# How Do We Sequence Antibiotics?

*Brute-Force Algorithms*

Phillip Compeau and Pavel Pevzner.

*Bioinformatics Algorithms: an Active Learning Approach*

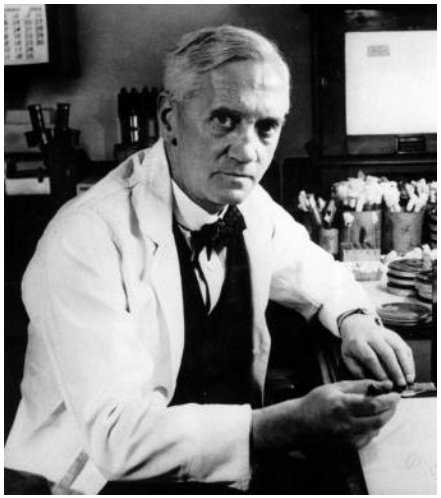
# Outline

- **The Discovery of Antibiotics**
- How Do Bacteria Make Antibiotics?
- Sequencing Antibiotics by Shattering Them into Pieces
- A Brute Force Algorithm for Cyclopeptide Sequencing
- Cyclopeptide Sequencing with Branch-and-Bound
- Adapting Sequencing for Spectra with Errors
- From 20 to More than 100 Amino Acids
- The Spectral Convolution Saves the Day
- The Truth about Spectra

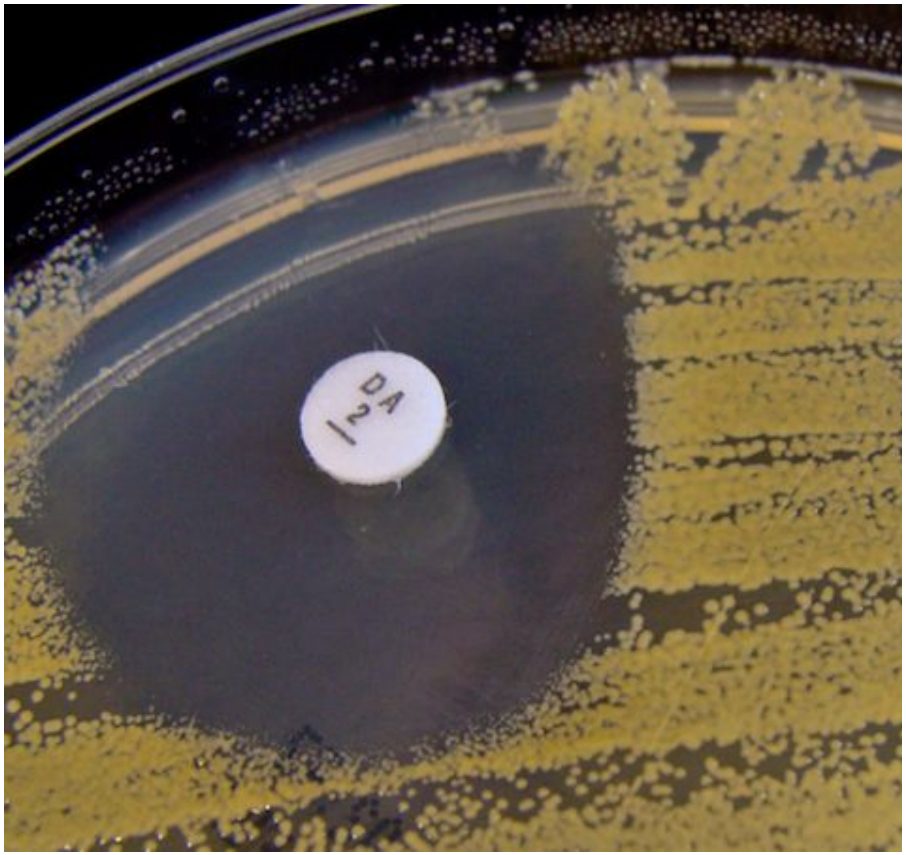
# Discovery of Penicillin (1928)

**Lessons:**

1. Keep a messy lab.
2. Take vacations.
3. Science = mistakes.



Alexander Fleming




Courtesy: Nathan Reading

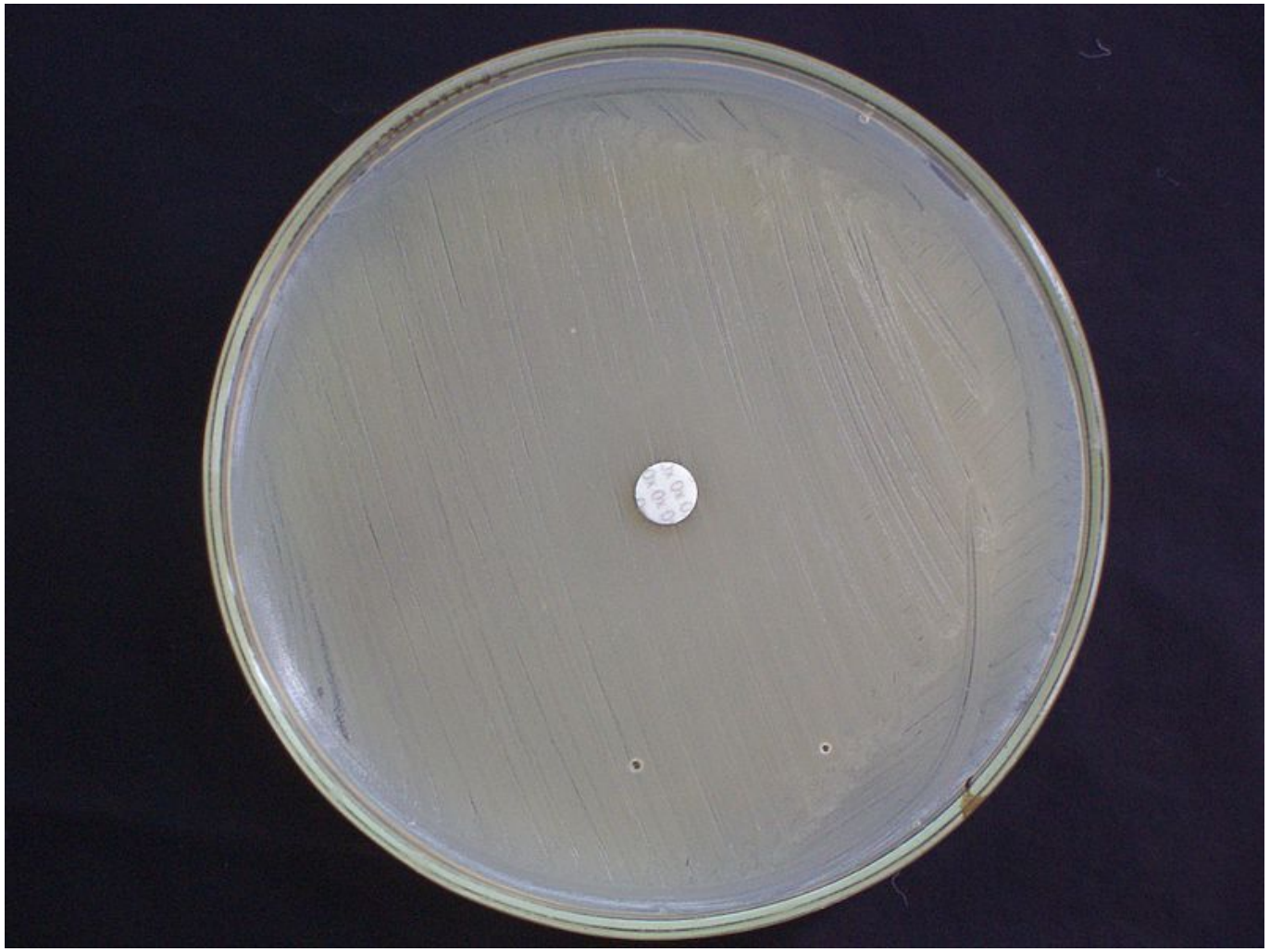
# 15 Years Later…

Antibiotics would be mass-produced for D-Day

# Would I Be Here without Antibiotics?

*Streptococcus pyogenes*

# The Rise of MRSA



Methicillin-resistant *Staphylococcus aureus*

# What are Antibiotics, Anyway?

**Antibiotic** □ "a substance that kills bacteria"

Occur naturally because of millions of years of evolutionary warfare

Produced by fungi (e.g., molds) and bacteria

Courtesy: Bios (Wikimedia)

# Antibiotics on the Molecular Level

We will study Tyrocidine B1, an antibiotic produced by *Bacillus Brevis*

Tyrocidine B1 is a "mini-protein" called a **peptide**: short string of amino acids

Valine Leucine Proline Phenylalanine Glutamine

**Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr**

**V K L F P W F N Q Y**

Lysine Phenylalanine Tryptophan Asparagine Tyrosine

# Questions

What makes antibiotics special as peptides?

How are antibiotics produced?

**How do we sequence antibiotics?**

`???-???-???-???-???-???-???-???-???-???`

# Questions

What makes antibiotics special as peptides?

How are antibiotics produced?

**How do we sequence antibiotics?**

**Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr**

# Outline

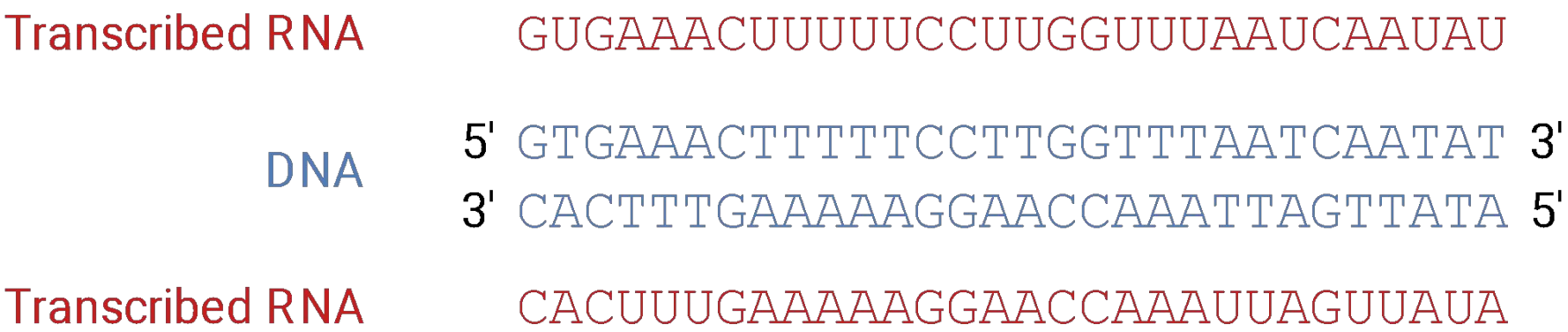
- The Discovery of Antibiotics
- **How Do Bacteria Make Antibiotics?**
- Sequencing Antibiotics by Shattering Them into Pieces
- A Brute Force Algorithm for Cyclopeptide Sequencing
- Cyclopeptide Sequencing with Branch-and-Bound
- Adapting Sequencing for Spectra with Errors
- From 20 to More than 100 Amino Acids
- The Spectral Convolution Saves the Day
- The Truth about Spectra

# How Are Proteins Produced?

DNA is **transcribed** into RNA

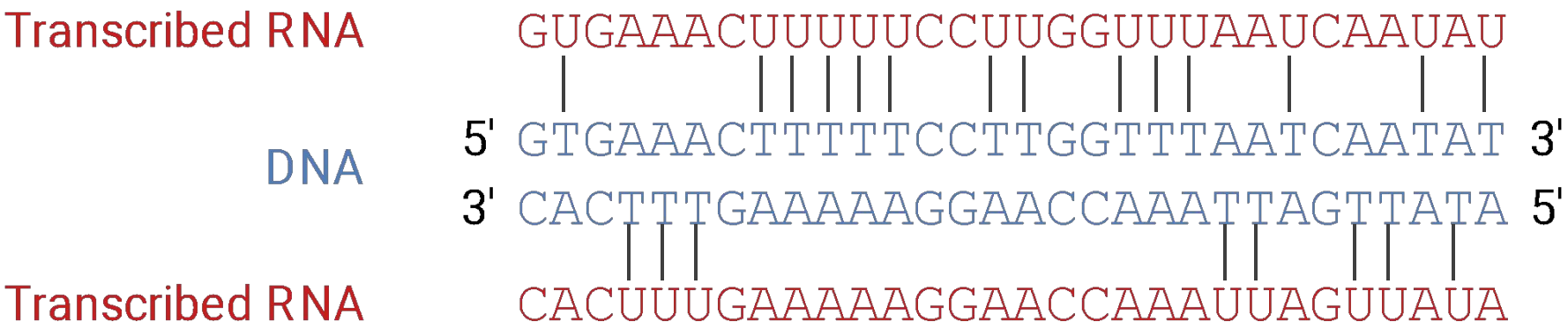
# How Are Proteins Produced?

DNA is **transcribed** into RNA



# How Are Proteins Produced?

DNA is **transcribed** into RNA




Replace T (thymine) with U (uracil)

# How Are Proteins Produced?

RNA is **translated** into peptides

# How Are Proteins Produced?

RNA is **translated** into proteins

**A**denine

**C**ytosine

4 nucleotides

**G**uanine

**U**racil

# How Are Proteins Produced?

RNA is **translated** into proteins

**A**denine

**C**ytosine

**G**uanine

**U**racil


| Amino acid | 3-letter code |
|---|---|
| Alanine | Ala |
| Cysteine | Cys |
| Aspartic acid | Asp |
| Glutamic acid | Glu |
| Phenylalanine | Phe |
| Glycine | Gly |
| Histidine | His |
| Isoleucine | Ile |
| Lysine | Lys |
| Leucine | Leu |
| Methionine | Met |
| Asparagine | Asn |
| Proline | Pro |
| Glutamine | Gln |
| Arginine | Arg |
| Serine | Ser |
| Threonine | Thr |
| Valine | Val |
| Tryptophan | Trp |
| Tyrosine | Tyr |

# How Are Proteins Produced?

RNA is **translated** into proteins

**A**denine

**C**ytosine

**G**uanine

**U**racil


**HOW?**

| Amino acid | 3-letter code |
|---|---|
| Alanine | Ala |
| Cysteine | Cys |
| Aspartic acid | Asp |
| Glutamic acid | Glu |
| Phenylalanine | Phe |
| Glycine | Gly |
| Histidine | His |
| Isoleucine | Ile |
| Lysine | Lys |
| Leucine | Leu |
| Methionine | Met |
| Asparagine | Asn |
| Proline | Pro |
| Glutamine | Gln |
| Arginine | Arg |
| Serine | Ser |
| Threonine | Thr |
| Valine | Val |
| Tryptophan | Trp |
| Tyrosine | Tyr |

# How Are Proteins Produced?

## Can We Translate 2 Nucleotides at a Time?

AA GA
AC GC
AG GG
AU GU
CA UA
CC UC
CG UG
CU UU


| Amino acid | 3-letter code |
|---|---|
| Alanine | Ala |
| Cysteine | Cys |
| Aspartic acid | Asp |
| Glutamic acid | Glu |
| Phenylalanine | Phe |
| Glycine | Gly |
| Histidine | His |
| Isoleucine | Ile |
| Lysine | Lys |
| Leucine | Leu |
| Methionine | Met |
| Asparagine | Asn |
| Proline | Pro |
| Glutamine | Gln |
| Arginine | Arg |
| Serine | Ser |
| Threonine | Thr |
| Valine | Val |
| Tryptophan | Trp |
| Tyrosine | Tyr |

# How Are Proteins Produced?

## Can We Translate 2 Nucleotides at a Time?

AA GA
AC GC
AG GG
AU GU
CA UA
CC UC
CG UG
CU UU


**NO!**

| Amino acid | 3-letter code |
|---|---|
| Alanine | Ala |
| Cysteine | Cys |
| Aspartic acid | Asp |
| Glutamic acid | Glu |
| Phenylalanine | Phe |
| Glycine | Gly |
| Histidine | His |
| Isoleucine | Ile |
| Lysine | Lys |
| Leucine | Leu |
| Methionine | Met |
| Asparagine | Asn |
| Proline | Pro |
| Glutamine | Gln |
| Arginine | Arg |
| Serine | Ser |
| Threonine | Thr |
| Valine | Val |
| Tryptophan | Trp |
| Tyrosine | Tyr |

# How Are Proteins Produced?

## Can We Translate 3 Nucleotides at a Time?

| | | | |
|---|---|---|---|
| AAA | CAA | GAA | UAA |
| AAC | CAC | GAC | UAC |
| AAG | CAG | GAG | UAG |
| AAU | CAU | GAU | UAU |
| ACA | CCA | GCA | UCA |
| ACC | CCC | GCC | UCC |
| ACG | CCG | GCG | UCG |
| ACU | CCU | GCU | UCU |
| AGA | CGA | GGA | UGA |
| AGC | CGC | GGC | UGC |
| AGG | CGG | GGG | UGG |
| AGU | CGU | GGU | UGU |
| AUA | CUA | GUA | UUA |
| AUC | CUC | GUC | UUC |
| AUG | CUG | GUG | UUG |
| AUU | CUU | GUU | UUU |

