CPSC 445 Algorithms for Bioinformatics

Academic Year 2010–2011 Winter Term 2

Learning goals and how to prepare for the final exam

Some general comments:

- note that no aids will be allowed in the final exam (this implies no books, calculators or notes)
- the duration of the final exam is 2 1/2 hours (150 minutes)
- your course notes, the homework sheets and the exercise sheet on evolutionary models and evolutionary trees are all you need to prepare for the final exam
- make sure you know the precise definitions of all terms that we introduced in the course and in the homework, e.g. pairwise sequence alignment, RNA secondary structure, pseudo-knotted RNA secondary structure, hidden Markov model, Viterbi matrix, state path, likelihood, log-odds ratio, stochastic context-free grammar, transversion etc (these are only a few examples)
- make sure you know and understand the examples introduced in the course and in the homework well
- make sure you have understood the basic ideas behind the different theoretical concepts and check that you know how to apply your knowledge to calculate small examples
- most exam questions will typically require a few lines of calculation and a few lines of explanation; the main purpose is to check that you can apply your knowledge to small examples and that you have understood the key ideas behind the theoretical concepts
- when preparing for the exam, make sure that you can apply old ideas to new examples and in a different context (e.g. we have seen the forward algorithm for pair hidden Markov models, you should be able to realize how we can come up with a forward algorithm for hidden Markov models).
- you may find that additional, selective reading in the "Biological Sequence Analysis" book is useful to deepen your knowledge, but it is not required in order to do well in your final exam (once again, please note that the notation in the book may deviate from ours in the course)
- preparing for the final exam together with a colleague is a good way to check that you can actually explain the course material

- you are welcome to make use of my and the TA's office hours to clarify your remaining questions

0. Biology background

You should know

- the central dogma for eukaryotes and prokaryotes, their similarities and differences, the meaning of transcription, splicing and translation
- how the different parts of protein-coding genes are called and how they are processed in the central dogma

1. Sequence annotation

- You should know how to translate known features about a system (e.g. the dishonest casino, CpG islands) into a hidden Markov model. In particular, you should know how to set up the states of the model, the connections between the state and the emission and transition probabilities of a hidden Markov model. You should know which role the Start state and the End state play in a hidden Markov model. You should know which constraints the emission and transition probabilities in a hidden Markov model have to fulfill.
- You should know how to assign an overall probability to any state path in a hidden Markov model for a given input sequence.
- You should know how a hidden Markov model can be used in conjunction with the Viterbi algorithm to predict an annotation for a given input sequence.
- You should know how the Viterbi algorithm and the traceback algorithm work, what the Viterbi and the pointer matrix are and how the Viterbi path is calculated and how it can be converted into a sequence annotation.
- You should know what more memory-efficient versions of the Viterbi algorithm are and for which kind of predictions they can be used (see homework).
- You should know how to compare two hypotheses and should know how to calculate log-odds ratios.
- You should know how log-odds ratios can be used as binary classifiers to test whether one of two hypotheses is correct.
- You should know how the performance of predictions can be evaluated in terms of sensitivity and positive predictive value and how to choose a threshold value in order to optimize the prediction performance for different purposes (e.g. in order to get a high sensitivity or high positive predictive value).

- You should know more complex variants of hidden Markov models and how they can be defined.
- You should know what the genetic code is (but don't memorize it), that the codon frequencies can be different for different organisms and how the relative frequency of seeing different amino-acids in proteins and the codon frequencies can be used to set up the emission probabilities of a hidden Markov for gene prediction in prokaryotes.
- You should know which sequence signals are typically captured by a hidden Markov model for gene prediction (course and homework) and which features of the model are species-specific.
- Examples: Dishonest casino, CpG islands, predicting protein-coding genes in prokaryotes, predicting protein-coding genes in eukaryotes (course and homework)

2. Sequence alignment

- You should know how pair hidden Markov models can be used to simultaneously align and annotate two input sequences of the same or a different biological type.
- You should know how pair hidden Markov models are defined and how we can set up their states, the connections between the states and their emission and transition probabilities from information about large sets of known alignments. You should know which constraints the transition and emission probabilities have to fulfill.
- You should know how the Viterbi algorithm and the traceback algorithm can be generalized from hidden Markov models to pair hidden Markov models.
- You should know that any state path in a pair hidden Markov model can be converted into an alignment between the two input sequences, an annotation of the first input sequence and an annotation of the second input sequence.
- You should know what more complex variants of pair hidden Markov models are and how to set up their parameters.
- You should know how log-odds ratios for hypotheses testing can be used in the context of pair hidden Markov models.
- You should know what the forward algorithm can be used for and should be able to come up with a forward algorithm for hidden Markov models given the forward algorithm for pair hidden Markov models.
- You should be able to imagine how the concept of pair hidden Markov models can be extended in order to simultaneously align and annotate N rather than only two input sequences.

