NETRECONSTRUCT Release Notes
Kappa Version
(Reconstruction Application for Phylogenetic Networks)

Monique M. Morin

Department of Computer Science
University of New Mexico

April 4, 2008

Abstract

NETRECONSTRUCT is an application designed for the purpose of reconstructing phylogenetic networks which contain a single diploid hybrid. The underlying reconstruction algorithm is a three-phased approach addressing different subtrees which are contained in a phylogenetic network. Internal nodes are reconstructed and assigned sequences using such existing techniques as parsimony, neighbor joining, and Fitch small parsimony. Options are available for altering certain aspects of the reconstruction algorithm and a summary report is provided. This document provides an overview of the software, installation, and execution instructions including a sample input file.

Contents

1 Introduction

This document is intended to accompany the NETRECONSTRUCT software which is currently available for download at http://www.cs.unm.edu/∼morin/ and is released under GNU General Public License (GNU GPL).[1]

The software is an application for reconstructing phylogenetic networks with diploid hybrids. This initial version of the code is written to address one diploid hybrid event in the network.

This software is command line driven, written in C, and developed in a Debian Linux environment. Its operation on/in other platforms/environments has not been tried. Theoretically, interested parties should be able to use this software on any Linux/Unix platform which has a standard C/C++ compiler.

A paper describing this software is currently under preparation for submission. Until a journal or conference publication is accepted, please cite NETRECONSTRUCT by referencing this document which is a technical report for the Department of Computer Science, University of New Mexico.

Caveat: This Network Reconstruction software (NETRECONSTRUCT, NR) is currently available as source from http://www.cs.unm.edu/∼morin/. The software is constantly under development as part of a research effort and is offered only "As-Is." There is no guarantee, written or implied, of the software being bug-free or reliable and no liability related to this software will be accepted.

2 Overview

NETRECONSTRUCT is a three-phase algorithm for reconstructing a phylogenetic networks with a single diploid hybrid. Each phase constructs a subtree that is then incorporated to the final overall network topology. Sequences are assigned to internal nodes using the Fitch small parsimony algorithm. [2] Currently the subtrees are created by PHYLIP [1], which is a common tree reconstruction tool.

[1] see http://www.gnu.org/licenses for license details

2.1 Input

NETRECONSTRUCT requires a strictly specified input file which includes such information as extant taxa and identification of taxa believed to be impacted by the historic hybrid.[2] While not always possible, it is known that biologists in certain fields have "markers" which lead them to them to suspect which taxa from a given set are descendants of a hybrid event. This application is intended for such situations. The user is required to identify these "hybrid-impacted extant taxa" in addition to outgroups (one for the network and one for the "hybrid subtree" though the same taxon can be used for each). Extant taxa are listed with identification numbers in the input file. These values are external values and separate internal values are generated. The final report contains both values for easy reference.

2.2 Reconstruction Algorithm

Phase one of the algorithm is to reconstruct what is referred to as the "hybrid subtree." Using the inputted information of which extant taxa were impacted by the hybrid, a subtree, rooted at the hybrid node is reconstructed.

Figure 1 reflects the topology of this first phase. The impacted extant taxa are shown inside the dashed box. The greyed taxon is the extant taxa which was identified by the user as the outgroup for the hybrid subtree. The historical structure is reconstructed by PHYLIP (see 2.3 for details). The next step is to assign sequences to these nodes using the Fitch small parsimony approach.[3]

The next step in this phase is to remove the outgroup, and add parent nodes to

[2] A random number seed is specified on the command line for repeatability purposes. The random numbers used in this code are generated by the source code known as **Mersenne Twister**. [3] This random number generator is known for its high periodicity. The creators provided .c and .h files for incorporation with other code.

[3] Only the preliminary phase of Fitch is implemented as the final phase allows the generation of all parsimonious assignments and adds realism to internal nodes provided certain evolutionary assumptions hold. Our need for realistic sequences is at the root level of the hybrid subtree (which Fitch calls the ultimate node) and these base options are not changed by the final phase.

