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

Vpisna številka / Student ID: _____

Osnove bioinformatike / Introduction to Bioinformatics

2. izpitni rok / Second Examination Period

15. februar 2012 / Februar 15, 2012

Naloga / Exercise	1	2	3	4	Vsota / Sum
Vrednost / Max	6	10	5	30	51
Točk / Points					

[6] 1. Na GenBanku smo našli naslednje zaporedje:

GAAGAATGGGGATAATAAAACATTGAAAAGGTGCTACGTGAAAAATGACCAGTACTTCATAATACA

Na zaporedju označi vse možne bralne okvire na tem delu DNA (z njihovo smerjo, začetkom in koncem). Pri tem upoštevaj standardno kodno tabelo (pričetek=`ATG`, konec={`TAA`,`TAG`,`TGA`}).

Consider the above nucleotide sequence from GeneBank. On the sequence, mark all possible open reading frames (start, end, direction). Use the standard coding table (start=`ATG`, end={`TAA`,`TAG`,`TGA`}).

[10] 2. Spodaj je podana implementacija algoritma “forward” (uporabljali smo ga pri izračunih verjetnosti opaženega niza):

```
def forward(s, hmm):
    t, e = hmm

    # Hidden states
    zh = set()
    for h, tmpd in e.iteritems():
        zh.add(h)

    zh = [0] + list(zh)

    # Create table
    f = [{} for i in range(len(s)+1)]

    # Initialize i = 0; f_0(0) = 1; f_k(0) = 0 for k > 0
    for k in zh:
        f[0][k] = 0
    f[0][0] = 1.0

    # Recursion (i=1..L):
    for i in range(1, len(s)+1):
        for l in zh:
            sum_val = sum([f[i-1][k]*t[k].get(l, 0.0) for k in zh])
            f[i][l] = e.get(l, {}).get(s[i-1], 0.0)*sum_val

    # P(x)
    ps = sum([f[len(s)][k] for k in zh])
    return f, ps

T = { 0: {'p': 0.5, 'g': 0.5},
      'p': {'p': 0.95, 'g': 0.05},
      'g': {'g': 0.9, 'p': 0.1} }

E = { 'p': {'C': 0.5, 'G': 0.5},
      'g': {'C': 0.9, 'G': 0.1} }

print forward('CCCCCCGGGC', (T,E))
```

Popravi funkcijo `forward` tako, da bo interno uporabljala logaritmirane vrednosti, kličemo pa jo lahko še vedno z istimi parametri. Popravke lahko tudi označiš kar na zgornji kodi (piši lepo in čitljivo!).

V Pythonu sta funkciji `log` in `exp` v modulu `math`. Na voljo imate še naslednjo funkcijo:

```
def log_sum(v1, v2):
    nv1 = max(v1, v2)
    nv2 = min(v1, v2)
    return nv1 + math.log(1.0+math.exp(nv2-nv1))
```

On the previous page please find the implementation of "forward" algorithm. The algorithm was used in estimation of probability of observed sequence. Correct this implementation such that it will use the logarithms to replace products with sums, and thus increase the numerical stability of the procedure. Function's call (arguments) should remain the same, and given the same input, the original and your new function should return the same values. Mark the changes in the code (write legibly!). Python includes functions `log` and `exp` in `math` module. You can use also the following:

```
def log_sum(v1, v2):
    nv1 = max(v1, v2)
    nv2 = min(v1, v2)
    return nv1 + math.log(1.0+math.exp(nv2-nv1))
```

