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 typical 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 operate directly on diverse data structures, including trees and graphs. Genetic programming (GP) [1] operates on parse trees and is the most common 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]. There is no general schema theory for arbitrary tree representations and crossover operators.

Genetic algorithms (GAs) can infer evolutionary relationships between organisms using simple binary tree representations [4,5,6]. A GA searches the space of valid phylogenetic trees, using the common optimality criteria of maximum parsimony or maximal likelihood. Typically, such programs use different crossover operators, but all of them operate on the same underlying binary tree structure. GAs typically achieve mixed results in phylogenetic inferencing. Most biologists use simpler hill-climbing approaches with success. The lack of real-world adoption of GAs in this domain is likely due to the difficulty in establishing and quantifying the effectiveness of a GA to efficiently assembling building blocks on tree structures. Additionally, the GA theory in this domain remains lacking. An empirical technique for quantifying the destructiveness of a tree-based operator may help elucidate the mechanisms by which building blocks are identified, recombined, and disrupted.

The most ubiquitous metric is the Robinson-Foulds (RF) approach [7] shown in Fig. 1. RF acknowledges that internal branches partition any 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 satisfies the triangle inequality.

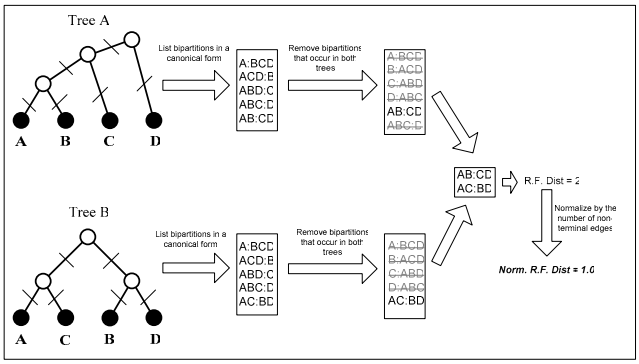


Figure 1. Computing normalized Robinson-Foulds distances.

Building blocks in any tree representation are mainly topological features such as subtrees. RF estimates the distance between the topologies of two trees by roughly counting shared topological features. Using RF, we quantify shared topology between crossover parents and offspring. We expect that highly-disruptive crossover results in offspring which do not share significant common topology with its parent trees. RF estimates the quantity of building block disruption due to crossover by measuring topological distance between each child and its parents. We

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.
ACM 1-59593-186-4/06/0007.

compute a distribution of RF distances by generating a crossover neighborhood and computing the RF distance between each parent and every child in the neighborhood. A summary of this distribution provides an estimate of crossover destructiveness, since it quantifies the topological distances between parents and all their children under the specific crossover operator. Overly destructive crossover operators are easily identified by summarizing or visualizing distributions of Robinson-Foulds distances.

2. METHODS

We randomly generated several pairs of parent binary trees, each with ten leaf nodes. For each pair of parents, we exhaustively applied Lewis's crossover operator at every valid crossover and insertion point and generated the neighborhood of all possible offspring to the two parent trees under Lewis's crossover. We then calculated the normalized Robinson-Foulds distance from every child to each of the original parent trees. This resulted in two distributions: one for each parent. Both distributions were summarized and plotted as a histogram.

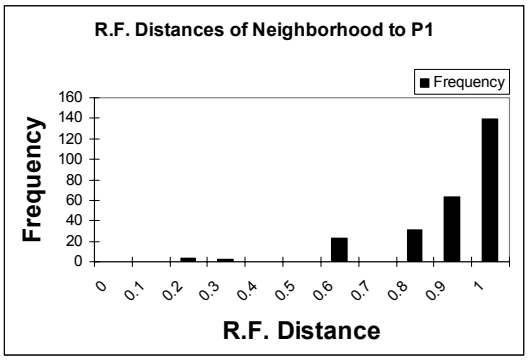
In addition, we generated a set of randomly constructed trees. We computed the distribution of RF distances from each of the randomly generated trees to a single random tree. This distribution represents the worst-case for which the neighborhood should have no significant topological similarity to the original tree. The distribution of distances from trees generated by highly-destructive recombination should approach this worst-case distribution. Thus, this worst-case distribution is used as a form of experimental control and distribution comparison.

3. RESULTS

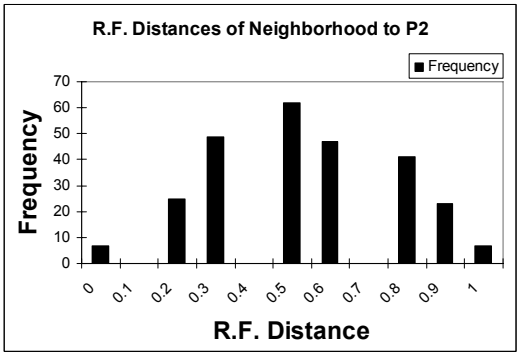
We discovered unexpected details about the underlying bias and overall destructiveness of Lewis's crossover operator. The two distributions (neighborhood vs. parent) distributions differ significantly in mean, variation, and overall shape as shown in Fig. 2. We conclude that Lewis's crossover operator is consistently more destructive to one parent tree versus another. 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 most of the topology (i.e. building blocks) of one parent and destroys almost all of the building blocks from another parent. This asymmetrical disruption occurs stems from the fact 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. We end up pruning the majority of one tree and leaving the majority of another tree. This reduces Lewis's crossover to a very local search relative to one tree.

4. ACKNOWLEDGMENTS

We thank Jason Evans from the University of Idaho for fruitful related 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.



(a)



(b)

Figure 2. Lewis's operator asymmetrically disrupts one parent (a) more than another (b).

5. REFERENCES

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

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

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

[4] 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

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

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

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