- Examples: pair-HMM for aligning two DNA sequences, pair-HMM for mapping a protein to a genome sequence, pair-HMM corresponding to a random model

3. RNA structure prediction

- You should know what the defining features of RNA structures are as we observe them in living organisms.
- You should know on which level of detail we study RNA structures and why this is permitted.
- You should know what stochastic context-free grammars (SCFGs) are, in which way the production rules correspond to features of RNA secondary structures and how the transition and emission probabilities can be set up and which constraints they have to fulfill. You should be aware of the differences between the constraints for the transition and emission probabilities of hidden Markov models and those of SCFGs.
- You should know why we introduced the Chomsky normal form and how to convert the production rules of a context-free grammar into Chomsky normal form.
- You should be able to explain how the CYK-algorithm and the corresponding traceback algorithm work and which quantities they derive. You should know that any derivation tree in an SCFG can be converted into a (structural) annotation of the input sequence.
- You should be able to explain the main differences and similarities between the Viterbi algorithm for hidden Markov models and the CYK-algorithm for SCFGs.
- You should be able to identify a derivation tree in a given SCFG for a given input sequence and a known RNA structure. You should know that any derivation tree for a given SCFG and given input sequence can be assigned an overall probability.
- You should be able to set up the production rules of an SCFG given information on the sequences and their annotation that are to be modeled.
- You should know what pseudo-knotted RNA secondary structures are and that they cannot be modeled using SCFGs.
- You should be able to explain what un-ambiguous SCFGs are and why we aim to deal with them rather than with ambiguous SCFGs. You should know how to prove that an SCFG is ambiguous.
- You should know how the inside algorithm works, what the similarities to the forward algorithm for hidden Markov models and the CYK-algorithm for SCFGs are and which quantity the inside algorithm calculates.

- You should be able to explain that there a two possible strategies for RNA structure prediction (the minimum free energy approach and SCFG-based approaches) and what their respective features are.
- You should be able to explain the main idea behind comparative SCFG-based methods for RNA structure prediction that take a fixed multiple-sequence alignment as input.
- Examples: two simple examples of context-free grammars, Pfold grammar, Nussinov algorithm

4. Phylogenetic trees and taking evolution into account

- You should be able to give examples of how biological sequences can evolve over time.
- You should be able to derive how the number of edges and nodes in a binary, rooted or binary, un-rooted tree is related to the number of leaf nodes in the tree.
- You should be able to derive how the number of tree topologies in binary, un-rooted trees depends on the number of leaf nodes. You should be able to derive the number of rooted tree topologies from a given tree topology of an un-rooted, binary tree.
- You show be able to explain what evolutionary models are, how they capture information on evolutionary events, what the rate matrix and the substitution matrix are and what their matrix entries mean and how these two matrices are related. You should know how to derive equilibrium frequencies and what they mean. You should know the characteristic features of the Jukes-Cantor model and the Kimura model. You should be able to check if a substitution matrix is multiplicative and what this actually means. Likewise, you should be able to check if a given evolutionary model is time-reversible and what this means. You should be able to explain why these properties are strategically important.
- You should be able to calculate the likelihood $P(\mathcal{X}|A,T)$ for a small, given binary, rooted tree T and a given un-gapped, multiple sequence alignment A whose set of sequences $\mathcal{X}$ correspond to the sequences at the leaf nodes of tree T. You should know how this calculation would have to be modified if we know the sequences for some internal tree nodes or if some of the leaf nodes correspond to gaps in the alignment.
- You should be able to explain how the Felsenstein algorithm works and which quantity it calculates. You should be able to check whether you may use a given evolutionary model in conjunction with the Felsenstein algorithm and you should know which information about the evolutionary model the Felsenstein algorithm requires.
- You should be able to outline the general strategy of finding a maximum-likelihood tree for a given evolutionary model which is time-reversible and

whose substitution matrix is multiplicative and know the corresponding theorem, its proof and the implications of this theorem.

- You should be able to explain how evolutionary information can be captured in order to improve the prediction of RNA structures. In particular, you should know how PFOLD works (input, output, SCFG, evolutionary models and all algorithms), in particular which evolutionary models it is based on and how they are used to derive the emission probabilities of the SCFG underlying PFOLD. You should know how the evolution of basepairs is modeled, i.e. what the key features of corresponding evolutionary model are, and how the emission probabilities for pairs of alignment columns are derived. Check that you know how the Felsenstein algorithm (as we introduced it) can be adapted to calculate the likelihood for a given pair of alignment columns. Make sure you know how gaps in the alignment are dealt with in PFOLD. You should be able to list and explain the advantages and dis-advantages of using this kind of comparative approach for RNA structure prediction.
- You should be able to explain how evolutionary information can be captured in order to improve the prediction of protein-coding genes in eukaryotes. In particular, you should know how EVOGENE works (input, output, HMM, evolutionary models and all algorithms), in particular which evolutionary models are employed, what their key features are and how they are employed to derive the emission probabilities of the HMM underlying EVOGENE. Check that you know how the Felsenstein algorithm (as we introduced it) can be adapted to calculate the likelihood for a given set of three consecutive alignment columns. Make sure you know how gaps in the alignment are dealt with in EVOGENE. As for PFOLD, you should be able to discuss the advantages and dis-advantages of using this kind of comparative approach for the prediction of protein-coding genes.
- Examples: Jukes-Cantor model, Kimura model