Bioinformatics

Course	Essential Protein Structure and Function
Confidence	Somewhat Confident
Next Review	@May 1, 2024
Last Edited	@May 1, 2024 3:38 PM

Aims and objectives

- Describe the importance of bioinformatics
- Discuss the importance of sequence and structural data and databases
- Give examples of databaks and show the difference between a databank and a search tool
- Describe how Bioinformatics is used in genome sequencing and structural biology
- Describe some of the current problems with data resources
- How sequence alignment works
- Current approaches (Protein Language Models and Machine Learning)

Bioinfomatics

- the application of computers to problems in Biology
- =Computational Biology
- Multiple Bioinformatics fields

Background

- Biology is overwhelmed with data (Petabytes of BNA reads data, 3 billion protein sequences in metagenome databases, 800 million models in AlphaFoldDB / DSMAtlas)
- Human genome 3.2Gbp = 3200 Mbp

- DNA sequencing data doubles every year, with the price per base dropping significantly
- New sequencing methods like Nanopore can be used to sequence DNA samples directly from the field
- AlphaFold models can be generated on a laptop or a Jupter notebook, usually less than 30 secs per structure
- Structure data (from NMR and Xray crystallography) doubles every six years

Data storage

- Database
  - A structured collection of data, with some tool enabling it to be queried
- Databank
  - A collection of data (normally in simple text files) without a fixed assciated query tool
- Knoledgebase
  - A collection of Databases and Databank with shared annotations and metadata

Databanks

- Primary - raw data with some additional information or metadata
  - PDB, GenBank, UniProtKB, SwissProt
- Secondary - derived from primary sources (signatures, motifs, patterns)
  - Patterns can be used to identify significant regions or active sites in a protein sequence directly. Needs to be tested afterwards, but useful to fish candidates in large genomes
    - E.g. Protein kinase C phosphorylation site [ST]-X-[RK]
    - N-linked glycosylation site: N-{P}-[ST]-{P}
    - Kringle domain: [FY]-C-[RH]-[NS]-X(7,8)-[WY]-C
  - Able to identify important and significant biological functions directly from the sequence without having to go through experiments

- BLAST, FASTA, UniProt, RefSeq

Databases

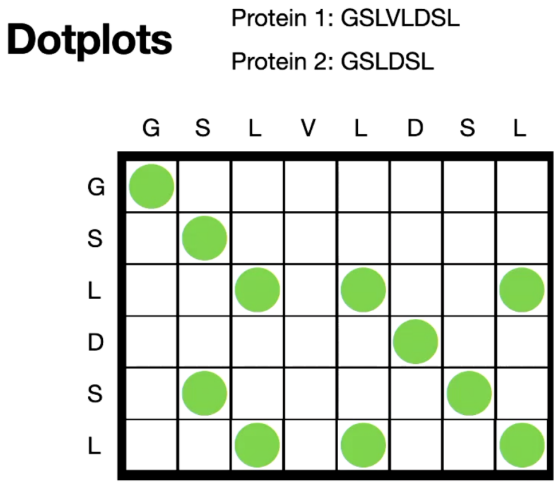
- NCBI BLAST
- InterPro (InterProScan)
- PROST
- Blast2Vec

Comparing sequences

Early approaches

- DNA sequence - Protein sequence (- Protein structure) - Protein function
- Homology based transfer or Genome assembly
  - Alignment of sequence

Dotplots



- Gaps
  - To align the first 3 dots with the rest, move it 2 grids to the right, introducing 2 gaps
- Mark every single time when there is a shared position between the two
  - Perfect diagonal if two proteins are matched
- Simple identity scoring (1 match, 0 mismatch)
- More complex scoring schemes - substitution matrices
- Amino acid similarity, size, chemistry derived from analyses of aligned homologous proteins - check for observed substitutions
  - First substitution matrix by Dayhoff (1978) then BLOSUM (Henikoff & Henikoff, 1992)

From dot plot to automated alignments - Dynamic Programming

- First applied to sequence alignment by Needleman and Wunsh in 1970
- Needleman and Wunsh algorithm
- Finds the best path through the matrix
- Start from one corner, while accumulating scores
- Used to solve the ‘travelling salesman’ problem (or other problems like traffic light sequencing)
- Global alignments vs Local alignments
    - Instead of align both structures together (global alignment) it starts to align the small sequences and slowly expand the alignment across the whole structure
    - Conservation of domains in evolution: Local>>Global
    - Thus, detect domains that conserved
    - Prosite contains patterns which are characteristic of protein families or functional sites

