Geometric models of biological phenomena

The American Institute of Mathematics

This is a hard–copy version of a web page available through http://www.aimath.org
Input on this material is welcomed and can be sent to workshops@aimath.org
Version: Wed Jun 18 16:51:55 2003

Table of Contents

Chapter A: Afternoon Discussion Sessions

Discussion session notes by Francis Su.

A.1 Introduction to Discussion Session Goals

This document presents an outline of some of content of the afternoon discussions. In the interest of producing these notes as quickly as possible, no attempt has been made to track down every reference or attribute every comment to a participant. Also, as the scribe, I summarized what I understood from the discussion, so attributed comments are, in some cases, paraphrases. (Thus, you may wish to check with the attributed before making any consequent attribution.)

Each afternoon session was centered around some theme. The purpose of the sessions was to try to identify problems that people might want to pursue, and identify which problems are most important to work on. Also, with participants from many different academic communities (biology, combinatorics, topology, statistics), it was important for each community to hear perspectives from the others.

A.2 Sunday Discussion on Biological Issues

The discussion was moderated by John Luecke. Susan Holmes opened the discussion by suggesting that we might ask the biologists: what mathematical or biological questions related to phylogenetic trees are most important to biologists? She also invited clarifications about things that have been addressed in talks earlier in the day.

Q. What properties of phylogenetic trees make them different from random trees? Is there some kind of structure that makes them different? (Epstein)

- Phylogenetic trees don't look like trees generated by random branching processes. (Huelsenbeck)

Q. What are the characteristics that a "distance between trees" should have that biologists would want? (Luecke)

- Luecke: For instance, the Billara-Holmes-Vogtmann (BHV) metric measures distances along geodesics in tree space, and the trees change as you move along the geodesic, but does this have a reasonable biological intepretation?
- Sober: that does not have biological meaning. Biologists just want to know how to tell if trees are similar or not.
- Felsenstein: you are assuming we (biologists) want a notion of distance between trees? (Scribe: this appears to have been in contrast to just a notion of similarity. This elicited remarks from mathematicians about the fruitfulness of using a metric to tell how similar two trees are.)
- Huelsenbeck pointed out several ways to measure how different trees are, e.g., involving the contraction/expansion of edges to get from one to the other, or differences in length of edges, or squared branch length, etc. Some discussion of the Robinson-Foulds distance ensued.
- Luecke: so computational simplicity would be one important consideration for a distance.
- Huelsenbeck: Any distance notion on phylogenetic trees should have a good theoretical foundation.

Q. Is there a difference between some real distribution of trees on tree space rather than some random distribution? (Holmes)

- Evans: this will help us to construct meaningful confidence sets on trees.
- Felsenstein: Tree spaces are weird. For a cloud of trees, we want to characterize where in this space those trees are.

Q. Problem: make notions of distances accessible to biologists, so they might use them. (Huelsenbeck)

- Some comments were made about the use of methods of Kuhner-Felsenstein, weighted Robinson-Foulds, in biology.
- Diaconis pointed out that in his work on non-standard data structures, the notions of distances are very useful and basic. In doing exploratory analyses, they can be useful in statistical tasks. So, what do biologists think of these distances?
- Holmes pointed out that how fast MCMC methods converge depends on the geometry of the space, and the metric chosen.
- Huelsenbeck suggested that a useful task for mathematicians would be to make some sense of the uses of these distances to biologists? (such as in an article pitched to biologists).

