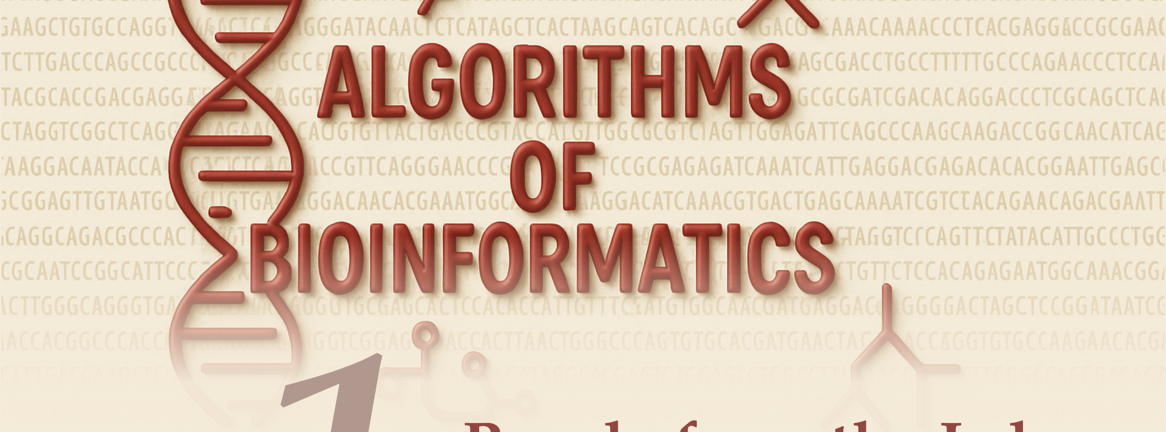


1 Puzzle from the Lab

16 October 2025

Prof. Dr. Sebastian Wild

CS594 (Winter 2025-26)
Philipps-Universität Marburg

version `2025-10-16 20:30`

Outline

1 Puzzle from the Lab

1.1 Protein Sequencing

1.2 The Turnpike Problem

1.3 Backtracking Algorithm

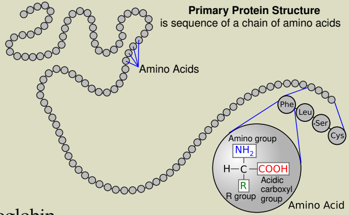
1.4 A Pseudopolynomial Algorithm

1.5 Back to the Lab

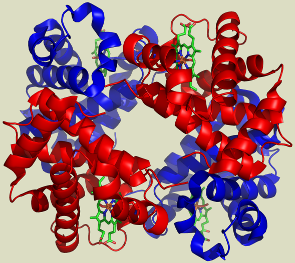
1.1 Protein Sequencing

Proteins: The Workhorses of the Cell

- **What are they?** Chains of amino acids, folded into specific 3D shapes. The shape determines the function.
- **What do they do?** Almost everything!
  - They act as *enzymes* (catalyzing chemical reactions)
  - provide structural support (cell walls, muscles!),
  - transport molecules (e. g., *hemoglobin*),
  - send signals (some *hormones*, e. g., *insulin*)
  - and more



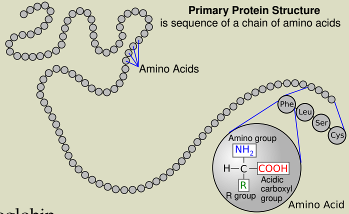
3D Structure of hemoglobin



https://commons.wikimedia.org/wiki/File:1GZX_Haemoglobin.png

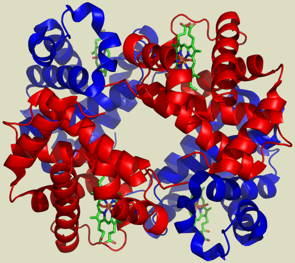
Proteins: The Workhorses of the Cell

- **What are they?** Chains of amino acids, folded into specific 3D shapes. The shape determines the function.