Speeding up dynamic programming - FASTA and BLAST

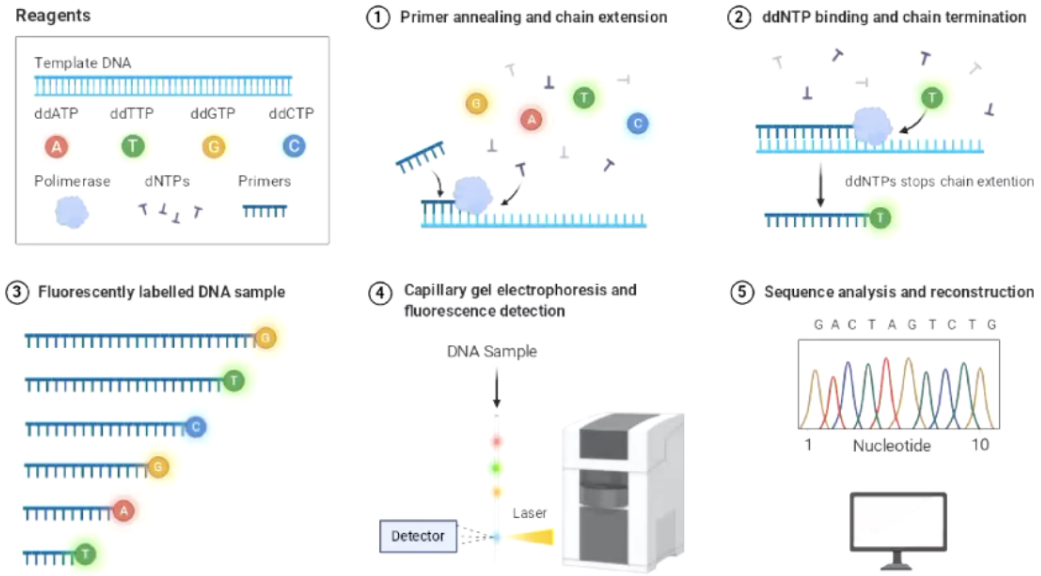
- Approximate fast methods (heuristics)
    - Approximate the actual sequence instead of searching for every exact match
- Index the search database by finding locations of short ‘words’
- Take ‘words’ from the query protein and search the database index
- Look for multiple matches and then extend them
- Very similar approach for genome assembly (k-mers to assemble genome contigs)
    - Take a small fragment of DNA and find which portions of genome that can be expand around

Genome sequencing and assembly

- How is DNA sequenced?
    - Sanger: di-deoxy chain termination (up to 2k bp per run)

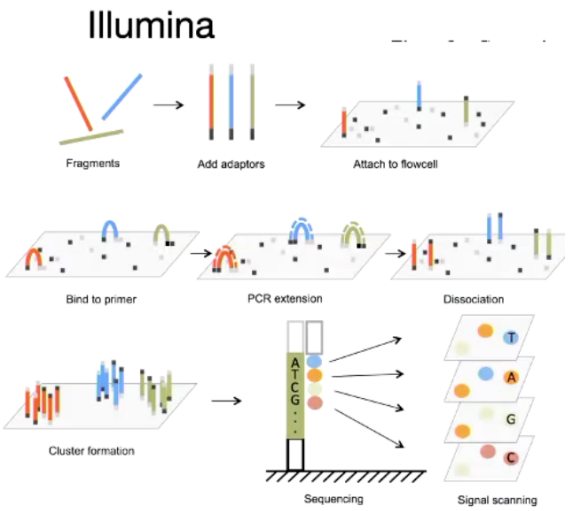
        - Electrophoresis
    - Next-gen: 500bp (Illumina)
        - Break out DNA into fragments
        - Detect flashes and reconstruct sequences
    - Long reads (PacBio)
    - Nano pre (Oxford)
    - How do we assemble an entire genome?

