COSC 348: Computing for Bioinformatics

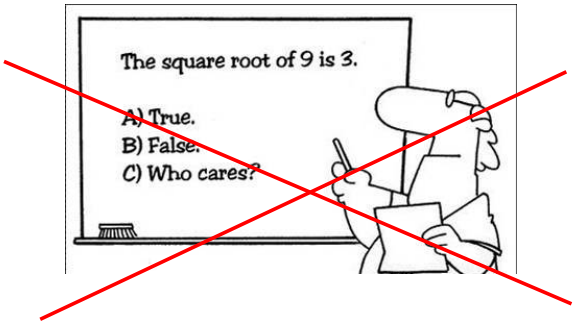
Lecture 26:
Revision and topics for the final exam

Lubica Benuskova, Ph.D.

http://www.cs.otago.ac.nz/cosc348/

What will be examined ??????????

- There will be **no** multi-choice questions.



Bioinformatics computing we've learned about:

- **Basic concepts** of molecular biology (4 lectures + 1 lab (the 2nd lab))
  - DNA, RNA, transcription, translation
  - Genetic code (properties, characteristics), mutations in DNA/RNA
  - Protein structure inference (lecture 24)
- **Sequence analysis** algorithms (5 lectures + 3 labs)
  - Exact string matching algorithms
  - Global and local sequence alignment
  - Discovery of biologically significant motifs
- **Hidden Markov Models** (HMM) (3 lectures + 2 labs)
  - How to construct the HMM from the sequence alignment
  - Applications of HMM

Bioinformatics computing we've learned about (cont'd)

- **Construction of phylogenetic trees** (5 lectures + 3 labs)
  - Construction of phylogenetic trees based on parsimony (i.e. tree cost)
  - Computing the tree cost (Fitch and Sankoff method)
  - algorithm for building a tree structure (adding the leaves and HTUs)
  - Greedy algorithm for building the optimal tree and its variants
  - optimisation of the cost (genetic algorithm, simulated annealing)
  - Construction of phylogenetic trees based on clustering
  - Principle of agglomerative, bottom-up, clustering method

Bioinformatics computing we've learned about (cont'd)

- **Analysis of microarray data** (5 lectures + 2 labs)
  - Steps in microarray data analysis
  - Gene selection (why we do it and how, t-test, SNR)
  - Clustering (K-means method, why we do clustering)
  - classification (principles, methods – SVM, MLP, k-nearest neighbours)
  - generalisation based on microarray data, what is its goal?
- **Modelling gene regulatory networks** (3 lectures + 2 labs)
  - Boolean networks (principles, assumptions, properties)
  - Weight matrices (principles, assumptions, properties)
  - Ordinary differential equations (principles, assumptions, properties)

Structure of the final exam

- Time allowed: **3 hours**
- No supplementary material is provided for the examination:
  - I.e., no reference books, notes, or other written/spoken material allowed.
- Bring the **CALCULATOR**. No restriction on the model of calculator, but no communication capability. They can be inspected.
- There are **5 questions, 12 marks each**. Answer **ALL 5** questions.
  - Each question has a number of sub-questions (individually marked)
- When a question asks for a numerical answer a working of how that answer was obtained is required.

Examples of questions from basic notions

- What is DNA? What is RNA? What constitutes a gene ?
- How are DNA and RNA related? What is a genetic code?
- What is coding and non-coding DNA for?
- What's transcription? What's translation? What's splicing?
- What is gene expression ? What is a transcription factor?
- How is the gene expression regulated?

Example questions: exact string matching

- A question may concern any of the 5 algorithms used to find occurrences of a pattern, **pat**, in a text, **txt**, which we covered in lecture 4:
  - Describe how the algorithm works on an example of txt and pat.
  - What is its worst-case complexity in terms of the length of pat and the length of txt?
- If a text contains 1000 characters, and a pattern 10 characters, what is the minimum number of character comparisons which might ensure that the pattern **does not** occur in the text?
  - Which exact string algorithm come closest to achieving this bound ?
  - Why (basically, describe the algorithm)?