- **What do they do?** Almost everything!
  - They act as *enzymes* (catalyzing chemical reactions)
  - provide structural support (cell walls, muscles!),
  - transport molecules (e. g., *hemoglobin*),
  - send signals (some *hormones*, e. g., *insulin*)
  - and more

3D Structure of hemoglobin



https://commons.wikimedia.org/wiki/File:1GZX_Haemoglobin.png

⇝ Target of many activities across bioinformatics

  - analyzing amino acid sequence
  - predicting structure (AlphaFold)
  - study interaction networks
  - design new proteins as potential drugs
  - . . .

Amino Acids

Amino acid	3-letter code	Molecular formula	Mass (Da)
Alanine	Ala	C_3H_5NO	71.03711
Cysteine	Cys	C_3H_5NOS	103.00919
Aspartic acid	Asp	$C_4H_5NO_3$	115.02694
Glutamic acid	Glu	$C_5H_7NO_3$	129.04259
Phenylalanine	Phe	C_9H_9NO	147.06841
Glycine	Gly	C_2H_3NO	57.02146
Histidine	His	$C_6H_7N_3O$	137.05891
Isoleucine	Ile	$C_6H_{11}NO$	113.08406
Lysine	Lys	$C_6H_{12}N_2O$	128.09496
Leucine	Leu	$C_6H_{11}NO$	113.08406
Methionine	Met	C_5H_9NOS	131.04049
Asparagine	Asn	$C_4H_6N_2O_2$	114.04293
Proline	Pro	C_5H_7NO	97.05276
Glutamine	Gln	$C_5H_8N_2O$	128.05858
Arginine	Arg	$C_6H_{12}N_4O$	156.10111
Serine	Ser	$C_3H_5NO_2$	87.03203
Threonine	Thr	$C_4H_7NO_2$	101.04768
Valine	Val	C_5H_9NO	99.06841
Tryptophan	Trp	$C_{11}H_{10}N_2O$	186.07931
Tyrosine	Tyr	$C_9H_9NO_2$	163.06333

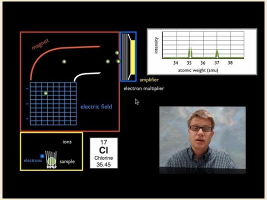
Compeau & Pevzner, *Bioinformatics Algorithms*, `https://cogniterra.org/lesson/29925/step/1`

- **Dalton (Da):** unit of molecular mass.
- **1 Da** = $\frac{1}{12}$ of a carbon-12 atom $\approx 1.66 \times 10^{-27}$ kg.
  - We will use rounded integer weights
- **Monoisotopic mass:** sum of atomic masses of most abundant isotopes.
- Only shows 20 *proteinogenic* amino acids (those encoded in DNA)

Protein Sequencing

How to determine the sequence of amino acids in a protein?

- indirect option: via *genes*
    - … we will come back to that
    - not always possible (e. g., for *non-ribosomal peptides*)
- (more) direct option: *mass spectrometry*
    1. Shatter (many copies) molecule into pieces
    2. Measure *spectrum* of particle masses* (which masses occur how often)



Mass Spectrometry
`https://youtu.be/mBT73Pesiog`

⇝ from this, reconstruct what the molecule was!?

1.2 The Turnpike Problem

Turnpike Problems

▶ The Sopranos Opening
`https://youtu.be/mJpNmYeooQE`

Turnpike Problems

▶ The Sopranos Opening
`https://youtu.be/mJpNmYeooQE`

⇝ Turnpike = toll road

- typically, price for road $\propto$ length of segment on road
- Can enter and leave at any pair of exits

Ideal Spectra

Back to mass spectrometry . . .

Simplifying assumptions

- perfect integer molecular weights, no isotopes
- all breakpoints realized
- multiplicities of weights correctly observed
- no contamination

Ideal Spectra

Back to mass spectrometry . . .

