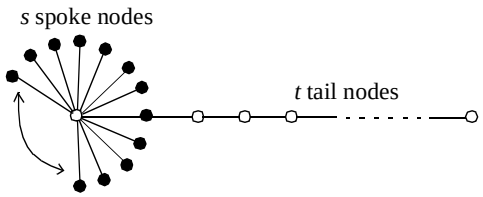

Figure 1: Example 1

and is also very large. For a given chromosomal tree structure T, there might be some constant c such that $dp(H) \leq c \cdot rel\Delta(H)$ for every schema $H \subseteq T$, but there is no constant c that will work for every tree structure. We have proved the following.

Proposition 1: There is no constant c for which the relation $dp(H) \leq c \cdot rel\Delta(H)$ will hold for every schema $H \subseteq T$ in every tree structured chromosome T.

Given any tree-structured chromosome T, and schema $H \subseteq T$, there is a constant c_H which depends upon H and for which $dp(H) \leq c_H \cdot rel\Delta(H)$, as we now set about showing.

Let T and H be given. Let a and b be two elements of H at a maximal distance from one another, so that the unique path between a and b has length $dist(a, b) = \Delta(H)$. Let m be the middle node on this path. Note that the distance from m to either a or b is at least as big as $\lfloor \Delta(H) / 2 \rfloor$, the floor of $\Delta(H) / 2$.

Lemma 1: Let T, H, a, b, and m be as in the preceding paragraph. The distance between m and any element of H is at most $\lceil \Delta(H)/2 \rceil$, the ceiling of $\Delta(H) / 2$.

Proof: By way of contradiction, assume there is an element $d \in H$ such that $dist(m, d) > \lceil \Delta(H) / 2 \rceil$. Consider the path P from a to d. Let S be the set of nodes common to P and the path between a and b. Set S cannot consist only of a alone (otherwise, the concatenation of P and the path from a to b must be the unique path from d to b, and its length exceeds $dist(a, b) = \Delta(H)$, a contradiction). Also set S must consist of consecutive nodes along the path from a to b (if not consecutive, then tree T contains a cycle, a contradiction). Let v_d be the last of the consecutive nodes of S. There are two cases.

Case I: v_d is between a and mid node m: In this case, it is at v_d that the path from d to a first overlaps the path from a to b. It follows that the unique path from d to b must pass through v_d and m. Therefore $dist(d, b) = dist(d, m) + dist(m, b) > \lceil \Delta(H) / 2 \rceil + \lfloor \Delta(H) / 2 \rfloor = \Delta(H)$, but that makes d too far from b, a contradiction.

Case II: v_d is between m and b: In this case it is the path from d to a which must pass through m, and this time we can conclude $dist(d, a) > \Delta(H)$, a contradiction.

Since both cases are impossible, the lemma now follows. ■

Recall our notation that $T_H \subseteq T$ is the smallest subtree of T which contains H, and the fact that the leaves of T_H must be elements of H. Denote by H_e those elements of H which are leaves of T_H (subscript e stands for extreme). Notation $|H_e|$ means the number of elements in the set H_e.

Proposition 2: The number of edges in T_H is at most $|H_e| \cdot \left\lceil \frac{\Delta(H)}{2} \right\rceil$.

Proof: Let node m be as above. Every edge of T_H is on a path between m and a leaf of T_H. Therefore the number of edges is bounded by the number of leaves times the maximum length of a path from m to a leaf. ■

The bound given in this proposition is a tight one, as can be seen from Example 1's schema H, which is the wheel of s spoke nodes.

Now we can obtain the inequality to which we alluded prior to Lemma 1 above.

Proposition 3: Let T be a tree-structured chromosome. For a schema $H \subseteq T$ there is a number c_H which depends upon H and for which $dp(H) \le c_H \cdot rel\Delta(H)$.

Proof: $dp(H) = n_{sep} / n_T$, where n_{sep} equals the number of edges that separate H, which equals the number of edges in T_H, and n_T equals the total number of edges in T. Our result follows from observing that $\frac{n_{sep}}{n_T} \le \frac{|H_e| \cdot \left\lceil \frac{\Delta(H)}{2} \right\rceil}{n_T} \le \frac{|H_e| \cdot \left\lceil \frac{\Delta(H)}{2} \right\rceil}{\Delta(T)} \approx \frac{|H_e|}{2} \cdot rel\Delta(H)$. Factor $c_H \approx \frac{|H_e|}{2}$ depends on the size of H_e. ■

Can we obtain an inequality of the form $dp(H) \le rel\Delta(H)$ for (most) schemas in some restricted class of tree-structured chromosomes? Eventually below, we will. Example 1 above suggests that its foible is its many spoke nodes, that is, is the fact that the hub node has degree s, which we can make arbitrarily large. Perhaps matters improve if the tree's nodes are of bounded degree. But we shortly will see this restriction is not yet enough. We begin with a definition.