64 3-mers → 20 amino acids

| Amino acid | 3-letter code |
|---|---|
| Alanine | Ala |
| Cysteine | Cys |
| Aspartic acid | Asp |
| Glutamic acid | Glu |
| Phenylalanine | Phe |
| Glycine | Gly |
| Histidine | His |
| Isoleucine | Ile |
| Lysine | Lys |
| Leucine | Leu |
| Methionine | Met |
| Asparagine | Asn |
| Proline | Pro |
| Glutamine | Gln |
| Arginine | Arg |
| Serine | Ser |
| Threonine | Thr |
| Valine | Val |
| Tryptophan | Trp |
| Tyrosine | Tyr |

# How Are Proteins Produced?

## Can We Translate 3 Nucleotides at a Time?

| | | | |
|---|---|---|---|
| AAA | CAA | GAA | UAA |
| AAC | CAC | GAC | UAC |
| AAG | CAG | GAG | UAG |
| AAU | CAU | GAU | UAU |
| ACA | CCA | GCA | UCA |
| ACC | CCC | GCC | UCC |
| ACG | CCG | GCG | UCG |
| ACU | CCU | GCU | UCU |
| AGA | CGA | GGA | UGA |
| AGC | CGC | GGC | UGC |
| AGG | CGG | GGG | UGG |
| AGU | CGU | GGU | UGU |
| AUA | CUA | GUA | UUA |
| AUC | CUC | GUC | UUC |
| AUG | CUG | GUG | UUG |
| AUU | CUU | GUU | UUU |


**YES!**

| Amino acid | 3-letter code |
|---|---|
| Alanine | Ala |
| Cysteine | Cys |
| Aspartic acid | Asp |
| Glutamic acid | Glu |
| Phenylalanine | Phe |
| Glycine | Gly |
| Histidine | His |
| Isoleucine | Ile |
| Lysine | Lys |
| Leucine | Leu |
| Methionine | Met |
| Asparagine | Asn |
| Proline | Pro |
| Glutamine | Gln |
| Arginine | Arg |
| Serine | Ser |
| Threonine | Thr |
| Valine | Val |
| Tryptophan | Trp |
| Tyrosine | Tyr |

# How Are Proteins Produced?

**Codon:** A triplet (3-mer) of nucleotides

**Genetic Code:** assignment of codons to amino acids to make proteins

# How Are Proteins Produced?

**Codon:** A triplet (3-mer) of nucleotides

**Genetic Code:** assignment of codons to amino acids to make proteins

# How Are Proteins Produced?

**Codon:** A triplet (3-mer) of nucleotides

**Genetic Code:** assignment of codons to amino acids to make proteins

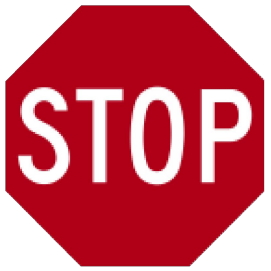
# How Are Proteins Produced?

**Codon:** A triplet (3-mer) of nucleotides

**Genetic Code:** assignment of codons to amino acids to make proteins

**Stop Codons**

UAA
UAG
UGA



# Central Dogma of Molecular Biology

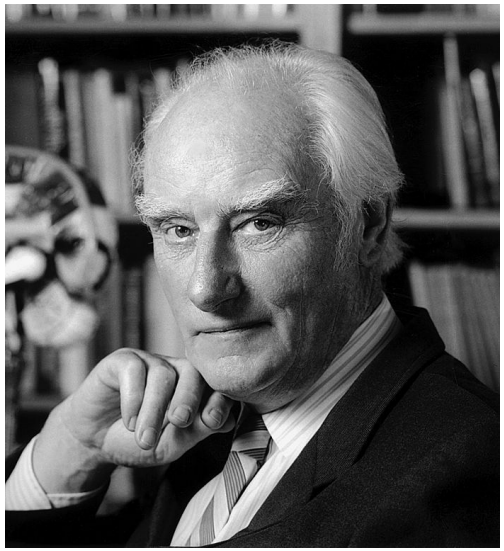
DNA

# Central Dogma of Molecular Biology

# Central Dogma of Molecular Biology

# Central Dogma of Molecular Biology

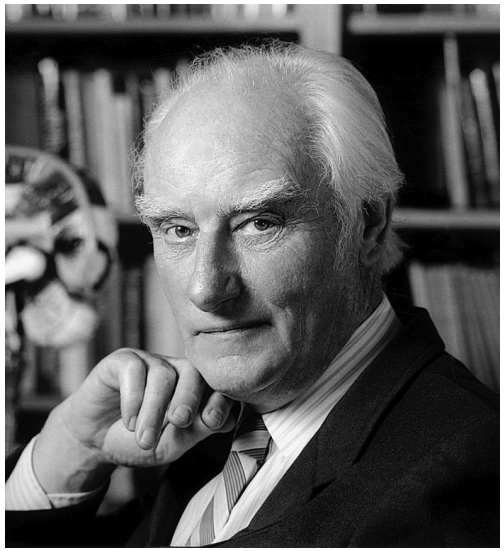
DNA —Transcription→ RNA —Translation→ Protein



Francis Crick

# Central **Dogma** of Molecular Biology

DNA —Transcription→ RNA —Translation→ Protein



Francis Crick

*"As it turned out, the use of the word dogma caused almost more trouble than it was worth. Many years later Jacques Monod pointed out to me that I did not appear to understand the correct use of the word dogma, which is* ***a belief that cannot be doubted****."*—Francis Crick

# Searching for Tyrocidine B1


**Goal:** Find a 30-mer in the *Bacillus brevis* genome that transcribes and translates into Tyrocidine B1 (peptide of length 10).

# Searching for Tyrocidine B1

*Thousands* of different 30-mers could translate into Tyrocidine B1.

# Searching for Tyrocidine B1

*Thousands* of different 30-mers could translate into Tyrocidine B1.

**GTTAAATTATTTCCTTGGTTTAATCAATAT**

**ValLysLeuPheProTrpPheAsnGlnTyr**

# Searching for Tyrocidine B1

*Thousands* of different 30-mers could translate into Tyrocidine B1.

GTTAAATTATTTCCTTGGTTTAATCAATAT

**ValLysLeuPheProTrpPheAsnGlnTyr**

**GTCAAGCTTTTCCCCTGGTTCAACCAGTAC**

**ValLysLeuPheProTrpPheAsnGlnTyr**

# Searching for Tyrocidine B1

*Thousands* of different 30-mers could translate into Tyrocidine B1.

GTTAAATTATTTCCTTGGTTTAATCAATAT

**ValLysLeuPheProTrpPheAsnGlnTyr**

GTCAAGCTTTTCCCCTGGTTCAACCAGTAC

**ValLysLeuPheProTrpPheAsnGlnTyr**

**GTAAAACTATTTCCGTGGTTCAATCAATAT**

**ValLysLeuPheProTrpPheAsnGlnTyr**

# Searching for Tyrocidine B1

*Thousands* of different 30-mers could translate into Tyrocidine B1.

**GTTAAATTATTTCCTTGGTTTAATCAATAT**

**GTCAAGCTTTTCCCCTGGTTCAACCAGTAC**

**GTAAAACTATTTCCGTGGTTCAATCAATAT**

# Searching for Tyrocidine B1

*Thousands* of different 30-mers could translate into Tyrocidine B1.

GTTAAATTATTTCCTTGGTTTAATCAATAT

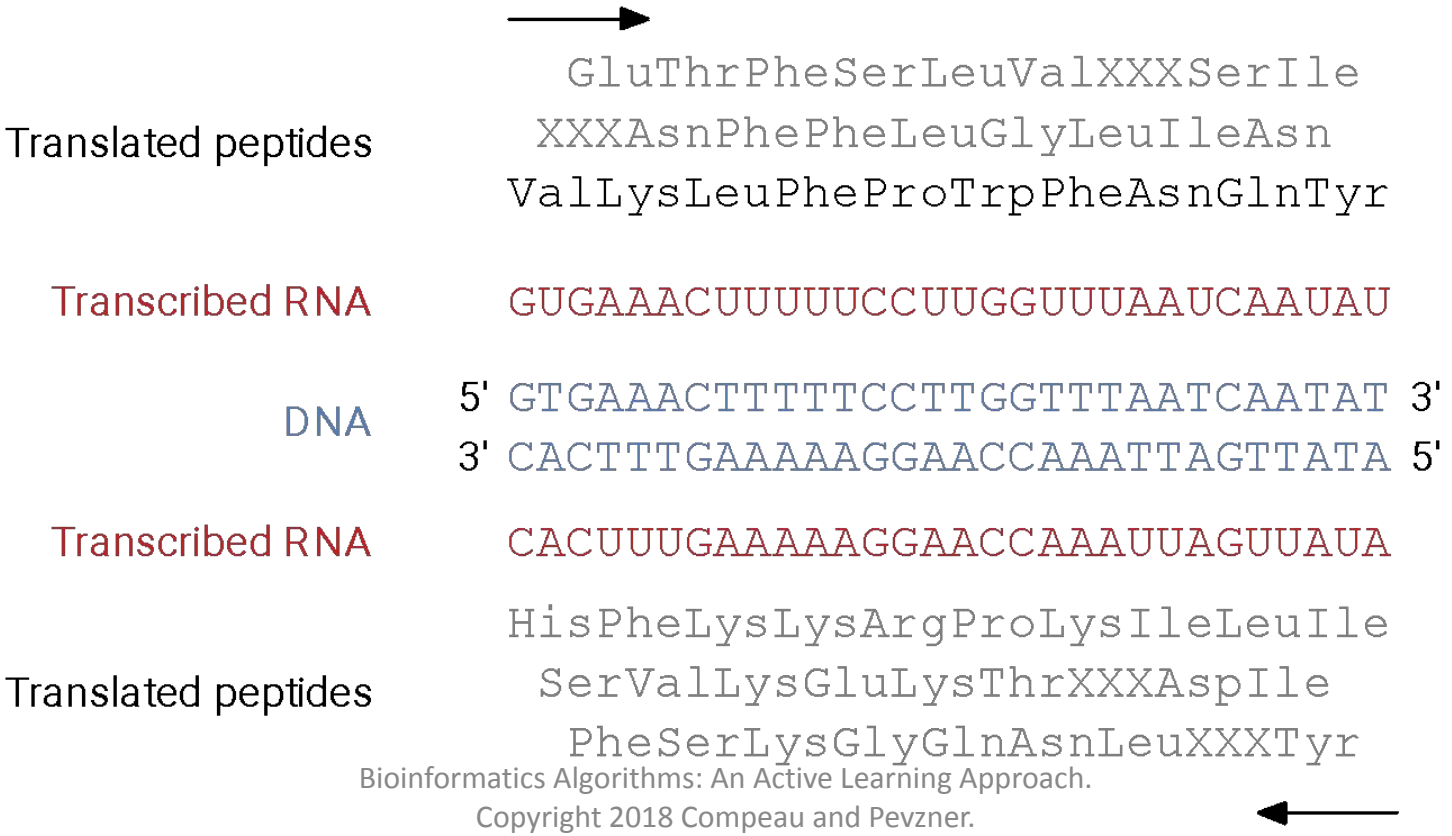
GTCAAGCTTTTCCCCTGGTTCAACCAGTAC

GTAAAACTATTTCCGTGGTTCAATCAATAT

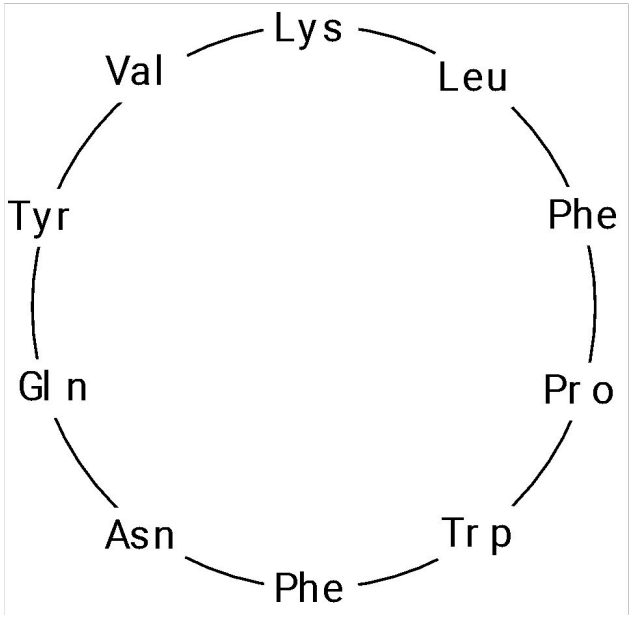
And they are not very similar…

# Searching for Tyrocidine B1

Translation can start anywhere in the genome;
6 different **reading frames**



# Tyrocidine B1 is **Cyclic**




*Ten different* linear representations:

**Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr**

Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr-Val

…

Tyr-Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln

# Searching for Tyrocidine B1

How many 30-mers in the *Bacillus brevis* genome encode a **linear representation** of Tyrocidine B1?

# Searching for Tyrocidine B1

How many 30-mers in the *Bacillus brevis* genome encode a **linear representation** of Tyrocidine B1?

...processing...

# Searching for Tyrocidine B1

How many 30-mers in the *Bacillus brevis* genome encode a **linear representation** of Tyrocidine B1?

...processing...

...processing...

# Searching for Tyrocidine B1

How many 30-mers in the *Bacillus brevis* genome encode a **linear representation** of Tyrocidine B1?

...processing...
...processing...
...processing...

# Searching for Tyrocidine B1

How many 30-mers in the *Bacillus brevis* genome encode a **linear representation** of Tyrocidine B1?

...processing...
...processing...
...processing...
...processing...

# Searching for Tyrocidine B1

How many 30-mers in the *Bacillus brevis* genome encode a **linear representation** of Tyrocidine B1?

...processing...
...processing...
...processing...
...processing...
...processing...

# Searching for Tyrocidine B1

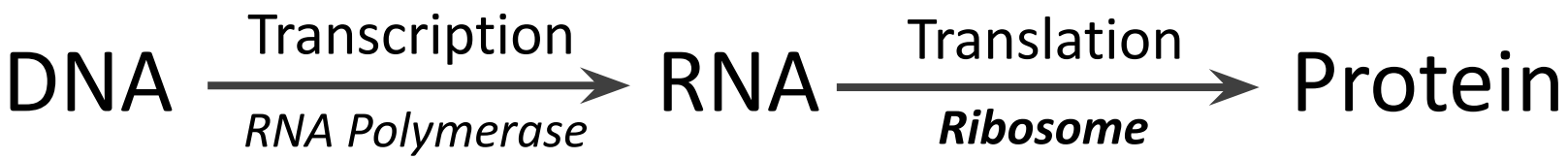
How many 30-mers in the *Bacillus brevis* genome encode a **linear representation** of Tyrocidine B1?

...processing...
...processing...
...processing...
...processing...
...processing...

**NONE!?**

# Dodging the Dogma

# Dodging the Dogma



# Dodging the Dogma

DNA —Transcription (*RNA Polymerase*)→ RNA —Translation (***Ribosome***)→ Protein

1963: Edward Tatum inhibits the ribosome in *Bacillus brevis.*

Edward Tatum

# Dodging the Dogma

DNA —Transcription (*RNA Polymerase*)→ RNA —Translation (***Ribosome***)→ Protein

1963: Edward Tatum inhibits the ribosome in *Bacillus brevis.*

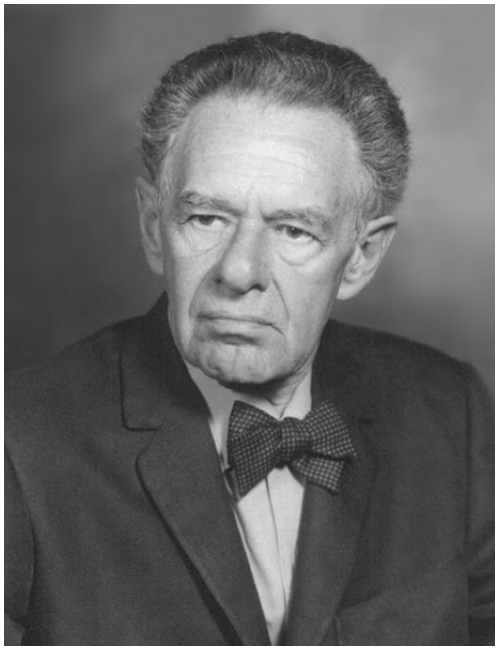
Production of some peptides, including tyrocidines, continues!

Edward Tatum

# Dodging the Dogma

DNA —Transcription (*RNA Polymerase*)→ RNA —Translation (***Ribosome***)→ Protein

1969: Lipmann shows tyrocidines are **non-ribosomal peptides** (**NRPs**).



