## DIMACS Working Group on Reticulated Evolution
## September 22, 2004

Summary of the Discussion on the future directions and collaborations between biologists and computer scientists to better understand reticulated evolution

- Creation of a repository of benchmark data to test different methods:
  - Create a web page for the repository.
  - Identify the problems that need reticulation representation
  - Each benchmark dataset should be accompanied by a thorough annotation, including type of the data, format of the data, description of how the data was obtained etc.
  - Data presented at this meeting could be deposited as initial benchmark data.
  - Coordinate with people who develop simulation packages and make them aware of benchmark data
  - Output of benchmark results for different methods could be posted to the web page too. Challenge: It is difficult to compare outputs of different programs (e.g., outputs from splits decomposition and neighbor-joining trees)
- Creation of a "universal" format to represent networks (such as Newick format for phylogenetic trees). And, as a consequence, creation of a network viewer (possibility to utilize developed graph-viewing software).
- Development of more realistic models of evolution
- Formulate questions mathematically and involve theoretical mathematicians to solve them.
- Development of methods should be an iterative communication between biologists and computer scientists: biologists need to communicate what they want and computer scientists need to make clear what assumptions were made, how the results are calculated and what range of dataset size the methods are appropriate/tested for.
- Computer packages should explicitly state all assumptions that were made.
- Newly developed reticulation packages should always have command-line interface (for automated analyses of multiple datasets); programs with graphical user interface (GUI) should have both options available in any given program.
- Development of statistical tools to evaluate networks, especially tools for reliability estimation. Report both false negatives and false positives. Explore potential for systematic and sampling errors. Provide tools to compare networks generated by different methods. Provide tools to merge two or more networks into a "consensus network" (i.e., extract network features common to all networks generated with different methods).
- Modeling networks is still a field in its infancy and biologists should use caution when using these new tools before the field has achieved some level of confidence in its methods.
- "Biologists need to work on determining what kinds of values we should give to various parameters. For example the diploid rate of hybridization or polyploidization versus regular speciation. We need to start to gather data to

answer these questions. In a given hybridization event, who are the parents? Given two species of plant, I can't predict whether they will hybridize. This would help modelers."

- using known networks to test the reliability of the methods (eg on Randy's orchid phylogeny or the epidemiological data mentioned by Keith) is essential before applying the tools to larger scale questions with unknown answers.

## Discussion 1: What needs to be done on the math/computer science side for better understanding reticulation?

TW (Tandy Warnow): We need methods to determine when the patterns we see in the genetic data are the result of reticulate evolution versus other sources. We need models of evolution that are sufficiently realistic so that you can see patterns in the data. If we rely on real data, we already know that there's reticulation (ie a hybrid lineage), but we need to get to the point where we can infer reticulation (versus other sources of variation) out of data in which we don't already know it's there.

Unknown older guy in blue shirt: Everyone works with their own source of data. It would be good if we could work together more, to have some benchmark of data or examples where people can compare different calculations, to do parallel work. Maybe somebody from DIMACS can coordinate something like this.

DG (Dan Gusfield): We should collect pointers on how to try different methods

KC (Keith Crandall): it would be helpful if someone develops a new method, they can directly compare it to 15 methods already done. Someone asks, simulated or real data? KC responds, depends on whether you want to know the answer.

DG: I looked at some longitudinal data at Los Alamos and it didn't take me long before I realized I couldn't deal with it.

RL (Randy Linder): The data needs to be annotated in some way.

PL (Pierre Legendre): We could contribute data from our own studies, whether distance matrices or real data, or link to our own websites. (need to share this with?) people who produce simulation tools, or they will repeat what other people are doing.

LN (Luay Nakhleh): re producing simulation tools, we have seen this in this meeting – for example, our tripartition methods you can't apply to SplitsTree because it's a graphical representation of data. You can't compare SplitsTree output (network) versus neighbor-joining (evolutionary output).

PL: Every simulation method makes assumptions, but it would be interesting to look at the results (with different assumptions?)