Sanger sequencing



- Still use nowadays to do DNA tests and paternity tests

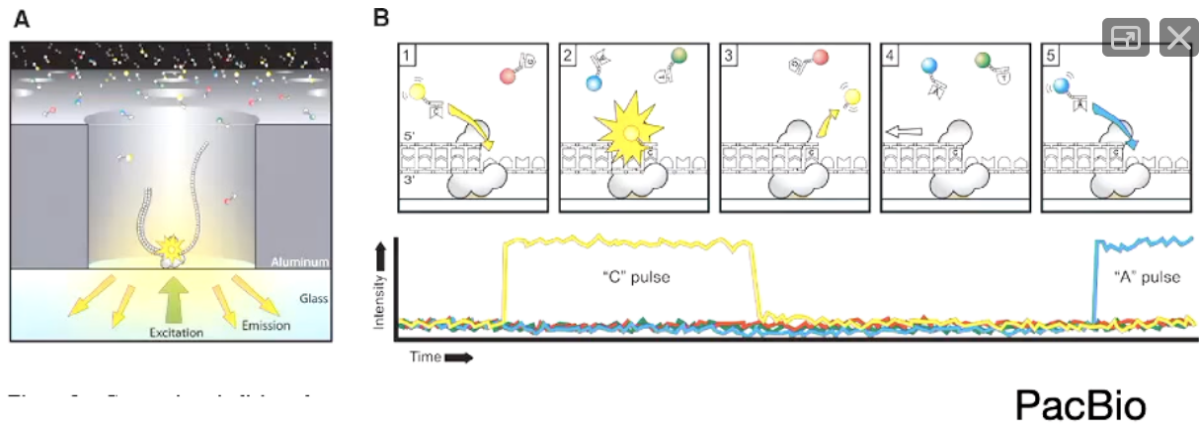
Ilumina



- Break DNA into fragments and attach them to a flow cell
- Cyclic addition of fluorescently labeled nucleotides to a growing DNA strand, followed by imaging to detect the incorporated nucleotide
- Detect the flash and reconstruct the sequence
- High accuracy
- Only 500bp at a time

PacBio

- Instead of attaching DNA to particular fixed positon, attaching DNA polymerase and DNA goes through it
- Involves the sequencing of single DNA molecules immobilized on a solid surface. As the DNA polymerase incorporates nucleotides, the emitted light is captured and analyzed to determine the DNA sequence.
- PacBio sequencing offers long read lengths, making it especially useful for resolving complex genomic regions, detecting structural variations, and characterizing repetitive sequences. It also enables the detection of DNA modifications, such as methylation, directly from the sequencing data.



Nanopore sequencing

- Works by passing a DNA or RNA molecule through a nanopore—a tiny hole in a membrane—and measuring changes in electrical current as nucleotides pass through the pore
    - Proteins expand when different nucleotides flow through
- Advantage
    - Real-time Sequencing
    - Long Read Lengths
        - don’t need to break DNA into individual pieces to put in a polymerase unlike the other methods
    - Portability

Reference mapping vs de-novo assembly

Reference mapping

- To compare the DNA to a particular sequence, e.g. a particular disease gene
- The easy way: A reference is available, mapping short contigs on a reference genome

Scaffold guided and de-novo

E.g. Completely new bacteria

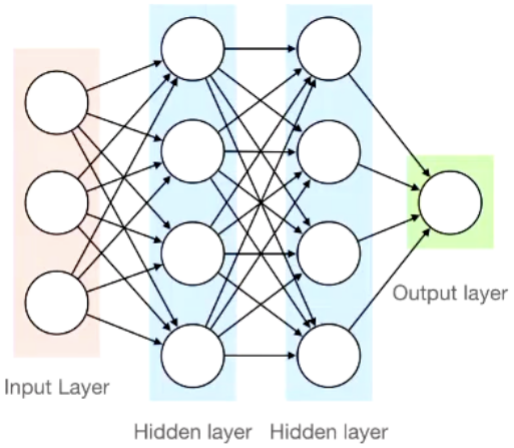
- More complicated
- Fuzzy matches (errors in sequencing)
- Apply a confidence score and minimum overlaps
- Problems with sequence repeats
- Current solutions: machine-learning based assemblers, and multiple sequencing techniques

Machine learning

- A general calss of software which learns from examples and is then able to make predictions
- “Train” a learning method with examples of real transcription start sites, intron/exon boundaries, sequence composition, etc.

