Supplementary test
Introduction to Bioinformatics, 2014

Name, Neptun code:
I am attending the exam today: yes/no

The point value of each question is listed in parentheses.

Total points: 100

1. Give a narrow and broad **definition of bioinformatics**! (4)

Narrow: Science of biological data. Mostly molecular biology. Description, management, interpretation.

Broad: Science of biological knowledge. All computer applications in (molecular) biology including modeling. (simulation of behavior)

2. Describe **structured** and **unstructured** data representation! (4)

Structured: - We known the internal structure in terms of Entities and Relationships
- Information-rich, allows detailed comparisons
- need alignment for comparison
- example: sequences, graphs

unstructered: - We know nothing about internal structu
- Only the properties are known (global desc
can be discreet or continuous.

3. Fill in the **missing** elements! (4)

	Entities	Relationships

- Best described as vecto

modeling. (simulation of behavior)

2. Describe **structured** and **unstructured** data representation! (4)

Structured: - We known the internal structure in terms of Entities and Relationships
- Information-rich, allows detailed comparisons.
- need alignment for comparison
- Example: sequences, graphs

unstructered: - We known nothing about internal structure
- Only the properties are known (global descr.)
can be discreet or continuous.
- Best described as vectors
Sometimes large number of dimensions
- Vector operations are fast

3. Fill in the **missing** elements! (4)

Models	Entities	Relationships
Molecules	Atoms	Atomic interactions (chemical bonds)
Pathways	Enzymes	chemical reactions
Genetic networks	Genes	Co-regulation
Protein structure	Atoms	chemical bonds
Folds	Cα atoms	Peptide bond

4. Define **Levenshtein edit-distance** and **Hamming distance**! (4)

Levenshtein: edit distance between character string defined as a sum of costs assigned to matches, replacements and gaps.
2 strings do not need to be of the same length

number of exchanges necessary to turn one

3. Fill in the **missing** elements! (4)

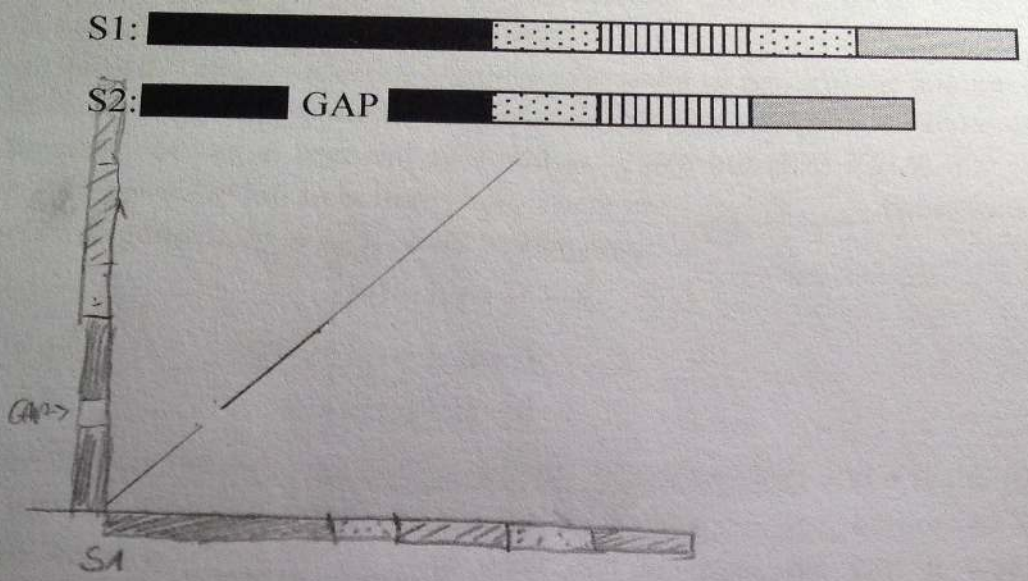
Models	Entities	Relationships
Molecules	Atoms	Atomic interactions (chemical bonds)
Pathways	Enzymes	chemical reactions
Genetic networks	Genes	co-regulation
Protein structure	Atoms	chemical bonds
Folds	Cα atoms	Peptide bond

discrete or continuous.
- Best described as vectors sometimes large number of dimensions
- Vector operations are fast