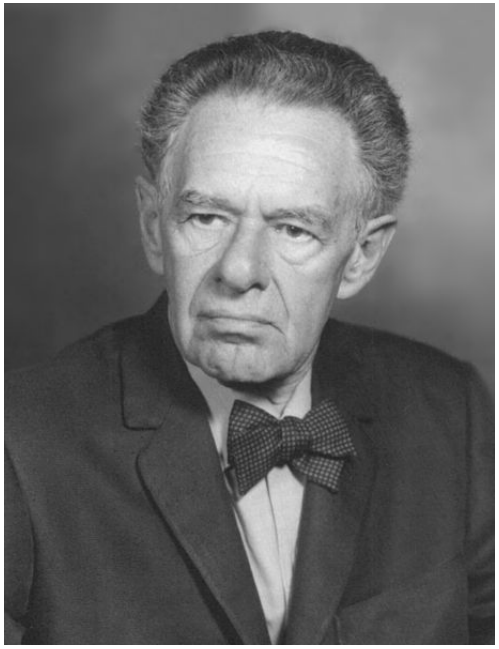
Fritz Lipmann

# Dodging the Dogma

DNA —Transcription (*RNA Polymerase*)→ RNA —Translation (***Ribosome***)→ Protein

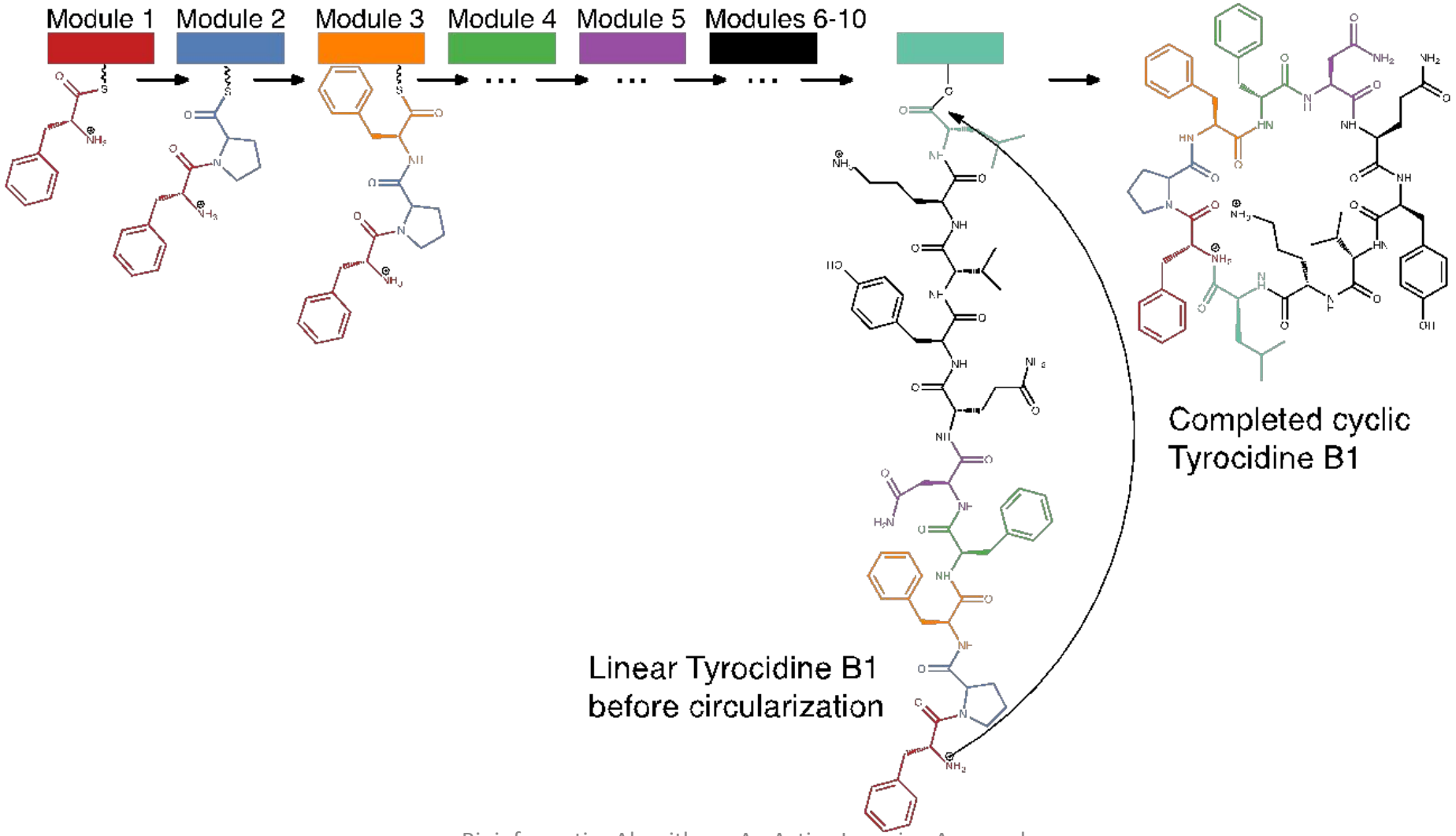
1969: Lipmann shows tyrocidines are **non-ribosomal peptides** (**NRPs**).

NRPs are synthesized not by the ribosome but by **NRP synthetase**.



Fritz Lipmann

# NRP Synthetase Adds One Amino Acid at a Time



# Outline

- The Discovery of Antibiotics
- How Do Bacteria Make Antibiotics?
- **Sequencing Antibiotics by Shattering Them into Pieces**
- A Brute Force Algorithm for Cyclopeptide Sequencing
- Cyclopeptide Sequencing with Branch-and-Bound
- Adapting Sequencing for Spectra with Errors
- From 20 to More than 100 Amino Acids
- The Spectral Convolution Saves the Day
- The Truth about Spectra

# The Mass Spectrometer

Finding an NRP “hidden” in the *Bacillus brevis* genome will not work for sequencing NRPs.

# The Mass Spectrometer

Finding an NRP "hidden" in the *Bacillus brevis* genome will not work for sequencing NRPs.

**Mass spectrometer:**
"expensive molecular scale"

# How Do We Measure Molecular Weight?

**1 Dalton** (Da) ≈ mass of proton/neutron

Mass of molecule ≈ sum of protons/neutrons

# How Do We Measure Molecular Weight?

**1 Dalton** (Da) ≈ mass of proton/neutron

Mass of molecule ≈ sum of protons/neutrons

Mass of Glycine ($C_2H_3ON$) ≈ **12** · **2**

# How Do We Measure Molecular Weight?

**1 Dalton** (Da) ≈ mass of proton/neutron

Mass of molecule ≈ sum of protons/neutrons

Mass of Glycine ($C_2H_3ON$) ≈ $12 \cdot 2 + 1 \cdot 3$

# How Do We Measure Molecular Weight?

**1 Dalton** (Da) ≈ mass of proton/neutron

Mass of molecule ≈ sum of protons/neutrons

Mass of Glycine ($C_2H_3$**O**N) ≈ $12 \cdot 2 + 1 \cdot 3 +$ **16**

# How Do We Measure Molecular Weight?

**1 Dalton** (Da) ≈ mass of proton/neutron

Mass of molecule ≈ sum of protons/neutrons

Mass of Glycine ($C_2H_3ON$) ≈ $12 \cdot 2 + 1 \cdot 3 + 16 + \mathbf{14}$

# How Do We Measure Molecular Weight?

**1 Dalton** (Da) ≈ mass of proton/neutron

Mass of molecule ≈ sum of protons/neutrons

Mass of Glycine ($C_2H_3ON$) ≈ $12 \cdot 2 + 1 \cdot 3 + 16 + 14$

≈ 57 Da

# How Do We Measure Molecular Weight?

**1 Dalton** (Da) ≈ mass of proton/neutron

Mass of molecule ≈ sum of protons/neutrons

Mass of Glycine ($C_2H_3ON$) ≈ $12 \cdot 2 + 1 \cdot 3 + 16 + 14$

≈ 57 Da

Actual mass: 57.02 Da **Integer mass:** 57

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | K | Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

# Integer Mass Table

Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | K | Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

What is the mass of Tyrocidine B1? (VKLFPWFNQY)

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | **V** | T | C | I | L | N | D | K | Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | **99** | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (**V**KLFPWFNQY)

**V**

**99**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | **K** | Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | **128** | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (V**K**LFPWFNQY)

V **K**

99+**128**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | **L** | N | D | K | Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | **113** | 114 | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (VK**L**FPWFNQY)

V K **L**

99+128+**113**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | K | Q | E | M | H | **F** | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | **147** | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (VKL**F**PWFNQY)

V K L **F**

99+128+113+**147**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | **P** | V | T | C | I | L | N | D | K | Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | **97** | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (VKLF**P**WFNQY)

V K L F **P**

99+128+113+147+**97**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | K | Q | E | M | H | F | R | Y | **W** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | **186** |

## What is the mass of Tyrocidine B1? (VKLFP**W**FNQY)

V K L F P **W**

99+128+113+147+97+**186**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | K | Q | E | M | H | **F** | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | **147** | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (VKLFPW**F**NQY)

V K L F P W **F**

99+128+113+147+97+186+**147**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | **N** | D | K | Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | **114** | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (VKLFPWF**N**QY)

| V | K | L | F | P | W | F | **N** |
|---|---|---|---|---|---|---|---|
| 99 | 128 | 113 | 147 | 97 | 186 | 147 | **114** |

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | K | **Q** | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | **128** | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (VKLFPWFN**Q**Y)

V K L F P W F N **Q**

99+128+113+147+97+186+147+114+**128**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | K | Q | E | M | H | F | R | **Y** | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | **163** | 186 |

## What is the mass of Tyrocidine B1? (VKLFPWFNQ**Y**)

V K L F P W F N Q **Y**

99+128+113+147+97+186+147+114+128+**163**

# Integer Mass Table

## Contains masses of all 20 amino acids

| G | A | S | P | V | T | C | I | L | N | D | K | Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 113 | 114 | 115 | 128 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

## What is the mass of Tyrocidine B1? (VKLFPWFNQY)

V K L F P W F N Q Y

99+128+113+147+97+186+147+114+128+163 = **1322**

# Integer Mass Table

Note that two amino acid pairs have equal mass:

| G | A | S | P | V | T | C | **I** | **L** | N | D | **K** | **Q** | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | **113** | **113** | 114 | 115 | **128** | **128** | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

# Integer Mass Table

Note that two amino acid pairs have equal mass:

| G | A | S | P | V | T | C | **I/L** | N | D | **K/Q** | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | **113** | 114 | 115 | **128** | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

# Integer Mass Table

Note that two amino acid pairs have equal mass:

| G | A | S | P | V | T | C | I/L | N | D | K/Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **57** | **71** | **87** | **97** | **99** | **101** | **103** | **113** | **114** | **115** | **128** | **129** | **131** | **137** | **147** | **156** | **163** | **186** |

We move from 20 amino acids □ 18 integer masses

# How the Mass Spectrometer Works

**NQEL**

**NQEL**

**NQEL**

**NQEL**

**NQEL**

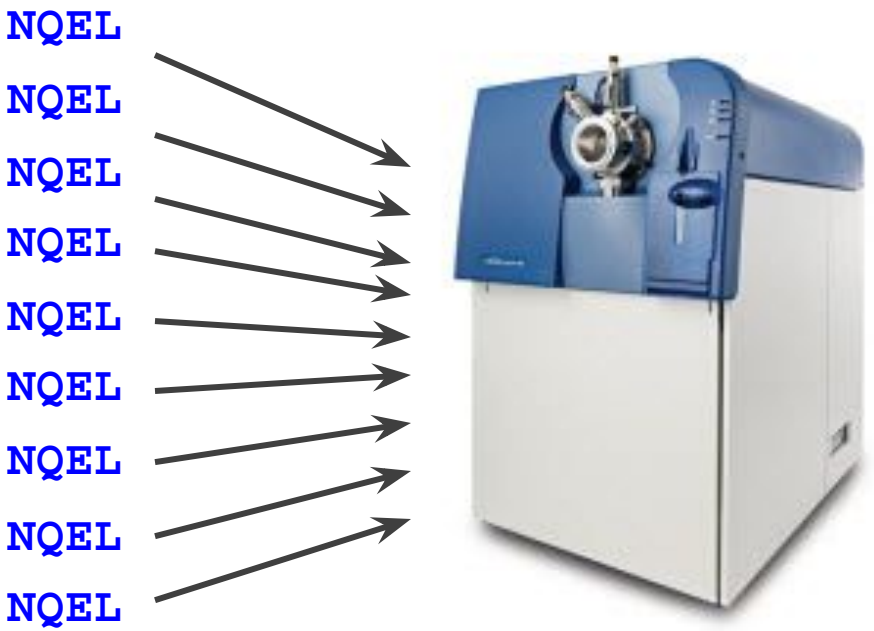
**NQEL**

**NQEL**

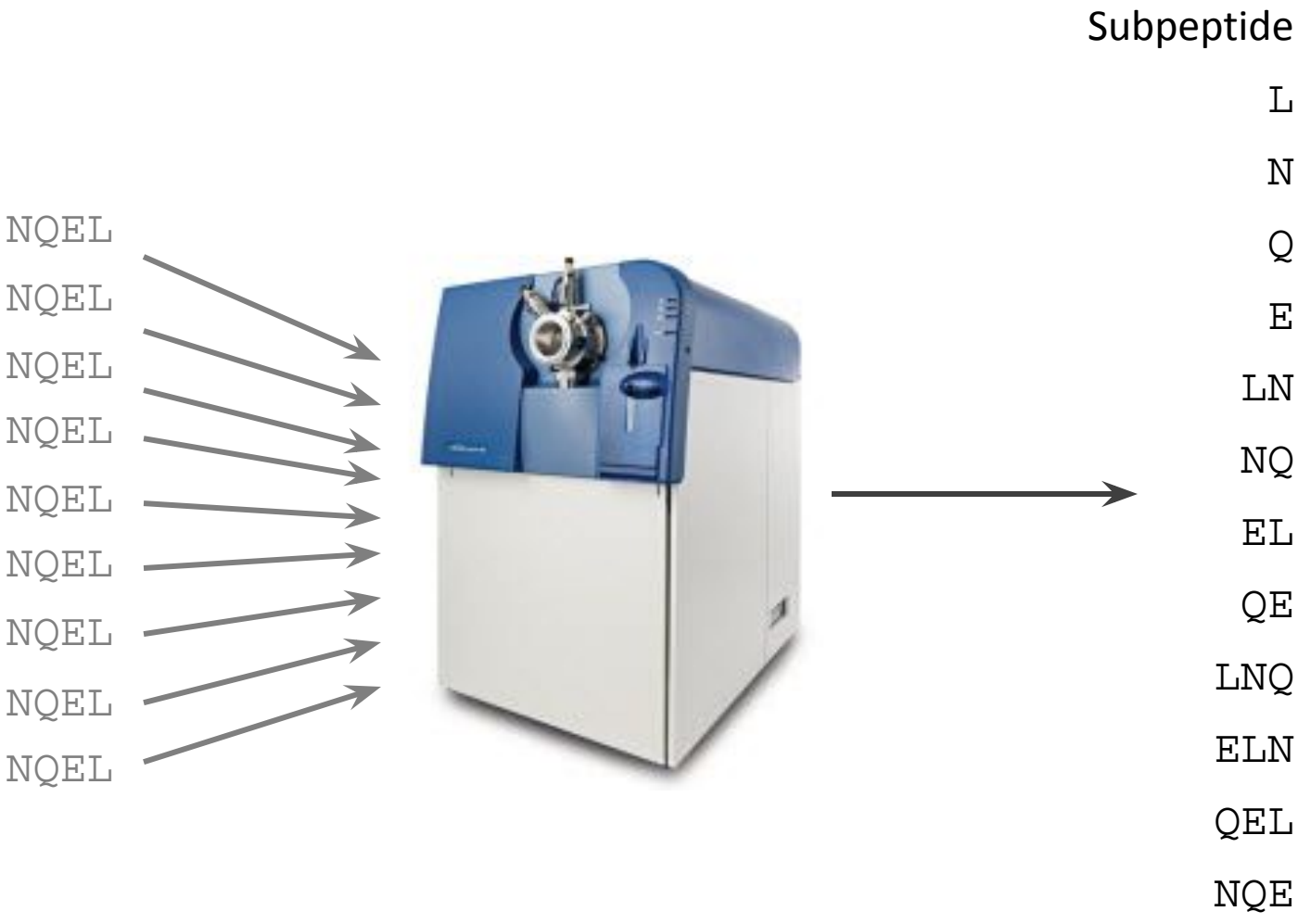
**NQEL**

**NQEL**

# How the Mass Spectrometer Works



# How the Mass Spectrometer Works



# How the Mass Spectrometer Works

NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL

| Subpeptide | Mass |
|---|---|
| L | 113 |
| N | 114 |
| Q | 128 |
| E | 129 |
| LN | 227 |
| NQ | 242 |
| EL | 242 |
| QE | 257 |
| LNQ | 355 |
| ELN | 356 |
| QEL | 370 |
| NQE | 371 |

# How the Mass Spectrometer Works

NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL

| Subpeptide | Mass |
|---|---|
| L | 113 |
| N | 114 |
| Q | 128 |
| E | 129 |
| LN | 227 |
| NQ | 242 |
| EL | 242 |
| QE | 257 |
| LNQ | 355 |
| ELN | 356 |
| QEL | 370 |
| NQE | 371 |

# How the Mass Spectrometer Works

NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL
NQEL

| Subpeptide | Mass |
|---|---|
| L | 113 |
| N | 114 |
| Q | 128 |
| E | 129 |
| LN | 227 |
| NQ | **242** |
| EL | **242** |
| QE | 257 |
| LNQ | 355 |
| ELN | 356 |
| QEL | 370 |
| NQE | 371 |

# Toward a Computational Problem

**Theoretical spectrum**: mass of *every possible* subpeptide, plus 0 and the mass of the peptide.

| Subpeptide | Mass |
|---|---|
| L | **113** |
| N | **114** |
| Q | **128** |
| E | **129** |
| LN | **227** |
| NQ | **242** |
| EL | **242** |
| QE | **257** |
| LNQ | **355** |
| ELN | **356** |
| QEL | **370** |
| NQE | **371** |
| NQEL | **484** |
| "" | **0** |

# Toward a Computational Problem

**Theoretical spectrum**: mass of *every possible* subpeptide, plus 0 and the mass of the peptide.

**Peptide**

**NQEL**

**Spectrum**

| Subpeptide | Mass |
|---|---|
| L | **113** |
| N | **114** |
| Q | **128** |
| E | **129** |
| LN | **227** |
| NQ | **242** |
| EL | **242** |
| QE | **257** |
| LNQ | **355** |
| ELN | **356** |
| QEL | **370** |
| NQE | **371** |
| NQEL | **484** |
| “” | **0** |

# Toward a Computational Problem

**Theoretical spectrum**: mass of *every possible* subpeptide, plus 0 and the mass of the peptide.

**Peptide**
**NQEL**

EASY →

**Spectrum**

| Subpeptide | Mass |
|---|---|
| L | **113** |
| N | **114** |
| Q | **128** |
| E | **129** |
| LN | **227** |
| NQ | **242** |
| EL | **242** |
| QE | **257** |
| LNQ | **355** |
| ELN | **356** |
| QEL | **370** |
| NQE | **371** |
| NQEL | **484** |
| "" | **0** |

# Toward a Computational Problem

**Theoretical spectrum**: mass of *every possible* subpeptide, plus 0 and the mass of the peptide.

| Subpeptide | Mass |
|---|---|
| L | **113** |
| N | **114** |
| Q | **128** |
| E | **129** |
| LN | **227** |
| NQ | **242** |
| EL | **242** |
| QE | **257** |
| LNQ | **355** |
| ELN | **356** |
| QEL | **370** |
| NQE | **371** |
| NQEL | **484** |
| "" | **0** |

**Peptide** ← **Spectrum**

????

