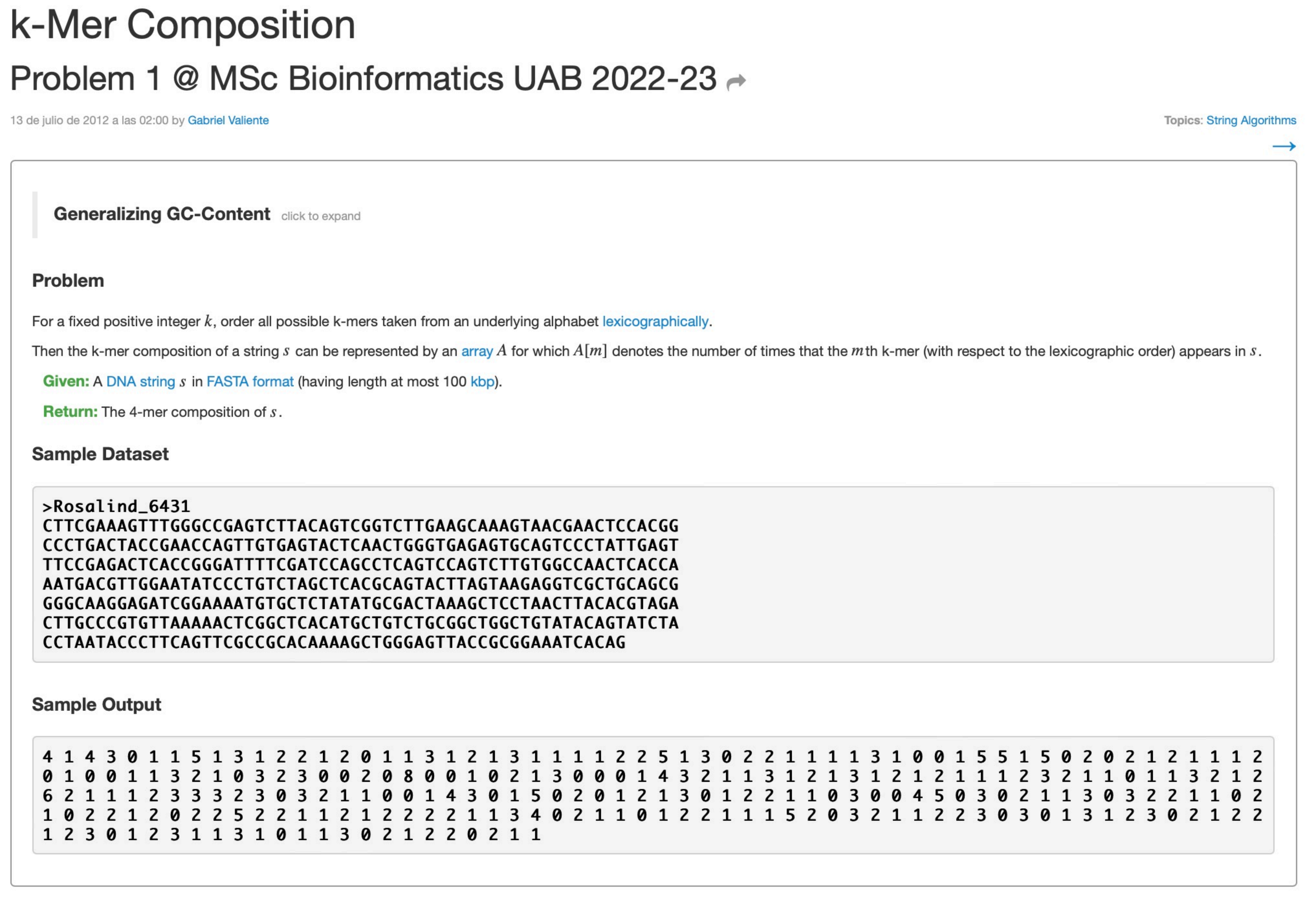


k-Mer Composition

Problem 1 @ MSc Bioinformatics UAB 2022-23

13 de julio de 2012 a las 02:00 by Gabriel Valiente

Topics: String Algorithms

→

Generalizing GC-Content click to expand

Problem

For a fixed positive integer k, order all possible k-mers taken from an underlying alphabet lexicographically.

Then the k-mer composition of a string s can be represented by an array A for which $A[m]$ denotes the number of times that the mth k-mer (with respect to the lexicographic order) appears in s.