Example questions: alignments

- What is an alignment score?
  - Why do we need it?
  - How do we (they) construct it?

	C	T	A	G
C	+2	+1	-1	-1
T	+1	+2	-1	-1
A	-1	-1	+2	+1
G	-1	-1	+1	+2

- Using the brute force method and a given substitution matrix (scoring scheme) with the constant gap penalty g = −10, find the best global pair-wise alignment between the sequences ATGGCG and ATGAG.

Example questions: alignments

- Describe the Needlemann-Wunsch algorithm for computing optimal global alignments and illustrate it with the example of aligning CATT and CTG if matches score 8, mismatches −4, and insertions or deletions −10.

S		C	A	T	T
	0	-10	-20	-30	-40
C	-10	?			
T	-20				
G	-30				

T		C	A	T	T
	done	left	left	left	left
C	up	?			
T	up				
G	up				

Motif discovery

```
HEM13   CCCATTGTTCTC
HEM13   TTTCTGGTTCTC
HEM13   TCAATTGTTTAG
ANB1    CTCATTGTTGTC
ANB1    TCCATTGTTCTC
ANB1    CCTATTGTTCTC
ANB1    TCCATTGTTCGT
ROX1    CCAATTGTTTTG
```

N aligned sequences of DNA

```
        YCHATTGTTCTC
```

Extract consensus sequence

```
A 002700000010
C 464100000505
G 000001800112
T 422087088261
```

Write letter count (profile table)

Extract sequence logo

Scoring strings with a Profile

Given a profile matrix **P** =

A	0.5	0.8	0.3	0
C	0.1	0	0.5	0.6
T	0.1	0.2	0	0
G	0.3	0	0.2	0.4

Note: in calculations of scores replace zeros with 0.1

Find the P-most probable 4-mer in this sequence :

CTATAAACCTT

Hidden Markov models

- Task: given these sequences how would you infer the underlying HMM?
- What are HMM good for?
- If you have several HMMs how would you decide which one is the best model for your sequence?

```
ctataaacgttacatc
atagcgattcgactga
cagcccagaaccctcc
cggtataccttacatc
tgcattcaatagctta
tatcctttccactcac
ctccaaatcctttaca
ggtcatcctttatcct
```

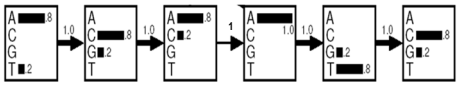

Inference/training of HMM based on alignment

```
1) A C A A T G
2) T C A A T C
3) A C A A G C
4) A G A A T C
5) A C C A T C
```

First we perform global alignment of *n* sequences, we assume there are as many states as letters

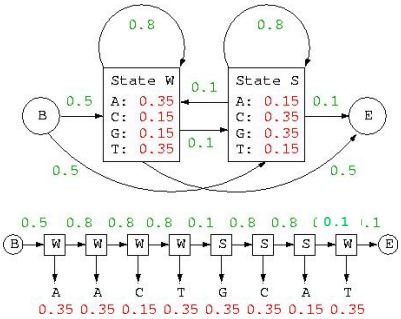
Observation probability of each letter at a given position is derived from the frequency. If these frequencies are the same at several positions, then we can collapse two or more states into one.

Transition probability: in our simple case $P(X_t \mid X_{t-1}) = 1.0$



Which HMM is the best for the query sequence?

- The ideal HMM is a *minimal* model against which all the query sequences will have the highest scores compared to any other HMMs.



Score = product of concrete emission probabilities

Examples of questions for phylogeny

- Example questions:
  - What kind of tree do we use to represent a phylogeny? Name all the parts and how they are related.
  - Why might this model fail to represent reality?
  - Parsimony approaches are based on minimising some criterion. What are those criteria?
  - Can we be sure what criterion Nature uses? Can we be sure that Nature generates optimal trees? If so, how; if not, why would we want them?
  - What assumption is the clustering method based on?
  - Put in contrast the parsimony and clustering method in phylogenetic tree construction.

