Estimating the Destructiveness of Crossover on Binary Tree Representations

Luke Sheneman
University of Idaho
Moscow, ID 83843
+1 (208) 882-3506
sheneman@cs.uidaho.edu

James A. Foster
University of Idaho
Moscow, ID 83843
+1 (208) 885-7062
foster@uidaho.edu

ABSTRACT

In some cases, evolutionary algorithms represent individuals as binary trees with *n* leaves and *n-1* internal nodes. When designing a crossover operator for a particular representation and application, it is desirable to quantify the operator's destructiveness in order to estimate its effectiveness at using building blocks. For the case of binary tree representations, we present a novel approach for empirically estimating the destructiveness of any crossover operator by computing and summarizing the distribution of Robinson-Foulds distances from the parent to the entire neighborhood of possible children. We demonstrate the approach by quantifying the destructiveness of a popular tree-based crossover operator as applied to the problem of phylogenetic inferencing. We discuss the benefits and limitations of the destructiveness metric.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Trees; I.2.8 [**Problem Solving, Control Methods, and Search**]: Graph and tree search strategies

General Terms

Algorithms, Performance, Design

Keywords

Crossover, Destructiveness, Robinson-Foulds, Trees

1. INTRODUCTION

Evolutionary algorithms can operate directly on many data structures, including trees and graphs. Traditional genetic programming (GP) [1] operates on parse trees and is the most cited example of tree-based evolutionary computation. Recent theoretical work formally establishes the ability of a GP to find and exploit building blocks for at least a narrow range of representations and crossover operators [2,3]. The considerable difficulty in formalizing a schema theory for arbitrary crossover operators explains why there is no general theory for tree-based genetic algorithms (GAs).

Genetic algorithms can infer evolutionary relationships between organisms using simple binary tree representations [4,5,6]. A phylogeny expresses the evolutionary relationships between a set of organisms, and this relationship is often represented as a bifurcating (i.e. binary) tree. The GAs search the space of valid phylogenetic trees, using the common optimality criteria of maximum parsimony or maximal likelihood. The previously cited examples of phylogenetic inferencing GAs use distinct and novel crossover operators, yet all of them operate on the same underlying binary tree structure. Typically genetic algorithms achieve mixed results in phylogenetic inferencing. Most biologists use simpler hill-climbing approaches with acceptable levels of success. The lack of real-world adoption of GAs in this domain is due to the difficulty in understanding the GA mechanics of assembling higher-order building blocks from lower-order tree sub-structures. Additionally, the theoretical mechanisms by which such a tree-based GA is capable of efficiently searching phylogenetic tree space are far from understood. However, by establishing a technique for empirically quantifying the destructiveness of a particular tree-based operator, one may be able to elucidate the mechanisms by which building blocks are constructed and disrupted.

Several metrics measure the distance between any two trees. The most ubiquitous tree distance metric is the Robinson-Foulds (RF) approach [7] as shown in Fig 1. RF acknowledges that every internal branch partitions a tree into two sets of terminal nodes. RF enumerates all bipartitions for both trees, throws out duplicate bipartitions, and counts the remaining unique bipartitions. The RF distance is simply the number of remaining non-duplicate bipartitions. Robinson-Foulds is a true mathematical measure of distance, largely because it is commutative and satisfies the triangle inequality [8]. In other words, the distance from A to B is equal to the distance from B to A, and the direct distance from A to C is always less than the distance from A to B to C.

The building blocks for tree representations are topological features (e.g. sub-trees and other patterns). Robinson-Foulds estimates the distance between two binary tree topologies. Thus, RF is a reasonable measure of the destructiveness of crossover operators working on binary tree representations.

GECCO'06, July 8–12, 2004, Seattle, WA, USA.