4. Define **Levenshtein edit-distance** and **Hamming distance**! (4)

Levenshtein: edit distance between character string defined as a sum of costs assigned to matches, replacements and gaps.
2 strings do not need to be of the same length

Hamming: number of exchanges necessary to turn one string of bits or characters into another one.
2 strings are of identical length and no alignment is done
- The exchanges in character strings can have different costs, stored in a lookup table.
In this case the value of the Hamming distance will be the sum of costs, rather than the number of exchanges



6. Calculate the **log odds ratio** of residues Alanine (A) and Valine (V) using the following multiple alignment! (4)

A	V	F	R	V	G
V	T	F	V	A	A
A	V	V	S	A	A
L	V	A	N	V	H

$g(A) = 7 \quad g(V) = 8 \quad g(A \wedge V) = 5$

$m(A/V) = \frac{g(A \wedge V)}{g(A) \cdot g(V)} = \frac{5}{7 \cdot 8} = \frac{5}{56}$

$M(A \wedge V) = \log(m(A/V)) = -1{,}05$

7. Define the following terms: (3)

running time: Defined as number of steps to be carried out.

V T F V A A
A V V S A A
L V A N V H

$$m(A/V) = \frac{-(A\wedge V)}{g(A)\cdot g(V)} = \frac{5}{7\cdot 8} = \frac{5}{56}$$

$$M(A/V) = \log(m(A/V)) = -1{,}05$$

7. Define the following terms: (3)

running time: Defined as number of steps to be carried out.

average case: using an average input problem-dependent, not trivial

worst case: inputs leading to an upper limit of running time / memory

8. Fill in the **missing** words! (9)

_____ algorithms find the best solution by constructing the solution from the stored optimal solutions of sub-problems. The Needleman-Wunsch algorithm is a well-known solution for global sequence alignment. In order to perform the alignment, a matrix is created which allows us to compare the two sequences. The score *F(i,j)* is the score of the best alignment between the initial segment $x_{1\ldots i}$ and $y_{1\ldots j}$. We initialize $F(0,0) = 0$. Then we proceed to fill the matrix from top left to bottom right. There are 3 possible ways that the best score *F(i,j)* of an alignment up to x_i, y_j could be obtained:

$$F(i,j) = \text{max}\begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

where $s(x_i,y_j)$ is the score value for aligning x_i and y_j and *d* is the penalty for inserting a _____. The equation is applied repeatedly until we reach the bottom right corner. As we fill in the *F(i,j)* values we also keep a _____ in each cell showing the cell it was derived from. Using these _____ we can _____ the alignment of our sequences.

Smith-Waterman algorithm is a solution for local sequence alignment. The algorithm is similar to global alignment, there are two main differences. First, at calculating F matrix an extra possibility is added allowing *F(i,j)* to take the value 0 if all other options are negative _____. If the best alignment up to som

solution for global sequence alignment. In order to perform the alignment, a matrix is created which allows us to compare the two sequences. The score $F(i,j)$ is the score of the best alignment between the initial segment $x_{1...i}$ and $y_{1...j}$. We initialize $F(0,0) = 0$. Then we proceed to fill the matrix from top left to bottom right. There are 3 possible ways that the best score $F(i,j)$ of an alignment up to x_i, y_j could be obtained:

$$F(i,j) = \text{max}\begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

where $s(x_i,y_j)$ is the score value for aligning x_i and y_j and d is the penalty for inserting a gap. The equation is applied repeatedly until we reach the bottom-right corner. As we fill in the $F(i,j)$ values we also keep a _____ in each cell showing the cell it was derived from. Using these _____ we can _____ the alignment of our sequences.

Smith-Waterman algorithm is a solution for local sequence alignment. The algorithm is similar to global alignment, there are two main differences. First, at calculating F matrix an extra possibility is added allowing $F(i,j)$ to take the value 0 if all other options are negative. Taking this option corresponds to starting a new _____. If the best alignment up to some point has a _____ score, it is better to start a new one rather than extend the old one. The second change is that now an alignment can end anywhere in the matrix, so at _____ instead of taking the value in the bottom-right corner, we look for the largest value $F(i,j)$ in the whole matrix, and start the backtrack from there.