Example phylogeny question

- Construct the rooted phylogenetic tree for the 3 species below. Calculate its Fitch cost and infer the characteristics of HTUs based on these characteristics: excessive body hair (present, absent); brain size (small, medium, large) and picking the nose (present, absent).

Australopithecus Afarensis

Homo Erectus

Homo Sapiens Sapiens

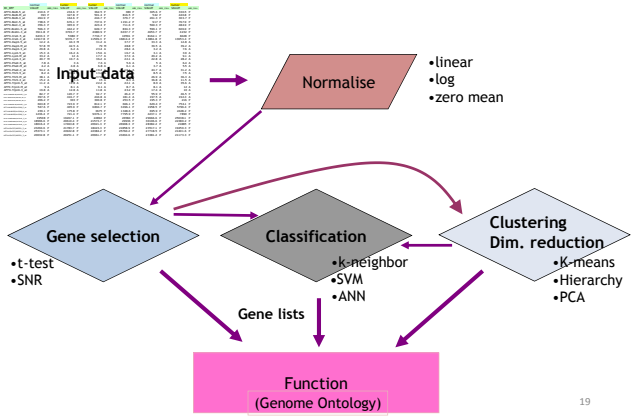
Evolutionary time

Examples of questions for clustering

- We introduced two methods of clustering:
  - Describe three ways to define a distance between two clusters, given objects in those clusters and the object/object distance matrix. Why defining a distance measure is the key for clustering?
  - Explain how bottom-up clustering works. Illustrate your explanation by showing what happens to any example data
  - Explain how the K-means clustering works. Illustrate your explanation by showing what happens to any example data

	A	B	C	D
A	0	4	3	6
B	4	0	1	2
C	3	1	0	5
D	6	2	5	0

Flowchart of microarray analysis



Gene regulatory network: Boolean network

	t1	t2	t3	t4	t5	t6
Gene A	1	1	0	0	0	0
Gene B	1	1	1	0	0	0
Gene C	0	1	1	1	0	0

- Find the set of Boolean functions describing the state of genes A, B and C in time t+1 based on states of A, B, C in time t.
- These Boolean functions have to hold for every transition, i.e.:
  - C(t+1) = B(t)
  - B(t+1) = A(t)
  - A(t+1) = B(t) & –C(t)

Weight matrix of gene interactions

- Write down the equation for GRN modelled by the weight matrix. What does the weight matrix **W** represent?
- Regulatory interactions between genes are modeled with a weight matrix **W** such that

$$g_i(t+1) = \sigma\,[\sum_j w_{ij}\, g_j(t)]$$

  - w_{ij} = Regulatory influence of gene *j* on gene *i* is assumed to be constant
  - $g_j(t)$ = Expression level of gene j at time t (mRNA level)
  - σ is a nonlinear saturation function, usually sigmoid

- If all the gene expression levels are given for time *t* and the gene regulatory network matrix **W** is given, how would you calculate all gene expressions at time *t*+1?

Preparation for the final exam worth 60%

- Lecture notes and lab notes.
- I will not ask anything that was **not** covered in the lectures or labs.
- An excellent source are the past exams.
- All past exams from 2004 to 2013 are available at: http://www.otago.ac.nz/library/exams

Which years/questions you should look at:

- 2010-2013: all questions (note: the paper was not offered in 2014).
- 2009: Questions 1, 2, 3, 4, 5 (except 5a & 5b), 6 (except 6b), 7.
- 2008: Questions 1, 2 (except 2e), 3, 4, 5 (except 5c), 6 (except 6a), 7, 9.
- 2007: Questions 1 (except c), 2 (except 2b), 7 (except 7a), 8, 9 (except 9b), 10.
- 2006: Questions 6, 7, and 8 (except 8a), 9a.
- No questions from 2005 and 2004.

Strategy

- You will notice there are overlaps between the exam questions.
- Identify those overlaps and focus on the common topics, which occur repeatedly, and study for these topics.
- Create small study groups, communicate with each other, exchange answers, email me or email me to arrange a meeting in person.
- FAREWELL & GOOD LUCK !