Given: A DNA string s in FASTA format (having length at most 100 kbp).

Return: The 4-mer composition of s.

Sample Dataset

```
>Rosalind_6431
CTTCGAAAGTTTGGGCCGAGTCTTACAGTCGGTCTTGAAGCAAAGTAACGAACTCCACGG
CCCTGACTACCGAACCAGTTGTGAGTACTCAACTGGGTGAGAGTGCAGTCCCTATTGAGT
TTCCGAGACTCACCGGGATTTTCGATCCAGCCTCAGTCCAGTCTTGTGGCCAACTCACCA
AATGACGTTGGAATATCCCTGTCTAGCTCACGCAGTACTTAGTAAGAGGTCGCTGCAGCG
GGGCAAGGAGATCGGAAAATGTGCTCTATATGCGACTAAAGCTCCTAACTTACACGTAGA
CTTGCCCGTGTTAAAAACTCGGCTCACATGCTGTCTGCGGCTGGCTGTATACAGTATCTA
CCTAATACCCTTCAGTTCGCCGCACAAAAGCTGGGAGTTACCGCGGAAATCACAG
```

Sample Output

```
4 1 4 3 0 1 1 5 1 3 1 2 2 1 2 0 1 1 3 1 2 1 3 1 1 1 1 2 2 5 1 3 0 2 2 1 1 1 1 3 1 0 0 1 5 5 1 5 0 2 0 2 1 2 1 1 1 2
0 1 0 0 1 1 3 2 1 0 3 2 3 0 0 2 0 8 0 0 1 0 2 1 3 0 0 0 1 4 3 2 1 1 3 1 2 1 3 1 2 1 2 1 1 1 2 3 2 1 1 0 1 1 3 2 1 2
6 2 1 1 1 2 3 3 3 2 3 0 3 2 1 1 0 0 1 4 3 0 1 5 0 2 0 1 2 1 3 0 1 2 2 1 1 0 3 0 0 4 5 0 3 0 2 1 1 3 0 3 2 2 1 1 0 2
1 0 2 2 1 2 0 2 2 5 2 2 1 1 2 1 2 2 2 2 1 1 3 4 0 2 1 1 0 1 2 2 1 1 1 5 2 0 3 2 1 1 2 2 3 0 3 0 1 3 1 2 3 0 2 1 2 2
1 2 3 0 1 2 3 1 1 3 1 0 1 1 3 0 2 1 2 2 0 2 1 1
```

Overlap Graphs

Problem 2 @ MSc Bioinformatics UAB 2022-23

2 de julio de 2012 a las 02:00 by Rosalind Team

Topics: Graph Algorithms

←

→

A Brief Introduction to Graph Theory

click to expand

Problem

A graph whose nodes have all been labeled can be represented by an **adjacency list**, in which each row of the list contains the two node labels corresponding to a unique edge.

A **directed graph** (or digraph) is a graph containing **directed edges**, each of which has an orientation. That is, a directed edge is represented by an arrow instead of a line segment; the starting and ending nodes of an edge form its **tail** and **head**, respectively. The directed edge with tail v and head w is represented by (v, w) (but *not* by (w, v)). A **directed loop** is a directed edge of the form (v, v).

For a collection of strings and a positive integer k, the **overlap graph** for the strings is a directed graph O_k in which each string is represented by a node, and string s is connected to string t with a directed edge when there is a length k suffix of s that matches a length k prefix of t, as long as $s \neq t$; we demand $s \neq t$ to prevent directed loops in the overlap graph (although directed cycles may be present).

Given: A collection of DNA strings in FASTA format having total length at most 10 kbp.

Return: The adjacency list corresponding to O_3. You may return edges in any order.