More comments about tree space topology and branch lengths:

- Huelsenbeck: are there versions of tree space that biologists might *not* be interested in, due to technical conditions? For instance, if you give a program an alignment, the branches have to be short enough so that you can align them.
- Felsenstein: In many cases in biology we are interested in grappling with, such as origins of mammals, many of the orders seem to have popped up rather quickly. We could be interested in (a) the branch length and (b) the topology.
  When the orders diverged, important things were happening very quickly, such as morphological changes. But the molecules you are studying may not have been involved. If you are interested in those morphological changes, it is important to know what order the branching occurred.
  On the other hand, in some other questions, topology many not be as important as branch length. It really depends on the question you are asking.
- Sober: A third issue is important: the character states of the interior nodes of trees. You might want to know what is the sequence of changes that occurred on some branch? for instance in some calculation, you may want to integrate over all possible states of interior nodes.
  Felsenstein pointed out that these states on interior nodes are sometimes overinterpreted— they are often viewed as actual states, rather than estimates.
- Penny: in biology, knowing the "true" tree may only be a starting point of the real investigation— some other aspect that the biologist is truly interested in. Huelsenbeck gave an example of an evolution biologist interested in sexual selection. Although she made a phylogeny (what she thought was the best tree), she wasn't primarily interested in the phylogeny, but in further questions about sexual selection.

Q. How to pick a tree (or tree average) from a set of trees resulting from data?

- Epstein: in bacteria dna fragments, each one gives a tree. Which one to use? Need a tree distance to analyze the resulting 22 different trees. The distance he used was the BHV metric, but he suspects one obtains much the same answer with many different tree metrics.
- Evans: this is similar to Mallow's model in statistics, where you have a distribution on permutations, centered on a permutation, dropping off in some radial sense. Here, you have a central tree, and then the probability that you observe some other tree dies off as you move farther away.
- Penny: one approach to identify "bad" portions of a tree is to try to "identify the guilty taxon", by succedssively removing taxa and then see whether this stabilizes the tree significantly.

Q. How to understand or deal with residuals, such as non-tree like data?

- Holmes: gave an example in which a friend has a reference tree, and 8 different plumage trees. In tree space, if the plumage tree differs from the reference dna tree in some direction, is that present in the way these things actually evolved?
  In statistics, when doing regression, you compare data points to a fitted line, and ignore the ones that are way off. Here, in tree space, there's a similar question for non-tree like data: how much did I have to bend the data to make it into a tree?
- Sober: how do you decide how much off is too much, before you are worried?
  (Scribe's note: this seems to require a notion of distance not just on tree space, but some larger space— such as on the space of DNA sequences— in which tree space is embedded. We discussed embedding questions on Wednesday.)

More comments on distances:

- Diaconis told a story where distances fit data remarkably well. Perception psychologist Roger Shepard studied the visual system, by showing subjects a configuration of blocks, then some other configuration, and then asked them: are they are the same? He found that the time it took people to decide was the geodesic in the three dimensional rotation group. The data is quite remarkable, and gave straight line plots. We know what it means to measure the distance in the rotation group (but why the brain should know about this is some fascinating subject in itself.)
- Similarly, does a distance between trees have to have some interpretation in terms of evolution, in order to be the "right" notion of distance?
- Epstein: For instance, given dna or amino acid sequences, one for each taxon, this data produces a tree. As sequences change or evolve one nucleotide at a time, the trees will change. What path does this trace out in tree space?
  (Scribe's note: this appears to require a space of trees that includes trees with many different numbers of leaves, so that one can speak of how a tree evolves as species split off from one another.)
- In response to a question, Felsenstein mentioned a few sources of variation in trees: (a) statistical error. (b) coalescence: take a gene copy in three species, and think of copies ancestral to these. The copies do not come together instantly, but have some stochastic chance of mixing, and they may come in some random order that conflicts with a species tree. (c) horizontal gene transfer.
  (Scribe's note: a more detailed discussion of this topic can be found in Tuesday's discussion.)

