PLNT4610 BIOINFORMATICS

# MID-TERM EXAMINATION

08:30 - 9:45 Tuesday, October 22, 2013

Answer any combination of questions totalling to exactly 100 points. If you answer questions totalling more than 100 points, answers will be discarded at random until the total points are less than or equal to 100. This exam is worth 20% of the course grade.

Hand in this question sheet along with your exam book. All questions must be answered in the exam book. The exam sheets will be shreded after the exam.

---

1. (5 points)

XhoII recognizes the following restriction site: 5'R^GATCY3'

Re-write the following as a double-stranded sequence, showing where cuts occur, and for each 5' and 3' end in the restriction site, label the coordinate of each end.

```
          10          20
      |     |     |     |
```

**5'CGGGATCGATAGATCCGGAATTC3'**

R = purine, Y = pyrimidine

2. (15 points)

Given four sequences, show the steps that T-COFFEE would perform to create a multiple protein alignment.

```
A) CMEKEKYE
B) CVKKHKKI
C) CVKKHKKI
D) CIQEDKFE
```

a) Draw the guide tree, based on visual inspection of the four sequences for similarity.

b) Write out the pairwise alignments, based on the guide tree

c) Write out the complete alignment, based on the pairwise alignments and the guide tree.

To make your job easier, just score alignments by considering perfect amino acid matches, rather than taking into account a scoring matrix. Remember, the goal of an optimal alignment is to maximize the similarity scores while minimizing the number of gaps added.

3. (10 points) Suppose that you wanted to do exhaustive pairwise similarity comparisons between very large sequences using the Smith-Waterman algorithm ie. global sequence alignment by dynamic programming. Consider the fact that the entire similarity matrix has to be stored in random access memory (RAM). If a typical PC has about 10 Gb (gigabytes) of RAM, what would be the maximum length of sequences that could be aligned?

Assumptions:

- Each bp in a sequence can be represented in a single charcter (A,G,C,T), which is a single byte.
- the memory taken up by software, the operating system etc. is negligible
- both sequences are the same length

4. (10 points) Suppose you wanted to create a dataset that would accurately sample sequences among different major taxonomic groups. Based on the data in the table below, what are some of the problems with creating such a dataset? Can you think of a strategy that would help you overcome these problems?

| taxon | estimated number of species | percentage of species | number of sequences in NCBI UniGene | percentage of sequences |
|---|---|---|---|---|
| insects | 830000 | 69.2 | 239944 | 12.7 |
| molluscs | 110000 | 9.2 | 40311 | 2.1 |
| other animals | 100000 | 8.3 | 216337 | 11.4 |
| arachnids | 60000 | 5.0 | 26582 | 1.4 |
| crustaceans | 50000 | 4.2 | 95901 | 5.1 |
| vertebrates | 50000 | 4.2 | 1275236 | 67.3 |
| total | 1200000 | | 1894311 | |

estimates from Stoeckle et al. Barcoding Life Illustrated.

http://barcoding.si.edu/PDF/BLIllustrated26jan04v1-3.pdf

5. (15 points) Create a table, similar to the one at right, that tells the time efficiency for each of the following tasks:

Choose from the following formulas for efficiency (one of them is NOT a correct answer):
**$O(mn)$, $O(k^2 2^k n^k)$, $O(n^4)$, $O(n)$, $O(n^2)$**

| task | time efficiency |
|---|---|
| translate DNA to protein | |
| multiple sequence alignment | |
| sequence database search | |
| comparing a sequence with itself | |
| comparing two different sequences | |

6. (10 points) For the following pairwise alignment, calculate the similarity score, using the BLOSUM45 scoring matrix provided.

```
CAGT_CDS1          T  I  F  L  M  M  L  L  V  F
GMU12150_CDS1      T  I  C  V  L  L  L  L  L  V
```

