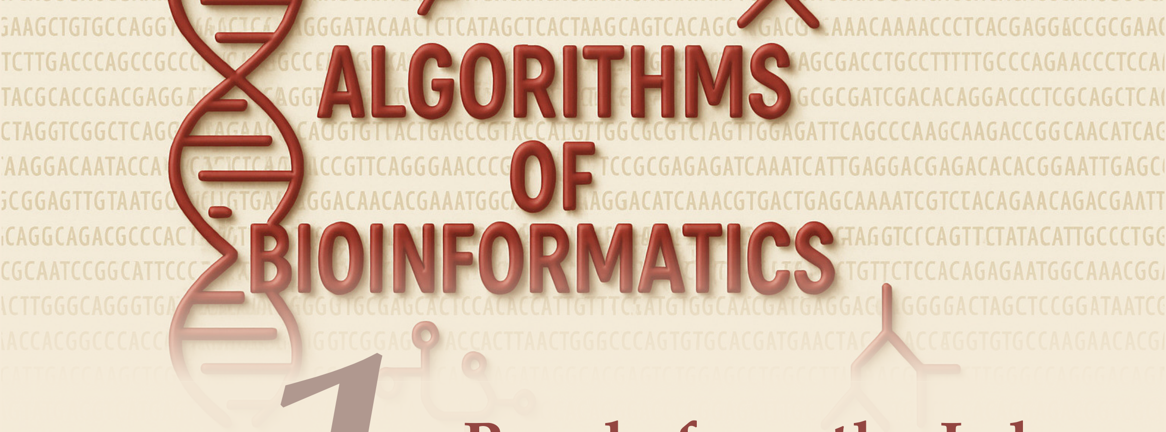


1

Puzzle from the Lab

16 October 2025

Prof. Dr. Sebastian Wild

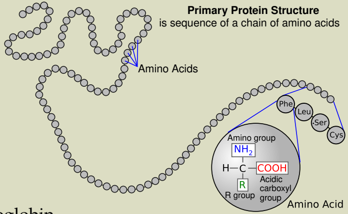
CS594 (Winter 2025-26)
Philipps-Universität Marburg
version `2025-10-23 11:58`

Outline

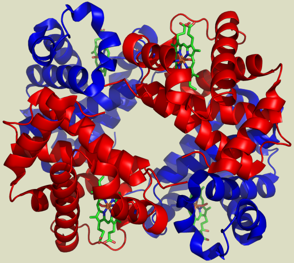
1.1 Protein Sequencing

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



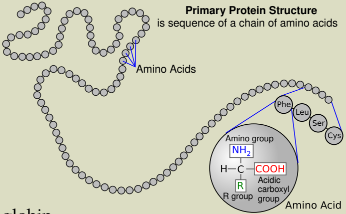
3D Structure of hemoglobin



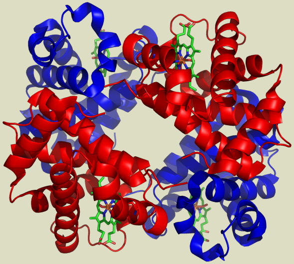
`https://commons.wikimedia.org/wiki/File:1GZX_Haemoglobin.png`

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

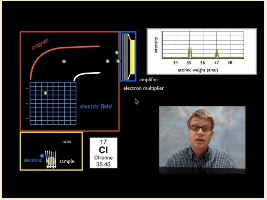
Compeau & Pevzner, *Bioinformatics Algorithms*, `https://cogniterra.org/lesson/29925/step/1`

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

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 $S[i] = P[0] + \cdots + P[i-1]$.

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 $\Delta(\mathrm{prefSum}(P)) = D$ (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:

③ ⑤ ① 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 $\Delta(\mathrm{prefSum}(P)) = D$ (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