TW: Its not about assumptions, but about what's reconstructed. They're not comparable, they don't mean the same thing. The purpose of simulation is to determine how well your method works, but we don't have a way to compare.

PL: If you compare lateral gene transfer, its not the same as reticulate events a the recombination level

BM (Bernard Moret): There's hybridization, lateral gene transfer, host-parasite effects, recombination – it would cost $10 million, you can't do all this. There are different boundary conditions. If we try to plan collaborations to cover everything it may not go anywhere. (I missed a lot of this): Large NSF ITR grant trying to put together CIPRES - Cyberinfrastructure for Phylogenetic Research (www.phylo.org). It is tree-based not network based.

RL: Good for CIPRES, talk about standards for programs to set up. We need cooperation not just collaboration, Its good to build simulators so they communicate. For example, include population genetic aspects into network simulations for phylogenetics.

BM: How to represent networks is still not known. For example a pedigree needs a different representation than reticulate evolution

DG: A (common?) language for representing networks is critical, Short of that, maybe a viewer.

KC: Once you get the language, a viewer is easy.

RL: (name) is doing nice work on this. Networks are really hard to make planar. Its hard to know how to represent in 2-dimensional space.

DG: (name) does a moderately good job at drawing networks. Someone could write a simple interface for your simulation program.

LN: We need to make programs run on command line, not just graphical user interface (GUI). Its ok if you want to run it a few times but not for simulations you have to run 1000 times. Allow for both possibilities.

SH (Samuel Hanelman): Any biologist studying a protein will be interested in consistency in that one method rather than comparison of different methods. They'll want computational efficiency.

TW: I think most people are not alarmed when UPGMA gives a different answer than Maximum Parsimony.

SH: If they're packaged together, there needs to be a way for the end user to be able to evaluate consistency between the methods.

TW: How do you evaluate consistency?

SH: Neighbor-joining trees are different from UPGMA trees but you can still find features in common, within a reasonable range.

BM: There are so many parameters. Looking at Boris' network and it shows he trusted the TIME measure. When we start comparing different assumptions about data we trust and don't trust… we barely trust tree-based methods, I'm not sure it's a goal for the short, or even medium-term. Biologists need to know what assumptions computer scientists made in their programs. What are the things biologists might need to ask.

RL: One of the good things about collaboration is showing aspects of biology to CS people. We want to constantly have that kind of communication so biologists don't look at some fully developed tool without correct biological assumptions.

KC: It's a problem in phylogenetics, people estimate a phylogeny and then start telling stories - A lot of statistical tools use phylogeny to test other hypotheses. (it's a problem if the phylogeny is not correct). We need statistics to test networks. Estimate population genetics parameters, genetic divergence given a network, dN/dS ratio etc. We need to do something with networks other than telling stories.

RL: We need to educate our colleagues about the programs. They need to understand the underlying models for the various software packages.

DG: If biologists found a phylogeny not totally resolved, and if the package is based on particular assumptions, they can test if it makes sense. For example, with maximum parsimony, if someone invests time to do parsimony people who look at the output can decide if MP is a good method.

BM: Just because a tool gives an answer that appears to give a lot of detail doesn't mean biologists should trust that answer. We need healthy skepticism and honest advertisement of the tools.

DG: There may be many ways to do computation, but the semantics of a package often in program without a clear understanding of what's in the package.

PG (Peter Gogarten): It would be nice to have reliability. For example, bootstrap values. Biologists have infinitely long sequences that they can use but phylogenetically useful information decays rather quickly, there are systematic errors, sampling errors.

DG: re scale: strive for a large scale. You have methods that work on a small scale but are useless on a large scale. Re algebra – I'm depressed or impressed.. there's no real mathematics being used here. Everything we're seeing is seat of the pants, ad hoc. I keep thinking there's a big world of mathematics out there and its not being used to address these questions.

TW: I'd slightly disagree. It doesn't matter how good the math is if we don't know the problem, or what the questions are. We need to ask the right questions, and then find out what the math appropriate for the questions are.