**Blosum45 Amino Acid Similarity Matrix**

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 7 | | | | | | | | | | | | | | | | | | |
| P | -2 | 9 | | | | | | | | | | | | | | | | | |
| D | -1 | -1 | 7 | | | | | | | | | | | | | | | | |
| E | -2 | 0 | 2 | 6 | | | | | | | | | | | | | | | |
| N | 0 | -2 | 2 | 0 | 6 | | | | | | | | | | | | | | |
| H | -2 | -2 | 0 | 0 | 1 | 10 | | | | | | | | | | | | | |
| Q | -2 | -1 | 0 | 2 | 0 | 1 | 6 | | | | | | | | | | | | |
| K | -2 | -1 | 0 | 1 | 0 | -1 | 1 | 5 | | | | | | | | | | | |
| R | -2 | -2 | -1 | 0 | 0 | 0 | 1 | 3 | 7 | | | | | | | | | | |
| S | 0 | -1 | 0 | 0 | 1 | -1 | 0 | -1 | -1 | 4 | | | | | | | | | |
| T | -2 | -1 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | 2 | 5 | | | | | | | | |
| A | 0 | -1 | -2 | -1 | -1 | -2 | -1 | -1 | -2 | 1 | 0 | 5 | | | | | | | |
| M | -2 | -2 | -3 | -2 | -2 | 0 | 0 | -1 | -1 | -2 | -1 | -1 | 6 | | | | | | |
| V | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -2 | -2 | -1 | 0 | 0 | 1 | 5 | | | | | |
| I | -4 | -2 | -4 | -3 | -2 | -3 | -2 | -3 | -3 | -2 | -1 | -1 | 2 | 3 | 5 | | | | |
| L | -3 | -3 | -3 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | -1 | 2 | 1 | 2 | 5 | | | |
| F | -3 | -3 | -4 | -3 | -2 | -2 | -4 | -3 | -2 | -2 | -1 | -2 | 0 | 0 | 0 | 1 | 8 | | |
| Y | -3 | -3 | -2 | -2 | -2 | 2 | -1 | -1 | -1 | -2 | -1 | -2 | 0 | -1 | 0 | 0 | 3 | 8 | |
| W | -2 | -3 | -4 | -3 | -4 | -3 | -2 | -2 | -2 | -4 | -3 | -2 | -2 | -3 | -2 | -2 | 1 | 3 | 15 |
| C | -3 | -4 | -3 | -3 | -2 | -3 | -3 | -3 | -3 | -1 | -1 | -1 | -2 | -1 | -3 | -2 | -2 | -3 | -5 | 12 |
| | G | P | D | E | N | H | Q | K | R | S | T | A | M | V | I | L | F | Y | W | C |

7. (10 points) The Standard genetic code is shown below. If you think about comparing a DNA sequence verses a DNA database, or comparing a protein sequence vs. a protein database, how does the genetic code affect

a) the sensitivity of the search
b) the time required to do the search

```
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      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      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      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
```

8. (10 points) Answer the following questions about the table below:

a) By random chance alone, what is the probability that an amino acid chosen from one protein will match a given amino acid from another protein?

b) By random chance alone, what is the probability that a nucleotide from one DNA sequence will match a nucleotide from another DNA sequence?

c) When comparing two amino acid sequences for similarity, if you use a k value of 3, how much would you expect to speed up the search?

d) Typically, proteins are only a few hundred amino acids long. How might that affect the actual speedup of the algorithm, given a k value of 3?

e) When comparing two DNA sequences, what is the probability a 20 base segment from one sequence will match a given 20 base segment from another sequence? Express the answer as an exponential number ie. scientific notation.

| Table 2. | Avg. dist. between k-matches $\frac{1}{p^k}$ | | | |
|---|---|---|---|---|
| Prob. of a match (**p**) | **k**= 2 | 3 | 4 | 5 |
| 0.050 | 400 | 8000 | | |
| 0.075 | 178 | 2370 | | |
| 0.100 | 100 | 1000 | | |
| 0.150 | 44 | 296 | | |
| 0.200 | 25 | 125 | | |
| | | | | |
| 0.250 | 16 | 64 | 256 | 1024 |
| 0.300 | 11 | 37 | 123 | 412 |
| 0.350 | 8 | 23 | 67 | 190 |
| 0.450 | 5 | 11 | 24 | 54 |
| 0.600 | 3 | 5 | 8 | 13 |
| 0.700 | 2 | 3 | 4 | 6 |
| 0.900 | 1 | 1 | 1 | 2 |

9. (5 points) What is the sequence complexity of the DNA molecules below?

```
5'GCACTGCACTGCACT3'
3'CGTGACGTGACGTGA5'
```