# Toward a Computational Problem

**Theoretical spectrum**: mass of *every possible* subpeptide, plus 0 and the mass of the peptide.


| Subpeptide | Mass |
|---|---|
| L | 113 |
| N | 114 |
| Q | 128 |
| E | 129 |
| LN | 227 |
| NQ | 242 |
| EL | 242 |
| QE | 257 |
| LNQ | 355 |
| ELN | 356 |
| QEL | 370 |
| NQE | 371 |
| NQEL | 484 |
| “” | 0 |

# Toward a Computational Problem

**Theoretical spectrum**: mass of *every possible* subpeptide, plus 0 and the mass of the peptide.


**Cyclopeptide Sequencing Problem**:
*Reconstruct a cyclic peptide from its theoretical spectrum.*

113
114
128
129
227
242
242
257
355
356
370
371
484
0

# Outline

- The Discovery of Antibiotics
- How Do Bacteria Make Antibiotics?
- Sequencing Antibiotics by Shattering Them into Pieces
- **A Brute Force Algorithm for Cyclopeptide Sequencing**
- Cyclopeptide Sequencing with Branch-and-Bound
- Adapting Sequencing for Spectra with Errors
- From 20 to More than 100 Amino Acids
- The Spectral Convolution Saves the Day
- The Truth about Spectra

# Brute Force Cyclopeptide Sequencing

The mass of the entire peptide is usually known.

**Algorithm:**

1. Generate all **peptides** with given mass (1322).
2. Form their theoretical spectra.
3. Look for matches with the given **spectrum**

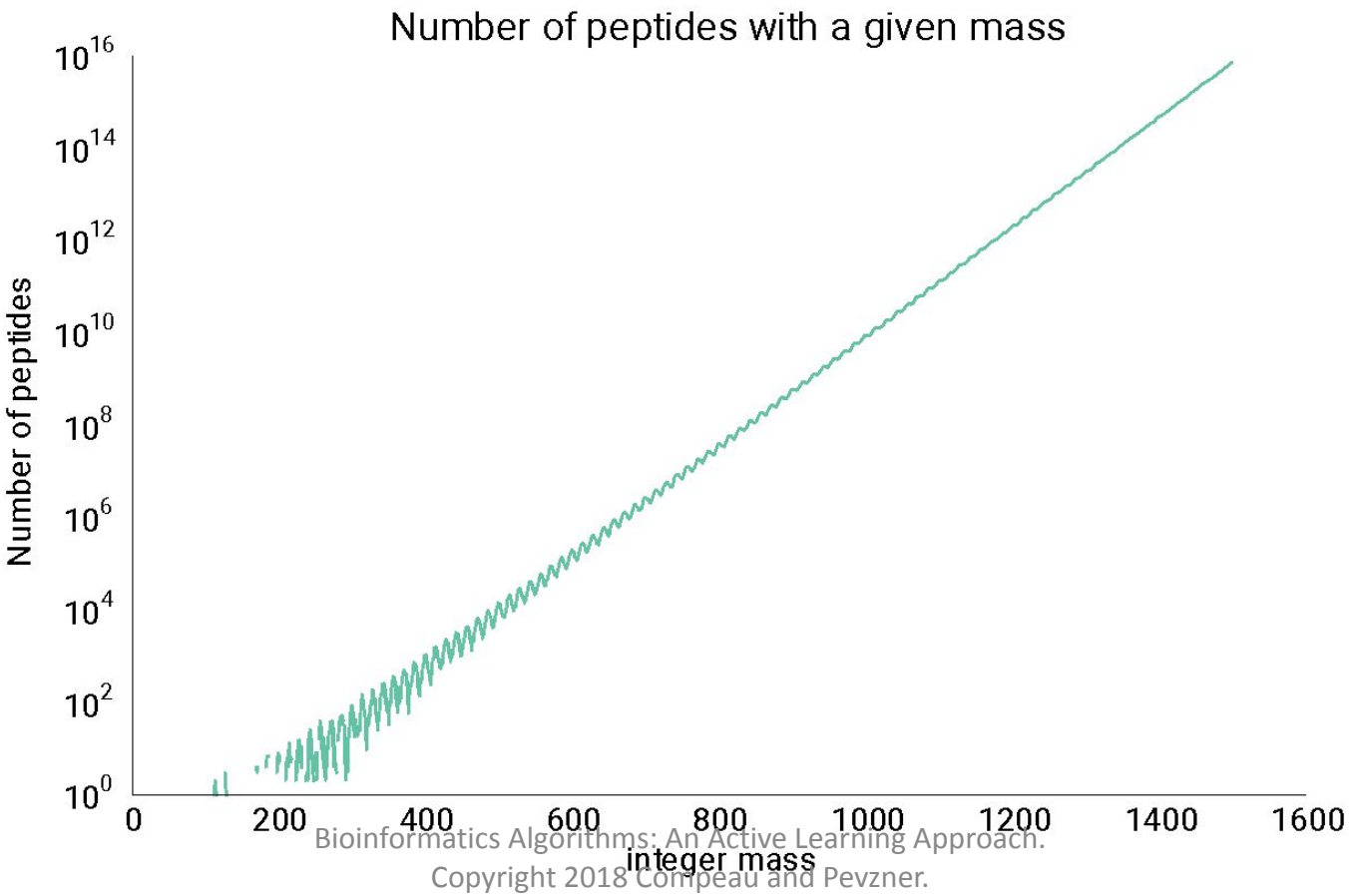
**Brute Force Algorithm**: "Try all" candidate solutions.

# Brute Force Cyclopeptide Sequencing

How many peptides have integer mass = 1322?

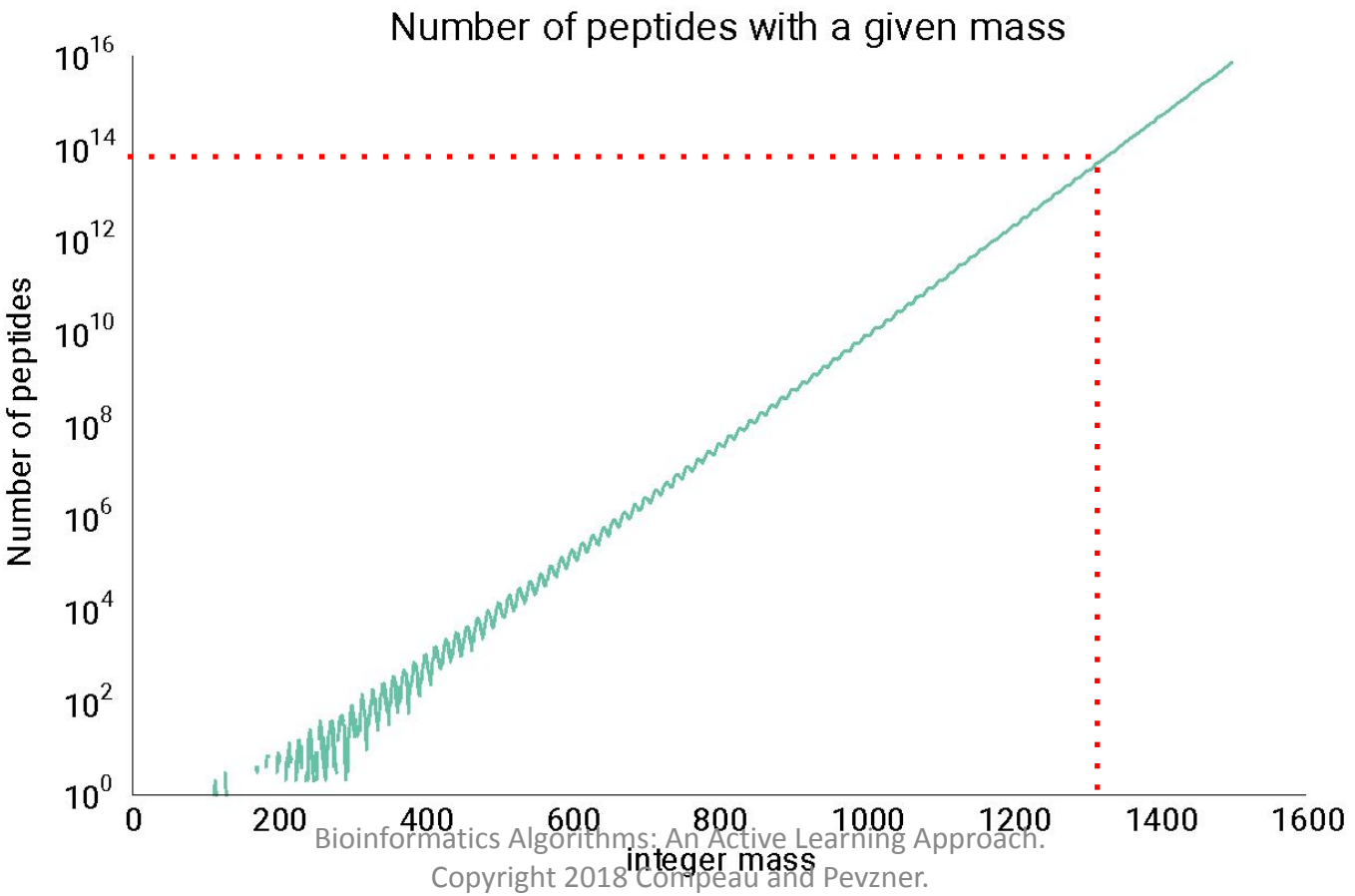
# Brute Force Cyclopeptide Sequencing

How many peptides have integer mass = 1322?



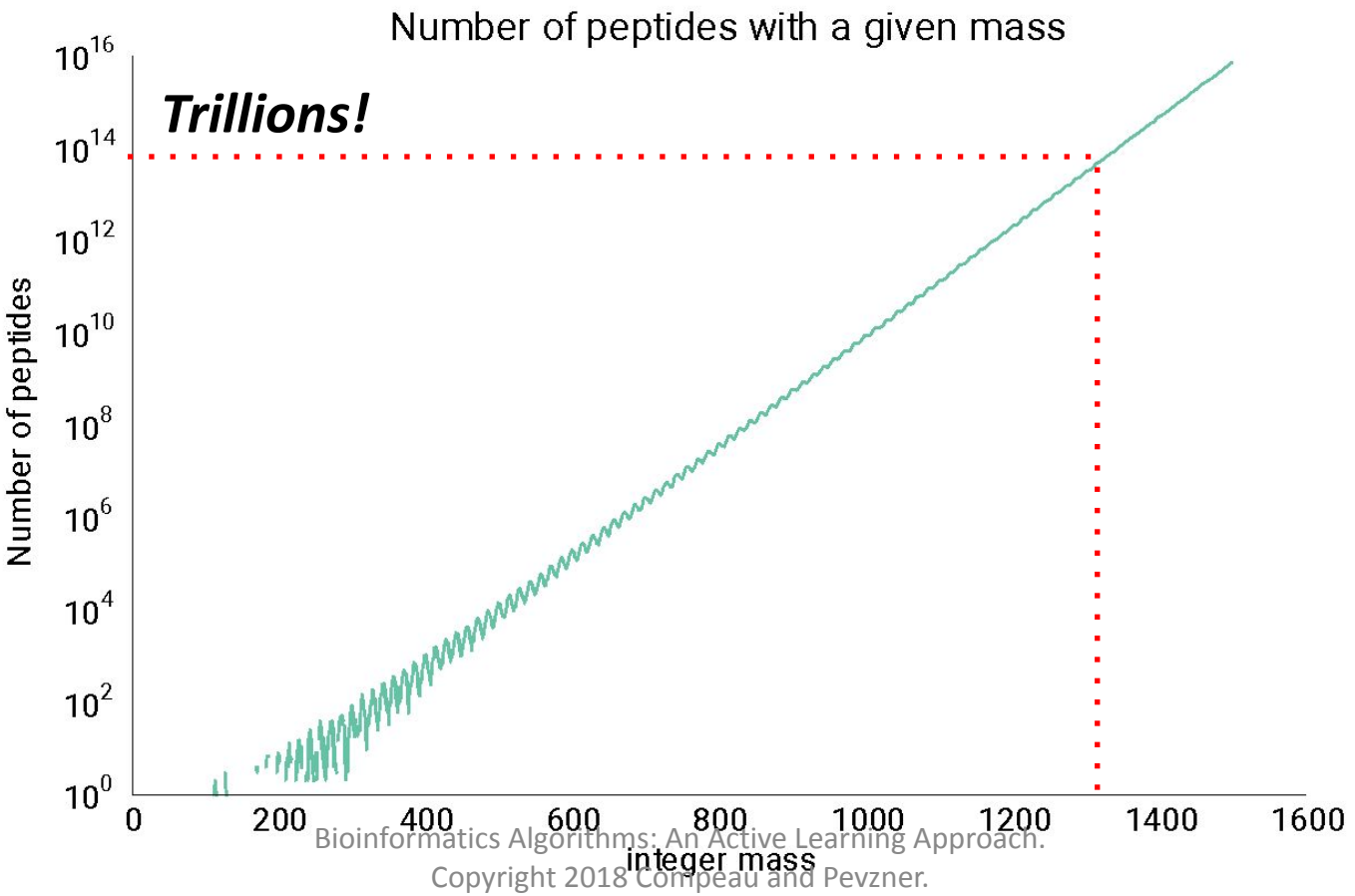
# Brute Force Cyclopeptide Sequencing

How many peptides have integer mass = 1322?



# Brute Force Cyclopeptide Sequencing

How many peptides have integer mass = 1322?



# Why Is Brute Force So Bad?

**114-128-129-113**
N Q E L

**101-131-115-137**
T M D H

# Why Is Brute Force So Bad?

**114-128-129-113** **101-131-115-137**

N Q E L T M D H

Total Mass: **484**

# Why Is Brute Force So Bad?

**114-128-129-113**

N Q E L

Total Mass: **484**

**101-131-115-137**

T M D H

Total Mass: **484**

# Why Is Brute Force So Bad?

**114-128-129-113**
N Q E L

Total Mass: **484**

**101-131-115-137**
T M D H

Total Mass: **484**

These peptides are *completely different*.

# Why Is Brute Force So Bad?

**114-128-129-113**
N Q E L

Total Mass: **484**

101-131-115-137
T M D H

Total Mass: **484**

These peptides are *completely different*.