Q. Which is better: concatenating DNA sequences first or averaging trees later?

- Holmes: empirical studies show that if you take all the data and compute one tree, you generally do less well at estimating the tree than than if you take the fragments, use them, then average them. This comes from the CAT(0) property. The intuition behind it is if you average in a negatively curved space, you converge much faster than you should.
- St. John: pointed out an example with 20 simulated DNA sequences: if a is one part of the genome, and b is another part, then the trees obtained from using a combined with b do not overlap, and are somewhat in between, the trees obtained from a and from b alone. Felsenstein commented that the stochastic effects pushed a and b in different directions in tree space.
  She also noted that in studying hybridization (e.g., sunflowers), she was surprised that she didn't get more overlap, and often the right answer was with part of the data, not all the data.
- Felsenstein: Biologists are in disagreement about whether concatenating or averaging is better. Statistically, which is better?
- It was pointed out that a paper by Cunningham (1997) does comparisons. Also, work by Amit, and statistical literature on boosting and bagging.
- Billera: there's a notion of equivalent metrics in topology, and any of these will do. (In doing geometry, where issues like curvature come into play, the choice of metrics is important.) It may be the case that even though certain metrics are good enough for some purposes, it doesn't mean that other metrics aren't valid.

Q. What is the right notion of an "average" of trees?

- The biologists present agreed that this was a very interesting question for biologists.
- Vert: Two important ideas emerge when working with decision trees: (1) average decrease some variance, (2) averaging can leave the space. Boosting will leave the space, if we don't leave the space it's really not boosting. The averaging we are discussing here doesn't leave the space of trees.
- More discussion on averaging took place here. Holmes made a comment about the "non-associativity" of trees (related to considering trees of trees, when building trees one character at a time). Billera made some remarks about how averaging may be better than concatenation, but we have no guarantee that it is any good.
- Penny discussed the notion of a median tree: the tree that is closest on average to all the others.
- Felsenstein: other examples are the consensus tree, and majority rule consensus tree.

Q. What are good properties of averaging? (Luecke)

- Penny: would like a fully resolved binary tree, though he points out that some others might sacrifice that in favor of other features.
- Felsenstein posed a problem about averaging properties: given a bunch of trees from different genes, to study a question like: are chimps closer to humans than gorillas? Say 2/3 of the trees trees show humans with chimps, others show chimps with gorillas. Suppose in trees that group humans with chimps, the average branch length is 1, but the others don't have it. When you average... do you want it to be length .66 or length 1?

There was a question about how different are trees that result from diffrent averaging methods. It was pointed out (Huelsenbeck) that differences between methods for a single gene is much smaller than the differences across genes.

Q. What would be the distribution on trees that gives majority rule consensus as its average? (Holmes)

- Holmes: would like any tree average to be an expected value with respect to some distribution on tree space.

A.3 Monday Discussion on Combinatorial Issues

The disussion was moderated by Michelle Wachs. She started off the discussion by asking:

Q. What combinatorial questions arise in biology that might be of interest to biologists? Or, what problems motivated by biology would be of interest to mathematicians, to give us some interesting problems?

- Huelsenbeck: In putting priors on the set of all possible trees, we sometimes need to count trees with particular characteristics, such as ones with a particular kind of edge (split, bipartition), or a list of either this or that kind of edge. The constraints may be multiple or conflicting. The count is needed so that we can sum probabilities over these trees.
- Holmes pointed out that inclusion/exclusion issues are involved here. There was some discussion about NP-hard problems, but Billera pointed out that complexity issues are not that relevant for the basic question of how to count these things.

Q. Given a set S of k edges, and a subset J of S, how many leaf labeled binary trees have some subset J of S but not the edges in $S - J$? Do this for all J.

- Evans asked why we can't just use MCMC to solve this problem? Holmes pointed out that the combinatorics may give some insight.
- Another example of a question that arises: the number of trees that are distance d from a particular set of trees (e.g., using a distance such as nearest-neighbor interchange)?