4. BALL-LIKE TREES

Definition: A(n ordinary) tree T is termed a *ball-like tree of degree* k+1 *and radius* ρ, provided
(i) there is a distinguished node *cntr*, the center node;
(ii) all nodes of T are at a distance at most ρ from the center;
(iii) there is at least one node at a distance ρ from the center;
(iv) all nodes have degree at most k+1.

Now we introduce more terminology about such trees. The *parent-child relation* between nodes is the expected one: in general, a node is parent to the non-parental nodes to which it is adjacent, with center node *cntr* being the ultimate ancestor. Given a ball-like tree T of degree k+1 and radius ρ, its center node can have up to k+1 children, and other non-leaf nodes can have up to k children. The *level of a node* is its distance from the center. *Level* λ *of tree T* is the set of nodes at level λ. A *level is full* if it has the maximum number of nodes possible, which is $(k+1)k^{\lambda-1}$. *Tree T is full* if its every level is full. *Tree T is complete* if its every level is full except possibly level ρ. The leaves of a complete tree can only appear on levels ρ–1 and ρ. The number of nodes in a complete ball-like tree

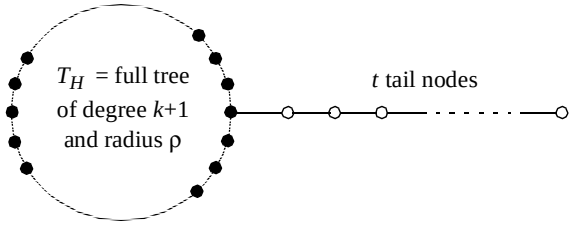

Figure 2: Example 2

T of degree k+1 and radius ρ is at most $1 + (k+1)\frac{k^{\rho}-1}{k-1}$ (the full case) and at least $2 + (k+1)\frac{k^{\rho-1}-1}{k-1}$ (complete but with only one node on level ρ). (The usual definition of a complete binary tree as seen in a Data Structures textbook also insists that the nodes on the bottom level are bunched together off to the left without gaps, but we will not need this stipulation.)

To obtain a general inequality of the form $dp(H) \le c \cdot rel\Delta(H)$ for some fixed constant c and arbitrary $H \subseteq T$, it is not enough to restrict to trees of bounded degree, as our next example shows. Example 2 is depicted in Figure 2. Subtree $T_H \supseteq H$ is a full ball-like tree of degree k+1 and radius ρ. Such T_H arises when H is any superset of the leaves of the full tree T_H. The rest of host chromosomal tree T is the tail depicted. Note that $\Delta(H) = 2\rho$ and $\Delta(T) = 2\rho + t$. The number of edges in a tree is always one less than the number of nodes, so the number of edges in tree T that disrupt H is the number of edges in T_H, which is $(k+1)\frac{k^{\rho}-1}{k-1}$. The total number of edges in tree T is t more than that. Now imagine that we let t equal $(k+1)\frac{k^{\rho}-1}{k-1}$; it follows that $dp(H) = 1/2$. Therefore the ratio $\frac{dp(H)}{rel\Delta(H)}$ equals $\frac{2\rho + t}{4\rho} = \frac{2\rho + (k+1)\frac{k^{\rho}-1}{k-1}}{4\rho}$, which can be made arbitrarily large by letting ρ become arbitrarily large. We have proved the following proposition.

Proposition 4: There is no constant c for which the relation $dp(H) \le c \cdot rel\Delta(H)$ will hold for every schema $H \subseteq T$ in every tree T of bounded degree.

5. COMPLETE (k+1)–ARY TREES

In this section we will obtain the inequality $dp(H) \le rel\Delta(H)$ for almost all schemas $H \subseteq T$, for a restricted class of tree-structured chromosomes T.

In the next proposition, we consider a limited range of values for k, namely, $2 \le k \le 7$, (i.e., trees of degree 3 through 8), and also a limited range of values for ρ, namely, $2 \le \rho \le 100$. The assumption is that the restricted ranges studied will indicate the general facts, and also will exhaust the types of branchy trees that are likely to arise in practice. The proof will reveal that the result appears to hold for arbitrary $k \ge 2$ and arbitrary $\rho \ge 2$. (Tree radius $\rho = 1$ amounts to trivialities.)