**How can we exclude the incorrect peptide?**

# Why Is Brute Force So Bad?

Spectrum of **TMDH**

| | |
|---|---|
| "" | 0 |
| T | 101 |
| D | 115 |
| M | 131 |
| H | 137 |
| TM | 232 |
| HT | 238 |
| MD | 246 |
| DH | 252 |
| TMD | 347 |
| DHT | 353 |
| HTM | 369 |
| MDH | 383 |
| **TMDH** | 484 |

Spectrum of **NQEL**

| | |
|---|---|
| "" | 0 |
| L | 113 |
| N | 114 |
| Q | 128 |
| E | 129 |
| LN | 227 |
| NQ | 242 |
| EL | 242 |
| QE | 257 |
| LNQ | 355 |
| ELN | 356 |
| QEL | 370 |
| NQE | 371 |
| **NQEL** | 484 |

# Why Is Brute Force So Bad?

Spectrum of **TMDH**

| | |
|---|---|
| "" | 0 |
| T | 101 |
| D | 115 |
| M | 131 |
| H | 137 |
| TM | 232 |
| HT | 238 |
| MD | 246 |
| DH | 252 |
| TMD | 347 |
| DHT | 353 |
| HTM | 369 |
| MDH | 383 |
| **TMDH** | 484 |

Their spectra completely disagree!

Spectrum of **NQEL**

| | |
|---|---|
| "" | 0 |
| L | 113 |
| N | 114 |
| Q | 128 |
| E | 129 |
| LN | 227 |
| NQ | 242 |
| EL | 242 |
| QE | 257 |
| LNQ | 355 |
| ELN | 356 |
| QEL | 370 |
| NQE | 371 |
| **NQEL** | 484 |

# Why Is Brute Force So Bad?

Spectrum of **TMDH**

| | |
|---|---|
| "" | 0 |
| T | 101 |
| D | 115 |
| M | 131 |
| H | 137 |
| TM | 232 |
| HT | 238 |
| MD | 246 |
| DH | 252 |
| TMD | 347 |
| DHT | 353 |
| HTM | 369 |
| MDH | 383 |
| **TMDH** | 484 |

Their spectra completely disagree!

*How can we use this?*

Spectrum of **NQEL**

| | |
|---|---|
| "" | 0 |
| L | 113 |
| N | 114 |
| Q | 128 |
| E | 129 |
| LN | 227 |
| NQ | 242 |
| EL | 242 |
| QE | 257 |
| LNQ | 355 |
| ELN | 356 |
| QEL | 370 |
| NQE | 371 |
| **NQEL** | 484 |

# A New Idea

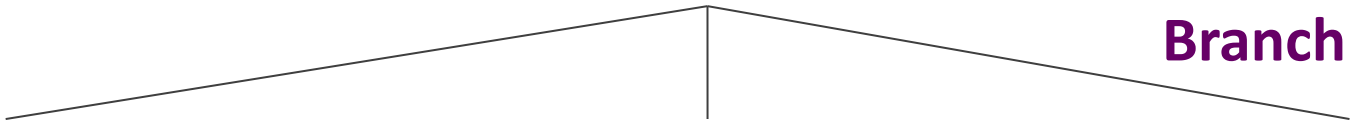
**Idea:** Let's slowly build up candidate solutions from smaller *linear* peptides.

We need to restrict the total number of linear peptides that we consider.

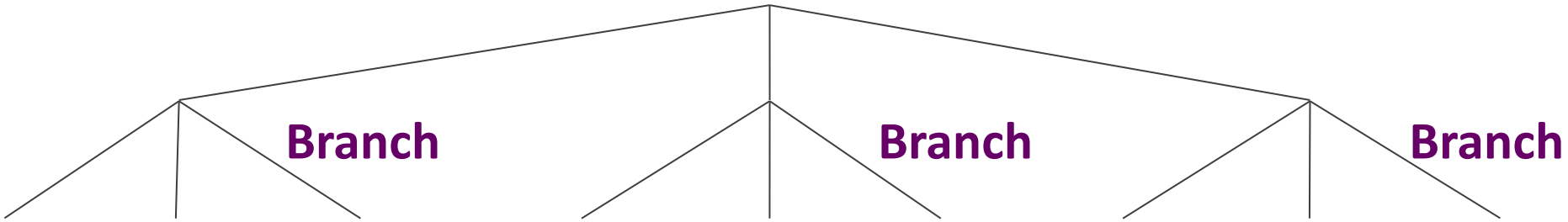
# Outline

- The Discovery of Antibiotics
- How Do Bacteria Make Antibiotics?
- Sequencing Antibiotics by Shattering Them into Pieces
- A Brute Force Algorithm for Cyclopeptide Sequencing
- **Cyclopeptide Sequencing with Branch-and-Bound**
- Adapting Sequencing for Spectra with Errors
- From 20 to More than 100 Amino Acids
- The Spectral Convolution Saves the Day
- The Truth about Spectra

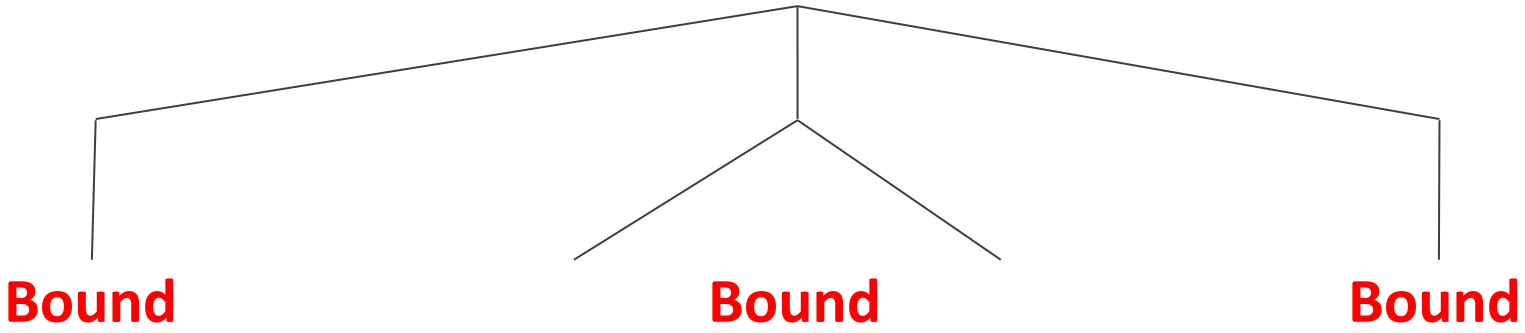
# **Branch**-and-**Bound** Algorithms



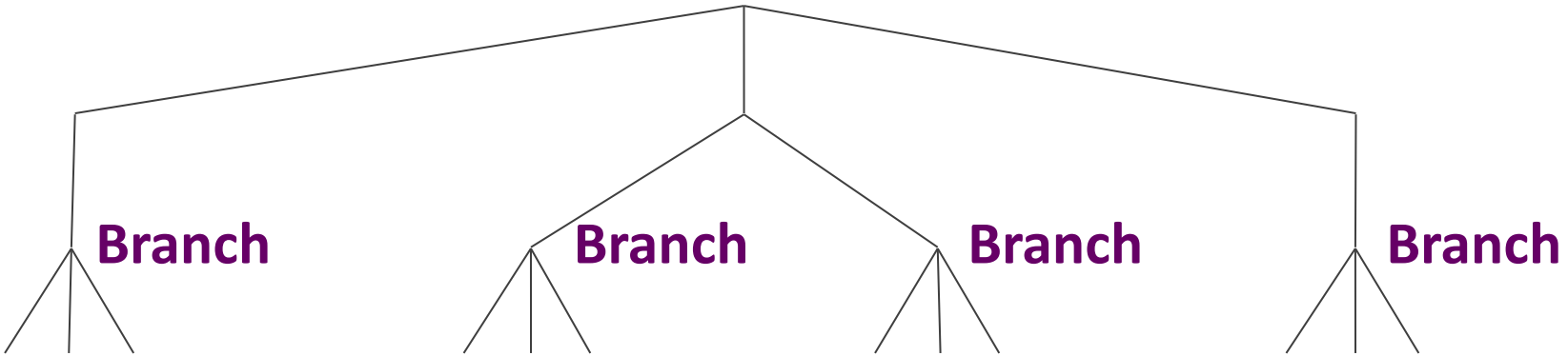
# **Branch**-and-**Bound** Algorithms



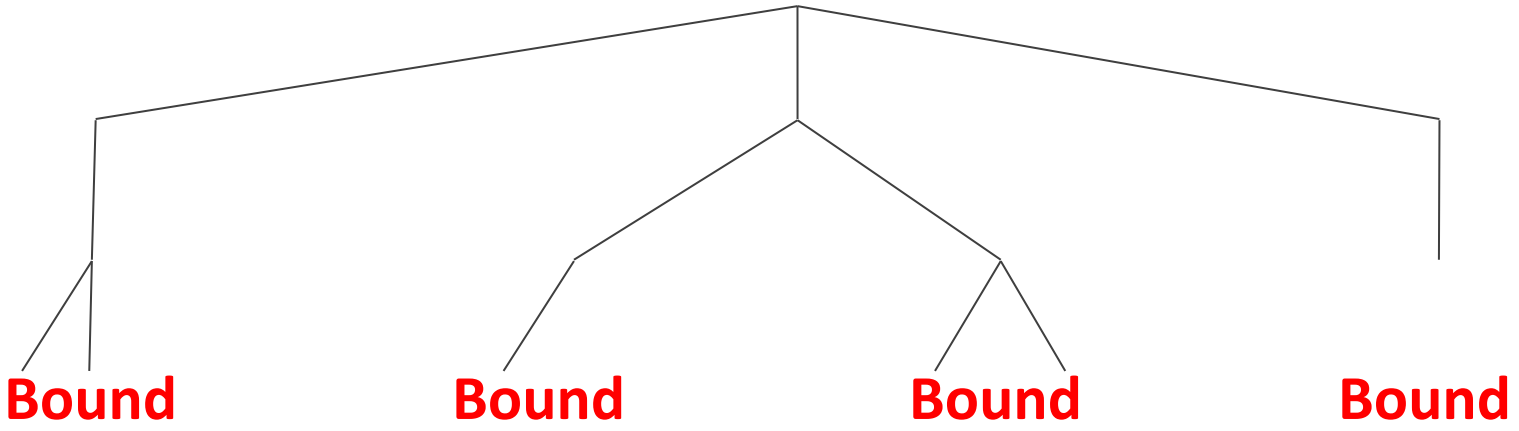
# **Branch**-and-**Bound** Algorithms



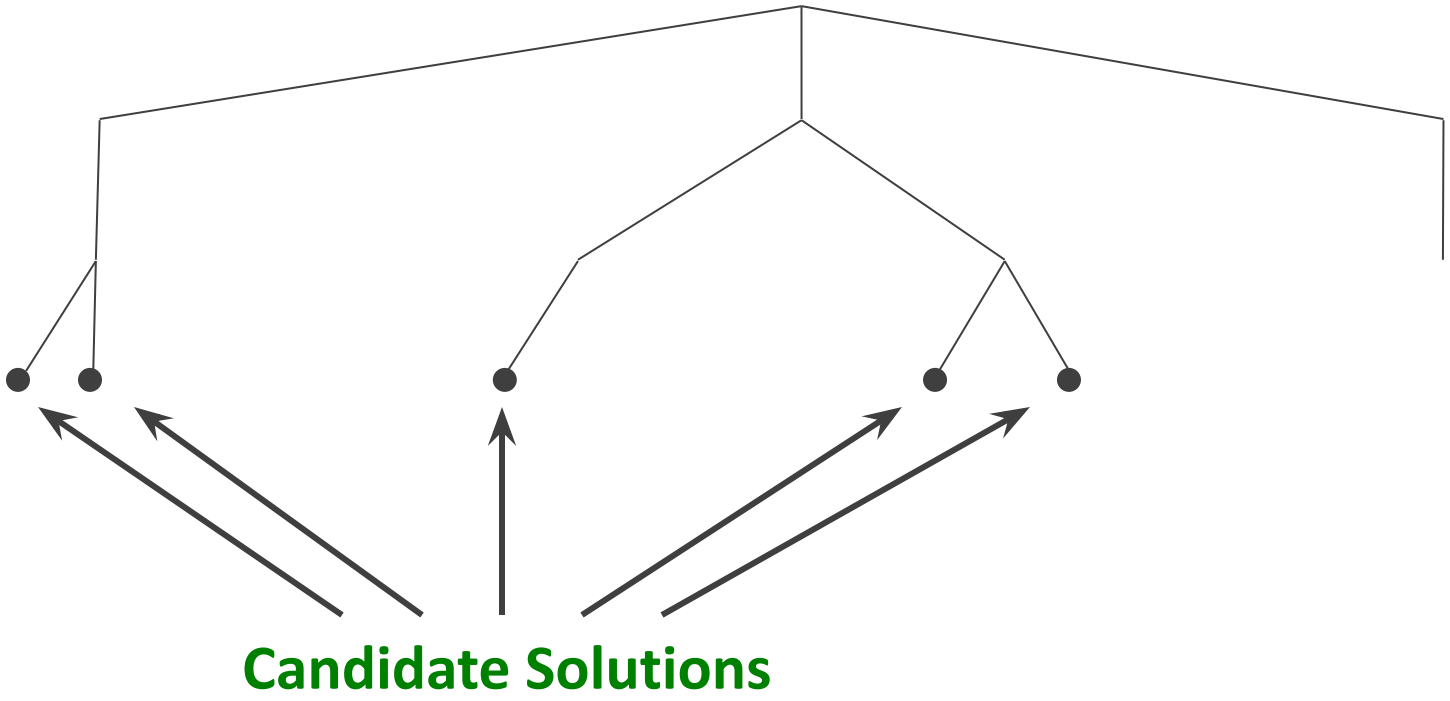
# **Branch**-and-**Bound** Algorithms



# Branch-and-Bound Algorithms



# **Branch**-and-**Bound** Algorithms



# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

Which amino acids have masses in *Spectrum*?

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

Which amino acids have masses in *Spectrum*?

| G | A | S | P | V | T | C | I/L | N | D | K/Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 114 | 115 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

Which amino acids have masses in *Spectrum*?

| G | A | S | P | V | T | C | I/L | N | D | K/Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 114 | 115 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

Which amino acids have masses in *Spectrum*?

| G | A | S | P | V | T | C | I/L | N | D | K/Q | E | M | H | F | R | Y | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | 71 | 87 | 97 | 99 | 101 | 103 | 113 | 114 | 115 | 128 | 129 | 131 | 137 | 147 | 156 | 163 | 186 |

We start with four "1-mer" peptides:

**P, V, T, C**

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

**Extend** these 1-mers into all possible 2-mers:

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

**Extend** these 1-mers into all possible 2-mers:

PA VA TA CA

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

**Extend** these 1-mers into all possible 2-mers:

| | | | |
|---|---|---|---|
| PA | VA | TA | CA |
| PC | VC | TC | CC |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

**Extend** these 1-mers into all possible 2-mers:

| | | | |
|---|---|---|---|
| PA | VA | TA | CA |
| PC | VC | TC | CC |
| PD | VD | TD | CD |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

**Extend** these 1-mers into all possible 2-mers:

| | | | |
|---|---|---|---|
| PA | VA | TA | CA |
| PC | VC | TC | CC |
| PD | VD | TD | CD |
| PE | VE | TE | CE |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

## Extend these 1-mers into all possible 2-mers:

| | | | |
|---|---|---|---|
| PA | VA | TA | CA |
| PC | VC | TC | CC |
| PD | VD | TD | CD |
| PE | VE | TE | CE |
| … | … | … | … |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

**Extend** these 1-mers into all possible 2-mers:

| | | | |
|---|---|---|---|
| PA | VA | TA | CA |
| PC | VC | TC | CC |
| PD | VD | TD | CD |
| PE | VE | TE | CE |
| … | … | … | … |
| PW | VW | TW | CW |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

**Extend** these 1-mers into all possible 2-mers:

| | | | |
|---|---|---|---|
| PA | VA | TA | CA |
| PC | VC | TC | CC |
| PD | VD | TD | CD |
| PE | VE | TE | CE |
| … | … | … | … |
| PW | VW | TW | CW |
| PY | VY | TY | CY |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

**Extend** these 1-mers into all possible 2-mers:

| | | | |
|---|---|---|---|
| PA | VA | TA | CA |
| PC | VC | TC | CC |
| PD | VD | TD | CD |
| PE | VE | TE | CE |
| … | … | … | … |
| PW | VW | TW | CW |
| PY | VY | TY | CY |

How can we **trim** this list?