Sample Dataset

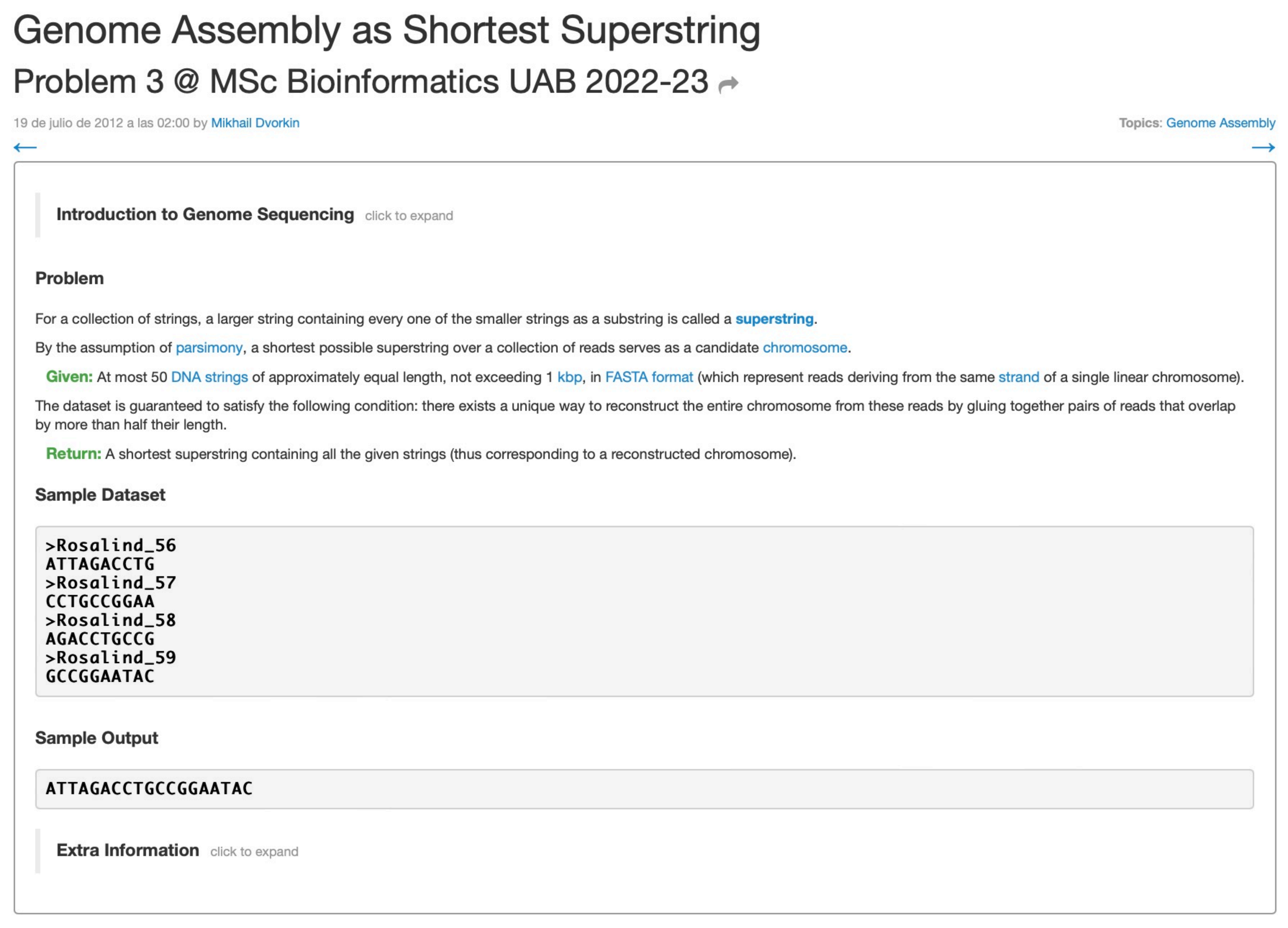
```
>Rosalind_0498
AAATAAA
>Rosalind_2391
AAATTTT
>Rosalind_2323
TTTTCCC
>Rosalind_0442
AAATCCC
>Rosalind_5013
GGGTGGG
```

Sample Output

```
Rosalind_0498 Rosalind_2391
Rosalind_0498 Rosalind_0442
Rosalind_2391 Rosalind_2323
```

Note on Visualizing Graphs

click to expand



Genome Assembly as Shortest Superstring

Problem 3 @ MSc Bioinformatics UAB 2022-23

19 de julio de 2012 a las 02:00 by Mikhail Dvorkin

Topics: Genome Assembly

Introduction to Genome Sequencing click to expand

Problem

For a collection of strings, a larger string containing every one of the smaller strings as a substring is called a **superstring**.

By the assumption of parsimony, a shortest possible superstring over a collection of reads serves as a candidate chromosome.

Given: At most 50 DNA strings of approximately equal length, not exceeding 1 kbp, in FASTA format (which represent reads deriving from the same strand of a single linear chromosome).