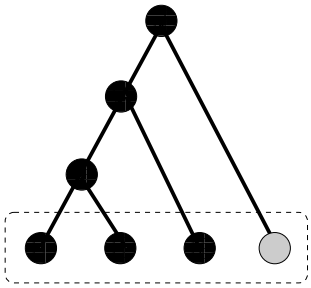

Figure 1: Initial topological result from phase 1 of the NETRECONSTRUCT algorithm.

the hybrid, including randomly splitting the sequences in each homologous chromosome to the two parents which is shown in Figure 2.

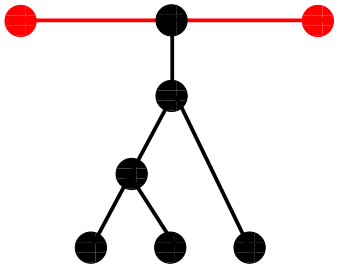

Figure 2: Final topological result from phase 1 of the NETRECONSTRUCT algorithm – hybrid subtree.

Phase two of the algorithm is when the "parental subtree" is reconstructed. As part of the input file, the method by which the extant taxa that will be descendants of the hybrid's parents is specified.[4] An external call is made to PHYLIP for

[4]The user specifies custom (where these taxa are listed individually at time of input), or extreme_custom (where two separate sets (one for each parent) are provided), or closest neighbor (where taxa whose closest neighbor is a taxon impacted by the hybrid are chosen by the program), or closest 2x hybrid (where double the number of hybrid impacted taxa which are the closest to the reconstructed hybrid node from phase 1 are chosen by the program). The default option is the closest neighbor approach and all scoring of "closeness" for these options is done by calculating the

reconstructing just this parental subtree (rooting with the outgroup identified for the network) and the result is shown in Figure 2.2. The final step of this phase is to assign each primary lineage (those lineages starting effectively at the root, but not the outgroup) to either the first or second parent of the hybrid. This association determines which clades will be affiliated with each parent and is determined by calculating an average hamming distance between the extant taxa of the clade and each parent.[5]

The next step is an intermediate phase which takes the decision of with which parent each primary lineage will affiliate and connects the hybrid and parental subtrees together, eliminating the parental root and outgroup from this last step. In this case, the subtree with two extant taxa were closest to the right parent/parent 2 and the single taxon subtree to the left parent/ parent 1. Figure 3 illustrates this stage.

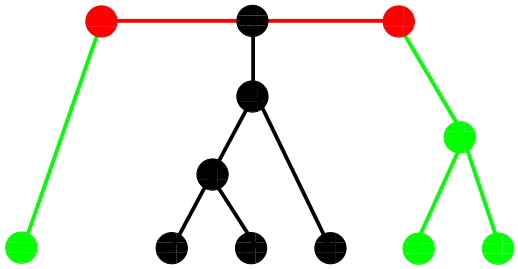

Figure 3: Final topological result from phase 2 of the NETRECONSTRUCT algorithm.

Finally PHYLIP is used to reconstruct a history for the remaining extant taxa AND the two parent nodes created earlier. Once that topology is known, the Fitch

total hamming distance. Note that the outgroup for the network is not eligible to become one of the parental offspring by definition because it must be used for rooting the network.

[5]Note the parent nodes only have a half set of sequences, that is why an average (not total) hamming distance is calculated.

small parsimony algorithm is used to assign the sequences for all but the hybrid subtree.[6] Our remaining structure looks something like the following Figure 4.