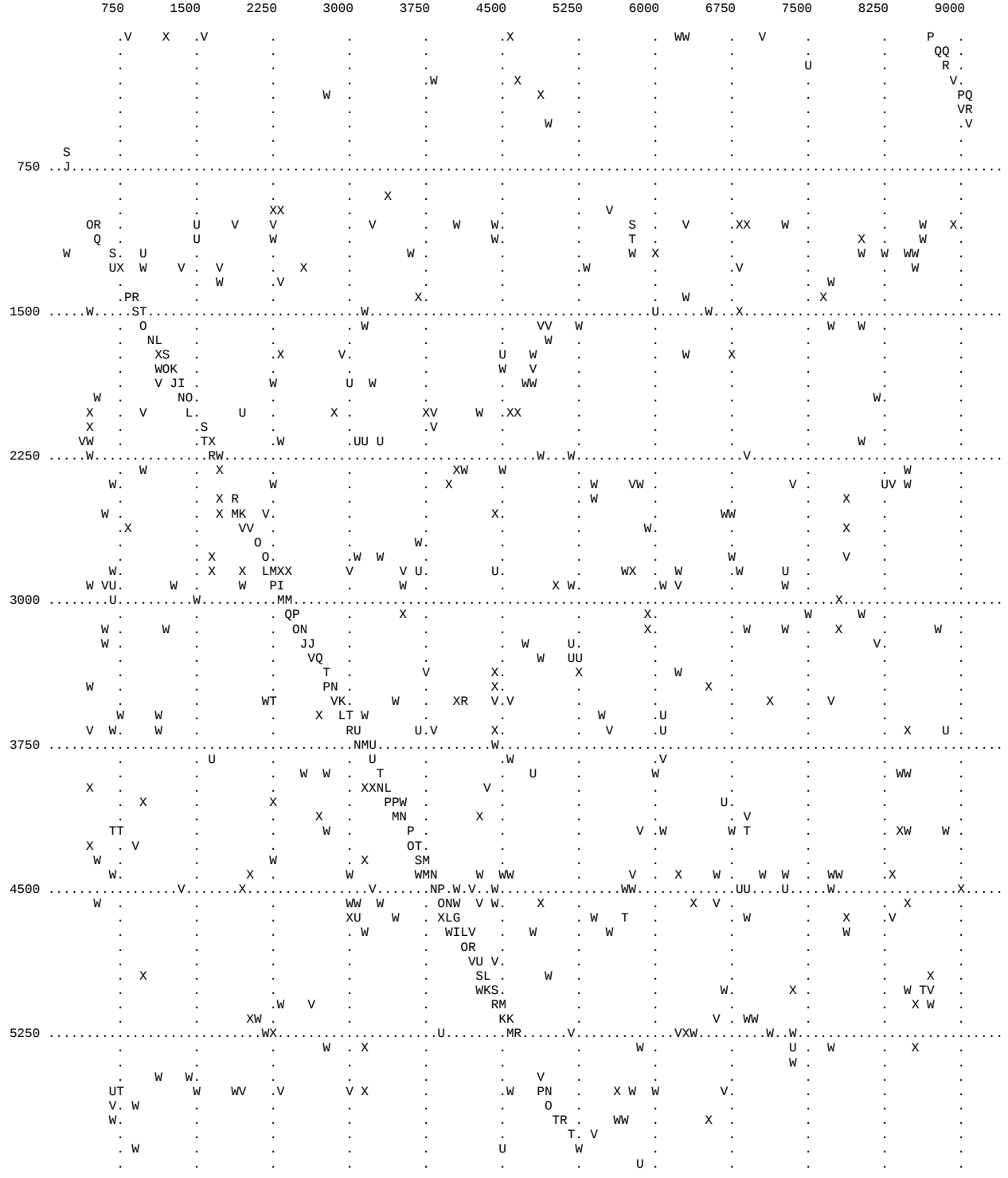
10. (5 points) Describe what is meant by the term "E-value", for BLAST and FASTA database searches.