KC: Mike Steele does some network stuff that's highly relevant. Its hard for me to walk into the math department and say hey fellas…

DG: 99% of what we think is ultimately going to be garbage. We have to go back and forth until something useful emerges.
[On that note, a favorite and relevant quote: I think and think for months and years. Ninety-nine times, the conclusion is false. The hundredth time I am right.-- Albert Einstein]

MJ (Mel Janowitz): I've contacted many biologists that I've interacted with over the years, and many are just not interested in reticulate evolution.

DG: My impression is the opposite. A lot of the meetings I go to half the talks are about networks.

RL: And there are a lot of talks about trees that should be about networks.

KC: At a recent Bioinformatics (DIMACS) conference, the theoreticians and population geneticists were really opposed to networks. They claimed that you can't establish any parameters on a network (this is sort of true but they just haven't been developed yet). We use the Robinson-Foulds score considering the network as a series of trees.

RL: It's like looking for keys under a lamp because that is where the light is.

KC: There are a lot of people using these networks to address phylogeographic questions.

## Discussion 2: What needs to be done at the biological level to better understand reticulation?

PL: Biologists could think of all the problems we have that require reticulated networks versus tree-based representations, and what assumptions are needed to model that network.

RL: Biologists need to work on determining what kinds of values we should give to various parameters. For example the diploid rate of hybridization or polyploidization versus regular speciation. We need to start to gather data to answer these questions. In a given hybridization event, who are the parents? Given two species of plant, I can't predict whether they will hybridize. This would help modelers.

RY (Robbie Young): You can get some of this from population genetics, for example how well a model predicts theta. You can build a model that mimics what's going on but isn't a true representation.

SH: The more degrees of freedom you have that's easier to do, you have to get more biologically realistic.

RY: You can penalize the addition of extra parameters.

KC: There may be cases of HGT people accept from an empirical standpoint, like Hillis generating a known network of phage to identify the evolutionary relationships between them. He was having to eliminate networks as noise! There's an increasing amount of epidemiological data where you know the linkages and time frames, for example who transferred the virus to whom. Sometimes people submit (sequence) data to Genbank or to Los Alamos but it's unlinked to the epidemiological data.

SH: There's an OMIM database with heritable birth defects.

PG: Organisms without a germ line have (more/less?) reticulations.

RY: Why can't you put ancestral recombination graphs to estimate network genealogies.

KC: You have to be able to relate them, to incorporate the whole machinery of coalescence.

RY: population genetics eliminates all that noise

RL: you need to know how parameter-rich your model can or should be.

PL: I wonder as a biologist if I should launch into computer networks for 500 species, because the methods are in their infancy.

RL: We have a known orchid phylogeny and I know what hybrids have been created by humans. We can test the predictions with this kind of data. If we can't solve simple cases well, then we can't apply it to larger problems.

KC: re: post-processing of network. If we're dong hypothesis testing of 500 sequences, we need stats to test rather than just visually looking at 500 networks.

RL: I've looked at some of the stuff in gene regulation networks and I have no understanding of how they have any confidence in the reliability in their methods.

TW: We tend to focus on false negatives, but high rates of false positives are not good. Computer scientists or mathematicians don't have any training in stats, for example in understanding Type I or Type II errors.

Unknown guy in blue shirt: idea: take one of the smallest genomes, <100 genes and determine which gene in this genome came from this other genome, which from HGT. Split into two groups: vertical (classical view) and horizontal. HGT may not need activation - it could be just to keep for the future or it could play an active role in the genome.

BM" A lot of methods are based on intergenomic comparisons. You would need >1 genome not just one. You need a whole population of genomes, then you can say this looks like gene duplication, etc.

PG: Biologists don't agree on anything. Every gene has a "gene most recent ancestor" and lived in several different organisms.

RL: We're veering dangerously close into discussing species concepts.

MJ & RL: Hopes everyone will use the listserv to continue conversation and collaboration, and that many of us can get back together again in the near future.