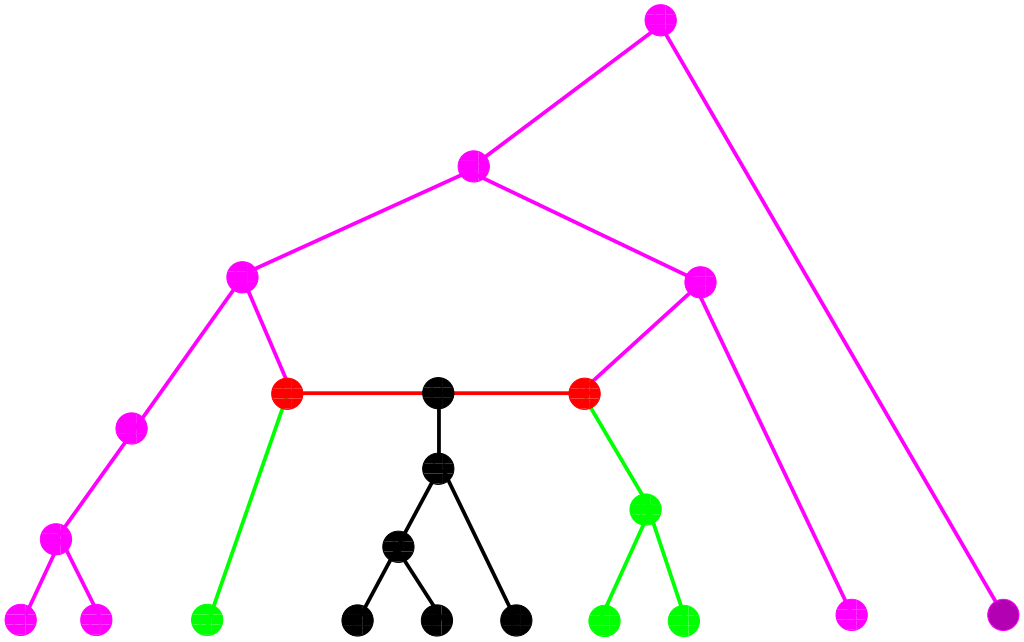

Figure 4: Final topological result from NETRECONSTRUCT. The patterned taxon is the outgroup for the network and each color is carried over from the prior figures representing the major steps.

2.3 PHYLIP Related Issues and Parameters

PHYLIP offers a variety of tree reconstruction routines. NETRECONSCTRUCT sets a threshold (9, not counting the outgroup, unless otherwise specified by the user) which determines whether the maximum parsimony routine (DNAPENNY) or the distance method of neighbor joining (NEIGHBOR) is used for the subtree reconstruction. (The rule of thumb from the PHYLIP documentation is that DNAPENNY

[6]The parents of the hybrid have half their sequences, but their remaining ones are assigned at this time.

works well for 10 or less taxa.) If DNAPENNY is used, it is very likely that there will be multiple, equally parsimonious trees returned, therefore the consensus routine (CONSENSE) using the extended majority rule algorithm is always used after the parsimony reconstruction. On the flip side, if NEIGHBOR is used, a distance matrix is required which is produced by DNADIST.

A critical input to the DNADIST routine of PHYLIP is which evolutionary model to use. PHYLIP has a choice of four main models (Jukes-Cantor (JC), F84, K2P, and LogDet) with a variety of parameters for each. As part of the input for NETRECONSTRUCT, the user can specify which of the four models to be given to DNADIST (if not specified, the default is Jukes-Cantor). These options however assume the standard parameter values provided by PHYLIP for factors such as base frequencies, and transition/transversion ratio. If a user wishes to use a customized set of parameters, "custom" can be entered as the `dna_dist_model` type and the name of a customized option file for DNADIST provided. (See section 4.2 for further details on how to request these options.)

As phylogenetic networks are rooted by definition, NETRECONSTRUCT treats all topologies as rooted. However, PHYLIP's neighbor joining provides *unrooted* trees. When these topologies are treated as rooted, it is common for the implied root to have three children, one of which is the outgroup. In order to remove any difficulty/confusion, an automatic outgroup adjustment is performed by NETRECONSTRUCT (which by default is turned on, but it can be turned off by a parameter in the input file). Specifically this adjusts the subtree topologies so the hybrid node has one offspring and that the overall root of the network has exactly 2 children. Note however that this does not prevent polytomies from occurring at other levels.