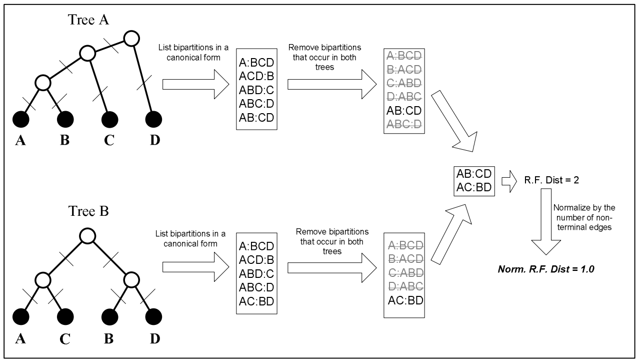

Tree A
List bipartitions in a canonical form
A:BCD
ACD:B
ABD:C
ABC:D
AB:CD
Remove bipartitions that occur in both trees
~~A:BCD~~
~~B:ACD~~
~~C:ABD~~
~~D:ABC~~
AB:CD
~~ABC:D~~
A
B
C
D
AB:CD
AC:BD
R.F. Dist = 2
Normalize by the number of non-terminal edges
Norm. R.F. Dist = 1.0
Tree B
List bipartitions in a canonical form
A:BCD
ACD:B
ABD:C
ABC:D
AC:BD
Remove bipartitions that occur in both trees
~~A:BCD~~
~~B:ACD~~
~~C:ABD~~
~~D:ABC~~
AC:BD
A
C
B
D

Figure 1. We Compute the normalized Robinson-Foulds distances between two 4-taxon trees in this example.

To use RF as a destructiveness estimate, we first generate a *crossover neighborhood*. A crossover neighborhood is the exhaustive set of child trees that result from applying a crossover operator to two parent trees in all possible ways. Therefore the trees in the crossover neighborhood are one crossover step away from their parents. We generate a distribution of Robinson-Foulds distances by computing the RF distance between every child to each of its parents. We estimate the destructiveness of the operator by analyzing and summarizing the RF distance distribution. For example, child trees that have very little topological resemblance to either of their parents have been transformed via crossover to such a degree that most meaningful building blocks have likely been destroyed. We can quickly identify extremely biased and destructive crossover operators by finding heavily-skewed distance distributions.

2. METHODS

We randomly generated several pairs of parent binary trees, each with ten leaf nodes. For each set of parents, we exhaustively applied Lewis's crossover operator [4] at every valid crossover and insertion point (see Fig 2) to generate the crossover neighborhood under Lewis's crossover operator. We subsequently calculated the normalized Robinson-Foulds distance from every child to each of the original parent trees. This resulted in two distributions: one distribution of distances of child to parent for each of the two parents. We summarized both distributions and visualized them as histograms.

For perspective and as a form of experimental control, we also generated a random tree and then generated a set of random trees of approximately the same magnitude as the neighborhood of children resulting from the application of Lewis's operator. We computed the distribution of Robinson-Foulds distances from each of the randomly generated trees to the original tree. This worst-case distribution represents the case wherein no systemic topological similarity exists between children and parents. Highly-destructive recombination produces distance distributions that approach this random, worst-case scenario. Therefore, this worst-case distribution serves as a coarse form of experimental control.

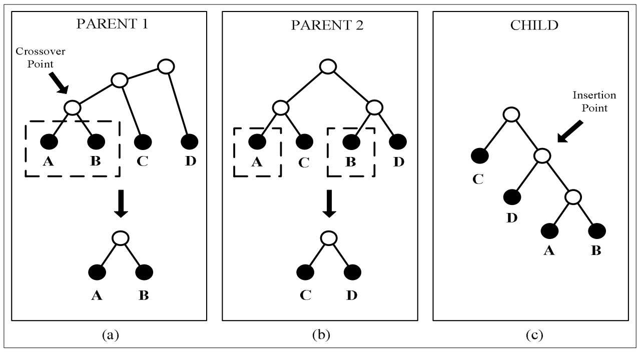


Figure 2. Lewis crossover takes two parent trees. A node known as a crossover point is identified in the first parent tree (a) and the remainder of the tree is pruned, leaving only the crossover point and all of its children nodes. From the second parent (b), all of the nodes which were selected from the first parent are pruned from the tree. In (c), the two remaining subtrees are recombined at an insertion point, creating a child tree which should share some topological characteristics of the parents.