Proposition 5: For $2 \le k \le 7$ and $2 \le \rho \le 100$, the inequality $dp(H) \le rel\Delta(H)$ holds for all schemas $H \subseteq T$ and all complete ball-like trees T of degree k+1 and radius ρ, with the exception of certain schemas H which contain atypically large numbers of fixed positions.

Proof: This result is the natural extension of Proposition 3 in Greene [3], and its proof.

Given a certain schema diameter value $\delta = \Delta(H)$, there are many schemas H which have that diameter. Some are large and some are small, and the same can be said for the enveloping subtree T_H of H. Now imagine the schema diameter value $\delta = \Delta(H)$ as a given. We will find an upper bound for fraction $dp(H)$, by calculating the most that its numerator can be, and then the least that its denominator can be and still exceed the numerator.

The numerator of $dp(H)$ can be as large as the number of edges in the largest possible enveloping subtree T_H. We introduce some notation: Let h_d be a fixed node of H at a maximal distance from T's center node *cntr*; let d be the distance between h_d and the center. Any two nodes if T_H are at most distance $\delta = \Delta(H) = \Delta(T_H)$ apart. So any node of T_H is at most distance δ from h_d. We will count how many nodes can possibly be in our chromosomal tree T, be no further from the center than h_d, and be at distance at most δ from h_d. Subtree T_H can be as large as that set of nodes.

Either schema diameter $\delta = \Delta(H)$ is even or it is odd. And either $\delta \le d$ or $\delta > d$. Thus there are four cases to consider. We will give full details for two of the cases and leave the details of the other two cases to the reader.

Case I: even $\delta \le d$ (see Figure 3 for guidance). Consider the path of length δ, consisting of the nodes that lead from h_d towards the center of T; denote the nodes on this path as $h_d = v_0, v_1, v_2, ..., v_\delta$. Subtree T_H could contain all the nodes in a full rooted k–ary tree, rooted at $v_{\delta/2}$ and having height $\delta/2$; the number of nodes in such a subtree is $1 + k + k^2 + ... + k^{\delta/2} = \frac{k^{\delta/2+1} - 1}{k - 1}$. Similarly, $v_{\delta/2+1}$ might have k–1 other children which are the roots of full rooted k–ary trees of height $\delta/2$–2, and T_H might contain all these subtrees; they would contribute $(k-1)(1 + k + k^2 + \ldots + k^{\delta/2-2}) = k^{\delta/2-1} - 1$ more nodes to T_H. Again similarly, $v_{\delta/2+2}$ might have k–1 other children which are the roots of full rooted k–ary trees of height $\delta/2$–3; these could contribute $k^{\delta/2-2} - 1$ more nodes to T_H. Continuing on towards the center node, $v_{\delta-1}$ might have k–1 other chil-

