Ime in priimek (s tiskanimi črkami) / Name (please print): _____

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

2. izpitni rok / Second Examination Period

10. februar 2014 / February 10, 2014

Naloga / Exercise	1	2	3	4	5	Vsota / Sum
Vrednost / Max	6	6	8	6	6	32
Točk / Points						

[6] 1. V danem zaporedju želimo poiskati vse možne bralne okvire in prevesti v zaporedje aminokislin na podlagi podane standardne tabele (za pričetek vzemite le `ATG`, za konec pa {`TAA`,`TAG`,`TGA`}). Poročajte le o proteinih z vsaj štirimi aminokislinami.

TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys
TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys
TTA L Leu	TCA S Ser	TAA * Ter	TGA * Ter
TTG L Leu i	TCG S Ser	TAG * Ter	TGG W Trp
CTT L Leu	CCT P Pro	CAT H His	CGT R Arg
CTC L Leu	CCC P Pro	CAC H His	CGC R Arg
CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CTG L Leu i	CCG P Pro	CAG Q Gln	CGG R Arg
ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile	ACC T Thr	AAC N Asn	AGC S Ser
ATA I Ile	ACA T Thr	AAA K Lys	AGA R Arg
ATG M Met i	ACG T Thr	AAG K Lys	AGG R Arg
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly

Given is a nucleotide sequence for which we would like to find all open reading frames. Assume `ATG` for a start codon, and {`TAA`, `TAG`, and `TGA`}) for the end codons. Report only on proteins with at least four aminoacids.

TCATGTTTATTCGTTCTACTTAAAGGTTAATCGTTCTGACACTCAGACATTTATGACTCACATTCGTGATTAAGGT

Page for your solutions. / Stran za vaše rešitve.

[6] 2. Given are two amino acid sequences

```
BCNJD
DNVBJ
```

and BLOSUM50 scoring matrix. Align the two sequences using the procedure for optimal local alignment. The gap penalty is equal to -5. The solution should include the alignment, dynamic programming table and final score.

Dani sta naslednji zaporedji aminokislin:

```
BCNJD
DNVBJ
```

ter matrika ocenjevalne funkcije (BLOSUM50). Poravnaj zaporedji z uporabo postopka za najboljšo lokalno poravnavo. Kazen za vrzel je enaka -5. Rešitev naj vsebuje poravnani zaporedji, tabelo delnih rešitev in končno oceno poravnave.

Subset of BLOSUM50:

```
   N  D  C  V  B  J  *
N  7  2 -2 -3  5 -4 -5
D  2  8 -4 -4  6 -4 -5
C -2 -4 13 -1 -3 -2 -5
V -3 -4 -1  5 -3  2 -5
B  5  6 -3 -3  6 -4 -5
J -4 -4 -2  2 -4  4 -5
* -5 -5 -5 -5 -5 -5  1
```

$$M_{i,j} = \max\left(M_{i-1,j} + \sigma(s_i, -), M_{i,j-1} + \sigma(-, t_j), M_{i-1,j-1} + \sigma(s_i, t_j), 0\right)$$

Page for your solutions. / Stran za vaše rešitve.

3. We are observing a cheating dice thrower. We know all the probabilities to define a Hidden Markov Model: probabilities of the thrower to choose between a fair (F) or a loaded (L) die, and the probabilities of the outcome of each die. From these and the observed sequence we can derive the tables using the forward and backward algorithms.

Observed sequence: 36513645

[2] (a) Fill-in the missing values in the forward and backward tables. (notation $e^x = \times 10^x$)

[4] (b) How many rolls were the result of the fair die according to posterior decoding? On what fraction does the predicted sequence agree with the true sequence? Assume the true sequence of hidden states is LLLFFFLL.

[2] (c) Answer in one or two sentences: How would you estimate the parameters of the Hidden Markov model (the transition and emission probabilities) if you were only given the observed sequence and the number of hidden states?

Opazujemo metanje igralne kocke nepoštenega metalca. Poznamo vse verjetnosti za opis metanja: verjetnost menjave med pošteno (F) in obteženo kocko (L), ter verjetnosti izidov posamezne kocke. Za opazovano zaporednje metov smo deloma izpolnili tabeli algoritmov *forward* in *backward*.

(a) Dopolnite tabeli algoritmnov *forward* in *backward* (notacija $e^x : \times 10^x$).

(b) Koliko metov je bilo izvedenih s pošteno kocko, če uporabimo *posterior decoding*? Kakšno je ujemanje med napovedanim in pravim zaporednjem skritih stanj, če vemo, da je pravo zaporedje LLLFFFLL?

(c) Odgovorite v enem ali dveh stavkih: kako bi ocenili parametre modela (verjetnosti prehodov med stanji in verjetnosti vidnih simbolov v posameznem stanju), če bi poznali le opazovano sekvenco ter število skritih stanj?

