Dynamic Programming

- Sequence Analysis
- The Sequence Alignment Problem
- Dynamic Programming Formulation

Imdad ullah Khan

Sequence Terminology

- Alphabet: A set of possible symbols/characters, $\Sigma = \{A, C, T, G\}$
- String or Sequence: ordered list of characters from Σ — *ATCCTGATCT*
- Length: Number of characters in the string, denoted by $len(S)$ or $|S|$
- Prefix: Consecutive first few characters

  A, *AT*, *ATC*, *ATCCTAG*, *ATCCTAGATCCT*
- Suffix: Consecutive last few characters

  T, *CT*, *CCT*, *ATCCT*, *ATCCTAGATCCT*
- Substring: Consecutive few characters from an index to another

  A, *CT*, *CCTAGATC*, *TAGA*
- Subsequence: few characters from the sequence in same order

  A, *CT*, *C T G TC*, *TA C T*

Sequences in Biology

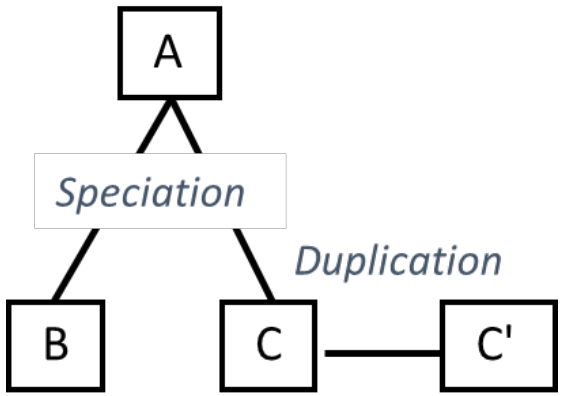
- DNA, RNA and proteins can be termed as molecular fossils as they encode the history of millions of years of evolution
- During evolution, molecular sequences accumulate random changes (mutations/variants) some of which provide a selective advantage or disadvantage, and some of which are neutral
- Sequences that are structurally and/or functionally important tend to be more conserved
- Such sequence conservation allows inference of evolutionary relatedness or homology (paralogs and orthologs)

Homology: Orthologs vs. Paralogs

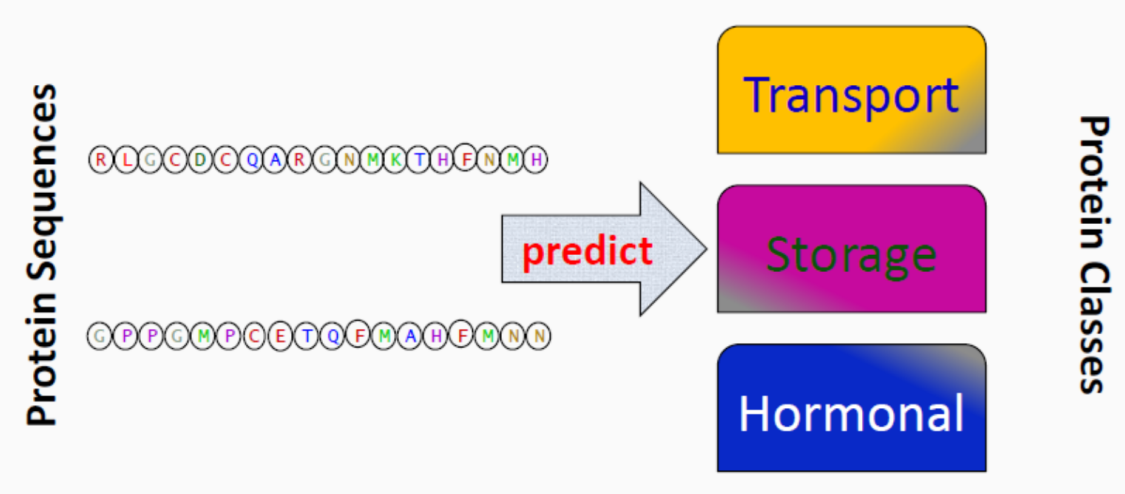
Homology is the existence of shared ancestry between a pair of structures, or genes, in different species

Two types of homologous sequences

- Orthologs
  - "same genes" in different species
  - result of common ancestry
  - corresponding proteins have "same" functions
  - e.g., human α-globin & mouse α-globin
- Paralogs
  - "similar genes" within a specie
  - result of gene duplication event
  - corresponding proteins may or may not have "same" functions
  - e.g., human α-globin & mouse β-globin



Sequence Analysis (Classification)