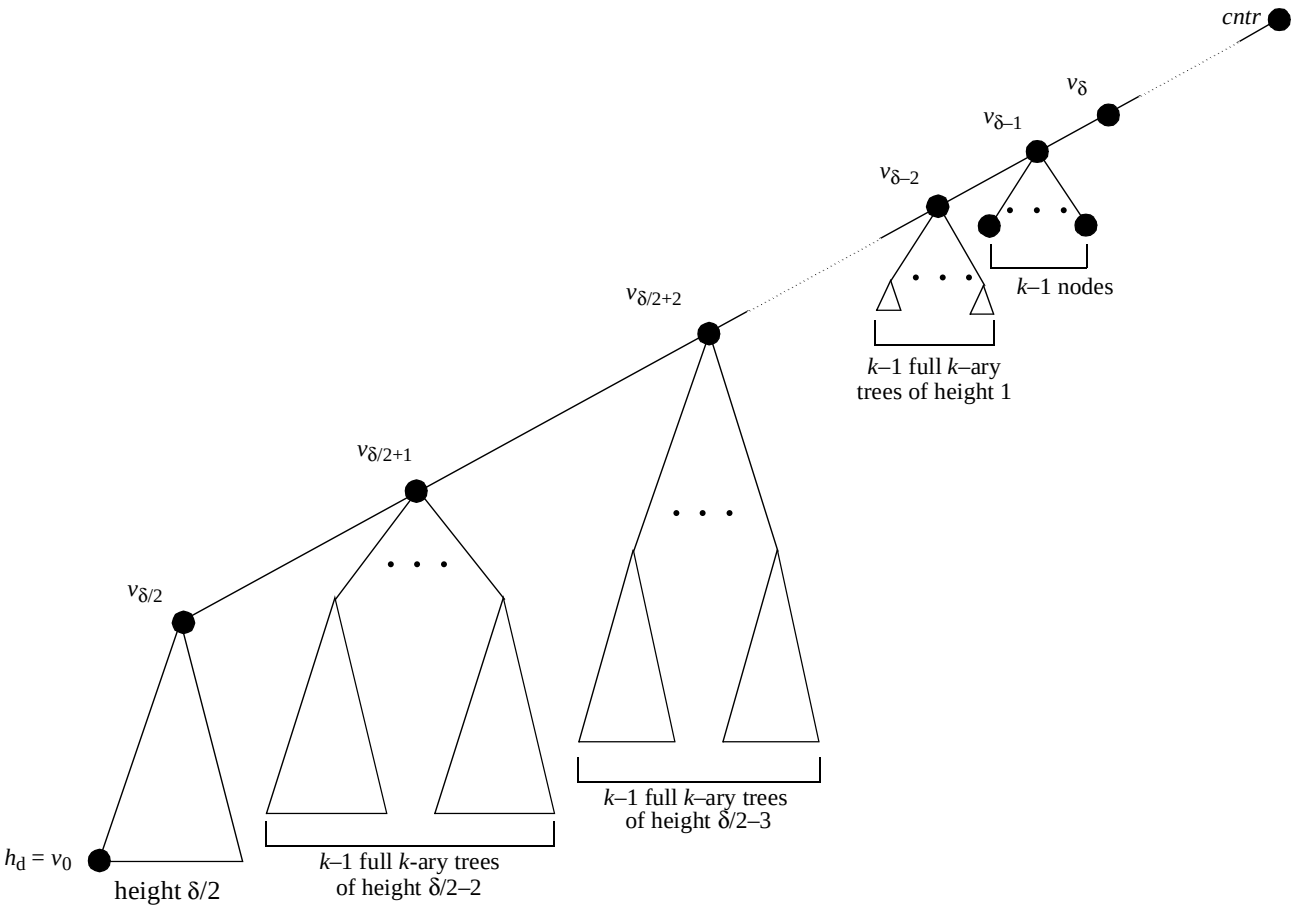

Figure 3: δ is even, $\delta \le d$.

dren which could belong to T_H. The δ/2 nodes $v_{\delta/2+1}$, $v_{\delta/2+2}$, ..., $v_{\delta-1}$, v_δ, together with the nodes in the trees just alluded to, altogether add up to

$$\frac{k^{\delta/2+1}-1}{k-1} + (k^{\delta/2-1}-1) + (k^{\delta/2-2}-1) + \ldots + (k-1) + \frac{\delta}{2}$$

$$= \frac{(k+1)k^{\delta/2}-2}{k-1}$$ nodes which might belong to T_H and in fact this constitutes the largest that T_H could be, given schema diameter δ. The number of edges in T_H is one less, or $\frac{(k+1)k^{\delta/2}-2}{k-1} - 1$.

Let N_{I} denote this last amount (subscript I is for Case I.)

Continuing with case I, we now consider host chromosomal tree *T*. If the distance *d* between *H*'s most outlying node h_{d} and *T*'s center node is less than the radius ρ of *T*, then tree *T*, to be complete and of radius ρ, can have as few as one node on level ρ, in which case *T* has $(k+1)\frac{k^{\rho-1}-1}{k-1} + 2$ nodes and therefore $(k+1)\frac{k^{\rho-1}-1}{k-1} + 1$ edges. But if distance *d* equals *T*'s radius ρ, then since we have allowed T_H to be as big as containing the full *k*–ary of height δ/2 rooted at node $v_{\delta/2}$, it follows that *T* will be required to have at least $k^{\delta/2}$ nodes on its farthest level ρ. Then *T* must have at least $k^{\delta/2} + (k+1)\frac{k^{\rho-1}-1}{k-1} + 1$ nodes and hence at least $k^{\delta/2} + (k+1)\frac{k^{\rho-1}-1}{k-1}$ edges. Ergo, $dp(H)$ is bounded above by $\frac{N_{\mathrm{I}}}{(k+1)\frac{k^{\rho-1}-1}{k-1}+1}$ if $d < \rho$, but bounded above by $\frac{N_{\mathrm{I}}}{k^{\delta/2} + (k+1)\frac{k^{\rho-1}-1}{k-1}}$, if $d = \rho$.

Since *T* is complete but not necessarily full, $\Delta(T)$ is either 2ρ or 2ρ–1; in either event, $rel\Delta(H) \geq \delta / 2\rho$.