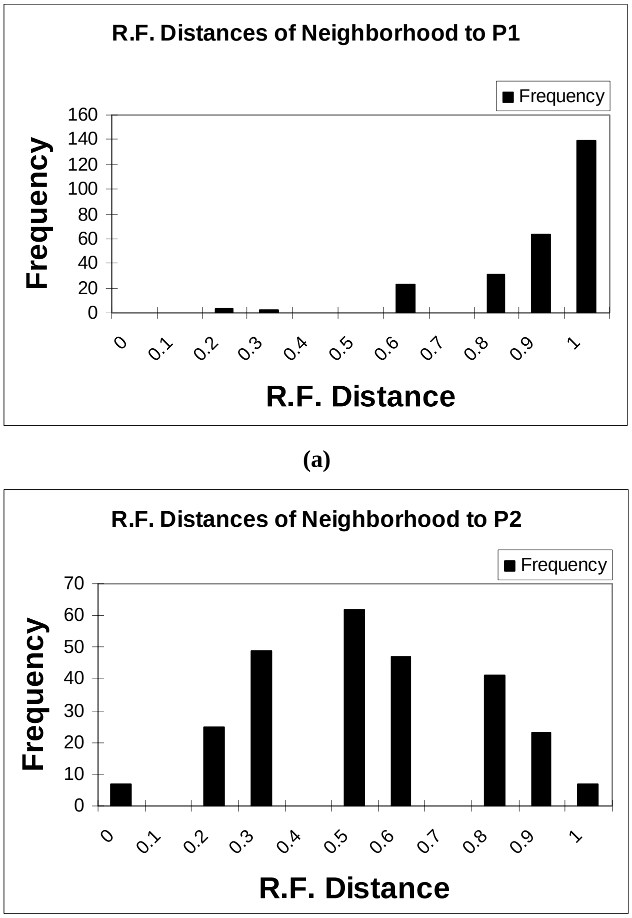
3. RESULTS

We discovered some surprising, yet easily explained details about the underlying mechanisms and overall destructiveness of Lewis's crossover operator. The two distributions of the crossover neighborhood vs. each parent tree were significantly different in mean, variation, and overall shape as shown in Fig. 3. The obvious conclusion to be reached from this result is that Lewis's crossover operator is consistently far more destructive to the first parent than the second. This result indicates a systematic bias in the way in which Lewis's crossover uses potential building blocks from each parent: Lewis's crossover operator preserves many of the topological features (i.e. building blocks) of the second parent and yet destroys almost all of the building blocks from the first parent.

This asymmetrical crossover destructiveness occurs for the simple reason that in binary trees, the majority of nodes are closer to the terminals of the tree. In fact, approximately half of all nodes are terminals, and one-quarter of all remaining nodes directly share an edge with a terminal. This introduces a dramatic bias in the way in which nodes are selected for crossover points from the first parent. A review of the source code for Lewis's GAML program indicates that this bias is not programmatically removed from his crossover algorithm. Because of this stochastic bias of choosing nodes near the leaves of the first parent as crossover points, Lewis's crossover preserves only single nodes or very small subtrees from the first parent. Lewis's crossover attaches these small features preserved from the first parent to a minimally pruned second parent. This results in children that almost always resemble one parent far more than the other.

4. DISCUSSION

A distribution of Robinson-Foulds distances gives an estimation of the destructiveness of a particular crossover operator. Different crossover operators have different amounts of destructiveness. Obviously, with a GA, every operator is destructive to some extent. Good GA designers must balance crossover destructiveness so that an operator is not so destructive that it throws away everything worthwhile from both parents. Likewise, an operator which is too preservative is equivalent to a small, local search step which is also unlikely to be effective.



(a)

(b)

Figure 3. This is a representative pair of distributions found when analyzing the destructiveness of Lewis's crossover operator. In each of our experimental runs, we found that in general the distances from the neighborhood to the first parent (a) were far greater (*mean=0.878, stdev=0.169*) than the distances of the neighborhood to the second parent (b) (*mean=0.487, stdev=0.234*). The shapes of the distributions were always consistent, with (a) being largely exponential in shape, while (b) was largely normal in shape.