Simplifying assumptions

- perfect integer molecular weights, no isotopes
- all breakpoints realized
- multiplicities of weights correctly observed
- no contamination

Definition 1.1 (Difference multiset)

molecular weights / distances / turnpike tolls

Given $P = P[0..n) \in \mathbb{N}_{\geq 1}^{n}$ a sequence of numbers,
define the *prefix sums* $S[0..n] = \mathrm{prefSum}(P[0..n))$ via $S[i] = P[0] + \cdots + P[i-1]$.

Ideal Spectra

Back to mass spectrometry . . .

Simplifying assumptions

- perfect integer molecular weights, no isotopes
- all breakpoints realized
- multiplicities of weights correctly observed
- no contamination

Definition 1.1 (Difference multiset)

molecular weights / distances / turnpike tolls

Given $P = P[0..n) \in \mathbb{N}^n_{\geq 1}$ a sequence of numbers,
define the *prefix sums* $S[0..n] = \mathrm{prefSum}(P[0..n))$ via $S[i] = P[0] + \cdots + P[i-1]$.

The *difference multiset* ΔS is the multiset

multiset

$$\Delta S \quad = \quad \{\!\!\{ S[j] - S[i] \,:\, 0 \leq i < j \leq n \}\!\!\}.$$

◀

Important: Keep duplicates / multiplicities of distances! $\rightsquigarrow$ $\big|\Delta S[0..n]\big| = \binom{n+1}{2}$

The Turnpike Problem

Definition 1.2 (Turnpike Problem)

Given: a multiset D with $|D| = \binom{n}{2}$

Goal: Find sequence P with $\Delta(\text{prefSum}(P)) = D$ (or state that no such P exists). ◀

The Turnpike Problem

Definition 1.2 (Turnpike Problem)

Given: a multiset D with $|D| = \binom{n}{2}$

Goal: Find sequence P with $\Delta(\text{prefSum}(P)) = D$ (or state that no such P exists).

Examples:

$\textcircled{3}\ \textcircled{5}\ \textcircled{1}\ \textcircled{2}$

$0 \le i < j \le n = 4$

1. $P_1 = [3, 5, 1, 2]$

$\rightsquigarrow$ $S_1 = [0, 3, 8, 9, 11]$

$\rightsquigarrow$ $D_1 = \Delta S_1 = \{\!\{1, 2, 3, 3, 5, 6, 8, 8, 9, 11\}\!\}$

$$\sum_{k=i}^{j-1} P(k)$$

The Turnpike Problem

Definition 1.2 (Turnpike Problem)

Given: a multiset D with $|D| = \binom{n}{2}$

Goal: Find sequence P with $\Delta(\text{prefSum}(P)) = D$ (or state that no such P exists).

3 5 1 2

3	3+5	3+5+1	3+5+1+2
5	5+1	5+1+2	
1	1+2		
2			

Examples:

1. $P_1 = [3, \underline{5, 1}, 2]$

$\rightsquigarrow$ $S_1 = [0, 3, 8, 9, 11]$

$\rightsquigarrow$ $D_1 = \Delta S_1 = \{\{1, 2, 3, 3, 5, \underline{6}, 8, 8, 9, 11\}\}$

2. $P_2 = [1, 1, 1, 1, 1]$

$\rightsquigarrow$ $S_2 = [0, 1, 2, 3, 4, 5]$

$\rightsquigarrow$ $D_2 = \Delta S_2 = \{\{1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5\}\}$

3. For $D = \{\{1, 1, 1\}\}$ no set S exists such that $D = \Delta S$

Any two points $a < b$ will give $\Delta(0, a, b) = \{\{a, b, b - a\}\}$ ↯ $a \neq b$

Clicker Question

$n = 3$

① ① ②

Suppose $\Delta S = \{\!\{1, 1, 2, 2, 3, 4\}\!\}$. What is S?