Combining facts, the inequality $dp(H) \leq rel\Delta(H)$ will hold, provided the next two inequalities hold:

$$\frac{N_{\mathrm{I}}}{(k+1)\left(\frac{k^{\rho-1}-1}{k-1}\right)+1} \leq \frac{\delta}{2\rho}$$, for even $\delta \leq d$ when $d < \rho$;

$$\frac{N_{\mathrm{I}}}{k^{\delta/2} + (k+1)\left(\frac{k^{\rho-1}-1}{k-1}\right)} \leq \frac{\delta}{2\rho}$$, for even $\delta \leq d$ when $d = \rho$.

We used a computer program to examine if or when these inequalities held, for *T* radius ρ in the range from 2 to 100 and *k* in the range from 2 to 7, and found the following results. There was only one class of failure. For every *k* in 2..7, the second inequality failed in the particular case that $d = 2$, $\rho = 2$, and $\delta = 2$. This is a failure of our upper bound on $dp(H)$ to be less than or equal to our lower bound on $rel\Delta(H)$. For instance, when $k = 4$, our upper bound on $dp(H)$ is 5/9, and our lower bound on $rel\Delta(H)$ is 2/4. Nonetheless, in fact the relation $dp(H) \leq rel\Delta(H)$ holds for values $d = 2$, $\rho = 2$, and $\delta = 2$, for any $k \geq 2$. Figure 4 illustrates what happens, exemplified by the choice of $k = 4$. Figure 4(a) shows, for $d = 2$, $\rho = 2$, and $\delta = 2$, the largest T_H and smallest $T \supseteq T_H$, and in this event, $dp(H)$ is 5/9 and $rel\Delta(H)$ is 2/3. Figure 4(b) shows what happens when we make *T* become larger by adding more nodes on level ρ. For such larger trees *T*, $dp(H)$ will be at most 5/10 and $rel\Delta(H)$ becomes 2/4, and therefore $dp(H) \leq rel\Delta(H)$.

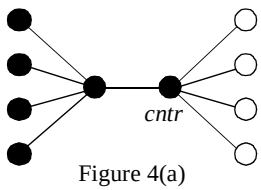

Figure 4(a)

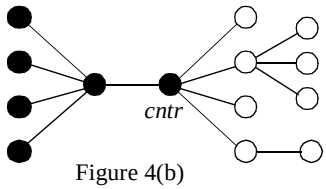

Figure 4(b)

Figure 4: *k* = 4, *d* = 2, ρ = 2, δ = 2.
Black nodes are fixed; white are unfixed.

That finishes Case I; now we can proceed to a next case.

Case II: odd $\delta > d$ (see Figure 5 for guidance). Note that since h_{d} is a fixed node of *H* at furthest distance from the center node *cntr*, it follows that $d \geq \lceil \delta/2 \rceil$. This time we consider the path of length *d*, consisting of the nodes that lead from h_{d} back to center node *cntr*; denote the nodes on this path as $h_{\mathrm{d}} = v_0, v_1, v_2, \ldots, v_d = cntr$. Subtree T_H could contain all the nodes in a full rooted *k*–ary tree of height $\lceil \delta/2 \rceil - 1$, rooted at $v_{\lfloor \delta/2 \rfloor}$; such a tree contributes $\frac{k^{\lceil \delta/2 \rceil}-1}{k-1}$ nodes to T_H. Similarly, $v_{\lceil \delta/2 \rceil}$ might have *k*–1 other children which are the roots of full rooted *k*–ary trees of height $\lceil \delta/2 \rceil - 2$, and T_H might contain all these subtrees; they would contribute $k^{\lceil \delta/2 \rceil - 1} - 1$ more nodes to T_H. Analogously, nodes can be contributed to T_H by groups of *k*–1 full rooted *k*–ary child trees, rooted at each of the nodes $v_{\lceil \delta/2 \rceil + 1}, \ldots, v_{d-1}$, and of respective heights $\lceil \delta/2 \rceil - 3, \ldots, \delta - d$. Finally, we note that the center node *cntr* might have (not *k*–1 but) *k* other children which are the roots of full rooted *k*–ary trees of height δ–*d*–1; those subtrees could contribute $k^{\delta-d} - 1 + \frac{k^{\delta-d}-1}{k-1}$ more nodes to T_H. Together with the $d - \lfloor \delta/2 \rfloor$ nodes $v_{\lceil \delta/2 \rceil}, \ldots, v_d$, we see that T_H could contain as many as $\frac{2k^{\lceil \delta/2 \rceil}-2}{k-1}$ nodes. Finally, T_H could contain as many as $N_{\mathrm{II}} = \frac{2k^{\lceil \delta/2 \rceil}-2}{k-1} - 1$

