Algorithms in Bioinformatics

CS 482

Bin Ma

LaTeXed by Sibelius Peng

Preface

Disclaimer Much of the information on this set of notes is transcribed directly/indirectly from the lectures of CS 482 during Winter 2022 as well as other related resources. I do not make any warranties about the completeness, reliability and accuracy of this set of notes. Use at your own risk.

For any questions, send me an email via `https://notes.sibeliusp.com/contact`.

You can find my notes for other courses on `https://notes.sibeliusp.com/`.

Sibelius Peng

Contents

1 Introduction

This course is officially titled as "Computational Techniques in Biological Sequence Analysis". However, this is an old title and this course has been offered for over approximately twenty years. It should actually be called "Algorithms in Bioinformatics". Around 1980, this area has a different name: Computational Biology. You may also hear another name: DNA computing. It is another area, and it does not solve biological problem. What is bioinformatics? It is biology + informatics. Biology is the reason, goal, purpose and informatics is the method.

Biology can be studied at different scales. In old days, biology tries to study organisms, namely living things, such as bacteria, animals. This is because people didn't have tools to study at a lower level at that time. Now people study organs and tissues. Then people can look into cell level, molecular level. At molecular level, DNA is chain of nucleotide bases. Protein is chain of amino acids.

There are a lot of public and free molecular data. There are tremendous amount of public biomolecule data and free software. For example:

- NCBI's sequence data bank: `https://www.ncbi.nlm.nih.gov/nuccore/NC_045512`
- PDB protein structure database: `https://www.rcsb.org/structure/6vxx`

Why do people do bioinformatics? The goal is to understand life at molecular level. Especially, for human health. For example, sequencing SARS-Cov2 genomes allowed people study the evolution of this virus. Study the structure of the spike protein and its interaction with the host cells. A lot of human diseases are related to genetics.

There are two areas of bioinformatics.

- Determine the molecule information, by analyzing the data produced by measuring instruments. Typically the data is in large scale and high throughput, which is hard for people to look at.
- Use the molecular data to make inference. For example, make prediction given the existing data.

Consider an example of genome sequencing. From wikipedia,

> The Human Genome Project (HGP) was an international scientific research project with the goal of determining the base pairs that make up human DNA, and of identifying, mapping and sequencing all of the genes of the human genome from both a physical and a functional standpoint. It remains the world's largest collaborative biological project. Planning started after the idea was picked up in 1984 by the US government, the project formally launched in 1990, and was declared complete on April 14, 2003. Level "complete genome" was

achieved in May 2021.

Bioinformatics played an essential role in analyzing the data and assemble the genome. Today one can sequence a human's genome with $<$ \$1000 in a couple of weeks. Bioinformatics is the key to utilize the NGS (next generation sequencing) data for genome sequencing. As such, today's cancer treatment starts to become personalized. And many new drugs now require gene sequencing as companion diagnostic.

1.1 Objectives of this course

- Know bioinformatics
  - Purpose and method
  - General topics
- Learn classic problems and algorithms in bioinformatics
- Learn wide-applicable computational techniques
  - String algorithms
  - Hidden Markov Model
  - Log likelihood ratio score
  - Statistical validation
  - A bit of machine learning

Some typical problems:

- Gene prediction problem
- Find the longest shared substring between human and mouse genomes. If we want to find similarities instead of exact matches, this will lead us to the homology search problem.
- Peptide Identification