The dataset is guaranteed to satisfy the following condition: there exists a unique way to reconstruct the entire chromosome from these reads by gluing together pairs of reads that overlap by more than half their length.

Return: A shortest superstring containing all the given strings (thus corresponding to a reconstructed chromosome).

Sample Dataset

```
>Rosalind_56
ATTAGACCTG
>Rosalind_57
CCTGCCGGAA
>Rosalind_58
AGACCTGCCG
>Rosalind_59
GCCGGAATAC
```

Sample Output

```
ATTAGACCTGCCGGAATAC
```

Extra Information click to expand

Constructing a De Bruijn Graph

Problem 4 @ MSc Bioinformatics UAB 2022-23

2 de julio de 2012 a las 02:00 by Mikhail Dvorkin

Topics: Genome Assembly

←

→

Wading Through the Reads click to expand

Problem

Consider a set S of $(k+1)$-mers of some unknown DNA string. Let S^{rc} denote the set containing all reverse complements of the elements of S. (recall from "Counting Subsets" that sets are not allowed to contain duplicate elements).

The **de Bruijn graph** B_k of order k corresponding to $S \cup S^{\text{rc}}$ is a digraph defined in the following way:

- Nodes of B_k correspond to all k-mers that are present as a substring of a $(k+1)$-mer from $S \cup S^{\text{rc}}$.
- Edges of B_k are encoded by the $(k+1)$-mers of $S \cup S^{\text{rc}}$ in the following way: for each $(k+1)$-mer r in $S \cup S^{\text{rc}}$, form a directed edge $(r[1:k], r[2:k+1])$.

Given: A collection of up to 1000 (possibly repeating) DNA strings of equal length (not exceeding 50 bp) corresponding to a set S of $(k+1)$-mers.

Return: The adjacency list corresponding to the de Bruijn graph corresponding to $S \cup S^{\text{rc}}$.

Sample Dataset

```
TGAT
CATG
TCAT
ATGC
CATC
CATC
```

Sample Output

```
(ATC, TCA)
(ATG, TGA)
(ATG, TGC)
(CAT, ATC)
(CAT, ATG)
(GAT, ATG)
(GCA, CAT)
(TCA, CAT)
(TGA, GAT)
```

Genome Assembly with Perfect Coverage

Problem 5 @ MSc Bioinformatics UAB 2022-23

14 de agosto de 2012 a las 02:00 by Mikhail Dvorkin

Topics: Genome Assembly, Graph Algorithms

←

→

Cyclic Chromosomes click to expand

Problem

A **circular string** is a string that does not have an initial or terminal element; instead, the string is viewed as a necklace of symbols. We can represent a circular string as a string enclosed in parentheses. For example, consider the circular DNA string (ACGTAC), and note that because the string "wraps around" at the end, this circular string can equally be represented by (CGTACA), (GTACAC), (TACACG), (ACACGT), and (CACGTA). The definitions of substrings and superstrings are easy to generalize to the case of circular strings (keeping in mind that substrings are allowed to wrap around).

Given: A collection of (error-free) DNA k-mers ($k \leq 50$) taken from the same strand of a circular chromosome. In this dataset, all k-mers from this strand of the chromosome are present, and their de Bruijn graph consists of exactly one **simple cycle**.

Return: A cyclic superstring of minimal length containing the reads (thus corresponding to a candidate cyclic chromosome).

Sample Dataset

```
ATTAC
TACAG
GATTA
ACAGA
CAGAT
TTACA
AGATT
```

Sample Output

```
GATTACA
```

Note click to expand

Assessing Assembly Quality with N50 and N75

Problem 6 @ MSc Bioinformatics UAB 2022-23

2 de julio de 2012 a las 02:00 by Mikhail Dvorkin

Topics: Genome Assembly

←

How Well Assembled Are Our Contigs? click to expand

Problem

Given a collection of DNA strings representing contigs, we use the **N statistic** NXX (where XX ranges from 01 to 99) to represent the maximum positive integer L such that the total number of nucleotides of all contigs having length $\geq L$ is at least XX% of the sum of contig lengths. The most commonly used such statistic is **N50**, although N75 is also worth mentioning.

Given: A collection of at most 1000 DNA strings (whose combined length does not exceed 50 kbp).

Return: N50 and N75 for this collection of strings.

Sample Dataset

```
GATTACA
TACTACTAC
ATTGAT
GAAGA
```

Sample Output

```
7 6
```

Extra Information click to expand