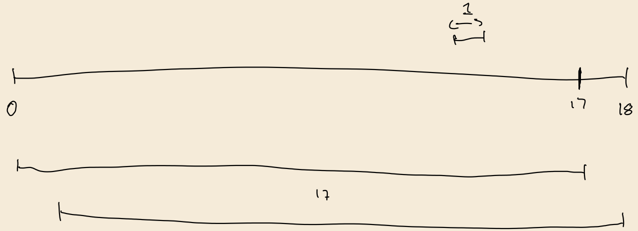
→ sli.do/cs594

1.3 Backtracking Algorithm

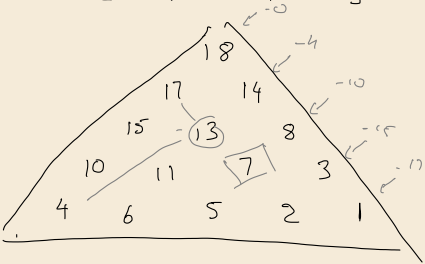
Systematic Solution

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

$S = [0, 4, 10, 15, 17, 18]$



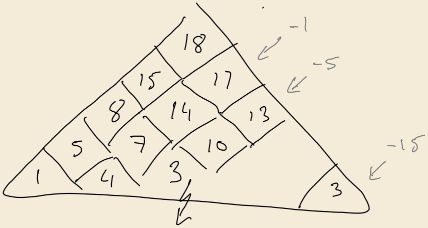
{{~~1~~, 2, ~~3~~, ~~4~~, ~~5~~, 6, ~~7~~, ~~8~~, ~~10~~, 11, ~~13~~, ~~14~~, ~~15~~, ~~17~~, ~~18~~}}.

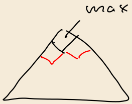


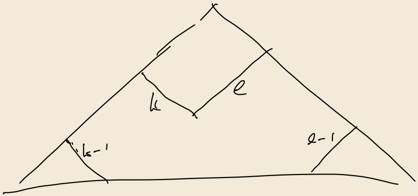
- pyramid is increasing

↗ ↖ along diagonals

⇒ second largest can go






k
ℓ
k-1
ℓ-1

Backtracking Turnpike

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

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

Backtracking Turnpike

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

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

- **Correctness**
  - After placing a few points in prefix sums S, 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($D$)
    $d := \max D$
    $S := \{0, d\}$ // sorted set of prefSums
    return turnpikeRec($S$, $D$)

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

- **Correctness**
  - After placing a few points in prefix sums S, 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(!)

Clicker Question

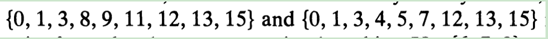
If D is a valid difference multiset, turnpikeBacktracking will find a sequence producing it.
Is the answer well-defined?

- **A** Yes
- **B** No
- **C** For most inputs
- **D** No idea
- **E** turnpikeBacktracking is too slow to tell.

→ `sli.do/cs594`

Clicker Question

If D is a valid difference multiset, turnpikeBacktracking will find a sequence producing it.
Is the answer well-defined?



{0, 1, 3, 8, 9, 11, 12, 13, 15} and {0, 1, 3, 4, 5, 7, 12, 13, 15}

S_1 S_2

$\Delta S_1 = \Delta S_2$

- **A** ~~Yes~~
- **B** No ✓
- **C** For most inputs ✓ – for random real P proven
- **D** ~~No idea~~
- **E** ~~turnpikeBacktracking is too slow to tell.~~

→ sli.do/cs594

1.4 A Pseudopolynomial Algorithm

Algebra to the Rescue

Few other algorithmic approaches known for the Turnpike Problem . . .

Algebra to the Rescue

Few other algorithmic approaches known for the Turnpike Problem . . .
but one seemingly magic one does!

Algebra to the Rescue

Few other algorithmic approaches known for the Turnpike Problem . . .
but one seemingly magic one does!

- Consider again $S = [0, 3, 8, 9, 11] \quad \rightsquigarrow \quad D = \Delta S = \{\!\{\mathbf{1, 2, 3, 3, 5, 6, 8, 8, 9, 11}\}\!\}$
- We can get all pairwise combinations (distances) via *convolutions*
  - Write $S(z) = \sum_{s \in S} z^s = z^{11} + z^9 + z^8 + z^3 + z^0$

Algebra to the Rescue

Few other algorithmic approaches known for the Turnpike Problem . . .
but one seemingly magic one does!

- Consider again $S = [0, 3, 8, 9, 11]$ $\rightsquigarrow$ $D = \Delta S = \{\!\{\mathbf{1, 2, 3, 3, 5, 6, 8, 8, 9, 11}\}\!\}$
- We can get all pairwise combinations (distances) via *convolutions*
  - Write $S(z) = \sum_{s \in S} z^s = z^{11} + z^9 + z^8 + z^3 + z^0$
  - Now observe that
    $$S(z) \cdot S(z^{-1}) = \left(\frac{1}{z^{11}} + \frac{1}{z^9} + \frac{1}{z^8} + \frac{1}{z^3} + 1\right)\left(z^{11} + z^9 + z^8 + z^3 + 1\right)$$

Algebra to the Rescue

Few other algorithmic approaches known for the Turnpike Problem . . .
but one seemingly magic one does!