Opazovano zaporedje: 36513645

	Viterbi	Forward
Initialisation:	$v_0(0) = 1$, $v_k(0) = 0$ for $k \neq 0$	$f_0(0) = 1, f_k(0) = 0$ for $k \neq 0$
$i = 1 \dots L$:	$v_l(i) = e_l(x_i) max_k(v_k(i-1)t_{kl})$	$f_l(i) = e_l(x_i) \sum_k f_k(i-1)t_{kl}$
Termination:	$P_v = \max_k(v_k(L))$	$P_f = \sum_k f_k(L)$

	Backward	Posterior decoding
Initialisation:	$b_k(L) = 1$ for all k	
$i = L-1 \dots 1$:	$b_k(i) = \sum_l t_{kl} e_l(x_{i+1}) b_l(i+1)$	$pd_k(i) = \frac{f_k(i) b_k(i)}{P_f}$
Termination:	$P_b = \sum_l t_{0l} e_l b_l(1)$	$P_{pd}(i) = \text{argmax}_k\ pd_k(i)$

$t_{0F} = 0.5$; $t_{0L} = 0.5$
$t_{FF} = 0.75$; $t_{FF} = 0.25$
$t_{LF} = 0.25$; $t_{LL} = 0.75$

$e_F(1) = 0.167$; $e_F(2) = 0.167$; $e_F(3) = 0.167$; $e_F(4) = 0.167$; $e_F(5) = 0.167$; $e_F(6) = 0.167$;
$e_L(1) = 0.0$; $e_L(2) = 0.0$; $e_L(3) = 0.1$; $e_L(4) = 0.2$; $e_L(5) = 0.3$; $e_L(6) = 0.4$;

Forward:

$$\begin{pmatrix} \mathbf{0} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & ? \\ \mathbf{F} & 0 & 8.33e^{-2} & 1.25e^{-2} & 2.53e^{-3} & 5.75e^{-4} & 7.18e^{-5} & 9.58e^{-6} & 1.68e^{-6} & ? \\ \mathbf{L} & 0 & 5.00e^{-2} & 2.33e^{-2} & 6.19e^{-3} & 0 & 1.44e^{-5} & 1.15e^{-5} & 2.20e^{-6} & ? \end{pmatrix}.$$

Backward:

$$\begin{pmatrix} \mathbf{0} & 0 & 7.19e^{-6} & 2.68e^{-5} & 1.34e^{-4} & 1.61e^{-3} & 1.29e^{-2} & 4.33e^{-2} & 2.33e^{-1} & 1 \\ \mathbf{F} & 0 & ? & 3.01e^{-5} & 2.01e^{-4} & 1.61e^{-3} & 9.63e^{-3} & 3.83e^{-2} & 2.00e^{-1} & 1 \\ \mathbf{L} & 0 & ? & 2.34e^{-5} & 6.69e^{-5} & 1.61e^{-3} & 1.61e^{-2} & 4.83e^{-2} & 2.67e^{-1} & 1 \end{pmatrix}.$$

4. You have identified a group of highly conserved genes (given in set C). Explain how to calculate the p-value for the enrichment of a Gene Ontology term T in your set C. Genes known to be associated to the Gene Ontology term T of interest are given in set T_{genes}. Use the hypergeometric distribution (formula below).

[3] (a) How should you set parameters m, n, and N of the hypergeometric distribution? What is k?

[2] (b) Write the formula to compute the p-value (the probability of finding such or better results at random).

[1] (c) Do lower or higher p-values correspond to more significant enrichments?

Odkrili ste skupino dobro ohranjenih genov C. Razložite, kako bi izračunali p-vrednost obogatenosti nekega pripisa T iz genske ontologije (Gene Ontology term) v odkriti skupini C. Množica vseh genov v skupini je T_{genes}. Uporabite hipergeometrijsko porazdelitev (spodnja formulo).

(a) Kako nastaviti parametre hipergeometrijske porazdelitve m, n, in N? Kaj je k?

(b) Napiši formulo za izračun p-vrednosti (verjetnost, da dobite tak ali boljši rezultat po naključju).

(c) Kakšne p-vrednosti pomenijo bolj značilno obogatenost skupine: manjše ali večje?

$$P(K = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \qquad \binom{N}{n} = \frac{N!}{(N-n)!n!}$$

Page for your solutions. / Stran za vaše rešitve.

[6] 5. Gene regulation networks can be inferred through epistasis, where one gene can block the other one in the pathway for a specific phenotype.

Given is a set of 9 experiments. We have observed a phenotype for the wild type organism (E1), different single mutants (E2 to E6), and different double mutants (E7 to E9). Genes were either knocked-out (e.g, B-) or over-expressed (e.g., A+). The phenotype can have tree values: n (decreased), 0, p (increased), where phenotype 0 is a wild-type phenotype.

ID	Gene 1	Gene 2	phenotype
E1			0
E2	A+		n
E3	B-		p
E4	C-		n
E5	D-		p
E6	E+		n
E7	B-	C-	n
E8	D-	C-	n
E9	E+	D-	p

Perform epistasis analysis on experimental data to derive and draw a gene network that fits the data. Indicate the type of influence by using appropriate arrow: $->$ or $-|$.

Pri gradnji genskih regulacijskih mrež iz fenotipskih podatkov o mutantih smo govorili o epistazi, pojavu, kjer en gen lahko blokira druge na regulacijski poti do fenotipa.

Razpredelnica podaja nabor devetih eksperimentov, kjer smo opazovali fenotip pri nemutiranem organizmu (E1), enojnih (E2 do E6) in dvojnih mutantih (E7 do E9). Gene smo pri eksperimentih ali izničili (npr. B-) ali jih čezmerno izrazili (npr. A+). Opazovani fenotip smo zajeli kvalitativno z vrednostmi n (znižan), 0, p (povečan). Fenotip divjega osebka je 0.

Na podlagi podatkov izvedite analizo epistaze ter tako zgradite gensko mrežo. Tip vpliva genov prikažite s pravilnim tipom povezave: $->$ ali $-|$.

Page for your solutions. / Stran za vaše rešitve.

Page for your solutions. / Stran za vaše rešitve.