Exam: **Computational life science, 15.0 hp**
Date: **2010.11.08**
Time: **9-15**
Place: **Rotundan, Universum**
Contacts: **Torgeir R. Hvidsten (tel. 786 5248)**
Anna Linusson (tel. 786 6890)
Patrik Ryden (tel. 786 6332)

Pocket calculator and dictionary allowed. No other aids (books, notes, etc.). Answers can be given in English or Swedish.

NB: Answer the three tasks on separate paper sheets. This way the three teachers can correct in parallel and you will get the results back faster.

Task 1 (30%) Statistics

a) 7.5%
Write down the three laws of statistics.

b) 7.5%
What are the advantages and disadvantages of using parametric and non-parametric tests?

c) 7.5%
Construct an approximate 95% confidence interval for mu (the true mean in the population) for an experiment with 81 observations with the observed sample mean equal to 13 and the sample standard deviation equal to 9. Can the null hypothesis (H0: mu=5, H1:mu≠5) be rejected at the 5% significance level?

d) 7.5%
Which of the following statements are true?

a. The correlation is always between -1 and 1.
b. ANOVA is a non-parametric test for comparing the true means in several populations.
c. The null hypothesis is rejected when the p-value is lower than the test's significance level.
d. The power of a test increases with the sample size.
e. The standard deviation is always lower than the median.

Task 2 (35%) Bioinformatics

a) (10%) Perl and programming

1. What is the difference between an array and a hash in Perl?

2. Write a Perl program that computes the sum of all elements in a list, e.g.

Input: (5,2,2,9,5,4), Output: 27

What can you say about the time complexity of your program?

3. What is the difference between call-by-value and call-by-reference when using procedures in Perl?

b) (10%) Sequence alignment

1. Consider the following scoring matrix δ for sequences in a six letter alphabet

	A	**B**	**M**	**O**	**S**	**T**	**-**
A	1	-1	-1	-2	-2	-3	-1
B		1	-1	-1	-2	-2	-1
M			2	-1	-1	-2	-1
O				1	-1	-1	-1
S					1	-1	-1
T						2	-1
-							-1

and the two sequences:

***v* = MOAT**

***w* = BOAST**

Fill out the dynamic programming table for a *local* alignment between ***v*** and ***w*** under the scoring matrix δ. Draw arrows in the cells to store the backtrack information. What is the score for the optimal alignment and what alignment(s) does this score correspond to?

2. Explain the general idea behind PSI-BLAST. How does PSI-BLAST compare to BLAST in terms of sensitivity and specificity. Explain.

c) (5%) Multiple alignments

Describe a greedy algorithm for constructing multiple alignments. Why is this algorithm greedy? Can you construct a simple example of three short sequences where this algorithm will not construct the optimal alignment?

d) (10%) Machine learning and function prediction

1. What is supervised and unsupervised learning? Briefly describe the general idea behind one supervised method and one unsupervised method.

2. Computational methods for protein function predictions can be based on one or more of these concepts:

1. Sequence alignment
2. Sequence-motifs
3. Structure alignment
4. Structure-motif

Shortly (2-3 sentences) describe the main idea behind a method based on each of these concepts. Then rank them by sensitivity with a short motivation.

Task 3 (35%) Chemometrics

a) (20%)

You have investigated three cake ingredients (flour, shortening, and egg powder) according to the matrix below.

1. What is the resolution of the fractional factorial design?
2. Make a detailed analysis of the confounding pattern and describe what other contributions that "hide" in the coefficients presented in the model below.
3. Calculate R^2 for the model.
    a. Interpret the results.
4. Calculate the pure error and the model error of your regression model.
    a. Use that information to judge the quality of the model. An F table for alpha = 0.05 is given on the last page of this exam.
5. Give four new experiments that could be used (together with the data in the matrix) to construct a model with no confounding.

X1	X2	X3	Y
flour	shorten	egg powder	taste
400	50	50	3,66
200	50	100	5,38
200	100	50	4,74
400	100	100	4,97
300	75	75	4,73
300	75	75	4,61
300	75	75	4,68

Model: y = 4.68 -0.37*X1 + 0.17*X2 + 0.49X3

b) (8%)

1. Explain the differences between fundamental models and empirical models.
2. How do you validate an empirical model?

c) (7%)

Give a geometrical explanation of principal component analysis.

F-table for alpha = 0.05: F(0.05,df1,df2)

df2/df1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	inf
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
inf	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00