9. How can you validate a multiple alignment using **3D structure** information? (2)

1. convert 3D structure into a series of secondary structure assignments, and write on top of the alignm

2. Maps conserved regions directly to 3D
eg. color the 3D with the entropy plot values
variable regions should maps to the surface of the

Smith-Waterman algorithm is a solution for local sequence alignment. The algorithm is similar to global alignment, there are two main differences. First, at calculating F matrix an extra possibility is added allowing $F(i,j)$ to take the value 0 if all other options are negative. Taking this option corresponds to starting a new _____. If the best alignment up to some point has a _____ score, it is better to start a new one rather than extend the old one. The second change is that now an alignment can end anywhere in the matrix, so at _____ instead of taking the value in the bottom-right corner, we look for the largest value $F(i,j)$ in the whole matrix, and start the backtrack from there.

9. How can you validate a multiple alignment using **3D structure** information? (2)

1. convert 3D structure into a series of secondary structure assignments, and write on top of the alignment

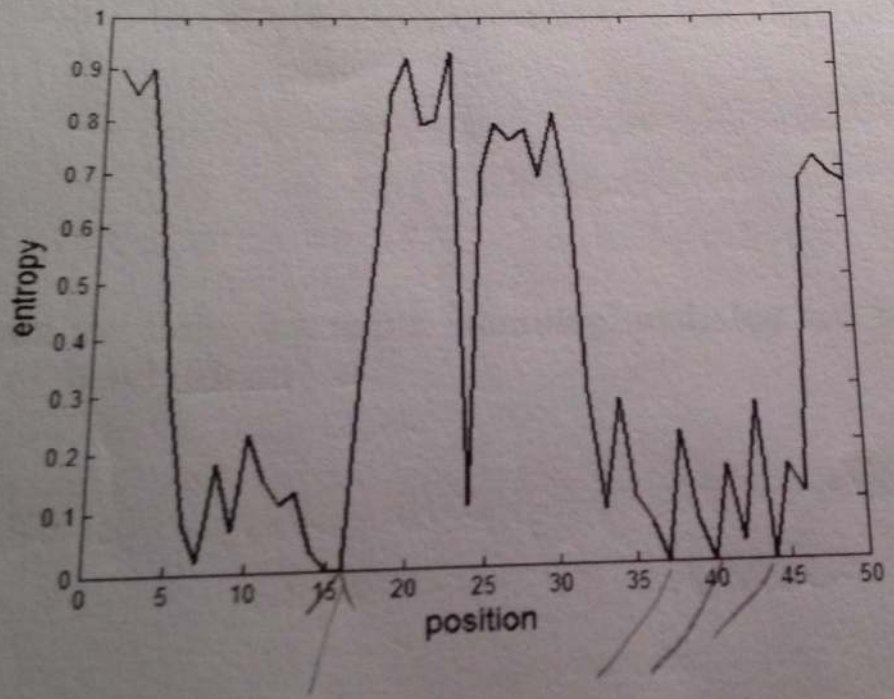
2. Maps conserved regions directly to 3D
eg. color the 3D with the entropy plot values
variable regions should maps to the surface of the protein, not in the buried parts

10. Which sequence has lower **complexity**? Support your statement by calculation! (3)
1. CGAGTAGCTCTGCTAA
2. GAGTGTCTTCTATTTG

Complexity: an empirical measure, proportional to the number of words (of arbitrary length) necessary to reproduce a sequence.

seq_1 seq_2 seq_3 seq_4 seq_5 seq_6 seq_7 seq_8 seq_9 seq_10
sequence name

12. What can you conclude from the multiple alignment that has generated the following normalized **entropy plot**? (2)



Best: 0
Worst: 2

dilages eset

conserved regions

WHAT THE FUCK?