Q. What are good codings for trees?

- Wachs described a bijection between phylogenetic trees and type B permutations whose left-to-right minimum is unbarred. (Type B permutations are permutations in which each number can be barred or unbarred.) This can be used as a coding for trees.
- There was a question about Penny's method for coding trees (in which each time you add an edge there are $2n - 1$ places to add it, so that you get $(2n - 1)!!$ of them). Billera said that this is the same as what was on the board.
- Penny mentioned the existence of another scheme for coding each tree by a unique number. Holmes pointed out that the biologists use a Newick notation (New Hampshire format) that uses lots of parentheses. Another coding is a matrix where the columns are edges, and the rows are vertices, and you put a 1 if the edge is an ancestor of the vertex. (Brooks used it in biology, and Graham-Winkler used for addressing a network).

- Holmes described a matching representation is one due to Diaconis-Holmes. When $n = 2$ there is one matching, when $n = 4$ there are 3 matchings, when $n = 6$ there are 15 matchings, corresponding to 4 leaf trees. The matchings are the sibling pairs. Example: $n = 6$. Call the numbers from 1 to 4 *available* numbers, and the other numbers *unavailable*. Given a matching $16 - 42 - 53$, pick the first available pair of numbers. By pigeonhole argument, there is some pair, that's a sibling pair in a tree. Then the smallest non-available number (5, in this case) is the parent, and the other member of the pair (3) is the sibling of 5. Continue.
- Diaconis pointed out their matching method is related to Billera's method using Prufer codes.

Q. Is there an efficient representation for coding trees? (Billera)

Q. Is there a nice neat notation for trees that is continuous when doing a random walk on trees? In other words, do small changes in notation correspond to close trees? (Diaconis)

- The type B permutation coding might be very amenable to doing a random walk. Bridson pointed out that this representation is related to subsets of the Cayley graph of the Weyl group.

Q. Are there questions that combinatorialists might have for biologists? Combinatorial structures that biologists might be interesting? (Wachs)

Q. If we took trees and instead of putting real numbers of them, putting discrete values, like 0-1, would that be of interest in biologists? (Diaconis)

- Huelsenbeck: this maybe useful in character-mapping, mapping characters that are on or off. Perhaps some substitutions are more likely when characters are on.
- Epstein: it certainly seems reasonable to associate to each edge a vector, rather than a real number.
- Evans: as in Penny's talk, characters of each of the vertices can be thought of as elements of the Klein 4-group... the edges can represent differences of elements at the vertices.
- Wachs: what about k-ary trees?
- Penny: people find cycles useful for representing uncertainty. Referred to some who tried to define how tree-like the data was, something tree-like would have only small cycles.
- Holmes referred to a program called *splitstree* that makes such diagrams.
- Penny: biologist would like structures that represents distances accurately.
- Shareshian: described a space that arose in his work, a space in which $(n-2)!$ cycles correspond to associahedra.

A.4 Tuesday Discussion on Statistical Issues

This discussion was moderated by Ruth Charney.

Q. Diaconis suggested making a list of the kinds of noise, or sources of variation, that are implicit in estimating trees? Participant discussion led to the following list:

- alignment
- errors in sequencing, (though Epstein pointed out that the quality is steadily improving)

- misspecification of the model (process of going from the data to a tree): mutation rates, independence between sites, change of nucleotide composition
- variation within species (1 in 1000 genes)
- bias in corrections for distances (for distance based models)
- variation between fragments of same DNA (bias created by choice of fragments)
- selection varies in different parts of the genome
- gene identification (problems created by gene duplication and gene loss)
- hybridization (branching comes back together, structure is not a tree) and horizontal gene transfer
- optimization
- not enough data— length of sequences, number of taxa

Q. are these problems worth working on incrementally or all at once? Which ones are most important? (Diaconis)

- Penny: differences in nucleotide composition?

Q. Can we define the geometry (distance) on tree space to behave well with respect to a particular model?

- Vert: what are the implications of the noise in the definition of the tree space? Instead of defining the geometry of tree space beforehand, and then asking what properties it satisfies, would it be possible to define a tree space in terms in such a way that it has good properties (is stable, averages are at least as good as the trees themselves, etc.)
- Diaconis: this is perhaps related to statistical geometry or information geometry— using the Fisher information to define a Riemannian metric on the space of distributions.
- Epstein: an average is a summary statistic, but doesn't summarize everything we might want.
- Evans: information geometry example: using the upper half plane to parametrize normals on a line, the geometry that best reflects closeness of normals is hyperbolic geometry.