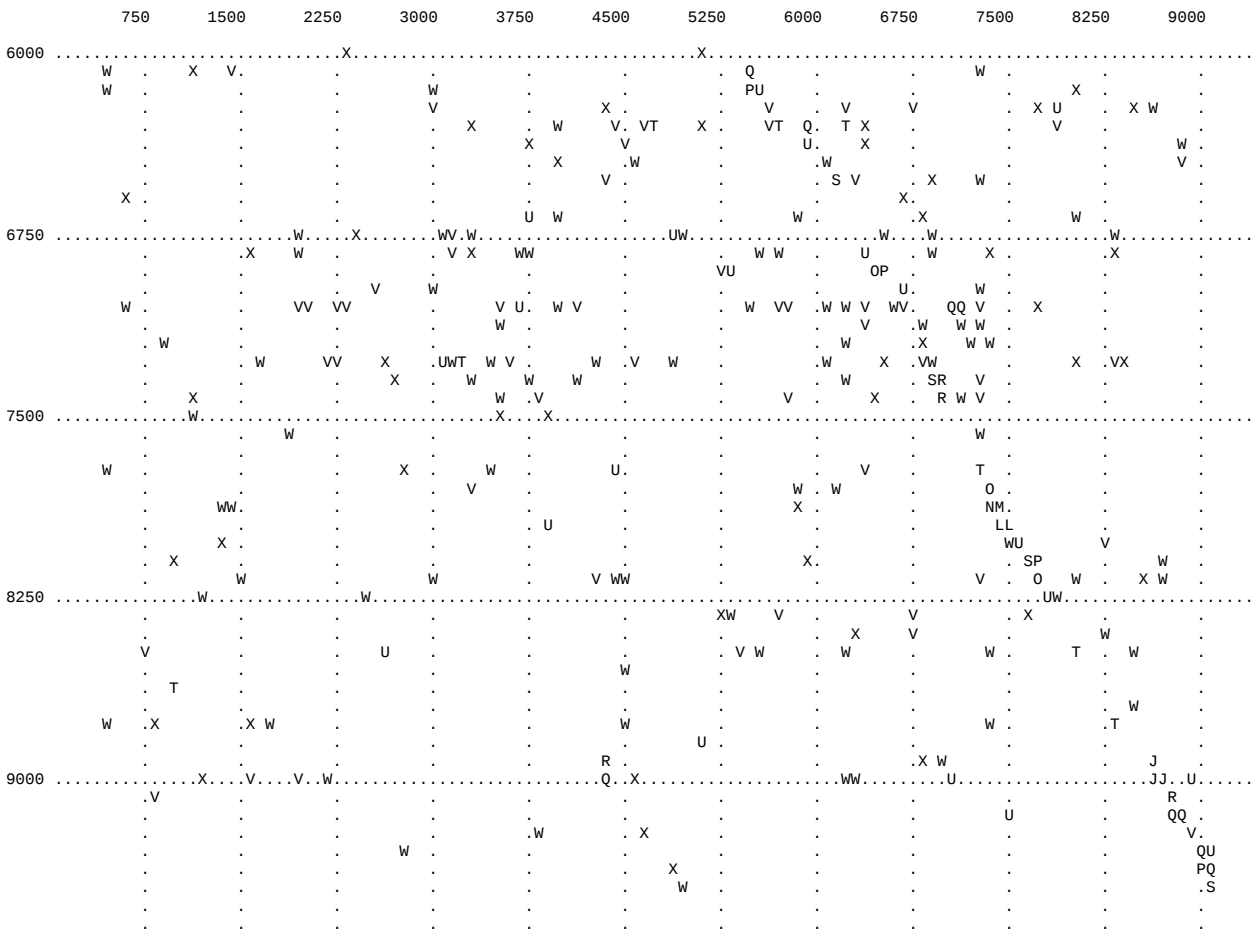
11. (10 points) TFASTA and TBLASTN use protein query sequences to search against DNA databases. How do these programs translate the sequences in the DNA databases into proteins? Suppose that you were searching a DNA database consisting of 100 billion nucleotides. How many amino acids would that correspond to?

12. (15 points) On the following pages, two RNA retroviruses, Human Immunodeficiency Virus (HIV) and Simian immunodeficiency virus (SIV) are compared in a dot-matrix similarity search. As well, an abbreviated annotation for the HIV sequence is also attached, showing the CDS (protein coding) features. For brevity, the corresponding features of SIV are not included.

a) Are all genes in HIV (ACCESSION NC_001802) also found in SIV (ACCESION NC_001549)?
b) What can you say about the intergenic regions of HIV, compared to SIV?
c) Are there any apparent major differences between the two genomes? What might be one possible explanation

```
D3HOM         Version  8/13/2001
X-axis: >NC_001802 (HIV)
Y-axis: >NC_001549 (SIV)
SIMILARITY RANGE:  25      MIN.PERCENT SIMILARITY:  55
SCALE FACTOR:    0.95      COMPRESSION:             75
```