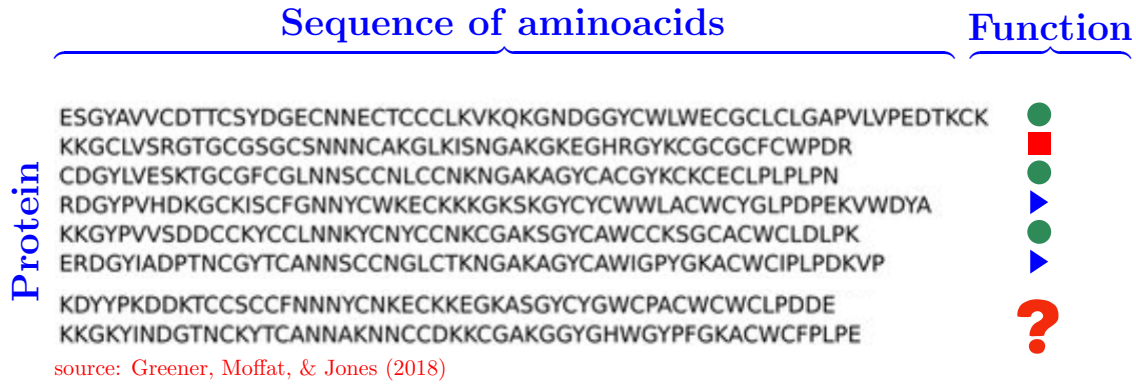


- Find homologous sequences (gene, protein)
- Protein 3d-Fold Recognition
- Protein Function identification
- ⋮

Sequence Similarity/Distance Measure

Gene and Protein Sequence Databases contain numerous entities with known and unknown functions

source: microbenotes.com



source: Greener, Moffat, & Jones (2018)

Genes/proteins with similar sequences have similar structure/functions

For a sequence with unknown function, find the "most similar"' sequence with known function and make a functional & evolutionary inference

Sequence Similarity/Distance Measure

Sequence similarity

- Can be used for spell checking and correction
- is used in Unix diff, svn/git, plagiarism detection
- Can be used for automatic music classification (music genre prediction, author identification)

Sequence Similarity: Edit Operations

Edit Operations:

Let $\Sigma^* = \Sigma \cup \{-\}$ for $- \notin \Sigma$ (a special symbol representing empty string)

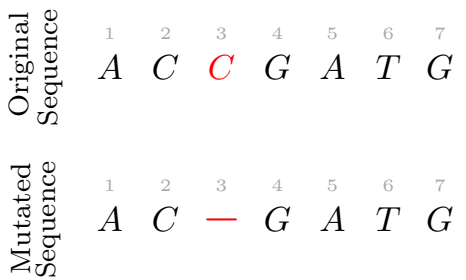
An edit operation is a triplet (x, i, y) with $(x, y) \in \Sigma^* \times \Sigma^*$ and integer i

$\triangleright\ (x, i, y) \neq (-, i, -)$

$$(x, i, y) \text{ is } \begin{cases} \text{deletion at } i\text{th index} & \text{if } \ x \neq - \wedge y = - \\ \text{insertion at } i\text{th index} & \text{if } \ x = - \wedge y \neq - \\ \text{substitution at } i\text{th index} & \text{if } \ x \neq - \wedge y \neq - \end{cases}$$

deletion at 3rd index

$(C, 3, -)$

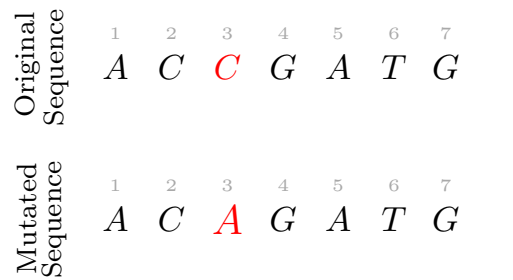


insertion at 4th index

$(-, 4, A)$

substitution at 3rd index

$(C, 3, A)$



Sequence Similarity: Edit Distance

Edit or Levenshtein Distance between S_1 and S_2 is the minimal number of delete, insert, and substitute operations needed to transform S_1 to S_2.

$S_1 = ACCCGAT$ and $S_2 = ACTGA$ have distance at most 4

$S_1 \underset{t}{\rightarrow} S_2$ (S_1 transformed to S_2): if there exists a sequence of edit operations on S_1 resulting in S_2

$$ACCCGAT \xrightarrow[(C,4,-)]{} ACCGAT \xrightarrow[(C,2,-)]{} ACGAT \xrightarrow[(T,5,-)]{} ACGA \xrightarrow[(-,5,G)]{} ACGAG$$

Different operations could have different costs (based on e.g. chemical properties of nucleotides or amino-acids)

▷ Leads to different notion of edit distances

Could also have more operations such as transposition, merge, split, etc.

Edit Distance: Applications

- Inspired by biological mutations in DNA and proteins
- Naturally many applications in molecular biology

  ▷ Protein homology detection (shared ancestry), protein fold prediction (3d structure), and many others
- Many applications in NLP and speech recognition

Spokesman	confirms		senior	government	adviser	was	shot	
Spokesman	said	the	senior		adviser	was	shot	dead
	sub	ins		del				ins

- Application in information extraction as Entity Recognition