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

PV is **consistent** with *Spectrum*:

Mass(P) = 97 Mass(V) = 99 Mass(PV) = 196

# **B**-&-**B** for Cyclopeptide Sequencing

*Spectrum* 0 **97** 97 **99** 101 103 **196** 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

PV is **consistent** with *Spectrum*:

Mass(P) = **97** Mass(V) = **99** Mass(PV) = **196**

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

PV is **consistent** with *Spectrum*:

Mass(P) = 97 Mass(V) = 99 Mass(PV) = 196

CD is **inconsistent** with *Spectrum*:

Mass(C) = 103 Mass(D) = 115 Mass(CD) = 218

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 **103** 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

PV is **consistent** with *Spectrum*:

Mass(P) = 97 Mass(V) = 99 Mass(PV) = 196

CD is **inconsistent** with *Spectrum*:

Mass(C) = **103** Mass(D) = **115** Mass(CD) = **218**

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

## List of consistent 2-mers:

PV PT PC VP VT
VC TP TV CP CV

# B-&-B for Cyclopeptide Sequencing

*Spectrum*

0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

List of consistent 2-mers:

PV PT PC VP VT
VC TP TV CP CV

**Expand**, then **Trim**…

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

## List of consistent 3-mers:

| | | | | |
|---|---|---|---|---|
| PVC | PVT | PTP | PTV | PCV |
| VPC | VPT | VTP | VCP | TPV |
| TPC | TVP | CPT | CPV | CVP |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

## List of consistent 3-mers:

| | | | | |
|---|---|---|---|---|
| PVC | PVT | PTP | PTV | PCV |
| VPC | VPT | VTP | VCP | TPV |
| TPC | TVP | CPT | CPV | CVP |

**Expand**, then **Trim**…

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

## List of consistent 4-mers:

PVCP PTPV PTPC PCVP VPTP
VCPT TPVC TPCV CPTP CVPT

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

## List of consistent 4-mers:

| | | | | |
|---|---|---|---|---|
| PVCP | PTPV | PTPC | PCVP | VPTP |
| VCPT | TPVC | TPCV | CPTP | CVPT |

**Expand**, then **Trim**…

# B-&-B for Cyclopeptide Sequencing

*Spectrum*

0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

## List of consistent 5-mers:

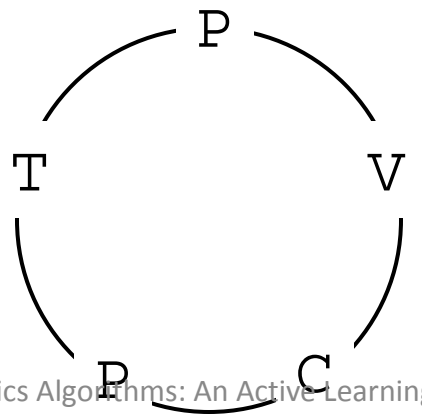
| | | | | |
|---|---|---|---|---|
| PVCPT | PTPVC | PTPCV | PCVPT | VPTPC |
| VCPTP | TPVCP | TPCVP | CPTPV | CVPTP |

# B-&-B for Cyclopeptide Sequencing

*Spectrum* 0 97 97 99 101 103 196 198 198 200 202
295 297 299 299 301 394 396 398 400 400 497

## List of consistent 5-mers:

PVCPT PTPVC PTPCV PCVPT VPTPC
VCPTP TPVCP TPCVP CPTPV CVPTP



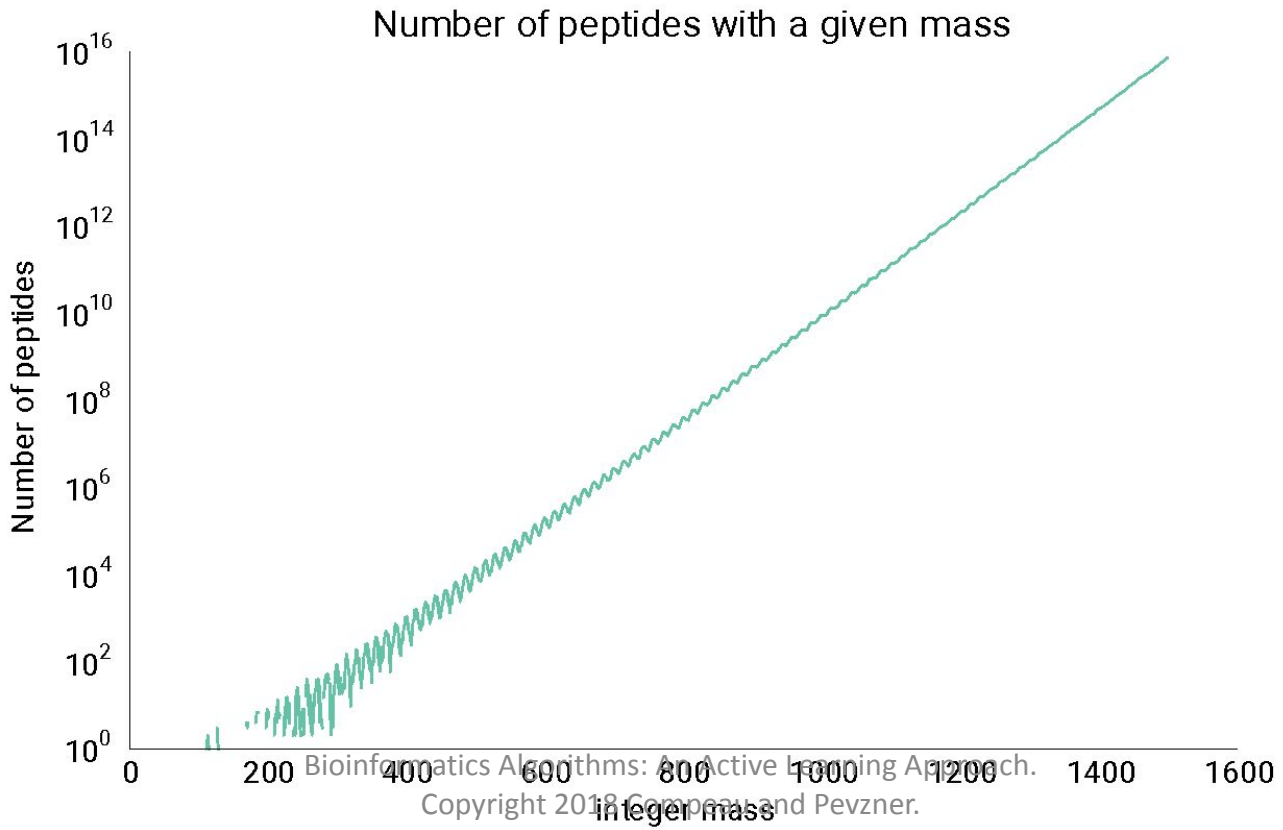

**ONE CYCLIC PEPTIDE!**

# **B**-&-**B** for Cyclopeptide Sequencing

1. Find all amino acids whose masses occur in *Spectrum*. Add to *List*.
2. **Extend** each peptide in *List* by each of 18 different amino acid masses.
3. **Trim** inconsistent peptides from *List*.
4. Return any peptides in *List* whose theoretical spectra match *Spectrum*.
5. Iterate Steps 2-4 until *List* is empty.

# Is This B-&-B Approach Efficient?

The brute force algorithm to cyclopeptide sequencing is **exponential**.



# Is This B-&-B Approach Efficient?

B&B for Cyclopeptide Sequencing *may* be exponential **on some dataset**…

# Is This B-&-B Approach Efficient?

B&B for Cyclopeptide Sequencing *may* be exponential **on some dataset**…

…but **in practice** it is very fast!

# Can We Go Home Now?

# Can We Go Home Now?

**NO!**

# Outline

- The Discovery of Antibiotics
- How Do Bacteria Make Antibiotics?
- Sequencing Antibiotics by Shattering Them into Pieces
- A Brute Force Algorithm for Cyclopeptide Sequencing
- Cyclopeptide Sequencing with Branch-and-Bound
- **Adapting Sequencing for Spectra with Errors**
- From 20 to More than 100 Amino Acids
- The Spectral Convolution Saves the Day
- The Truth about Spectra

# From Theoretical to Noisy Spectra

**Experimental spectra** often produce errors.

# From Theoretical to Noisy Spectra

**Experimental spectra** often produce errors.

Consider the following spectra for NQEL:

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Theoretical:** | 0 | | 113 | 114 | 128 | 129 | 227 | 242 | 242 | 257 | | 355 | 356 | 370 | 371 | 484 |
| **Experimental:** | 0 | 99 | 113 | 114 | 128 | | 227 | | | 257 | 299 | 355 | 356 | 370 | 371 | 484 |

# From Theoretical to Noisy Spectra

**Experimental spectra** often produce errors.

Consider the following spectra for NQEL:

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Theoretical:** | 0 | | 113 | 114 | 128 | 129 | 227 | 242 | 242 | 257 | | 355 | 356 | 370 | 371 | 484 |
| **Experimental:** | 0 | **99** | 113 | 114 | 128 | | 227 | | | 257 | **299** | 355 | 356 | 370 | 371 | 484 |

**False masses:** present in experimental spectrum, absent from theoretical spectrum

# From Theoretical to Noisy Spectra

**Experimental spectra** often produce errors.

Consider the following spectra for NQEL:

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Theoretical:** | 0 | | 113 | 114 | 128 | **129** | 227 | **242** | **242** | 257 | | 355 | 356 | 370 | 371 | 484 |
| **Experimental:** | 0 | **99** | 113 | 114 | 128 | | 227 | | | 257 | **299** | 355 | 356 | 370 | 371 | 484 |

**False masses:** present in experimental spectrum, absent from theoretical spectrum

**Missing masses:** present in theoretical spectrum, absent from experimental spectrum

# We Need a New Algorithm

Currently: a peptide's theoretical spectrum must match the experimental spectrum **exactly**.

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Theoretical:** | 0 | | 113 | 114 | 128 | **129** | 227 | **242** | **242** | 257 | | 355 | 356 | 370 | 371 | 484 |
| **Experimental:** | 0 | **99** | 113 | 114 | 128 | | 227 | | | 257 | **299** | 355 | 356 | 370 | 371 | 484 |

# We Need a New Algorithm

Currently: a peptide's theoretical spectrum must match the experimental spectrum **exactly**.

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Theoretical:** | **0** | | **113** | **114** | **128** | 129 | **227** | 242 | 242 | **257** | | **355** | **356** | **370** | **371** | **484** |
| **Experimental:** | **0** | 99 | **113** | **114** | **128** | | **227** | | | **257** | 299 | **355** | **356** | **370** | **371** | **484** |

Instead: **score** a peptide on how many masses its spectrum **shares** with the experimental spectrum.

# We Need a New Algorithm

Currently: a peptide's theoretical spectrum must match the experimental spectrum **exactly**.

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Theoretical:** | **0** | | **113** | **114** | **128** | 129 | **227** | 242 | 242 | **257** | | **355** | **356** | **370** | **371** | **484** |
| **Experimental:** | **0** | 99 | **113** | **114** | **128** | | **227** | | | **257** | 299 | **355** | **356** | **370** | **371** | **484** |

Instead: **score** a peptide on how many masses its spectrum **shares** with the experimental spectrum.

*Score*(NQEL, *ExperimentalSpectrum*) = **11**

# Cut in a Golf Tournamnet

**Cut:** reduces field to only those players in contention.

| Golfer | Score |
|---|---|
| Cabrera | -6 |
| Woods | -4 |
| Watson | -1 |
| McDowell | -1 |
| Scott | +1 |
| Daly | +14 |

# Cut in a Golf Tournamnet

**Cut:** reduces field to only those players in contention.

| Golfer | Score |
|---|---|
| Cabrera | -6 |
| Woods | -4 |
| Watson | -1 |
| McDowell | -1 |
| Scott | +1 |
| Daly | +14 |

Keep top 3 players

# Cut in a Golf Tournamnet

**Cut:** reduces field to only those players in contention.

| Golfer | Score |
|---|---|
| Cabrera | -6 |
| Woods | -4 |
| **Watson** | **-1** |
| **McDowell** | **-1** |
| Scott | +1 |
| Daly | +14 |

Keep top 3 players

# Cut in a Golf Tournamnet

**Cut:** reduces field to only those players in contention.

| Golfer | Score |
|---|---|
| Cabrera | -6 |
| Woods | -4 |
| **Watson** | **-1** |
| **McDowell** | **-1** |
| Scott | +1 |
| Daly | +14 |

Keep top 3 players "with ties"

# Cut in a Golf Tournamnet

**Cut:** reduces field to only those players in contention.

| Golfer | Score |
|---|---|
| Cabrera | -6 |
| Woods | -4 |
| Watson | -1 |
| McDowell | -1 |

Keep top 3 players "with ties"

# LeaderboardCyclopeptideSequencing

1. Add “0-peptide” to *Leaderboard* as *LeaderPeptide*.
2. **Extend** each peptide in *Leaderboard* by each of 18 different amino acid masses.
3. **Cut** low-scoring peptides from *Leaderboard*. (Keep “top N with ties”)
4. Update *LeaderPeptide* if there is a higher scoring peptide in *Leaderboard* with mass = parent mass.
5. Eliminate all peptides with mass > parent mass.
6. Iterate 2-5 until *Leaderboard* is empty.
7. Return *LeaderPeptide.*

# Testing on a Tyrocidine B1 Spectrum

**Warning:** This method is a **heuristic**; it sacrifices precision and may miss the correct solution.

# Testing on a Tyrocidine B1 Spectrum

## *Spectrum*$_{10}$: 10% false/missing masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | 113 | 114 | 128 | 128 | 147 | 147 | 163 | 186 | 227 | 241 | 242 |
| 244 | 260 | 261 | 262 | 283 | 291 | 333 | 340 | 357 | 385 | 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 | 820 | 835 | 837 | 875 | 892 | 892 | 917 | 932 | 932 | 933 |
| 934 | 965 | 982 | 989 | 1030 | 1031 | 1309 | 1060 | 1061 | 1062 | 1078 | 1080 | 1081 | 1095 |
| 1136 | 1159 | 1175 | 1175 | 1194 | 1194 | 1208 | 1209 | 1223 | 1225 | 1322 | | | |

# Testing on a Tyrocidine B1 Spectrum

## $Spectrum_{10}$: 10% **false**/missing masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | 113 | 114 | 128 | 128 | 147 | 147 | 163 | 186 | 227 | 241 | 242 |
| 244 | 260 | 261 | 262 | 283 | 291 | 333 | 340 | 357 | **385** | 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 | **820** | 835 | 837 | 875 | 892 | 892 | 917 | 932 | 932 | 933 |
| 934 | 965 | 982 | 989 | **1030** | 1031 | 1309 | 1060 | 1061 | 1062 | 1078 | 1080 | 1081 | 1095 |
| 1136 | 1159 | 1175 | 1175 | 1194 | 1194 | 1208 | 1209 | 1223 | 1225 | 1322 | | | |

# Testing on a Tyrocidine B1 Spectrum

## *Spectrum*$_{10}$: 10% **false**/**missing** masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | **113** | 114 | **128** | 128 | 147 | 147 | 163 | 186 | 227 | 241 | 242 |
| 244 | 260 | 261 | 262 | 283 | 291 | 333 | 340 | 357 | **385** | **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 | **820** | 835 | 837 | 875 | 892 | **892** | 917 | 932 | 932 | 933 |
| 934 | 965 | 982 | 989 | **1030** | **1031** | 1309 | 1060 | 1061 | 1062 | 1078 | 1080 | 1081 | 1095 |
| 1136 | 1159 | 1175 | 1175 | 1194 | 1194 | 1208 | 1209 | 1223 | 1225 | 1322 | | | |

# Testing on a Tyrocidine B1 Spectrum

## $Spectrum_{10}$: 10% **false**/**missing** masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | | 114 | | 128 | 147 | 147 | 163 | 186 | 227 | 241 | 242 |
| 244 | 260 | 261 | 262 | 283 | 291 | 333 | 340 | 357 | **385** | | 389 | 390 | 390 |
| 405 | 430 | 430 | 447 | 485 | 487 | 503 | 504 | 518 | 543 | 544 | 552 | 575 | 577 |
| 584 | | 632 | 650 | 651 | 671 | 672 | 690 | 691 | 738 | 745 | 747 | 770 | 778 |
| 779 | 804 | 818 | 819 | **820** | 835 | 837 | 875 | 892 | | 917 | 932 | 932 | 933 |
| 934 | 965 | 982 | 989 | **1030** | | 1309 | 1060 | 1061 | 1062 | 1078 | 1080 | 1081 | 1095 |
| 1136 | 1159 | 1175 | 1175 | 1194 | 1194 | 1208 | 1209 | 1223 | 1225 | 1322 | | | |

# Testing on a Tyrocidine B1 Spectrum

## $Spectrum_{10}$: 10% **false**/**missing** masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | | 114 | | 128 | 147 | 147 | 163 | 186 | 227 | 241 | 242 |
| 244 | 260 | 261 | 262 | 283 | 291 | 333 | 340 | 357 | **385** | | 389 | 390 | 390 |
| 405 | 430 | 430 | 447 | 485 | 487 | 503 | 504 | 518 | 543 | 544 | 552 | 575 | 577 |
| 584 | | 632 | 650 | 651 | 671 | 672 | 690 | 691 | 738 | 745 | 747 | 770 | 778 |
| 779 | 804 | 818 | 819 | **820** | 835 | 837 | 875 | 892 | | 917 | 932 | 932 | 933 |
| 934 | 965 | 982 | 989 | **1030** | | 1309 | 1060 | 1061 | 1062 | 1078 | 1080 | 1081 | 1095 |
| 1136 | 1159 | 1175 | 1175 | 1194 | 1194 | 1208 | 1209 | 1223 | 1225 | 1322 | | | |

