Lecture 1: Introduction

Brief History of Bioinformatics

▼ What was the biomolecule first sequenced in 1955 and using which technology?

▼ proteins were sequenced first using sanger sequencing on bovine insulin.

▼ Which algorithm was first developed to compare two sequences of proteins in 1970?

▼ Needleman-Wunsch algorithm to compare a new amino acid sequence of protein with an older sequence to see how similar they are. It is called pairwise sequence alignment.

▼ Why do people care about aligning protein sequences?

▼ If two proteins have the same/similar sequence, they might have the similar structure and similar function.

▼ Which data repository contains 3D protein structures?

▼ Protein Data Bank or PDB.

▼ Which algorithm can give you a similarity score of your protein sequence compared to many other protein sequences much faster than a pairwise sequence algorithm?

▼ BLAST - Basic local alignment search tool.

▼ How does BLAST algorithm work? Describe briefly.

▼ BLAST works by breaking database sequences into k-mers (all possible sequences of k-nucleotides) and indexing the location of those k-mers on each entry in the database. When a new query sequence comes, it is also broken into k-mers and matched against indexes of all database sequence indexes previously stored. The highest scoring segment pairs that are statistically significant are pulled out and a more thorough alignment is run.

▼ Which experimental technique has been historically used to determine 3D protein structure?

- ▼ Crystallography

▼ How does Critical Assessment of Structure Prediction (CASP) competition work?

- ▼ Run every two years
- ▼ Take the latest collection of experimentally verified protein structures not yet published and ask computational researchers to make protein structure predictions.
- ▼ After submission, they would compare predictions to experimentally validated structures.

▼ How did the initial protein structure prediction algorithms work?

- ▼ Using presence of functional domains in a protein - small stretches of sequences that look very similar and usually have similar structure and function.

▼ Which database stores important functional domains in families of proteins?

- ▼ BLOCKS database

▼ Which metric is used as a measure of comparing protein structure prediction to the experimental structure?

- ▼ Global distance test (GDT)
- ▼ Greater the GDT, better the prediction.

▼ What’s the latest algorithm for protein structure prediction that is many fold better than previous algorithms and what is its prediction score?

- ▼ AlphaFold2 that came out in 2020 and has an average GDT of close to 90%

▼ What is the threshold GDT score above which a protein structure prediction is considered equivalent to an experimentally determined structure?

- ▼ Above a GDT score of 90%

▼ Which type of protein structures are still difficult to predict with AlphaFold2?

▼ Structures that contain multiple proteins in a complex are still unsolved.

▼ What was the first technique developed to measure the expression of a single gene in 1977?

▼ Northern blot

▼ What was the first technique developed to measure the expression of many genes (hundreds to thousands of genes through thousands to millions of surface probes) simultaneously in 1995? How did this technique work?

▼ Microarray technology

▼ probes were attached to a surface (membranes or glass slides) and cDNA with attached fluorescent molecules was produced from mRNA of different samples, hybridized to the probes on the surface and measured.

▼ What are the latest commercial technologies to measure gene expression of ~1000 marker genes at very low cost?

▼ RASL-Seq (RNA-mediated oligonucleotide annealing, selection, and ligation with next-generation sequencing)

▼ Luminex assay

▼ What is a Broad institute database containing gene perturbations of multiple cell types through luminex assay ?

▼ ConnectivityMap

▼ What is the technology developed in 1972 that is used to integrate a DNA fragment into a vector called?

▼ Recombinant DNA technology

▼ What was the first organism whose DNA was sequenced in 1977 and what was the name of the technique used to sequence it?

▼ yeast or Saccharomyces cerevisiae

▼ Sanger sequencing

▼ What was the key technology developed in 1985 that allowed researchers to amplify DNA in any sample by making its copies?

▼ Polymerase chain reaction or PCR

▼ What algorithm can be used to compare a gene sequence with many other sequences at the same time?

▼ BLAST

▼ How many chromosome pairs do humans have?

▼ 23 pairs

▼ In which year was the human genome sequencing first completed?

▼ Working draft finished in spring 2000

▼ Complete human genome in 2003

▼ What are the three generations of DNA sequencing technology?

▼ Sanger sequencing - using four tubes with different ddNTPs and a blot

▼ High throughput sequencing/Next generation sequencing - millions of sequences at the same time using cycles of fluorescent imaging and dye cleaving one nucleotide at a time.

▼ Single molecule sequencing - using polymerase or pore technology

▼ Which company is the market leader in high throughput genome sequencing technology?

▼ Illumina

▼ What is the capacity of current next generation sequencing machines?

▼ The current capacity is millions of reads per run, where a read is 150 base pairs long. The overall time taken is proportional to the size of a read.

Definitions

▼ What is the difference between bioinformatics and computational biology?

▼ Bioinformatics - creation of tools such as algorithms and databases that solve problems. The goal is to build useful tools that work on biological data. Bioinformatics is about engineering.

▼ Computational biology - study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems.

Computational biology is about discovery.

▼ What are the different stages of bioinformatics/computational biology work?

▼ Entry Stage: Use published tools to analyze data and generate hypotheses for experimentalists.

▼ Computational Biology Stage: Make biological discoveries from public data integration and modeling.

▼ Advanced Computational Biology Stage: Integrative studies on data from big consortia.

▼ Bioinformatics Stage: Develop algorithms and databases for data analyses on new technologies, data integration and reuse.

▼ Which broad fields are included under Bioinformatics and Computational Biology?

▼ Statistics, Computer science, Biology