750
1500
2250
3000
3750
4500
5250
6000
6750
7500
8250
9000
6000
6750
7500
8250
9000

```
GETOB          Version 1.3.2  13 Jun 2004
Please cite: Fristensky B. (1993) Feature expressions:
creating and manipulating sequence datasets.
Nucl. Acids Res. 21:5997-6003

NC_001802:CDS1
    join
        (
             336                 1637
             1637                 4642
        )


/gene="gag-pol"
/locus_tag="HIV1gp1"
/note="fusion protein consisting of the viral structural
proteins and enzymes; cleaved by the viral protease into
individual mature proteins; The processing products of the
Gag and Gag-Pol polyproteins were annotated with the help
of Pettit et al., 2003 and references therein; Pr160;
ribosomal slippage at slippery sequence tttttta
(1631..1637)"
/codon_start=1
/product="Gag-Pol"
//-----------------------------------------------
NC_001802:CDS2

     336         1838


/gene="gag"
/locus_tag="HIV1gp2"
/note="The processing products of the Gag and Gag-Pol
polyproteins were annotated with the help of Pettit et
al., 2003 and references therein"
/codon_start=1
/product="Pr55(Gag)"
//-----------------------------------------------
NC_001802:CDS3

     4587         5165


/gene="vif"
/locus_tag="HIV1gp3"
/note="p23; viral infectivity factor; viral accessory
protein important for virus replication in vivo"
/codon_start=1
/product="Vif"
//-----------------------------------------------
NC_001802:CDS4
    join
        (
             5105                5319
             5321                5396
        )

/gene="vpr"
/locus_tag="HIV1gp4"
/exception="artificial frameshift"
/note="p15; viral protein R; viral accessory protein
important for virus replication in vivo; involved in the
nuclear import of the HIV-1 preintegration complex;
induces G2 cell cycle arrest; influences mutation rates
during viral DNA synthesis; An artificial frameshift
eliminating the orf-disrupting nucleotide at position 5320
is introduced to obtain the typical HIV-1 Vpr protein
sequence. For this particular HIV-1 strain, HXB2, only a
short (78 amino acid long) variant of the Vpr sequence can
be obtained by translation of nucleotides 5105 through
5341 without the frameshift"
/codon_start=1
/product="Vpr"
//-----------------------------------------------
NC_001802:CDS5
    join
        (
             5377                5591
             7925                7970
        )

/gene="tat"
/locus_tag="HIV1gp5"
/note="p14; transcriptional activator; viral regulatory
protein required for virus replication; transactivates the
viral LTR promoter through interactions with cellular
transcription factors; associated with pathogenic effects
of the virus; the length of Tat varies depending on virus
strain or clade"
/codon_start=1
/product="Tat"
//-----------------------------------------------
NC_001802:CDS6
    join
        (
             5516                5591
             7925                8199
        )

/gene="rev"
/locus_tag="HIV1gp6"
/note="p19; regulator of expression of virion proteins;
prevents splicing of viral RNA; shuttles unspliced viral
RNA to the cytoplasm for expression of viral proteins and
incorporation of full length viral genomic RNA into
virions"
/codon_start=1
/product="Rev"
//-----------------------------------------------
NC_001802:CDS7
     5608         5856

/gene="vpu"
/locus_tag="HIV1gp7"
/note="p16; viral protein U; viral accessory protein
important for virus replication in vivo; promotes
degradation of CD4 and down-regulates cell surface
expression of MHC class I proteins; helps mediate
efficient virus particle release from infected cells;
reported to induce apoptosis by suppressing the nuclear
factor kappaB-dependent expression of antiapoptotic
factors; may attenuate the level of Env precursor(gp160)
biosynthesis; Vpu and gp160 are translated from different
reading frames of the same bicistronic mRNA"
/codon_start=1
/product="Vpu"
//-----------------------------------------------
NC_001802:CDS8
     5771         8341

/gene="env"
/locus_tag="HIV1gp8"
/note="gp160; envelope glycoprotein; envelope polyprotein;
cleaved by cellular proteases into mature proteins gp120
and gp41"
/codon_start=1
/product="Envelope surface glycoprotein gp160, precursor"
//-----------------------------------------------
NC_001802:CDS9
     8343         8963

/gene="nef"
/locus_tag="HIV1gp9"
/note="p27; negative factor; viral accessory protein;
important for virus replication in vivo; determinant of
HIV-1 pathogenesis; down-regulates cell surface CD4 and
MHC class I molecules; enhances virus infectivity through;
interactions with multiple cellular signaling proteins;
This particular nucleotide sequence has a premature stop
codon in place of a well-conserved tryptophan codon at
position 8712-8714 that truncates the HIV1 Nef protein
sequence to a 123 amino acids-long N-terminal portion (not
shown)"
/codon_start=1
/transl_except=(pos:8712..8714,aa:Trp)
/product="Nef"
```