Review for CS 581 midterm

Tandy Warnow

March 2, 2020

Introduction Recall that the midterm will cover everything you've learned in the class, including all the assigned reading from the textbook, the homework assignments, and the lectures given in the course. For example, recently we have been looking at figures comparing methods and talking about the trends – how model conditions impact methods, and how model conditions impact the relative trends between methods. This means you should be prepared to discuss trends you see in a figure on the midterm. You will also need to use the mathematical theory you've been exposed to and learned in the course to solve new problems (i.e., problems that do not appear in the textbook or in the lectures).

What follows is a list of topics that you should familiarize yourself with, in preparation for the midterm. Of course, the midterm is open book, but you can use the time before the midterm to review these technical concepts, so the midterm is easier to do. In general, you should know which computational problems are solvable in polynomial time and which ones are NP-hard; you should know what the stochastic models of evolution are and which ones are submodels of others; you should know what methods are statistically consistent and which ones aren't (and note that this depends on the model); you should be able to apply algorithms you've learned to data; you should be able to synthesize different concepts you've learned to address a new problem. Also, you should be able to express yourself clearly and simply.

Note that the midterm will contain several problems where only your answer is needed – no justification is requested, and justifying your answer will not affect the grade. Please note when this is the case, and for such problems only give an answer. For other problems where we ask you to justify your answer or prove a statement, make sure to justify your work.

For each chapter, I have indicated what we have covered, either because

it was assigned reading or it was covered in class anyway. However, please know that when I say that the assigned reading covered all of a given chapter, I do not include the "Further reading" section.

Chapter 1 (Brief introduction to phylogenetic estimation). Your assigned reading covered all of this Chapter.

1. CFN model
2. strict molecular clock, UPGMA, and inferring trees when evolution is clocklike
3. statistical consistency
4. RF, FP, and FN error rates
5. additive matrix
6. quartet trees, Four Point Method, Naive Quartet Method
7. using simulations to understand methods
8. branch support techniques (bootstrapping)
9. CFN distance corrections, dissimilarity matrices, triangle inequality
10. p-distances (normalized Hamming distances)
11. UPGMA, neighbor joining
12. maximum parsimony, maximum likelihood
13. positively misleading, statistical consistency

Chapter 2 (Trees). Your assigned reading covered all of this Chapter.

1. rooted and unrooted trees, binary trees, polytomies, branches, internodes
2. Newick notation for rooted and unrooted trees
3. clade representation of a rooted tree
4. bipartition encoding (also called the character encoding) of an unrooted tree
5. refinement (and saying that tree T refines tree T'), and "fully resolved trees"

6. bipartition compatibility
7. methods to construct a rooted tree from its set of clades
8. methods to construct an unrooted tree from its set of bipartitions
9. determining if a set of clades or bipartitions is compatible
10. the Hasse diagram for a partially ordered set
11. pairwise compatibility vs. setwise compatibility
12. strict consensus trees
13. tree error rates (FN, FP, RF)
14. number of binary trees on n leaves
15. rogue taxa (and why taxa are rogue taxa)
16. difficulties in rooting trees and outgroup selection
17. induced subtrees (i.e., homeomorphic subtrees)
18. some special trees: caterpillar tree, completely balanced trees, star tree

Chapter 3 Constructing trees from true subtrees). Your assigned reading covered all of this Chapter.

1. tree compatibility (rooted and unrooted versions)
2. compatibility supertree
3. triplet trees
4. ASSU algorithm
5. All Quartets Method
6. Quartet Tree Compatibility
7. constructing trees from quartet trees
8. Dyadic Closure of a set of quartet trees
9. short quartet trees

Chapter 4 (Constructing trees from qualitative characters). Your assigned reading covered all of this Chapter.

1. qualitative characters, character states
2. homoplasy, perfect phylogenies, compatible characters
3. directed characters
4. maximum parsimony
5. maximum compatibility
6. the Fitch algorithm for solving MP on a fixed tree
7. the Sankoff algorithm for solving MP on a fixed tree
8. transitions and transversions
9. constructing trees from compatible characters
10. parsimony informative characters
11. how missing data are handled

Chapter 5 (Distance-based tree estimation methods). Your assigned reading covered all of this Chapter.

1. triangle inequality, dissimilarity matrices
2. strict molecular clock, UPGMA, and inferring trees when evolution is clocklike
3. Additive matrices definition, and the Four Point Condition
4. Four Point Method
5. The Naive Quartet Method
6. The Buneman tree (aka the Q^* method)
7. Neighbor Joining
8. distance-based methods as functions (and their properties)
9. Optimization problems using distances
10. Minimum evolution methods
11. the safety radius and proofs of consistency
12. The Agarwala et al. method (3-approximation for the L_∞-nearest tree)

Chapter 6 (Consensus and Agreement Trees). Your assigned reading covered Chapter 6.1-6.2 and so did not look at the material on Agreement Trees.