Highest-scoring peptide: VKLFPWFNQY ☺

# A “Noisier” Tyrocidine B1 Spectrum

## $Spectrum_{25}$: 25% false/missing masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | 113 | 114 | 115 | 128 | 128 | 147 | 147 | 163 | 186 | 227 | 241 |
| 242 | 244 | 244 | 256 | 260 | 261 | 262 | 283 | 291 | 309 | 330 | 333 | 340 | 347 |
| 357 | 385 | 388 | 389 | 390 | 390 | 405 | 430 | 430 | 435 | 447 | 485 | 487 | 503 |
| 504 | 518 | 543 | 544 | 552 | 575 | 577 | 584 | 599 | 608 | 631 | 632 | 650 | 651 |
| 653 | 671 | 672 | 690 | 691 | 717 | 738 | 745 | 747 | 770 | 778 | 779 | 804 | 818 |
| 819 | 827 | 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 | | | | | | | |

# A “Noisier” Tyrocidine B1 Spectrum

## $Spectrum_{25}$: 25% **false**/missing masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | 113 | 114 | **115** | 128 | 128 | 147 | 147 | 163 | 186 | 227 | 241 |
| 242 | **244** | 244 | **256** | 260 | 261 | 262 | 283 | 291 | **309** | **330** | 333 | 340 | **347** |
| 357 | **385** | 388 | 389 | 390 | 390 | 405 | 430 | 430 | **435** | 447 | 485 | 487 | 503 |
| 504 | 518 | 543 | 544 | 552 | 575 | 577 | 584 | **599** | **608** | 631 | 632 | 650 | 651 |
| **653** | 671 | 672 | 690 | 691 | **717** | 738 | 745 | 747 | 770 | 778 | 779 | 804 | 818 |
| 819 | **827** | 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 | | | | | | | |

# A “Noisier” Tyrocidine B1 Spectrum

## *Spectrum*$_{25}$: 25% **false**/**missing** masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | 113 | 114 | **115** | 128 | 128 | 147 | 147 | 163 | 186 | 227 | 241 |
| 242 | **244** | 244 | **256** | 260 | 261 | 262 | 283 | 291 | **309** | **330** | 333 | 340 | **347** |
| **357** | **385** | 388 | 389 | 390 | 390 | 405 | **430** | **430** | **435** | 447 | 485 | 487 | 503 |
| 504 | 518 | **543** | 544 | 552 | 575 | 577 | 584 | **599** | **608** | 631 | 632 | 650 | 651 |
| **653** | **671** | 672 | 690 | 691 | **717** | 738 | 745 | **747** | 770 | **778** | 779 | 804 | 818 |
| 819 | **827** | 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 | | | | | | | |

# A “Noisier” Tyrocidine B1 Spectrum

## $Spectrum_{25}$: 25% **false**/**missing** masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | 113 | 114 | **115** | 128 | 128 | 147 | 147 | 163 | 186 | 227 | 241 |
| 242 | **244** | 244 | **256** | 260 | 261 | 262 | 283 | 291 | **309** | **330** | 333 | 340 | **347** |
| | **385** | 388 | 389 | 390 | 390 | 405 | | | **435** | 447 | 485 | 487 | 503 |
| 504 | 518 | | 544 | 552 | 575 | 577 | 584 | **599** | **608** | 631 | 632 | 650 | 651 |
| **653** | | 672 | 690 | 691 | **717** | 738 | 745 | | 770 | | 779 | 804 | 818 |
| 819 | **827** | 835 | 837 | 875 | 892 | 892 | 917 | 932 | 932 | 933 | 934 | 965 | 982 |
| 989 | | 1039 | 1060 | | 1062 | 1078 | 1080 | 1081 | 1095 | 1136 | 1159 | 1175 | 1175 |
| 1194 | 1194 | 1208 | 1209 | 1223 | | 1322 | | | | | | | |

# A “Noisier” Tyrocidine B1 Spectrum

## *Spectrum*$_{25}$: 25% **false**/**missing** masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | 113 | 114 | **115** | 128 | 128 | 147 | 147 | 163 | 186 | 227 | 241 |
| 242 | **244** | 244 | **256** | 260 | 261 | 262 | 283 | 291 | **309** | **330** | 333 | 340 | **347** |
| | **385** | 388 | 389 | 390 | 390 | 405 | | | **435** | 447 | 485 | 487 | 503 |
| 504 | 518 | | 544 | 552 | 575 | 577 | 584 | **599** | **608** | 631 | 632 | 650 | 651 |
| **653** | | 672 | 690 | 691 | **717** | 738 | 745 | | 770 | | 779 | 804 | 818 |
| 819 | **827** | 835 | 837 | 875 | 892 | 892 | 917 | 932 | 932 | 933 | 934 | 965 | 982 |
| 989 | | 1039 | 1060 | | 1062 | 1078 | 1080 | 1081 | 1095 | 1136 | 1159 | 1175 | 1175 |
| 1194 | 1194 | 1208 | 1209 | 1223 | | 1322 | | | | | | | |

Highest-scoring peptide: VKLFP**AD**FNQY ☹

# Outline

- The Discovery of Antibiotics
- How Do Bacteria Make Antibiotics?
- Sequencing Antibiotics by Shattering Them into Pieces
- A Brute Force Algorithm for Cyclopeptide Sequencing
- Cyclopeptide Sequencing with Branch-and-Bound
- Adapting Sequencing for Spectra with Errors
- **From 20 to More than 100 Amino Acids**
- The Spectral Convolution Saves the Day
- The Truth about Spectra

# From 18 to 100+ Amino Acids

# From 18 to 100+ Amino Acids

NRPs contain more **non-standard** amino acids because they are free from the Central Dogma.

# From 18 to 100+ Amino Acids

NRPs contain more **non-standard** amino acids because they are free from the Central Dogma.

Tyrocidine B

`Val-Orn-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr`

**Ornithine**: non-standard amino acid

# From 18 to 100+ Amino Acids

NRPs contain more **non-standard** amino acids because they are free from the Central Dogma.

Tyrocidine B

`Val-Orn-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr`

**Ornithine**: non-standard amino acid

Bioinformaticians assume *any integer* between 57 and 200 can act as the mass of an amino acid.

# Back to Noisy Spectra

## $Spectrum_{10}$: 10% **false**/**missing** masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | | 114 | | 128 | 147 | 147 | 163 | 186 | 227 | 241 | 242 |
| 244 | 260 | 261 | 262 | 283 | 291 | 333 | 340 | 357 | **385** | | 389 | 390 | 390 |
| 405 | 430 | 430 | 447 | 485 | 487 | 503 | 504 | 518 | 543 | 544 | 552 | 575 | 577 |
| 584 | | 632 | 650 | 651 | 671 | 672 | 690 | 691 | 738 | 745 | 747 | 770 | 778 |
| 779 | 804 | 818 | 819 | **820** | 835 | 837 | 875 | 892 | | 917 | 932 | 932 | 933 |
| 934 | 965 | 982 | 989 | **1030** | | 1309 | 1060 | 1061 | 1062 | 1078 | 1080 | 1081 | 1095 |
| 1136 | 1159 | 1175 | 1175 | 1194 | 1194 | 1208 | 1209 | 1223 | 1225 | 1322 | | | |

# Back to Noisy Spectra

## $Spectrum_{10}$: 10% **false**/**missing** masses

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 97 | 99 | | 114 | | 128 | 147 | 147 | 163 | 186 | 227 | 241 | 242 |
| 244 | 260 | 261 | 262 | 283 | 291 | 333 | 340 | 357 | **385** | | 389 | 390 | 390 |
| 405 | 430 | 430 | 447 | 485 | 487 | 503 | 504 | 518 | 543 | 544 | 552 | 575 | 577 |
| 584 | | 632 | 650 | 651 | 671 | 672 | 690 | 691 | 738 | 745 | 747 | 770 | 778 |
| 779 | 804 | 818 | 819 | **820** | 835 | 837 | 875 | 892 | | 917 | 932 | 932 | 933 |
| 934 | 965 | 982 | 989 | **1030** | | 1309 | 1060 | 1061 | 1062 | 1078 | 1080 | 1081 | 1095 |
| 1136 | 1159 | 1175 | 1175 | 1194 | 1194 | 1208 | 1209 | 1223 | 1225 | 1322 | | | |

Highest-scoring peptide: VKLFPWFN-**98**-**65**



# Outline

- The Discovery of Antibiotics
- How Do Bacteria Make Antibiotics?
- Sequencing Antibiotics by Shattering Them into Pieces
- A Brute Force Algorithm for Cyclopeptide Sequencing
- Cyclopeptide Sequencing with Branch-and-Bound
- Adapting Sequencing for Spectra with Errors
- From 20 to More than 100 Amino Acids
- **The Spectral Convolution Saves the Day**
- The Truth about Spectra

# Restricting Amino Acid Alphabet

**Goal:** reduce the number of amino acids that we need to consider.

# Restricting Amino Acid Alphabet

**Goal:** reduce the number of amino acids that we need to consider.

Recall the following spectrum for NQEL:

**Experimental:** 0 99 113 114 128 227 257 299 355 356 370 371 484

# Restricting Amino Acid Alphabet

**Goal:** reduce the number of amino acids that we need to consider.

Recall the following spectrum for NQEL:

**Experimental:** 0 99 113 114 128 227 257 299 355 356 370 371 484

Mass(**E**) = **129**, which is missing, but…

# Restricting Amino Acid Alphabet

**Goal:** reduce the number of amino acids that we need to consider.

Recall the following spectrum for NQEL:

**Experimental:** 0 99 113 114 **128** 227 **257** 299 355 356 370 371 484

Mass(**E**) = **129**, which is missing, but…

Mass(Q**E**) – Mass(Q) = **257** – **128** = **129**

# Restricting Amino Acid Alphabet

**Goal:** reduce the number of amino acids that we need to consider.

Recall the following spectrum for NQEL:

**Experimental:** 0 99 113 114 128 **227** 257 299 355 **356** 370 371 484

Mass(**E**) = **129**, which is missing, but…

Mass(**E**LN) – Mass(LN) = **356** – **227** = **129**

# Restricting Amino Acid Alphabet

**Goal:** reduce the number of amino acids that we need to consider.

Recall the following spectrum for NQEL:

**Experimental:** 0 99 113 114 128 227 257 299 **355** 356 370 371 **484**

Mass(**E**) = **129**, which is missing, but…

Mass(NQ**E**L) – Mass(LNQ) = **484** – **355** = **129**

# The Spectral Convolution

**Spectral convolution:** positive difference between every pair of masses in spectrum.

| | " " | false | L | N | Q | LN | QE | false | LNQ | ELN | QEL | NQE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 99 | 113 | 114 | 128 | 227 | 257 | 299 | 355 | 356 | 370 | 371 |
| 0 | | | | | | | | | | | | |
| 99 | 99 | | | | | | | | | | | |
| 113 | 113 | 14 | | | | | | | | | | |
| 114 | 114 | 15 | 1 | | | | | | | | | |
| 128 | 128 | 29 | 15 | 14 | | | | | | | | |
| 227 | 227 | 128 | 114 | 113 | 99 | | | | | | | |
| 257 | 257 | 158 | 144 | 143 | 129 | 30 | | | | | | |
| 299 | 299 | 200 | 186 | 185 | 171 | 72 | 42 | | | | | |
| 355 | 355 | 256 | 242 | 241 | 227 | 128 | 98 | 56 | | | | |
| 356 | 356 | 257 | 243 | 242 | 228 | 129 | 99 | 57 | 1 | | | |
| 370 | 370 | 271 | 257 | 256 | 242 | 143 | 113 | 71 | 15 | 14 | | |
| 371 | 371 | 272 | 258 | 257 | 243 | 144 | 114 | 72 | 16 | 15 | 1 | |
| 484 | 484 | 385 | 371 | 370 | 356 | 257 | 227 | 185 | 129 | 128 | 114 | 113 |

# The Spectral Convolution

**What are the most frequent elements between 57 and 200?**

| | " " | false | L | N | Q | LN | QE | false | LNQ | ELN | QEL | NQE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 99 | 113 | 114 | 128 | 227 | 257 | 299 | 355 | 356 | 370 | 371 |
| 0 | | | | | | | | | | | | |
| 99 | 99 | | | | | | | | | | | |
| 113 | 113 | 14 | | | | | | | | | | |
| 114 | 114 | 15 | 1 | | | | | | | | | |
| 128 | 128 | 29 | 15 | 14 | | | | | | | | |
| 227 | 227 | 128 | 114 | 113 | 99 | | | | | | | |
| 257 | 257 | 158 | 144 | 143 | 129 | 30 | | | | | | |
| 299 | 299 | 200 | 186 | 185 | 171 | 72 | 42 | | | | | |
| 355 | 355 | 256 | 242 | 241 | 227 | 128 | 98 | 56 | | | | |
| 356 | 356 | 257 | 243 | 242 | 228 | 129 | 99 | 57 | 1 | | | |
| 370 | 370 | 271 | 257 | 256 | 242 | 143 | 113 | 71 | 15 | 14 | | |
| 371 | 371 | 272 | 258 | 257 | 243 | 144 | 114 | 72 | 16 | 15 | 1 | |
| 484 | 484 | 385 | 371 | 370 | 356 | 257 | 227 | 185 | 129 | 128 | 114 | 113 |

# The Spectral Convolution

**What are the most frequent elements between 57 and 200?**

| 99 | 113 | 114 | 128 | 129 |
|---|---|---|---|---|
| V | L | N | Q | E |

# The Spectral Convolution

What are the most frequent elements between 57 and 200?

| 99 | **113** | **114** | **128** | **129** |
|---|---|---|---|---|
| V | **L** | **N** | **Q** | **E** |

**5 Most Frequent Elements in Convolution □**
**4 amino acids of NQEL!**

# ConvolutionCyclopeptideSequencing

1. Form spectral convolution of spectrum.

2. Take the *M most frequent* elements in the convolution (between 57 and 200).

3. Run **LeaderboardCyclopeptideSequencing**, forming peptides only on these *M* integers.

# Does It Really Work?

1. Take the convolution of $Spectrum_{10}$.

# Does It Really Work?

1. Take the convolution of $Spectrum_{10}$.

2. Pick $M$ = 10 most frequent elements.

147
128
97
113
114
186
57
163
99
145

# Does It Really Work?

1. Take the convolution of $Spectrum_{10}$.

2. Pick $M$ = 10 most frequent elements.

| | |
|---|---|
| 147 | F |
| 128 | K/Q |
| 97 | P |
| 113 | I/L |
| 114 | N |
| 186 | W |
| 57 | G |
| 163 | Y |
| 99 | V |
| 145 | |

# Does It Really Work?

1. Take the convolution of $Spectrum_{10}$.

2. Pick $M$ = 10 most frequent elements.

3. Run the algorithm…

| | |
|---|---|
| 147 | F |
| 128 | K/Q |
| 97 | P |
| 113 | I/L |
| 114 | N |
| 186 | W |
| 57 | G |
| 163 | Y |
| 99 | V |
| 145 | |

# Does It Really Work?

1. Take the convolution of $Spectrum_{10}$.

2. Pick $M$ = 10 most frequent elements.

3. Run the algorithm…

| | |
|---|---|
| **147** | **F** |
| **128** | **K/Q** |
| **97** | **P** |
| **113** | **I/L** |
| **114** | **N** |
| **186** | **W** |
| 57 | G |
| **163** | **Y** |
| **99** | **V** |
| 145 | |

**Winning peptide**: **VKLFPWFNQY** 😊

# Does It Really Work?

**ConvolutionCyclopeptideSequencing** even reconstructs Tyrocidine B1 from the "noisier" $Spectrum_{25}$.

# Does It Really Work?

**ConvolutionCyclopeptideSequencing** even reconstructs Tyrocidine B1 from the "noisier" $Spectrum_{25}$.

# Does It Really Work?

**ConvolutionCyclopeptideSequencing** even reconstructs Tyrocidine B1 from the "noisier" $Spectrum_{25}$.

**ONE MORE THING…**

# Outline

- The Discovery of Antibiotics
- How Do Bacteria Make Antibiotics?
- Sequencing Antibiotics by Shattering Them into Pieces
- A Brute Force Algorithm for Cyclopeptide Sequencing
- Cyclopeptide Sequencing with Branch-and-Bound
- Adapting Sequencing for Spectra with Errors
- From 20 to More than 100 Amino Acids
- The Spectral Convolution Saves the Day
- **The Truth about Spectra**

# The Truth About Spectra

$Spectrum_{25}$ is much less noisy than the spectra obtained in practice.

# The Truth About Spectra

$Spectrum_{25}$ is much less noisy than the spectra obtained in practice.