[5] 3. Jana je od sodelavke, biologinje Katarine, prejela nadvse zanimive podatke o izražanju 34,000 genov gozdnih jagod *Fragaria vesca*. V Katarininem eksperimentalnem vzorcu je 50 zdravih in 20 okuženih jagod. Katarina Jano prosi, če lahko ugotovi, ali je možno iz profilov genskih izrazov napovedati, katere jagode so zdrave in katere okužene. Da bi zmanjšala kompleksnost problema (podatkovna matrika vsebuje $34000 \cdot 70$ elementov in stolpec z razredom) se Jana odloči, da naprej za vse gene (atribute) oceni njihov informacijski prispevek, to je korelacijo med izrazom izbranega gena in razredom, nato pa izbere 10 najbolj informativnih genov. Na tako dobljenih podatkih (70 gozdnih jagod uvrščenih, ki so opisane z izrazi 10ih genov in razvrščene v 2 skupine) potem z 10-kratnim prečnim preverjanjem oceni napovedno točnost gozdov klasifikacijskih dreves. Rezultat je visoka povprečna klasifikacijska točnost, enaka 89%. Jana vsa navdušena sporoči Katarini, da je njen problem rešen in da lahko na osnovi izraznih profilov z visoko točnostjo prepoznava z virusom okužene gozdne jagode.

Jana, sicer navdušena bioinformatičarka, se je malce prenaglila in pri analitičnem postopku naredila eno večjo, pravzaprav kar hudo napako. Katero? Kako jo lahko odpravi oziroma kakšen bi bil pravilni postopek?

A biology student Katherine gave Jana, a student of Computer Science, some data on expression of 34.000 genes of wild strawberry *Fragaria vesca*. The data includes 70 samples, of which 50 are healthy strawberries and 20 are infected with bacteria. Katherine is asking Jana if she can use data mining to use gene expression profiles and from these infer which strawberries are infected and which not.

To lower the complexity (raw data is a matrix of size 34.000 by 70), Jana decided to estimate the information gain for each of the genes. She the picked 10 most informative genes, and used them in 10-fold cross validation to estimate the predictive power of classification trees. Estimated accuracy is 89%. Jana, excited by this high score, informs Katherine that her problem is solved and that she can easily determine the infection of the strawberries from expression of only 10 genes.

Jana's procedure includes an error, and her analytical procedure is not as conclusive as she thought. Why?

4. Genotipi 12-ih organizmov so podani s tremi SNP-i in razvrščeni v dve kategoriji:

SNP1	SNP2	SNP3	class
TT	TT	CC	case
GG	TT	AA	case
TT	AA	CC	case
TT	AT	CC	case
GG	AT	AA	case
GG	TT	AA	case
GG	TT	CC	control
TT	AA	AA	control
TT	AA	AA	control
GG	AT	CC	control
GG	TT	CC	control
TT	AT	AA	control

[10] (a) Določite par SNP-ov v največji interakciji. (Nasvet: pri tej nalogi ni potrebno številsko ovrednotiti interakcije. Če že hočeš, lahko, ampak bo vzelo veliko časa. Rajši nariši tabele s kombinacijami SNP-ov in v te vnesi porazdelitve razredov pri določeni kombinaciji, ter na ta način oceni, katera dva SNP-a sta najbolj v interakciji.)

[5] (b) Recimo, da imamo na voljo program, ki nam izračuna interakcijo dveh SNP-ov, $IG(X, Y; C)$. Opišite postopek, kako določimo statistično značilnost interakcije.

Genotypes of twelve organisms are given with three SNPs and are categorized into two different categories (see the table above).

[10] (a) Which of the two SNPs are in the highest interactions? (Advice: there's no need to compute the actual interaction. Instead, use the matrices to present the phenotype of all different combinations of two SNPs.)

[5] (b) Suppose there is a program to estimate SNP-SNP interactions, $IG(X, Y; C)$. Describe a procedure to estimate the statistical significance of the acquired interaction score.

$H(X) = -\sum_x p(x) \cdot \log_2 p(x)$

$H(X, Y) = -\sum_x \sum_y p(x, y) \cdot \log_2 p(x, y)$

$I(X; Y) = H(X) + H(Y) - H(X, Y)$

$IG(X, Y; C) = I(X, Y; C) - (I(X; C) + I(Y; C))$