2.4 Output

A summary report of the network is provided including a listing of all the nodes and their sequences. The user is cautioned that while the sequences represent a somewhat parsimonious solution, it may not be the *optimally* parsimonious as neighbor joining (a distance not maximum parsimony technique) is possibly used

by PHYLIP (see section 2.3) AND except for minimizing change, the internal sequence assignment does not make assumptions about preference for how the change occurs (i.e. parallel vs. backward changes) as the final rule phase of Fitch is not implemented. The output report also lists the threshold used for PHYLIP (determining whether dnapenny or neighbor joining is used) as well as the type of reconstruction (PHYLIP or RANDIP discussed in the input section below) and which model was specified to PHYLIP for the distance calculation work.

The topology aspect is reported using the *modified* Newick format. [4] The topology is reported twice – once using internal ids and once using external ids. Note that the external id value for all non-extant taxa is -1. Intermediate output files (mainly the results from PHYLIP are placed in the *execs* directory and are identified by names containing "hybrid subtree," "parental subtree", etc. There are also files starting with "rc" which are "run capture" files from the PHYLIP calls. These are useful to review if an unexpected problem occurs during the PHYLIP execution.

The parsimony score of the reconstructed network is also calculated and reported. This value is simply the measurement of how much change has occurred to each strand along the branch lengths of the network.

On a final note, the user should realize that while the provided example is binary in nature, the NETRECONSTRUCT code handles polytomies as it is possible for PHYLIP to return a structure that is not refined to the binary level, especially when the CONSENSE routine is used after NEIGHBOR.

3 Installation Guidelines

NETRECONSTRUCT is available as a compressed tar file. The following section discusses the details of code execution. It can be uncompressed and untarred in the normal ways. (e.g. unzip.....tar -x....)

When the code is compiled in the *source* directory (using the provided Makefile), the executable NR is automatically placed in the *execs* directory. There are multiple intermediate files which are created during NR's execution and these are

placed in the *execs* directory as well. These files are automatically overwritten with each fresh run of the code. NR requires the freely available software PHYLIP. [1] (See http://evolution.genetics.washington.edu/phylip.html for further information about PHYLIP.) and it needs to be executable from the *execs* directory.

4 Execution Guidelines

4.1 Execution Summary

The standard command line format for running NETRECONSTRUCT from the execs directory is:

- `./NR -i` *input_file* `-o` *output_file* `-r` *seed* `[>` *run_capture*`]`

For example:

- `./NR -i nr_simple.in -o nr_simple.out -r 821> nr_simple.run_capture`

The first six items after the executable are required; the user must specify the input and output files and a random number seed in this format. It is also recommended that STDOUT (standard out) is redirected as in the above format and example. All debugging and error messages are directed to STDOUT and it is best to capture these in a separate file. If an output file is not generated, the program most likely aborted due to some error. Refer to the run_capture file for messages.

4.2 Sample Input

The input file, which is specified on the command line, is a simple text file containing the parameters for a single run. A very simplified input file provided with the download is called `nr_simple.in` and is:

```
num_extant_total 6
num_hom_chromos 2
num_strands_per_hom_chrom 2
sequence_length 2
```

```
outgroup_taxon_id 5
mp_phylip_thresh 9
dna_dist_model jc
recon_type_PHYLIP
outgroup_adjustment 1
extant_taxa_listing
5
CT
CT
AT
AT
7
GT
GT
GT
GT
4
GT
AT
TT
TT
3
AG
GG
CG
AG
2
GG
GG
TG
```

```
GG
1
AC
TC
AC
GC
# subtree listing info
begin_subtree_listing
subtree_outgroup_id 5
num_hybrid_impacted_taxa 1
3
parental_offspring_id_method closest_neighbor
end_subtree_listing
```