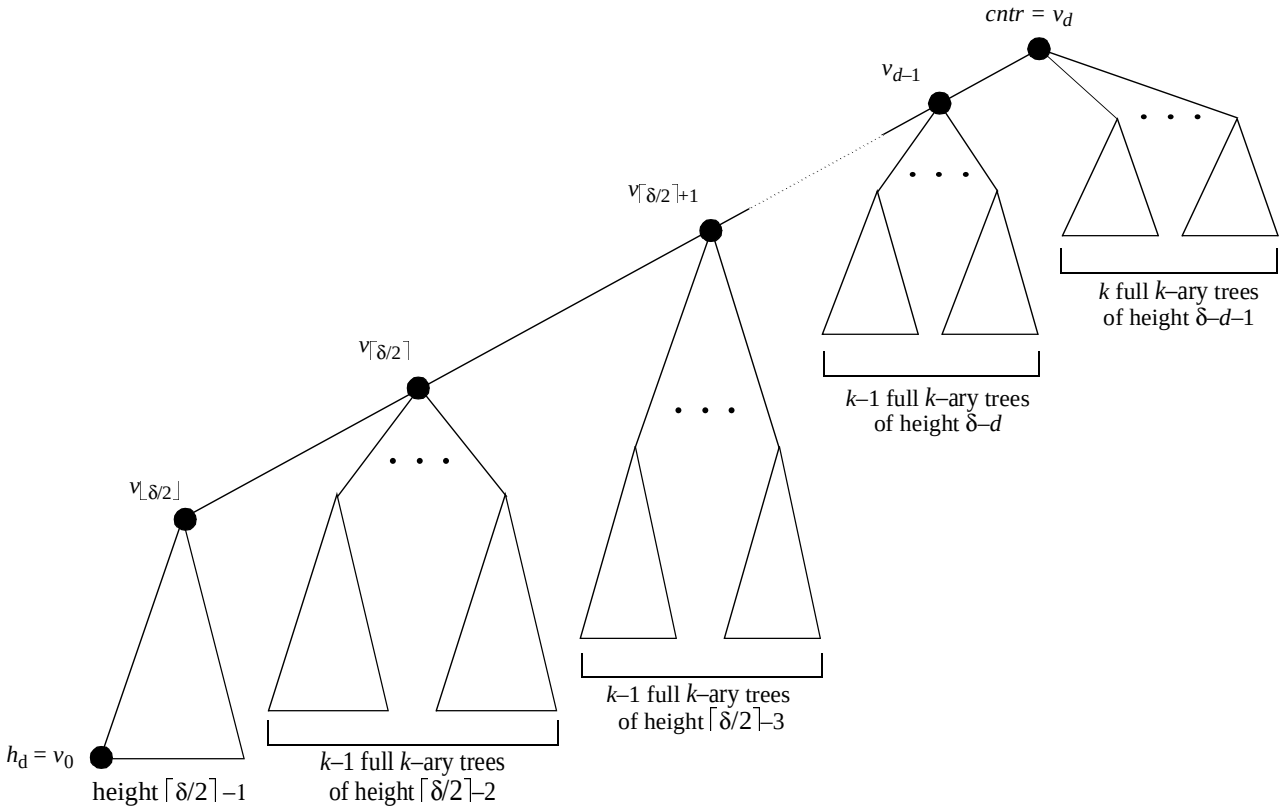

Figure 5: δ is odd, δ > *d*.

edges. Reasoning as in Case I, relation $dp(H) \le rel\Delta(H)$ will hold, provided the next two inequalities hold:

$$\frac{N_{\mathrm{II}}}{(k+1)\left(\frac{k^{\rho-1}-1}{k-1}\right)+1} \le \frac{\delta}{2\rho} \quad \text{, for odd } \delta > d \text{ when } d < \rho;$$

$$\frac{N_{\mathrm{II}}}{k^{\lfloor \delta/2 \rfloor} + (k+1)\left(\frac{k^{\rho-1}-1}{k-1}\right)} \le \frac{\delta}{2\rho} \text{, for odd } \delta > d \text{ when } d = \rho.$$

