1. Overview

Slides for the Lecture on
Foundations of Sequence Analysis
Winter 2022/2023
Stefan Kurtz

October 16, 2022

Application areas

- There are thousands of different algorithms and data structures for handling biological sequence data, many of which have originally been developed in the bioinformatics and computational biology community.
- However, in many cases methods applied to biological sequences are adapted from methods developed for sequences from other sources.
- For example, many methods for fast index based search in biological sequence databases are adapted from methods originally developed for indexing web-pages.
- For this reason, it is often a good idea to look at methods developed in computer science before developing an own method.
- Often the best and most widely used methods in Bioinformatics are developed by an interdisciplinary interplay of computer scientists, bioinformaticians and wet-lab scientists.

Application areas

- When it comes to analyzing sequences one common feature is the following: Usually sequences encoding experimental or natural information are almost always inexact.
- Thus similar sequences have, in many cases, the same or similar meaning or effect.
- For this reason, a main part of this lecture will be devoted to notions of similarity of sequences, and we will show how to handle these notions algorithmically

Figure 1: Sources of sequences

- molecular biology

DNA	...`aactacgt`...	4 nucleotides,	length: $\approx 10^3 - 10^9$
RNA	...`aacuacgu`...	4 nucleotides,	length: $\approx 10^2 - 10^3$
proteins	...`LISAISTLIEB`...	20 aminoacids,	length: $\approx 10^2 - 10^3$

  `L` = Leucine, `I` = Isoleucine, `S` = Serine, `A` = Alanine, etc.
- text processing, like abstract of papers, laboratory books, medical reports
- graphics: (r, g, b) vectors with $r, g, b \in [0, 255]$ for the intensity of the red, green, and blue color of a pixel.
- information transmission: sequence of bits, blockcodes
- phonetic spelling: e.g. english with 40 phonemes; japanese with 113 "morae" (syllables)
- spoken language: discretized measurements, multidimensional (specifying frequency and energy) on a dynamic time scale

Figure 1: Sources of sequences

- molecular biology

DNA	...`aactacgt`...	4 nucleotides,	length: $\approx 10^3 - 10^9$
RNA	...`aacuacgu`...	4 nucleotides,	length: $\approx 10^2 - 10^3$
proteins	...`LISAISTLIEB`...	20 aminoacids,	length: $\approx 10^2 - 10^3$

  `L` = Leucine, `I` = Isoleucine, `S` = Serine, `A` = Alanine, etc.
- text processing, like abstract of papers, laboratory books, medical reports
- graphics: (r, g, b) vectors with $r, g, b \in [0, 255]$ for the intensity of the red, green, and blue color of a pixel.
- information transmission: sequence of bits, blockcodes
- phonetic spelling: e.g. english with 40 phonemes; japanese with 113 "morae" (syllables)
- spoken language: discretized measurements, multidimensional (specifying frequency and energy) on a dynamic time scale

We focus on biological sequences.

Figure 2: Some problems relevant for sequences

1. sequence comparison: compare two sequences and show the similarities and differences. Example: how similar is gene A to gene B?
2. sequences matching: find all positions in a sequence where a pattern sequence occurs. Example: Where does the pattern `TATAA` occur in my genome?
3. approximate sequence matching: find all positions in a text where a sequence matches, allowing for errors in the match. Example: Map all my 10^8 short reads of length 100 bp to the mouse genome of $2.9 \cdot 10^9$ *bp*
4. dictionary matching: for a given word w find the word v in a given set of words with maximal similarity to w. Example: find the nearest neighbor of my oligomer of length 18 in the Protein Family Database.
5. structural pattern matching: find regularities in sequences, like repeats, tandems ww, palindromes, or unique subsequences. Example: find all maximal repeats of minimum length 100 with at least 97% identity in my genome.