Bioinformatics and Statistical Genetics **Final Exam**
Jan Graffelman and Gabriel Valiente **January 21, 2016**

Name: .....

Question:	1	2	3	4	5	6	7	8	9	10	Total
Points:	10	10	10	10	10	10	10	10	10	10	100
Score:											

1. Consider the following double-stranded DNA sequence.

```
ACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCATTCAACCATACCACTCCGAAC
CACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATC
CAACCCACTGCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATAC
TGTTCTTCTACCCACCATATTGAAACGCTAACAAATGATCGTAAATAACACACACGTGCT
TACCCTACCACTTTATACCACCACCACATGCCATACTCACCCTCACTTGTATACTGATTT
TACGTACGCACACGGATGCTACAGTATATACCATCTCAAACTTACCCTACTCTCAGATTC
```

1 point (a) How many reading frames does it have?

(a) _____

3 points (b) How many open reading frames does it have? Remember that the start codon is ATG and the stop codons are TAA, TAG and TGA.

(b) _____

4 points (c) What is the length of the longest open reading frame of the sequence?

(c) _____

2 points (d) In what reading frame would you expect translation to take place, assuming the sequence is eukaryotic?

(d) _____

4 points 2. (a) The number $R(n)$ of possible RNA secondary structures of length n is given by

$$R(n+1) = R(n) + \sum_{j=1}^{n-1} R(j-1)R(n-j)$$

Explain why.

4 points (b) The number $R(i,j)$ of possible RNA secondary structures from position i to position j of an RNA sequence is given by

$$R(i,j) = R(i,j-1) + \sum_{k=i}^{j-1} R(i,k-1)R(k+1,j-1)$$

Explain why.

2 points (c) The RNA sequence fragment GGGUGCUCAGUACGAGAGGAACCGCACCC has 8,622,571,758 possible secondary structures, only 789,564 of which are indeed feasible. Explain why.

3. Consider the following dynamic programming table for the edit distance computation of the sequences GCTTCCGGCTCGTATAATGTGTGG and TGCTTCTGACTATAATAG.

		0	T 1	G 2	C 3	T 4	T 5	C 6	T 7	G 8	A 9	C 10	T 11	A 12	T 13	A 14	A 15	T 16	A 17	G 18
	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
G	1	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
C	2	2	2	2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	3	3	2	3	2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T	4	4	3	3	3	2	1	2	3	4	5	6	7	8	9	10	11	12	13	14
C	5	5	4	4	3	3	2	1	2	3	4	5	6	7	8	9	10	11	12	13
C	6	6	5	5	4	4	3	2	2	3	4	4	5	6	7	8	9	10	11	12
G	7	7	6	5	5	5	4	3	3	2	3	4	5	6	7	8	9	10	11	11
G	8	8	7	6	6	6	5	4	4	3	3	4	5	6	7	8	9	10	11	11
C	9	9	8	7	6	7	6	5	5	4	4	3	4	5	6	7	8	9	10	11
T	10	10	9	8	7	6	7	6	5	5	5	4	3	4	5	6	7	8	9	10
C	11	11	10	9	8	7	7	7	6	6	6	5	4	4	5	6	7	8	9	10
G	12	12	11	10	9	8	8	8	7	6	7	6	5	5	5	6	7	8	9	9
T	13	13	12	11	10	9	8	9	8	7	7	7	6	6	5	6	7	7	8	9
A	14	14	13	12	11	10	9	9	9	8	7	8	7	6	6	5	6	7	7	8
T	15	15	14	13	12	11	10	10	9	9	8	8	8	7	6	6	6	6	7	8
A	16	16	15	14	13	12	11	11	10	10	9	9	9	8	7	6	6	7	6	7
A	17	17	16	15	14	13	12	12	11	11	10	10	10	9	8	7	6	7	7	7
T	18	18	17	16	15	14	13	13	12	12	11	11	10	10	9	8	7	6	7	8
G	19	19	18	17	16	15	14	14	13	12	12	12	11	11	10	9	8	7	7	7
T	20	20	19	18	17	16	15	15	14	13	13	13	12	12	11	10	9	8	8	8
G	21	21	20	19	18	17	16	16	15	14	14	14	13	13	12	11	10	9	9	8
T	22	22	21	20	19	18	17	17	16	15	15	15	14	14	13	12	11	10	10	9
G	23	23	22	21	20	19	18	18	17	16	16	16	15	15	14	13	12	11	11	10
G	24	24	23	22	21	20	19	19	18	17	17	17	16	16	15	14	13	12	12	11

1 point (a) What is the edit distance between the two DNA sequences?

(a) _____

5 points (b) Highlight the entries that correspond to alignments implicit in the dynamic programming table.

4 points (c) List the two alignments that correspond to the boundaries of the highlighted area.

4. Consider the phylogenetic trees with Newick strings

```
((Red_Panda,Racoon),Giant_Panda,(Spectacled_Bear,(Sloth_Bear
   ,(Polar_Bear,Brown_Bear))));
```

and

```
((Red_Panda,Giant_Panda),Racoon,(Brown_Bear,(Sloth_Bear,(
   Polar_Bear,Spectacled_Bear))));
```

1 point (a) Are they rooted or unrooted?

(a) _____

3 points (b) Draw the two trees.

3 points (c) What is their partition distance?

(c) _____

3 points (d) What is their nodal distance?

