Lezioni di Informatica Biomedica

Computational Visual Design (CVD-Lab), DIA, “Roma Tre”
University, Rome, Italy

Common Longest Subsequence

Sommario

Longest Common Subsequence 1

- Introduction
- FASTA Format
- Sequence alignment
- Dot Matrix of two sequences
- Introduction to dynamic programming

Contents

Longest Common Subsequence 1

- Introduction
- FASTA Format
- Sequence alignment
- Dot Matrix of two sequences
- Introduction to dynamic programming

The alignment of biological sequences

The alignment of biological sequences for the purpose of assessing the degree of similarity (and conservation, in the case of amino acid sequences) and the possibility of homology began in the early 1970's

- Since, it has become the core of numerous applications in sequence analysis including the functional annotation of genes and proteins, the analysis of protein domains, and the prediction of 3D structure due to homology
- Many sophisticated computational methods in molecular biology such as multiple alignments, profile analysis, and threading use a pair-wise sequence alignment as a subprocedure

(excerpt from web page at cBio@MSKCCMemorial Sloan Kettering Cancer Center, New York, NY)

Global alignment of sequences

A global alignment is the alignment of two sequences over their entire length

- The first dynamic programming algorithm for the global alignment of two sequences was introduced by Needleman and Wunsch (1970)

- A local alignment is the alignment of some portion of two sequences

- Smith and Waterman (1981) modified the Needleman-Wunsch method to calculate the score of the best alignment between two proteins

- Thus, the optimal local alignment between a pair of sequences involves a simple modification to the Needleman-Wunsch method in which only the highest-scoring sub-segments of the two sequences are aligned

Alignment programs

A number of programs have been written to rapidly search a database for the sequence in question, the query sequence

- The two most commonly used programs are BLAST (Altschul, et al
- 1990) and FASTA (Lipman and Pearson, 1985)
- These programs are an ideal starting point to determine whether a related sequence, or a family of sequences, already exists in a database
- The results from these programs will provide evidence of function, utility, and completeness of the gene product

Contents

Longest Common Subsequence 1

- Introduction
- FASTA Format
- Sequence alignment
- Dot Matrix of two sequences
- Introduction to dynamic programming

FASTA format

In bioinformatics, FASTA format is a

- Text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes
- The format also allows for sequence names and comments to precede the sequences
- The simplicity of FASTA format makes it easy to manipulate and parse sequences using text-processing tools and scripting languages like Python, Ruby, and Perl
- FASTA format description

Contents

Longest Common Subsequence 1

- Introduction
- FASTA Format
- Sequence alignment
- Dot Matrix of two sequences
- Introduction to dynamic programming

Background

Biomolecules are strings from a restricted alphabet

- Let Σ be an alphabet, a non-empty finite set.
- Elements of Σ are called symbols or characters.
- A string (or word) over Σ is any finite sequence of characters from Σ.
- For example, if $\Sigma = \{0, 1\}$, then 0101 is a string over Σ

Background

Biomolecules are strings from a restricted alphabet

DNA alphabet Length=4

- 4 nucleotides

Protein alphabet Length=20

- 20 amino acids

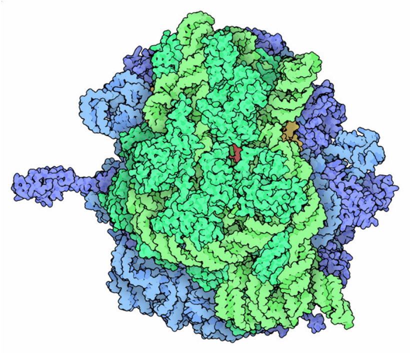
Shape determines function

```
RIBOSOME =
 "MARIAGVEIPRNKRVDVALTYIYG IGKARAKEALEKTGINPATRVK
    DLTEAEVVRLREYVENTWKLE GELRAEVAANIKRLMDIGCYR
    GLRHRRGLPVRGQRTRTNAR TRKGPRKTVAGKKKAPRK . . ."
```

1. After solving the structures of the individual small and large subunits, the next step in ribosome structure research was to determine the structure of the whole ribosome.
2. This work is the culmination of decades of research, which started with blurry pictures of the ribosome from electron microscopy, continued with more detailed cryoelectron micrographic reconstructions, and now includes many atomic structures.
3. These structures are so large that they don’t fit into a single PDB file — for instance, the structure shown here was split into PDB entries `2wdk` and `2wdl`.

Shape determines function

- Protein is a string (sequence of amino acids)
- Proteins do not stay linear in space
- Folding happens
- Folding determines overall 3-D shape
- Shape determines function



In 2000, structural biologists Venkatraman Ramakrishnan, Thomas A. Steitz and Ada E. Yonath made the first structures of ribosomal subunits available in the PDB, and in 2009, they each received a Nobel Prize for this work.