Q. What are residuals for trees, and how do we estimate the residuals? (Residuals measure how far each data point is from fitted tree.) Are there graphical methods for detecting departure from the model? (Penny)

- Penny also asked why there are so many definitions of maximum likelihood?
- Diaconis suggested that one topic probabalists and biologists may benefit from is work of Aldous and students, on analyzing rates of convergence of random walks on phylogenetic trees. Aldous responded by noting that the case they can analyze is not necessarily so useful to biologists (one where leaf is taken off and pushed somewhere else). For an n-leaf tree, the random walk needs n^2 random steps. A more realistic example is to cut somewhere, and attach whole subtree in a different place. It is believed that about $n^{3/2}$ steps is needed to mix this kind of chain. This could be useful in MCMC algorithms.

Q. Random walks on sets of trees (using tree rotations)– how fast does it converge?

- Once again, the question of how to measure distance in the space of trees emerged.

Q. What are appropriate meeasures on tree space?

- Glenn asked if there a need for a non-parametric likelihood on trees? Diaconis said that empirical likelihood related to the boostrap, so it could contribute.

Q. If exact distances aren't easy to compute, can we obtain any upper and lower bounds for distances (such as the BHV metric)? (Su)

A.5 Wednesday Discussion on Geometric Issues

This session was moderated by John Shareshian. The discussion started off with some comments about Riemannian metrics and whether the set of metrics on tree space would be useful to study.

- Vogtmann asked if this space may be too large to study?
- Forman noted that when geometers look at the space of all Riemannian metrics on a given space, they do it to try to find the "best" metric in some sense (e.g., such as one with constant curvature).

Q. Are there many useful or interesting metrics on tree space?

Q. Is the BHV metric the only one (up to scalars) of non-positive curvature? (Bridson)

- Charney: metrics are slightly different, but they should all be similar. If there is no reason that biologists suggest that one is better than another, then it makes sense to choose one that has useful properties, such as metrics of nonpositive curvature.
- Billera: for biologists, what metrics do you want to have on the orthants?

Q. What metrics "should" be used on the subspace obtained by fixing the tree type?

- St. John: uses the L^1 metric, though Felsenstein uses L^2.
- Penny: we use L^1 because then lengths scale with time.
- Charney: since the data is not completely in tree space, maybe we should be considering non-intrinsic metrics.
- Forman: want to choose metrics so that the statistical methods we are using are continuous.
- Flath: perhaps some edges are more important than others?
- Penny: we are not surprised when we get edges near leaves are accurate, but we are really interested in getting deep internal edges right.
- Penny: taxonomic studies, just care about the branch order (weight 1 on edges), but when we consider time studies, we do want the lengths.
- Forman: measuring the residuals seem to imply that embedding tree space in some larger space and considering an extrinsic metric would be important.
- Vogtmann noted that Felsenstein was talking about embedding tree space in Euclidean space (using the Robinson-Foulds metric).

Q. How does the Robinson-Foulds metric compare with the BHV metric? (Diaconis) What's the Lipshitz constant? (Bridson)

- Vert: Another distance (referred to by Felsenstein in his talk) is Kullback-Leibler distance (relative entropy) between probabilities. Fix a model. Each tree defines a probability on the set of assignments of letters (A,G,C,T) to leaves. Maximum likelihood corresponds to projection of the empirical measure of a data point to tree space according to this distance.

There was some explanation and discussion of this embedding. Diaconis noted that while there are many metrics on probability distributions to consider here, the Kullback-Liebler separation is good for maximum likelihood.

Q. Should the metric on tree space come from an intrinsic metric, or an extrinsic metric on a larger space in which tree space is embedded? Which is a more natural way to view tree space?