Some tips for creating input files:

- Use comment lines (those started with # to annotate data.
- Keep the input order the same as the sample file – there are lots of checks and abort situations because input data is not listed in the correct order.
- Utilize unique ids for the extant taxa, but though this is not enforced, however the results will be adversely affected.
- Each sequence must be on its own continuous line with no spaces.
- Outgroup for the network cannot be an extant taxa impacted by the hybrid or its parents – this is checked and enforced.
- Options for parental_offspring_id_method are 1) closest_neighbor (default if this line is not specified), 2) custom which then requires the listing of which extant taxa are impacted by the parents, 3) closest_2x_hybrid, and 4) extreme_custom. (The details of these options are discussed in the prior algorithm section.)

- The value for mp_phylip_thresh must be 2 or greater as it does not make sense for PHYLIP to run neighbor joining on anything less than three and as the outgroup counts as 1, the minimum here is 2.

- If the mp_phylip_thresh is set too high, PHYLIP may simply give up on the reconstruction – this results in an empty outtree file and messes up NR as it relies on the outtree file. The outfile from PHYLIP does note the problem, but that is not checked, and is overwritten. So the best approach is not to use anything over the default of 9.

- The outgroup_adjustment flag value must be 0 or 1. The flag controls how the outgroup for the hybrid and remaining trees are handled. (Outgroup for the parental subtree does not need to be addressed as the root is not used and the outgroup for it is automatically removed as the subtrees are split up.) If set to 1 (the internal default) the topology is adjusted so the final root has two children, the outgroup and the subtree – this is instead of the possible 3 or more children of the root in the polytomy case sometimes generated by neighbor joining as it technically reconstructs unrooted trees. For the hybrid subtree root, the outgroup is removed and the topology is adjusted to guarantee that the hybrid has exactly one immediate descendant lineage. Note however, in both cases, that this does not eliminate polytomies from occurring at other levels, just the first "root" level for the overall tree and the hybrid subtree.

- For analysis purposes only, it is possible to reconstruct the subtrees using a program called RANDIP which is included with the NR release. If the parameter recon_type_RANDIP is specified, subtrees will be constructed as repeatedly putting two taxa together (chosen randomly) as siblings until complete. The one piece of knowledge that is known is the outgroup and it is reconstructed as such – namely the last to be hooked into the reconstructed tree. This is needed to preserve the reconstruction algorithm of how the subtrees are put together in NR.

- The dna_model_dist parameter can take “jc”, “f84”, “k2p”, “LogDet”, and “custom” as arguments. The first four result in the default PHYLIP options for the respective models when DNADIST is called. If “custom” is specified, it must be followed by a file name and that file must be a valid script file that can be read by PHYLIP. Note that the input file for PHYLIP in this situation must be specified as “phylip_infile”. A sample file (custom_jc_standard) implementing the standard jukes cantor is provided with the download.

The output files resulting from this input are included in the download as nr_simple.out and simple_run_capture. See the output subsection in the overview discussion for further details.

5 Acknowledgements and Contact

This work has been supported by the National Science Foundation under grants IIS 01-21377, DEB 01-20709, and EF 03-31654.

The author of this software and document can be contacted by the following means:

morin@cs.unm.edu

OR

Monique Morin
University of New Mexico
Department of Computer Science
Mail stop: MSC01 1130
1 University of New Mexico
Albuquerque, NM 87131-0001
USA

References

[1] J. Felsenstein. *Phylogenetic Inference Package (PHYLIP), Version 3.6*.University of Washington, Seattle,2004.

[2] W.M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406-416, 1971.

[3] T. Matsumoto and T. Nishimura. Mersenne TwisterL A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. on Modeling and Computer Simulation*, 8(1):3-30, 1998.

[4] M.M. Morin, B.M.E. Moret, NetGen: Generating phylogenetic networks with diploid hybrids. Bioinformatics 22:1921-1923, 2006.