1 1

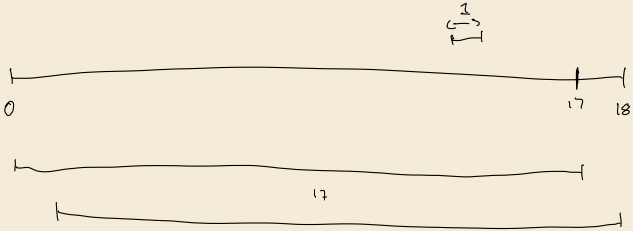
2 1 1

→ sli.do/cs594

1.3 Backtracking Algorithm

Systematic Solution

Consider $\Delta S = \{\!\{\underline{1}, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, \underline{17}, \underline{18}\}\!\}$.



Backtracking Turnpike

```
procedure turnpikeBacktracking($D$)
    $d := \max D$
    $P := \{0, d\}$ // sorted set
    return turnpikeRec($P$, $D$)

procedure turnpikeRec($P$, $D$)
    // Invariant: $\Delta P \subseteq D$
    if $\Delta P == D$
        return $P$
    $d := \max(D \setminus \Delta P)$
    // Option 1: Distance $d$ from left end
    $P' := P \cup \{d\}$
    if $\Delta P' \subseteq D$
        $R :=$turnpikeRec($P'$, $D$)
        if $R \neq$ NO_DIFFERENCE_MULTISET
            return $R$
    // else try Option 2: Distance $d$ from right
    $P' := P \cup \{(\max D) - d\}$
    if $\Delta P' \subseteq D$
        return turnpikeRec($P'$, $D$)
    // else: no option worked!
    return NO_DIFFERENCE_MULTISET
```

Backtracking Turnpike

```
procedure turnpikeBacktracking($D$)
    $d$ := $\max D$
    $P$ := $\{0, d\}$ // sorted set
    return turnpikeRec($P$, $D$)

procedure turnpikeRec($P$, $D$)
    // Invariant: $\Delta P \subseteq D$
    if $\Delta P == D$
        return $P$
    $d$ := $\max(D \setminus \Delta P)$
    // Option 1: Distance $d$ from left end
    $P'$ := $P \cup \{d\}$
    if $\Delta P' \subseteq D$
        $R$ := turnpikeRec($P'$, $D$)
        if $R \neq$ NO_DIFFERENCE_MULTISET
            return $R$
    // else try Option 2: Distance $d$ from right
    $P'$ := $P \cup \{(\max D) - d\}$
    if $\Delta P' \subseteq D$
        return turnpikeRec($P'$, $D$)
    // else: no option worked!
    return NO_DIFFERENCE_MULTISET
```

- **Correctness**
    - After placing a few points in prefix sums P, largest remaining distance must be measured from one endpoint.
    - Otherwise we are immediately missing a larger distance ↯
    - ⇝ only two checked options are possible
    - invariant explicitly checked for recursive calls
    - invariant at return guarantees correct answer

Backtracking Turnpike

```
procedure turnpikeBacktracking($D$)
    $d$ := max $D$
    $P$ := $\{0, d\}$ // sorted set
    return turnpikeRec($P$, $D$)

procedure turnpikeRec($P$, $D$)
    // Invariant: $\Delta P \subseteq D$
    if $\Delta P == D$
        return $P$
    $d$ := $\max(D \setminus \Delta P)$
    // Option 1: Distance $d$ from left end
    $P'$ := $P \cup \{d\}$
    if $\Delta P' \subseteq D$
        $R$ := turnpikeRec($P'$, $D$)
        if $R \neq$ NO_DIFFERENCE_MULTISET
            return $R$
    // else try Option 2: Distance $d$ from right
    $P'$ := $P \cup \{(\max D) - d\}$
    if $\Delta P' \subseteq D$
        return turnpikeRec($P'$, $D$)
    // else: no option worked!
    return NO_DIFFERENCE_MULTISET
```

- **Correctness**
  - After placing a few points in prefix sums P, largest remaining distance must be measured from one endpoint.
  - Otherwise we are immediately missing a larger distance ↯
  - ⇝ only two checked options are possible
  - invariant explicitly checked for recursive calls
  - invariant at return guarantees correct answer

- **Running time**
  - worst case: exponential! ⇝ *see tutorials*
  - not known whether problem is NP-hard(!)