We applied the Robinson-Foulds distance metric to the crossover operator that Lewis implemented in GAML. We found some surprising results which gave us insight not only into how effective we can expect Lewis's crossover operator to be in general, but it allowed us to trivially identify the underlying biases present in the crossover operator which would prevent it from behaving efficiently.

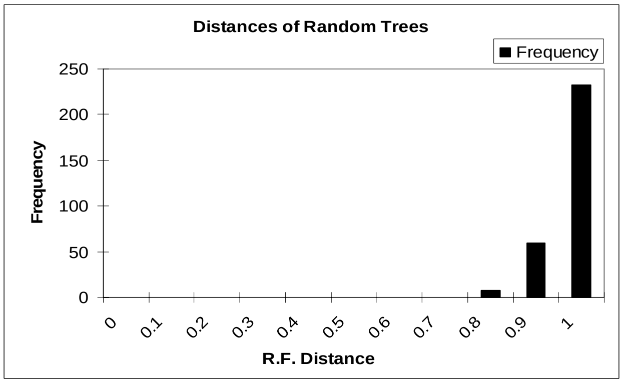


Figure 4. The distribution of Robinson-Foulds distances from one randomly generated tree to a large group of other randomly-generated trees. This distribution is useful as an experimental control, in that it shows the worst-case distribution of scores for an operator which is as maximally destructive as a random step. The mean R.F. distance was 0.964, with a standard deviation of 0.07.

Robinson-Foulds has limitations. The Robinson-Foulds distance estimate is heavily influenced by the topology of the trees being compared. For example, seemingly small, individual changes to pectinate (maximally deep) topologies can result in maximum normalized distances (i.e. RF distance of 1.0) and yet these large distances are not possible in one step when starting with balanced (perfect trees). Ideally, one would discard Robinson-Foulds and instead use a distance metric that estimates the minimum edit distance between tree topologies (i.e. the cost of transforming one tree to another via the transformation operator). Unfortunately, the general problem of determining the minimum edit distance is NP-complete. In addition, for our use in generating a distribution of distances between parents and all possible children, such an edit-distance approach is non-informative (all distances would be 1), thus using Robinson-Foulds as a topological distance estimate is a rational approach. Finally, since we are dealing with randomly generated trees which are, taken as a whole, closer to perfectly balanced than pectinate, the chances of dealing with the topology-oriented biases of Robin-Foulds are limited.

5. ACKNOWLEDGMENTS

We thank Jason Evans from the University of Idaho for fruitful discussions. Foster was partially funded by NIH NCRR 1P20 RR16448. Sheneman was funded by NIH P20 RR16454 from the INBRE Program of the National Center for Research Resources.

6. REFERENCES

[1] Koza, J. *Genetic Programming: On the programming of computers by means of natural selection*. MIT Press, Cambridge, Massachusetts, 1992.

[2] R. Poli and W.B. Langdon. Schema theory for genetic programming with one-point crossover and point mutation. *Evolutionary Computation*, 6(3): 232-252, 1998

[3] O'Reilly, U.M., Oppacher, F., The troubling aspects of a building block hypothesis for genetic programming. In L. Darrell Whitley and Michael D. Vose, editors, Foundations of Genetic Algorithms 3, pages 73-88, Estes Park, Colorado, USA, 31 July - 2 August 1994. (1995). Morgan Kaufmann

[4] Lewis, P.O., A Genetic Algorithm for Maximum Likelihood Phylogeny Inference Using Nucleotide Sequence Data, *Mol. Biol. Evol*. 15(3):277-283. 1998

[5] Matsuda, H. 1996. Protein phylogenetic inference using maximum likelihood with a genetic algorithm. Pp. 512–523 in L. Hunter and T. E. Klein, eds. Pacific Symposium on Biocomputing '96. World Scientific, London

[6] C. B. Congdon, "Gaphyl: An Evolutionary Algorithms Approach for the Study of Natural Evolution", Genetic and Evolutionary Computation Conference (GECCO-2002), New York, NY, July 2002

[7] Robinson, D.R., and Foulds, L.R. Comparison of phylogenetic trees. Mathematical Biosciences 53: 131-147, 1981.

[8] Felsenstein, J. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts, 2004