03-60-558
Computational Molecular Biology
Assignment #1
Deadline: October 29th, 2005

Dr. Alioune Ngom

September 29, 2005

The main objectives of this assignment are

- Understanding pair-wise alignment algorithm
- Exploring NCBI database

1 Counting Alignments

- As you know already, there is an exponential number of global alignments between any two strings.
- Write a BioPerl (or Perl, BioJava or Java) program, using dynamic programming, to counts the number of global alignments for two sequences of length m and n respectively. Use the *edit distance* as the scoring function.
- Hint: Observe the Needleman-Wunsch algorithm discussed in class, and find a recurrence formula. That is, you must write down a recurrence equation that counts the total number of global alignments $C(i, j)$, for prefixes $S_{1...i}$ and $T_{1...j}$ (of input sequences S and T). You should first define the bases cases $C(0, 0)$, $C(i, 0)$ and $C(0, j)$, and then develop the general case, $C(i, j)$. The value your program should return is $C(m, n)$, i.e. the total number of global alignments for the entire input sequences.
- Note: The input to your program can be as simple as two numbers, m and n, which represent the length of S and T. For instance

```
> perl CountAlignments 10 20
4,354,393,801
```

2 Generating Pairwise Alignments

1. Write a BioPerl (or Perl, BioJava or Java) program that *finds* all optimal global alignments between AAAG and ACG
2. Write a BioPerl (or Perl, BioJava or Java) program to find the best
   (a) Global alignment

(b) Local alignment

(c) End-space free alignment with Needleman-Wunsch algorithm

(d) End-space free alignment with Smith-Waterman algorithm

between AGATAGAAACTGATATATA and AGAAAAAGAGT using the following affine gap penalty function

- Match = +1
- Mismatch = −1
- Gap opening penalty = −2
- Gap extension penalty = −1

3 NCBI Sequence Databases [BONUS]

Visit the National Centre for Biotechnology Information (NCBI) web site (www.ncbi.nlm.nih.gov/). This is a major resource for bioinformatics data. One of the tools available to search the information is called Entrez. Go to the Entrez section of NCBIs web site. As you can see, Entrez allows to search the literature, genomes, protein structures, and several other databases. Go to Entrez Genome section. On the left hand side of the page, you will find links for each kingdom of life, Archa, Bacteria and Eukaryota, as well as viral genomes. We will limit this analysis to bacterial genomes because of size limitations. Go to Bacteria/Genome section. You will find a list of organisms (bacteria) for which the complete genomic sequence has been determined. For each genome, the web site lists the latin name of the organism (green link), e.g. Bacillus anthracis, and possibly the particular strain, e.g. Ames, followed by a unique identifier called the accession number (blue link), such as NC_003997, followed by the size of the genome and the date of completion.

1. How many complete bacterial genomes are available at NCBI web site?
2. Select your favorite bacteria, click on its accession number. This will bring a summary page. Give me the latin name of the selected organism, the size of its genome, expressed as the number of base pairs, as well as its accession number.
3. On the summary page, select the accession number, this will bring a page that contains information in GenBank format. It contains, amongst other things, a list of all the genes. Scrolling toward the bottom of the page, you will find the actual DNA sequence information. GenBank format is too verbose for the tasks at hand. Go back to the top of the page, select the file format FASTA and display the information again. The meta-information has been reduced to a single line, as you can see. Save this to a file, I suggest using the accession number, followed by the suffix .fa as the name of the file.
4. Write a computer program (or find an existing one) that reads a FASTA file, then counts and prints the the mono- and di-nucleotides frequencies. Beware of newline characters. What are the frequencies of A,C,G and T letters?
5. Which is the least frequent di-nucleotide?
6. Using the observed mono- and di-nucleotide frequencies, calculate probability of the string CGAT. Show your calculation.
7. Apply this program to at least one other bacterial genome. Are the frequencies approximately the same for these genomes? Give the accession number of a pair of genomes having the most distinct nucleotide distributions (you dont need to be exhaustive, simply try a few cases). Can you find clues explaining the differences?

8. How many times would you expect to find the (string) sequence TTGACA in the first selected genome? To answer the question, use the simplest assumption (model), i.i.d.. This model assumes that the positions are independent one from another and identically distributed.
9. Write a program (or find an existing one) that counts the number of occurrences of TTGACA in the selected genomic sequence. What is the observed frequency of TTGACA?