- Learns from real examples, then applies the trained system to make predictions

- Different types:
  - Neural networks
  - Support Vector Machines (SVMs)
  - Decision Trees
  - XGBOOST
  - Bayesian Classifiers
  - Transformers

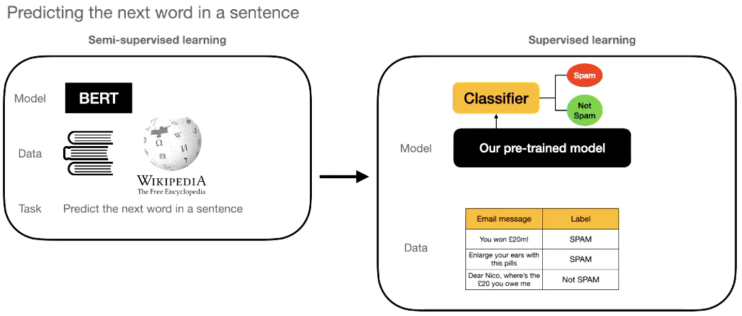


Neural Network

- Supervised learning
  - Spam filter, feed information to the machine
- Unsupervised learning
  - No information feeding but pick patterns
  - Need a lot of data
- Reinforcement learning
  - AlphaGo, play against itself and pick the winner
- Generative learning
  - Learn how to generate new data points that are similar to those in the training dataset

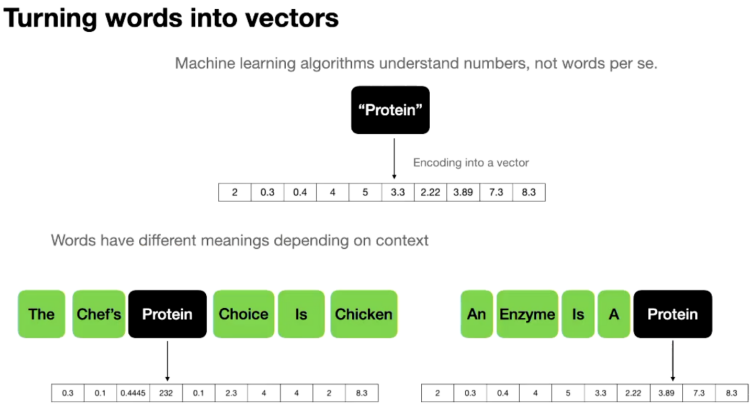
Semi-supervised learning

- Predict the next word in a sentence
- Model - BERT
- Data: e.g. wikipedia

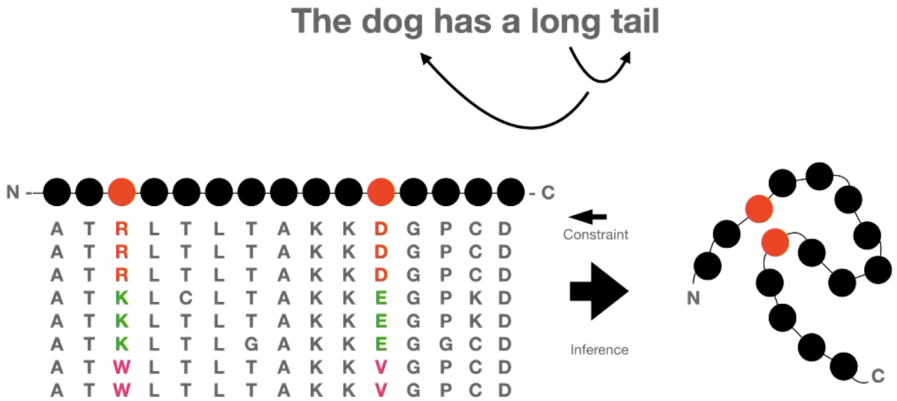


Embeddings

- A numerical representation of words or entities in a high-dimensional vector space.
- Embeddings are used to represent words or entities in a way that captures semantic relationships and similarities between them.



- Every single amino acid is a word in a sentence
- Learn what is the context of every amino acid in a particular protein sequence



- Seq2Vec: protein language models
- Training on large sequence datasets
  - Capture the diversity of protein ‘language’
  - Various databases: UniRef50, BFD, MGnify, CATH, Pfam
  - Years of computing hours
- Transformers reconstruct masked or currupted sequences