(d) _____

5. Consider the following sample of eleven DNA sequence fragments.

```
CCGCAATAATGGCGCTACCCCCACAAAAACGCACTAGACAGCCT
CCCCAATATGGGCGCTACCCCCACAAAAACGCACTAGACAGCCT
CCGCAATATGGGCGCTACCCCCCGGAATCTGCACTAGACAGTCA
CCGCAATATGGGCGCTGTCCCCCGGAATCTGCACTAGACTGCCT
CCGAAATAAGTCCGAGGTCCCCCGGAATCTGCACTAGCCAGCCT
CCCCAATATGGGCGCGACCCCCCGGAATCTGTATTCGCCAGCTT
CCCCAATATGGGCGCGACCCCCCGGAATCTGTCTCCGCCAGCCT
TGCAAATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
TGCAAATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
TGCAAATAAGTCGGCGACCCCCCGGAATCTGTCTCCGCGAGCCT
TGCAAATGAGGGCTCGACCCCCCGGGATCTGTCTCCGCCAGCCT
```

2 points (a) How many single nucleotide polymorphisms are there?

(a) _____

8 points (b) Do they have a phylogenetic tree? If yes, draw it. Otherwise, explain why.

6. Are the following statements true or false?

1 point (a) The inbreeding coefficient is a measure for linkage disequilibrium. ☐ T ☐ F

1 point (b) For a marker with k alleles there are $k(k+1)/2$ possible heterozygotes. ☐ T ☐ F

1 point (c) The nucleus of the body cells of human beings contain 44 autosomes. ☐ T ☐ F

1 point (d) If two alleles for the same marker are perfectly correlated, then the marker is in Hardy-Weinberg disequilibrium. ☐ T ☐ F

1 point (e) If the sample minor allele frequency of a bi-allelic marker is 0, then the sample does not contain any heterozygotes. ☐ T ☐ F

1 point (f) If a bi-allelic marker is co-dominant, then one cannot distinguish the heterozygote from one of the homozygotes. ☐ F ☐ F

1 point (g) If the expected heterozygosity for a bi-allelic marker is 0.5 then its two alleles are equally frequent. ☐ T ☐ F

1 point (h) If, for a particular marker, all genotypes in the sample are homozygotes, then the marker is called monomorphic. ☐ T ☐ F

1 point (i) Let there be 20 consecutive bi-allelic markers on the same chromosome, of which 18 are monomorphic. The maximum number of possible haplotypes in this situation is 4. ☐ T ☐ F

1 point (j) If a pair of individuals shares two alleles IBD for a locus, then they also share two alleles IBS for that locus. ☐ T ☐ F

7. A sample of individuals has been genotyped for an A/T polymorphism and the following genotype counts have been obtained:

AA	AT	TT
23	49	18

1 point (a) Calculate the total number of alleles in the sample.

1 point (b) Calculate the minor allele count.

1 point (c) Calculate the (relative) minor allele frequency.

1 point (d) Calculate the observed heterozygosity.

1 point (e) Calculate the expected heterozygosity.

2 points (f) Calculate the expected genotype frequencies under the assumption of Hardy-Weinberg equilibrium (HWE).

1 point (g) For the sample data, calculate the chi-square statistic for HWE, given by

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i and E_i are the observed and expected genotype frequencies.

1 point (h) What reference distribution does this statistic have under the null hypothesis?

1 point (i) In the light of the value of Q, do you think the null hypothesis should be rejected (you don't need to compute a p-value)?

8. In order to investigate whether a marker is in Hardy-Weinberg equilibrium or not, a permutation test can be used.

6 points (a) Explain in detail how a permutation test for HWE is performed.

2 points (b) Mention some alternative sample statistics that could be used in a permutation test for HWE.

2 points (c) Discuss advantages and disadvantages of the permutation test in comparison with other tests for HWE.

9. Consider the co-dominant test used in the context of a genetic association analysis of a single locus (a bi-allelic polymorphism) and binary response variable (diseased or not). Give an account of the theory of the co-dominant test by answering the following

questions:

2 points (a) Describe the original data table used in the co-dominant test.

1 point (b) Formally state the null hypothesis and alternative hypothesis of the test

1 point (c) What test statistic is typically used in a co-dominant test?

1 point (d) State the reference distribution of the test statistic

1 point (e) On which assumption(s) does the test rely?

2 points (f) It may occur that some genotype is very rare in the sample. Discuss the consequences for the co-dominant test and explain you would do in these circumstances.

2 points (g) Sketch a graph which depicts the risk profile for having the disease as a function of the genotype for the co-dominant test. Sketch in the same graph also the risk profiles for dominant and a recessive test.

10. Consider a bi-allelic polymorphism with alleles A and B, where we distinguish paternal alleles A and B from the corresponding maternal IBS alleles α and β. An AA man and an $\alpha\beta$ woman have children.

2 points (a) How many different types (k) of children can they have? Enumerate them.

3 points (b) Build the $k \times k$ table of all possible pairs of children and indicate for each pair how many IBD alleles and how many IBS alleles they share.

3 points (c) What is the probability vector for sharing 0, 1 or 2 IBD alleles for an FS pair of this couple? And what is the probability vector for sharing 0, 1 or 2 IBS alleles for an FS pair of this couple?

2 points (d) Also build $1 \times k$ vectors showing how many IBD and IBS alleles the father shares with each of the k possible children.