1. what are consensus trees used for?
2. majority consensus
3. median trees
4. greedy consensus (and the extended majority tree)
5. strict consensus
6. the "compatibility tree" of a set of compatible trees
7. asymmetric median tree (and why it was proposed)
8. the "characteristic tree"

Chapter 7 (Supertrees) Your assigned reading covered Chapter 7.1-7.7, but we also covered SuperFine (and the Strict Consensus Merger) during the class presentation.

1. What are supertrees used for?
2. compatibility supertrees
3. asymmetric median supertrees
4. Robinson-Foulds supertrees
5. quartet-based supertrees
6. FastRFS
7. MRP matrix, MRP, and MRL optimization problems
8. quartet-based supertree methods: Quartet Puzzling, Quartets MaxCut, Maximum Quartet Support Supertrees, etc.
9. Split-constrained (or Clade-constrained) Quartet Support Supertree method of Bryant and Steel
10. SuperFine (and its performance on data)

Chapter 8 (Statistical gene tree estimation methods). Your assigned reading covered everything but Chapter 8.3.

1. The standard DNA models of site evolution, from JC up to GTR (figure 8.1), all of which are time-reversible
2. The General Markov model, and tree estimation under this model
3. Extending to sequence evolution – assuming *i.i.d.* site evolution
4. The issue of heterotachy
5. Amino acid models
6. statistical identifiability
7. Markov property
8. computing the probability of a site pattern (and hence aset of sequences) at the leaves of a model tree (Felsenstein's Pruning Algorithm)
9. the maximum likelihood problem
10. Bayesian phylogenetics, and getting point estiamtes
11. methods (software) for constructing trees under these models
12. distance-based approaches (and distance corrections), and conditions sufficient to prove consistency
13. statistical consistency, inconsistency, and being positively misleading
14. the role of simulations in understanding methods
15. maximum parsimony is positively misleading, and how to prove this (long branch attraction, Felsenstein Zone Tree)
16. taxon sampling and its impact on tree estimation
17. parsimony informative sites, and using it to solve MP on small datasets
18. bootstrapping
19. sample complexity and absolute fast converging methods
20. The no-common-mechanism model, and how ML and MP perform

Chapter 9 (Multiple sequence alignment). Your assigned reading covered Chapter 9.1-9.15.

1. definition of true alignment, homology, homology pairs, indels
2. SPFN, SPFP, TC
3. computing edit distances using Needleman-Wunsch
4. optimization problems for MSA (sum-of-pairs, tree alignment, and generalized tree alignment)
5. sequence profilles
6. profile Hidden Markov Models (HMMs), and specifically "unadjusted" profiles, and the generic graphical structure of a profile HMM
7. building a profile HMM
8. computing probabilities of a string being generated by a profile HMM
9. using a profile HMM to compute an alignment
10. progressive alignment
11. consistency
12. divide-and-conquer methods, such as SATé and PASTA

Chapter 10 (Phylogenomics). Your assigned reading covered Chapter 10.1-10.7, but we also covered material related to this during the course lectures (e.g., MP-EST for species tree estimation under MSC, TreeMerge and GTM for improving scalability to large numbers of species, and ASTRAL-multi and FastMulRFS for species tree estimation under GDL). So in particular, look at the lectures during the course, including the lecture by Erin Molloy on TreeMerge and FastMulRFS.

1. Biological processes that create heterogeneity (e.g., HGT, ILS, gene duplication and loss, hybridization)
2. the MSC model of gene tree evolution
3. the anomaly zone and anomalous gene trees
4. theorems about unrooted quartet trees
5. concatenation analyses under the MSC (are they consistent?)
6. summary methods (e.g., SRSTE, SUSTE, ASTRAL, MP-EST, etc.)

7. site-based methods (e.g., SVDquartets)
8. co-estimation methods (e.g., *BEAST)
9. impact of gene tree estimation error on summary methods, and fixed-length statistical consistency
10. Improving scalability of co-estimation methods to large numbers of loci using BBCA
11. Improving scalability to large numbers of taxa using TreeMerge, GTM, and DACTAL
12. GDL (gene duplication and loss), orthologs, paralogs
13. Species tree estimation under GDL (e.g., gene tree parsimony, Phyldog, ASTRAL-multi, and FastMulRFS)

Chapter 11 (Designing methods for large-scale phylogeny estimation). Your assigned reading covered Chapter 11.1-11.3, 11.10.

1. tree search methods (local search strategies, such as NNI, SPR, TBR)
2. Bayesian MCMC
3. Different software packages for ML
4. disk-covering methods
5. triangulated graphs, perfect elimination schemes, recognizing if a graph is triangulated, etc.

Other. Most of what we have discussed is covered in the textbook, but some of what we have done is not in the textbook, and can be found instead in the posted lectures. Please make a point of reviewing the posted lectures to see what was covered there. Here are some examples:

1. Discuss trends observed in a figure
2. Disjoint Tree Mergers (covered during the lectures, not in the textbook) - including GTM and TreeMerge
3. FastMulRFS and ASTRAL-multi for species tree estimation under GDL
4. What we learn about MSA from data