- Consider again $S = [0, 3, 8, 9, 11] \rightsquigarrow D = \Delta S = \{\!\{\mathbf{1, 2, 3, 3, 5, 6, 8, 8, 9, 11}\}\!\}$
- We can get all pairwise combinations (distances) via *convolutions*
  - Write $S(z) = \sum_{s \in S} z^s = z^{11} + z^9 + z^8 + z^3 + z^0$
  - Now observe that

$$\begin{aligned} S(z) \cdot S(z^{-1}) &= \left(\frac{1}{z^{11}} + \frac{1}{z^9} + \frac{1}{z^8} + \frac{1}{z^3} + 1\right)\left(z^{11} + z^9 + z^8 + z^3 + 1\right) \\ &= z^{\mathbf{11}} + z^{\mathbf{9}} + \mathbf{2}z^{\mathbf{8}} + z^{\mathbf{6}} + z^{\mathbf{5}} + \mathbf{2}z^{\mathbf{3}} + z^{\mathbf{2}} + z^{\mathbf{1}} \end{aligned}$$

Algebra to the Rescue

Few other algorithmic approaches known for the Turnpike Problem . . .
but one seemingly magic one does!

- Consider again $S = [0, 3, 8, 9, 11]$ $\rightsquigarrow$ $D = \Delta S = \{\!\{\mathbf{1, 2, 3, 3, 5, 6, 8, 8, 9, 11}\}\!\}$
- We can get all pairwise combinations (distances) via *convolutions*
  - Write $S(z) = \sum_{s \in S} z^s = z^{11} + z^9 + z^8 + z^3 + z^0$
  - Now observe that

$$
\begin{aligned}
S(z) \cdot S(z^{-1}) &= \left(\frac{1}{z^{11}} + \frac{1}{z^9} + \frac{1}{z^8} + \frac{1}{z^3} + 1\right)\left(z^{11} + z^9 + z^8 + z^3 + 1\right) \\
&= \quad z^{\mathbf{11}} + z^{\mathbf{9}} + \mathbf{2}z^{\mathbf{8}} + z^{\mathbf{6}} + z^{\mathbf{5}} + \mathbf{2}z^{\mathbf{3}} + z^{\mathbf{2}} + z^{\mathbf{1}} \\
&\quad + \frac{1}{z^{11}} + \frac{1}{z^9} + \frac{2}{z^8} + \frac{1}{z^6} + \frac{1}{z^5} + \frac{2}{z^3} + \frac{1}{z^2} + \frac{1}{z} \ + 5z^0 \\
&= \sum_{s \in S} \sum_{t \in S} z^{s-t}
\end{aligned}
$$

Algebra to the Rescue

Few other algorithmic approaches known for the Turnpike Problem . . .
but one seemingly magic one does!

- Consider again $S = [0, 3, 8, 9, 11] \rightsquigarrow D = \Delta S = \{\{\mathbf{1, 2, 3, 3, 5, 6, 8, 8, 9, 11}\}\}$
- We can get all pairwise combinations (distances) via *convolutions*
  - Write $S(z) = \sum_{s \in S} z^s = z^{11} + z^9 + z^8 + z^3 + z^0$
  - Now observe that

$$
\begin{aligned}
S(z) \cdot S(z^{-1}) &= \left(\frac{1}{z^{11}} + \frac{1}{z^9} + \frac{1}{z^8} + \frac{1}{z^3} + 1 \right) \left(z^{11} + z^9 + z^8 + z^3 + 1 \right) \\
&= z^{\mathbf{11}} + z^{\mathbf{9}} + \mathbf{2}z^{\mathbf{8}} + z^{\mathbf{6}} + z^{\mathbf{5}} + \mathbf{2}z^{\mathbf{3}} + z^{\mathbf{2}} + z^{\mathbf{1}} \\
&\quad + \frac{1}{z^{11}} + \frac{1}{z^9} + \frac{2}{z^8} + \frac{1}{z^6} + \frac{1}{z^5} + \frac{2}{z^3} + \frac{1}{z^2} + \frac{1}{z} + 5 \\
&= \sum_{s \in S} \sum_{t \in S} z^{s-t} \\
&= \boxed{\sum_{\boldsymbol{d \in D}} \boldsymbol{z^d} + \sum_{d \in D} z^{-d} + |S|}
\end{aligned}
$$

not really a polynomial!