- Forman: studying the residuals seems to be very important. This larger space is part of what you are given in the data, and should be useful. Charney rephrased the question as: what is the right ambient space to embed tree space? While Bridson remarked that he viewed tree space as God-given, Forman asserted that he views the data as God-given.
- Vert remarked that the "average" of trees with high likelihood may not be high likelihood with the intrinsic metric.

Q. How do you deliver geometry to biologists? (Billera)

- Billera pointed out that interesting mathematics in biology seems to happen by serendipity— cases in which someone happened to know some mathematics or knew someone who did.
- Penny: notes that there is a role for mathematics to play. Also new majors in mathematical biology (at some schools) have arisen that will train students to think in both disciplines.
- Diaconis notes that as a result of this conference, and ensuing discussions, people might write an article for *Science*. This may bring biologists to mathematics.
- Snel suggests advocating a notion of average, and communicate why it is better. For instance, articles on concatenated alignments have appeared in Science.
- Huelsenbeck: write review articles that explain things clearly. Write for *Systematics Biology*— that's where stuff on trees would appear.
- Billera asked again how one spreads mathematical knowledge in the biological community, noting that courses offered at most institutions are either at too low or too high a level for biologists.

A.6 Thursday Open Discussion

This discussion was moderated by Lou Billera. Billera noted that, as it was the final discussion session, anything is a valid topic today. The discussion started off with some comments about the usefulness of the workshop:

- Billera began by remarking that the organizers worried that having 4 different groups of people at this workshop from different backgrounds could have been disastrous. But he noted that it didn't appear to be... and had found it interesting, and was pleased with how much discussion had taken place.
- Diaconis emphasized the usefulness of this workshop for generating problems to work on, and cited an example of residuals for discrete data analysis, as something worth following up, as well as the discussion in the geometry session about what are natural measures on tree space. He noted that the workshop was valuable because he couldn't think of another setting in which all these ideas from different arenas could all come together.

Q. Concatenation as "averaging"?

- Billera noted that in the discussion about averaging, he was struck with the idea that concatenation was an average in sequence space, an idea that he hadn't appreciated until this workshop. One question He noted that the conference had touched on many ways to go from sequences to trees, and one of the big questions was whether these methods were coherent with concatenation-averaging.
- Penny: it is worrisome that there are many different ways of concatenating, and how to handle the number of parameters?

Q. Likelihood contours?

- There was further discussion about confidence sets and likelihood contours, and how these might intersect transition points where the tree space branches in several directions. For instance, if a maximum likelihood process produces a tree with very short edge, then it would be near a transition. There was also discussion of whether the notion of curvature on tree space would be helpful.

Q. Multivariate "metrics"?

- Holmes: in the date that arises for trees, you get multivariate distances. We may want to speak of directions? is there anything like that in metric topology?
- Diaconis: there are some distances that take values in partially ordered sets, etc., distances as a vector, etc.

Q. Alternate representations (not trees)?

- Epstein asked if are there other notions besides trees that would be helpful?
- St. John: Networks? Trees with several non-tree portions.

Q. Random walks on trees (generate letters, prodcue new trees)

- Vert: One can imagine a random walks in this space of trees, by starting a tree, generate data, estimate a tree (by maximum likelihood). Do this number of times, get a random walk? Might be interesting to study this walk?

Q. Would it be useful to study 2-colored trees to model orthology and paralogy? (Shareshian)

During the session, Diaconis suggested that, as a follow up to this conference, people can now get together in small groups and communicate and start to work on a problem, whereas this wasn't true before the beginning of the week. AIM could facilitate such follow-up meetings.

There was some brief discussion about holding a follow-up conference to this one, since all the participants now have a solid foundation on which to try to work on some joint problems.

ARCC and the organizers (Lou Billera, Susan Holmes, Karen Vogtmann) were all thanked enthusiastically for a stimulating week of talks and discussions. The workshop was concluded over evening refreshments.