16. Given the following distance matrix draw a dendrogram based on the **UPGMA** method and briefly explain your calculation! (6)

	A	B	C	D
B	8			
C	6	8		
D	8	2	8	
E	4	8	6	8

17. Explain how **BLAST** algorithm works (what are the BLAST tricks). List the main computational steps and briefly describe them! (8)

→ twice

17. Explain how **BLAST** algorithm works (what are the BLAST tricks). List the main computational steps and briefly describe them! (8)

⇒ wait

18. List 5 applications of **next generation sequencing**! (5)

medical, microbiological

agricultural, such as only human cancer genes, only microbial genomes, etc

19. Compare traditional (**Sanger**) and **next generation sequencing** (list the advantages, disadvantages)! (4)

Sanger:

- Accurate
- Also works on few samples
- Expensive for data
- Small capital investment
- Slow

NSG:

- Less accurate
- Shorter reads
- Economical only with many samples
- ~1000 less expensive for data
- Large capital investment
- very fast

20. What is the order of the following steps/stages at **NGS** sequencing? (4)

(a) alignment (b) fragmented DNA (c) sample DNA (d) parallel sequencing (e) sequence (f) DNA fragments with sequencing adapters (g) PCR amplification

21. What are the future challenges of **NGS in clinical medicine**? Shortly explain at least 3 examples! (3)

caallenges for NGS: accuracy, interpretation, storage

Accuracy: 99.99% accuracy ×3 billion nucleotides
= 300 000 errors per patient
need for confirmation at present

- Expensive for data
- Small capital investment
- Slow

- Shorter reads
- Economical only with many samples
- ~1000 less expensive for data
- Large capital investment
- very fast

20. What is the order of the following steps/stages at **NGS** sequencing? (4)

(a) alignment (b) fragmented DNA (c) sample DNA (d) parallel sequencing (e) sequence (f) DNA fragments with sequencing adapters (g) PCR amplification

21. What are the future challenges of **NGS in clinical medicine**? Shortly explain at least 3 examples! (3)

caallenges for NGS: accuracy, interpretation, storage

Acceracy: 99.99% accuracy ×3 billion nucleotides
= 300 000 errors per patient
need for confirmation at present

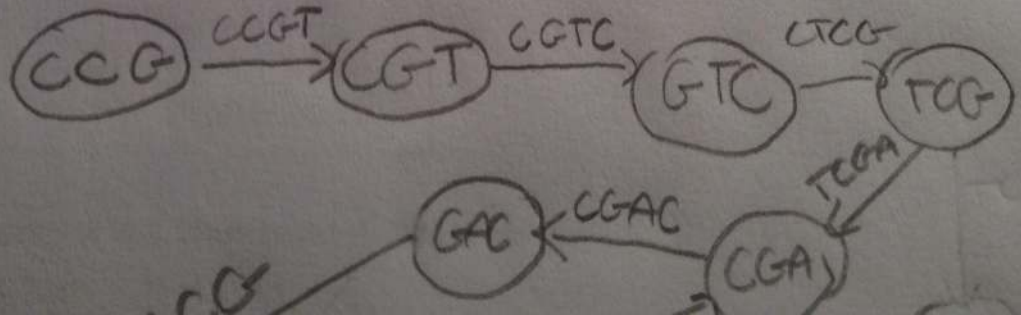
Interpretion of the variant we find

Storage and access in the medical record → each have ~4 million variant

22. Given the reads {CCGT, GAAA, AAAC, AACG, CGTC, GTCG, GACG, ACGA, TCGA, CGAC, CGAA}. What is the original sequence? Use the **de Bruijn graph** (k-mer = 3) based algorithm for the computation! Very briefly describe your computation! (6)

CCGTCGAAACGACG

CCGT
CGTC
GTCG
TCGA
CGAA
GAAA
AAAC
AACG



21. What are the future challenges of **NGS in clinical medicine**? Shortly explain at least 3 examples! (3)

challenges for NGS: accuracy, interpretation, storage

Accuracy: 99.99% accuracy × 3 billion nucleotides
= 300 000 errors per patient
need for confirmation at present

Interpretation of the variant we find

Storage and access in the medical record → each have ~6 million variant

22. Given the reads {CCGT, GAAA, AAAC, AACG, CGTC, GTCG, GACG, ACGA, TCGA, CGAC, CGAA}. What is the original sequence? Use the **de Bruijn graph** (k-mer = 3) based algorithm for the computation! Very briefly describe your computation! (6)

CCGT
CGTC
GTCG
TCGA
CGAA
GAAA
AAAC
AACG
ACGA
CGAC
GACG

CCGTCGAAACGACG és talált még a weboldal adta