Sequence ⇒ Structure ⇒ Function

- the amino acids in a protein sequence interact locally and establish hydrogen (and even covalent) bounds
- the interaction folds the protein in space and gives it a 3D structure
- the 3D structure determines the protein function
- each protein within the body has a specific function

Sequence alone does not reveal structure

Much less function ... So?

Nature does not solve the same problem twice (usually)

- Short sequence with a specific function (or shape) is called a domain
- The same domain appears in multiple proteins
- If we find the same domain in multiple proteins that provides a clue to function and/or structure

Sequence is easier to get than structure or function

How biologists study proteins

- To study the 3D structure of proteins is hard and expensive (NMR, x-ray crystallography)
- Analogously, the discovery of function through laboratory (in-vitro) and animal (in-vivo) experiments is difficult
- Therefore, few (tens of) thousands of proteins are understood in detail
- Many (i.e. millions) are known only by sequence

SEQUENCE ALIGNMENT SCENARIO

sequence of a new protein with unknown function

- Biologist discovers the sequence of a new protein with unknown function

- If sequence can be associated with a known protein sequence we have a clue about structure and/or function

- Vast quantities of sequence, structure, function info is deposited into public databases

- The new sequence should be compared to the database to find the more similar domains

Main Alignment Methods

- Dot Matrix
- Dynamic Programming
- BLAST, FASTA

Contents

Longest Common Subsequence 1

Introduction

FASTA Format

Sequence alignment

Dot Matrix of two sequences

Introduction to dynamic programming

Similarity of Sequences as homology of structures

homology ≡ any characteristic of biological organisms that is derived from a common ancestor

- Locating regions of similarity between two DNA or protein sequences
- Provide a lot of information about the function and structure of the query sequence
- Similarity of sequences indicates homology
- Two structures are called homologous if they represent corresponding parts of organisms which are built according to the same body plan
- The existence of corresponding structures in different species is explained by derivation from a common ancestor

Similarity relation

matrix picture of sequence similarity

A picture of the similarity of two sequences X, Y can be given by the graph of the similarity relation $S \subseteq X \times Y$ such that:

$$x_i \, S \, y_j \equiv (x_i, y_j) \in S \Longleftrightarrow x_i = y_j$$

By the way, the interesting part of the similarity relation S is given by its reflexive subsets

$$S_{i,j,k} = \{(x_i, y_j) \,|\, x_{i+\ell} = y_{j+\ell}, \quad \ell = 0, \ldots, k\}$$

with starting point (i, j) and length k

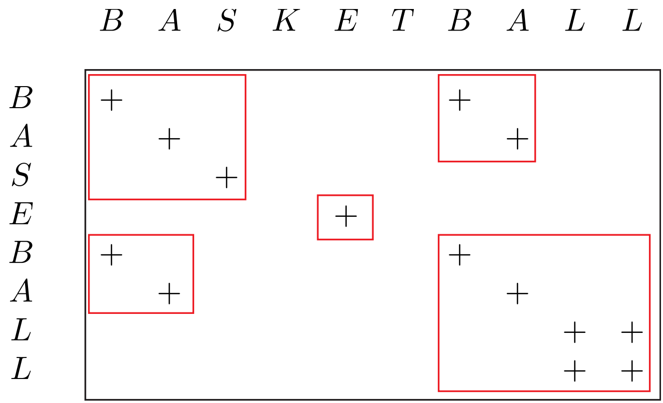
Similarity relation

matrix picture of sequence similarity

$$
\begin{array}{c|cccccccccc|}
 & B & A & S & K & E & T & B & A & L & L \\
\hline
B & + & & & & & & + & & & \\
A & & + & & & & & & + & & \\
S & & & + & & & & & & & \\
E & & & & & + & & & & & \\
B & + & & & & & & + & & & \\
A & & + & & & & & & + & & \\
L & & & & & & & & & + & + \\
L & & & & & & & & & + & + \\
\hline
\end{array}
$$