We used a computer program to examine if or when these inequalities held, for *T* radius ρ in the range from 2 to 100, and *k* in the range from 2 to 7, and found the following results. For every *k* in 2..7, the program invariably reported a failure when *d* = ρ and δ = 2*d* – 1. In fact, it can happen that *dp*(*H*) exceeds *rel*Δ(*H*) when *d* = ρ and δ = 2*d* – 1, as we now show by example, pictured in Figure 6. Let tree T_0 satisfy: its center *cntr* has one full rooted *k*–ary subtree S_0 of height ρ – 1, and has *k*–1 full rooted *k*–ary subtrees S_1, S_2, ..., S_k of height ρ–2. Let the fixed nodes of schema H be the leaves of T_0. Let tree *T* be T_0 but with an exceptional node appended to a former leaf of one of the shorter subtrees of *cntr*. The exceptional node will not be one of the fixed nodes of schema *H*. With respect to schema $H \subseteq T$, we have the following. $T_H = T_0$; the number of edges in T_H is $\frac{2k^{\rho}-k-1}{k-1}$; the number of edges in T is one more, or $\frac{2k^{\rho}-2}{k-1}$; hence, $dp(H) = \frac{2k^{\rho}-k-1}{2k^{\rho}-2}$. On the other hand, $rel\Delta(H) = \frac{2\rho-1}{2\rho}$. Since $dp(H)$ can be much closer to 1 than is $rel\Delta(H)$, we see that $dp(H)$ can exceed $rel\Delta(H)$. Let us also note that schema *H* is an atypically large one in host chromosomal tree *T*. The number of nodes in H is $k^{\rho-1} + k \cdot k^{\rho-2} = 2k^{\rho-1}$, whereas the number of nodes in T is $\frac{k^{\rho}-1}{k-1} + k \cdot \frac{k^{\rho-1}-1}{k-1} + 2 = \frac{2k^{\rho}+k-3}{k-1}$. So the ratio $\frac{\text{number of nodes in } H}{\text{number of nodes in } T} \approx \frac{(k-1)2k^{\rho-1}}{2k^{\rho}} = \frac{k-1}{k}$, which is half

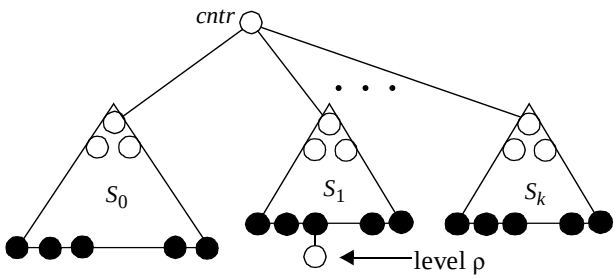

Figure 6: *d* = ρ and δ = 2*d*–1.
Black nodes are fixed nodes of *H*.

or more. We typically think of schemas (or building blocks) as being smaller than that. That ends our analysis of Case II.

Case III: odd $\delta \le d$. In this case, the numerator for our upper bound on $dp(H)$ is $N_{\text{III}} = 2\left(\frac{k^{\lceil \delta/2 \rceil} - 1}{k-1}\right) - 1$. Inequality $dp(H) \le rel\Delta(H)$ will hold provided

$$\frac{N_{\text{III}}}{(k+1)\left(\frac{k^{\rho-1}-1}{k-1}\right)+1} \le \frac{\delta}{2\rho} \quad \text{, for odd } \delta \le d \text{ when } d < \rho;$$

$$\frac{N_{\text{III}}}{k^{\lfloor \delta/2 \rfloor} + (k+1)\left(\frac{k^{\rho-1}-1}{k-1}\right)} \le \frac{\delta}{2\rho} \text{, for odd } \delta \le d \text{ when } d = \rho.$$

We used a computer program to examine if or when these inequalities held, for T radius ρ in the range from 2 to 100, and k in the range from 2 to 7, and found the following results. No failures at all were reported.

Case IV: even $\delta > d$. In this case, the numerator for our upper bound on $dp(H)$ is $N_{\text{IV}} = \frac{(k+1)k^{\delta/2} - 2}{k-1} - 1$. Inequality $dp(H) \le rel\Delta(H)$ will hold provided

$$\frac{N_{\text{IV}}}{(k+1)\left(\frac{k^{\rho-1}-1}{k-1}\right)+1} \le \frac{\delta}{2\rho} \quad \text{, for even } \delta > d,\ d < \rho;$$

$$\frac{N_{\text{IV}}}{k^{\delta/2} + (k+1)\left(\frac{k^{\rho-1}-1}{k-1}\right)} \le \frac{\delta}{2\rho} \quad \text{, for even } \delta > d,\ d = \rho.$$