but $z^{11} \cdot S(z) \cdot S(z^{-1})$ is

$|D| = \binom{|S|}{2}$

Factoring Polynomials

- The expanded product depends only on D
  - ⇝ can be constructed from the input
- Use polynomial factorization to check if it can be written as a product $S(z)S(z^{-1})$
  - this can be done in *pseudopolynomial time* (poly time only if all number is in $O(n^c)$ in value
    - a polynomial of degree d with integer coefficients represented with b bits can be factored over the integers in time $O(\mathrm{poly}(d, b)) = (db)^{O(1)}$
    - *Lenstra-Lenstra-Lovász (LLL) algorithm*

$z^{1000000} + 50z^{10} + 5$

  - polynomial running time in terms of $n = |D|$, but exponential in $b = \log(\max D)$
    b is the number of bits in the occurring numbers

d can be expontial

in input encoding size

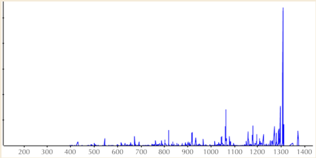
of a polynomial

1.5 Back to the Lab

Ideal vs. Real Spectra

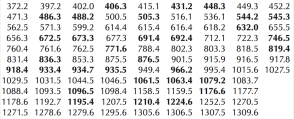
Real protein sequencing tasks unfortunately need additional work . . .

Actual spectrum



Compeau & Pevzner, *Bioinformatics Algorithms*, Fig. 4.13
`https://cogniterra.org/lesson/29918/step/2?unit=22015`

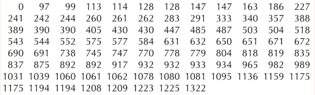
Values of peaks



372.2	397.2	402.0	**406.3**	415.1	**431.2**	**448.3**	449.3	452.2
471.3	**486.3**	**488.2**	500.5	**505.3**	516.1	536.1	**544.2**	**545.3**
562.5	571.3	599.2	614.4	615.4	616.4	618.2	**632.0**	655.5
656.3	**672.5**	**673.3**	677.3	**691.4**	**692.4**	712.1	722.3	**746.5**
760.4	761.6	762.5	**771.6**	788.4	802.3	803.3	818.5	**819.4**
831.4	**836.3**	853.3	875.5	**876.5**	901.5	915.9	916.5	917.8
918.4	**933.4**	**934.7**	**935.5**	949.4	**966.2**	995.4	1015.6	1027.5
1029.5	1031.5	1044.5	1046.5	**1061.5**	**1063.4**	**1079.2**	1083.7	
1088.4	1093.5	**1096.5**	1098.4	1158.5	1159.5	**1176.6**	1177.7	
1178.6	1192.7	**1195.4**	1207.5	**1210.4**	**1224.6**	1252.5	1270.5	
1271.5	1278.6	1279.6	1295.6	1305.6	1306.5	1307.5	1309.6	

Compeau & Pevzner, *Bioinformatics Algorithms*, Fig 4.14
`https://cogniterra.org/lesson/29918/step/3?unit=22015`

Ideal Spectrum



0	97	99	113	114	128	128	147	147	163	186	227
241	242	244	260	261	262	283	291	333	340	357	388
389	390	390	405	430	430	447	485	487	503	504	518
543	544	552	575	577	584	631	632	650	651	671	672
690	691	738	745	747	770	778	779	804	818	819	835
837	875	892	892	917	932	932	933	934	965	982	989
1031	1039	1060	1061	1062	1078	1080	1081	1095	1136	1159	1175
1175	1194	1194	1208	1209	1223	1225	1322				

Compeau & Pevzner, *Bioinformatics Algorithms*, Fig 4.7
`https://cogniterra.org/lesson/29912/step/5?unit=22009`

Complications:

- **inaccuracy** of “weights”
- weights are actually *mass/charge ratios* (often not so bad)
- missing/**missed peaks**
- **false peaks**, e. g., from contamination

Dealing with Real Spectra

Typical situation in bioinformatics!

- Inaccuracies in the data
    - can sometimes be cleaned
    - or avoided with better lab techniques
    - or averaged out by producing more repetitions
    - and/or be worked around by **better algorithms**!
- For example, we can
    - Find *best fitting* sequence instead of Yes/No (robust algorithms)
    - Use further domain knowledge (range of molecular weights of amino acids!)

⇝ Must deal with possibilities of incorrect results

- learn how to judge
- learn how to communicate shortcomings of methods clearly