Similarity relation

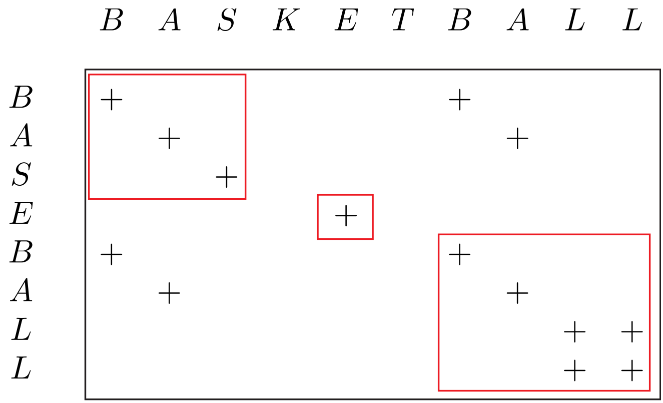
matrix picture of sequence similarity



$$\begin{array}{c|cccccccccc}
 & B & A & S & K & E & T & B & A & L & L \\
\hline
B & + & & & & & & + & & & \\
A & & + & & & & & & + & & \\
S & & & + & & & & & & & \\
E & & & & & + & & & & & \\
B & + & & & & & & + & & & \\
A & & + & & & & & & + & & \\
L & & & & & & & & & + & + \\
L & & & & & & & & & + & +
\end{array}$$

Similarity relation

drop out the reflexive subset that are non maximal[1]



$$\begin{array}{c|cccccccccc}
 & B & A & S & K & E & T & B & A & L & L \\
\hline
B & + & & & & & & + & & & \\
A & & + & & & & & & + & & \\
S & & & + & & & & & & & \\
E & & & & & + & & & & & \\
B & + & & & & & & + & & & \\
A & & + & & & & & & + & & \\
L & & & & & & & & & + & + \\
L & & & & & & & & & + & +
\end{array}$$

[1] if we (i.e. that are contained within another reflexive subset)

Similarity relation

finally project the maximal reflexive subrelations in one (or both) starting sequence

getting the Longest Common Subsequence

$$\boxed{B \quad A \quad S} \ \boxed{E} \ \boxed{B \quad A \quad L \quad L}$$

Contents

Longest Common Subsequence 1

- Introduction
- FASTA Format
- Sequence alignment
- Dot Matrix of two sequences
- Introduction to dynamic programming

Introduction to dynamic programming

Bellman optimality principle

> *Principle of Optimality: An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.*

Richard Bellman, 1957. *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Optimal substructure

necessary condition

necessary condition for optimality associated with the mathematical optimization method known as dynamic programming

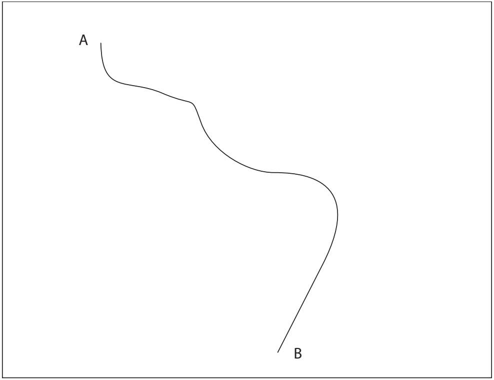
It breaks a dynamic optimization problem into simpler subproblems

In computer science, a problem that can be broken apart like this is said to have optimal substructure

Optimal substructure

a global optimal policy

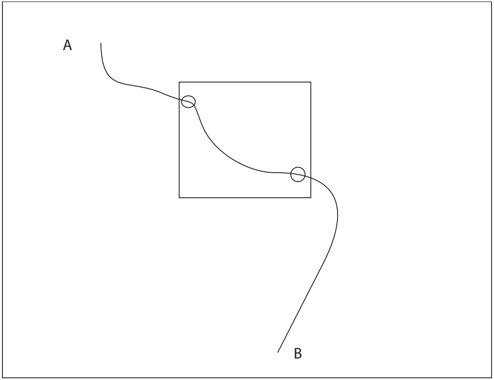
The (optimal) solution of a problem with optimal substructure is made by composition of (optimal) solutions to subproblems, each having in turn optimal substructure



Optimal substructure

a global optimal policy

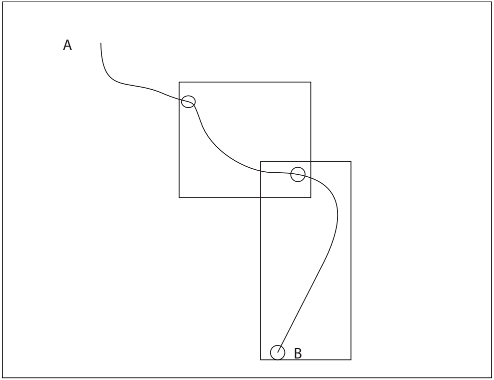
The (optimal) solution of a problem with optimal substructure is made by composition of (optimal) solutions to subproblems, each having in turn optimal substructure



Optimal substructure

a global optimal policy

The (optimal) solution of a problem with optimal substructure is made by composition of (optimal) solutions to subproblems, each having in turn optimal substructure



Optimal substructure

a local optimal policy

The (optimal) solution of a problem with optimal substructure is made by composition of (optimal) solutions to subproblems, each having in turn optimal substructure