We used a computer program to examine if or when these inequalities held, for T radius ρ in the range from 2 to 100, and k in the range from 2 to 7, and found the following results. There were three classes of failures. (1) The second inequality invariably fails when $d = \rho$ and $\delta = 2\rho$. But these values imply that $rel\Delta(H) = 1$, and so it is certainly as large as the probability $dp(H)$. Thus, there really is not a failure of the relation $dp(H) \le rel\Delta(H)$ for this class of report. (Our upper bound on $dp(H)$ this time evaluates to a number greater than 1, so is too generous.) (2) Only for $k = 2$, the second inequality fails for the values $d = 3$, $\rho = 3$, and $\delta = 4$. Similar to our analysis of the failure reported in Case I, in fact there is no failure of relation $dp(H) \le rel\Delta(H)$ for these values. (3) Invariably the first inequality fails when $d = \rho - 1$ and $\delta = 2d$. This can give a genuine failing of inequality $dp(H) \le rel\Delta(H)$. Now we can have trees for which $dp(H) = \frac{(k+1)k^{\rho-1} - k - 1}{(k+1)k^{\rho-1} - 2}$, whereas $rel\Delta(H) = \frac{2\rho - 2}{2\rho - 1}$. Such trees again feature schemas H which are atypically large.

■

6. CONCLUSIONS & FUTURE WORK

We have investigated schema disruption probability $dp(H)$ when chromosomes are structured as the nodes of a tree T. We have sought the existence of upper bounds for $dp(H)$ of the form $rel\Delta(H)$, or form $c \cdot rel\Delta(H)$ for some constant c. Three of the results in this paper are the following. There is no constant c for which relation $dp(H) \le c \cdot rel\Delta(H)$ holds for arbitrary schema H in arbitrary tree T. For a particular H, there is a number c_H which depends upon H and for which $dp(H) \le c_H \cdot rel\Delta(H)$. For tree arity k in the range 2..7, and tree radius ρ in the range 2..100, relation $dp(H) \le rel\Delta(H)$ holds for all schemas $H \subseteq T$ and all complete ball-like trees T of degree k+1 and radius ρ, with the exception of certain schemas H which contain atypically large numbers of fixed positions. Finally, a lesson to be drawn is that the more ball-like and full-ish are our tree-structured chromosomes, the more likely are our GAs to work well.

As future work we will investigate chromosomes with other alternative structures, such as grids and tori, and investigate the existence of upper bounds on $dp(H)$ for schemas in such chromosomes.

7. REFERENCES

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

[2] Greene, William A. (2000). "A Non-Linear Schema Theorem for Genetic Algorithms," in Whitley, D. *et al*. (Eds.) *Proceedings of the Genetic and Evolutionary Computation Conference* (GECCO-2000), pages 189-194. Morgan Kaufmann Publishers, San Francisco, CA.

[3] Greene, William A. (2004) "Schema Disruption in Chromosomes that are Structured as Binary Trees", in K. Deb *et al*. (eds.), *Proceedings of the Genetic and Evolutionary Computation Congress*, June 26-30, 2004, Seattle, WA (pages 1197-1207). Springer Verlag LNCS 3102, New York, NY.

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

[5] Koza, John R. (1992). *Genetic Programming: On the Programming of Computers by Natural Selection.* MIT Press, Cambridge, MA

[6] O'Reilly, Una-May (1995). *An Analysis of Genetic Programming*. PhD thesis, Carleton University, Ottawa-Carleton Institute for Computer Science, Ottawa, Ontario, Canada, 22 September 1995.

[7] O'Reilly, Una-May, and Franz Oppacher (1995). "The Troubling Aspects of a Building Block Hypothesis for Genetic Programming", in Whitley, D. and Vose, M. D. (eds.) *Foundations of Genetic Algorithms 3*. Morgan Kaufmann Publishers, San Francisco.

[8] Poli, Riccardo, and William Langdon (1998). "Schema Theory for Genetic Programming with One-Point Crossover and Point Mutation". *Evolutionary Computation* 6(3), pages 231-252. MIT Press, Cambridge, MA.

[9] Rosca, Justinian P. (1997). "Analysis of Complexity Drift in Genetic Programming", in Koza, John R. *et al*. (eds) *Genetic Programming 1997: Proceedings of the Second Annual Conference* (pages 286-294). Morgan Kaufmann Publishers, San Francisco, CA.

[10] Whigham, Peter A. (1995) "A Schema Theorem for Context-Free Grammars", in *1995 IEEE Conference on Evolutionary Computation*, Vol 1, pages 178-181. IEEE Press.