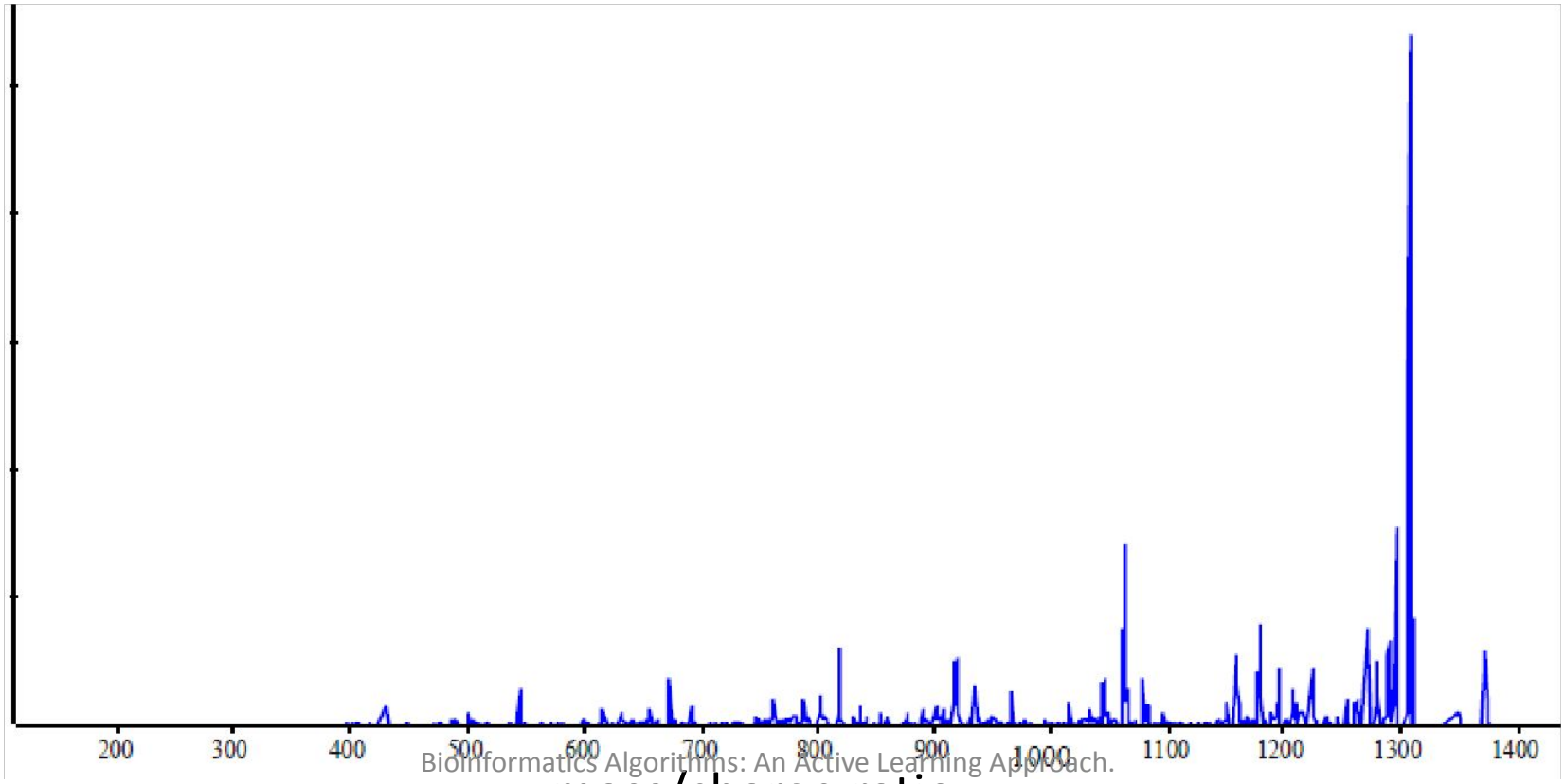
Also, the mass spectrometer doesn't simply "weigh" peptide fragments.

# The Truth About Spectra

1. Ionize the peptide fragments.
2. Sort fragments using electromagnetic field.
3. Measure **mass/charge ratio** of each fragment.
4. Determine **intensity** (# of ions) at each mass/charge ratio.

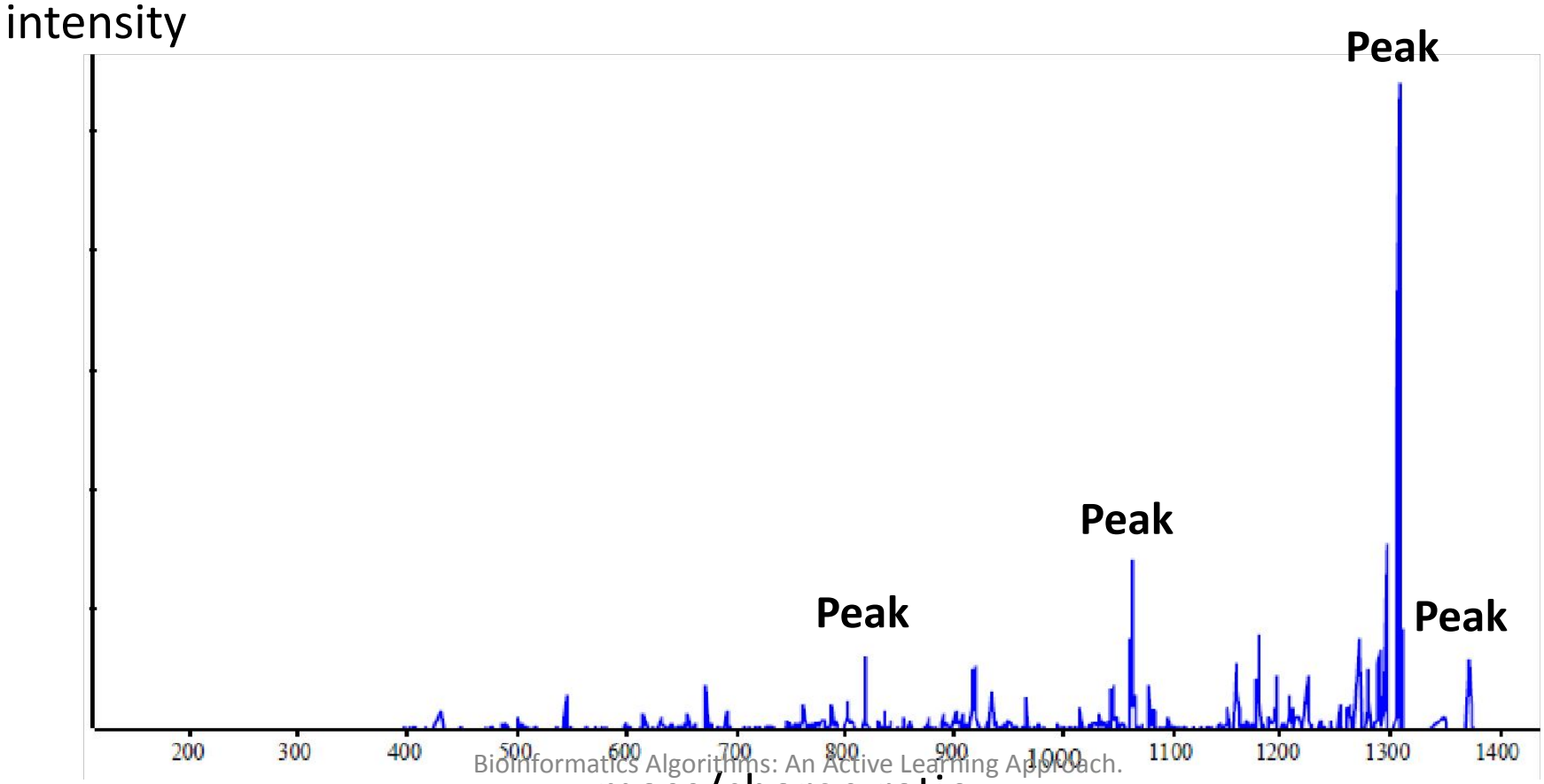
# A Real Tyrocidine B1 Spectrum

**Spectrum**: graph of intensity vs. mass/charge ratio



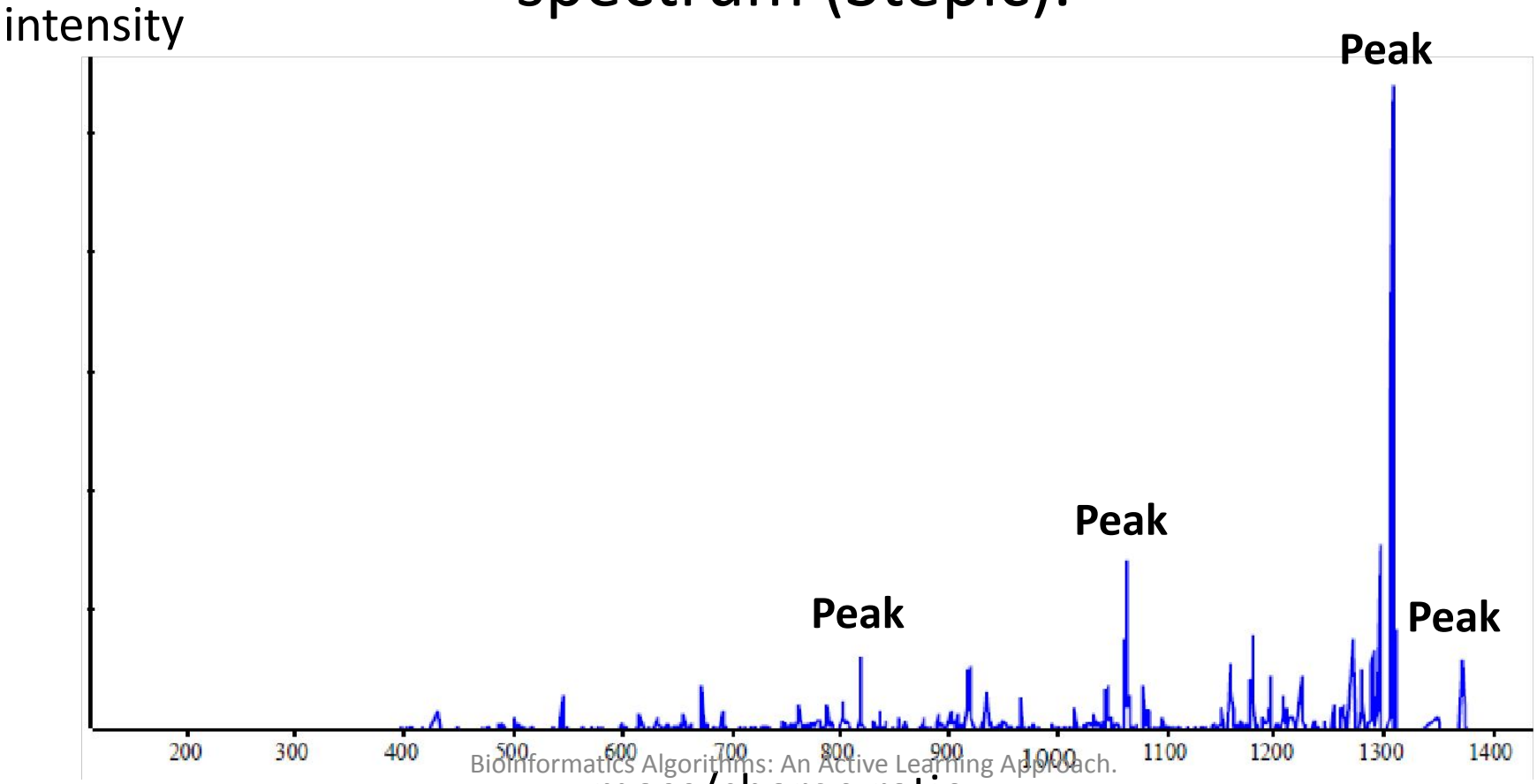
# A Real Tyrocidine B1 Spectrum

**Spectrum**: graph of intensity vs. mass/charge ratio



# A Real Tyrocidine B1 Spectrum

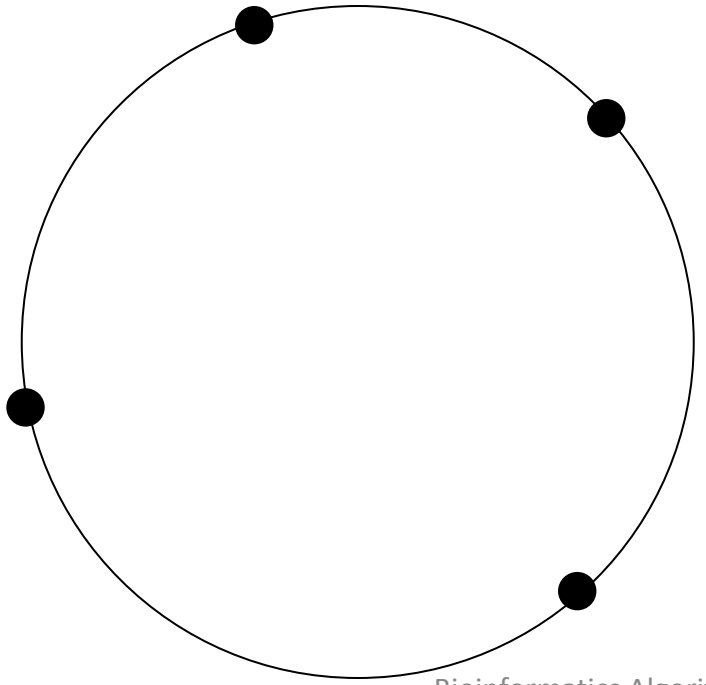
**Challenge**: Reconstruct a peptide from real spectrum (Stepic).



# Open Problems

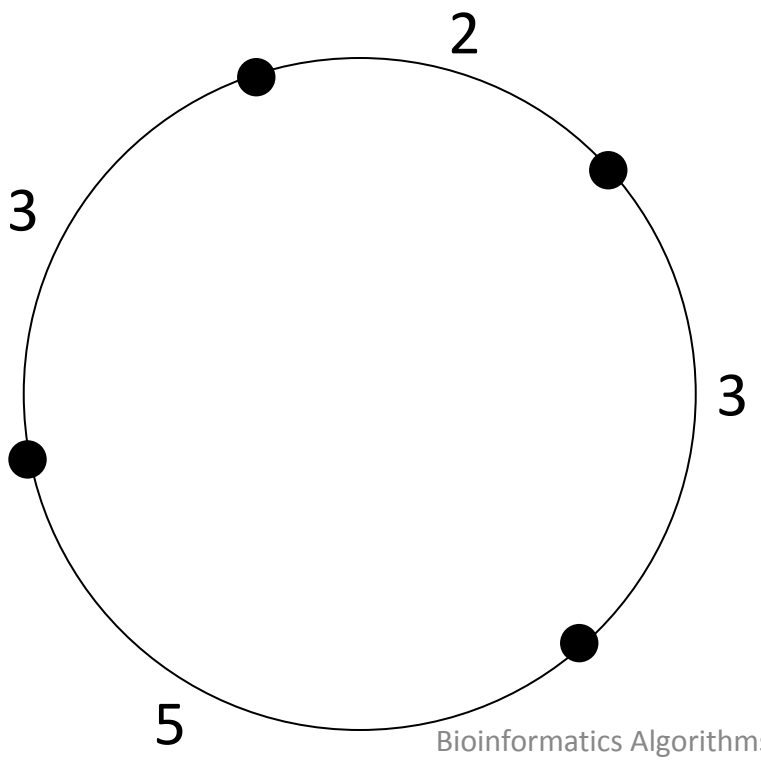
# Beltway and Turnpike Problems

Given a collection of points on a circle, it is easy to find the pairwise distances between them.



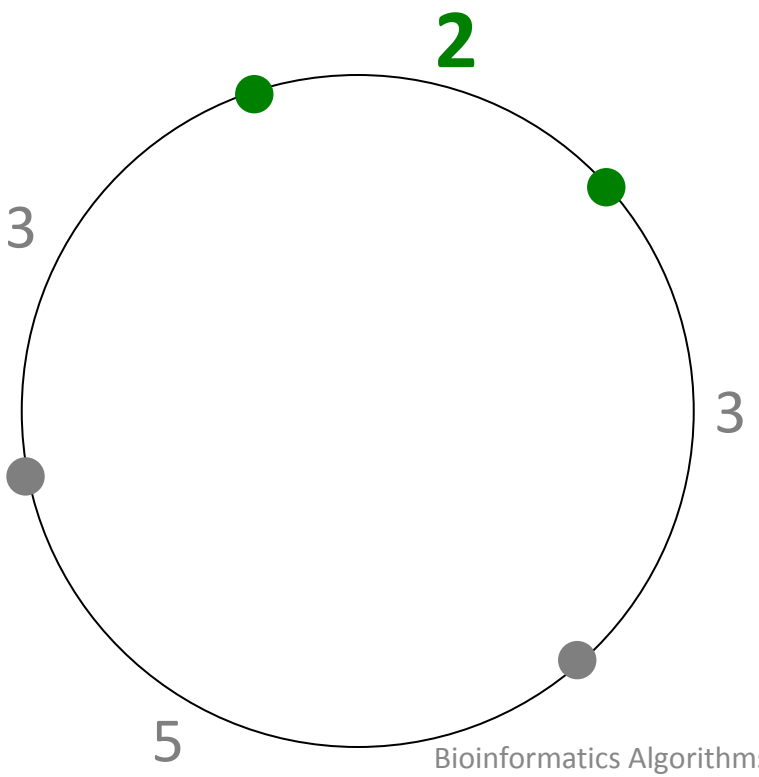
# Beltway and Turnpike Problems

Given a collection of points on a circle, it is easy to find the pairwise distances between them.



# Beltway and Turnpike Problems

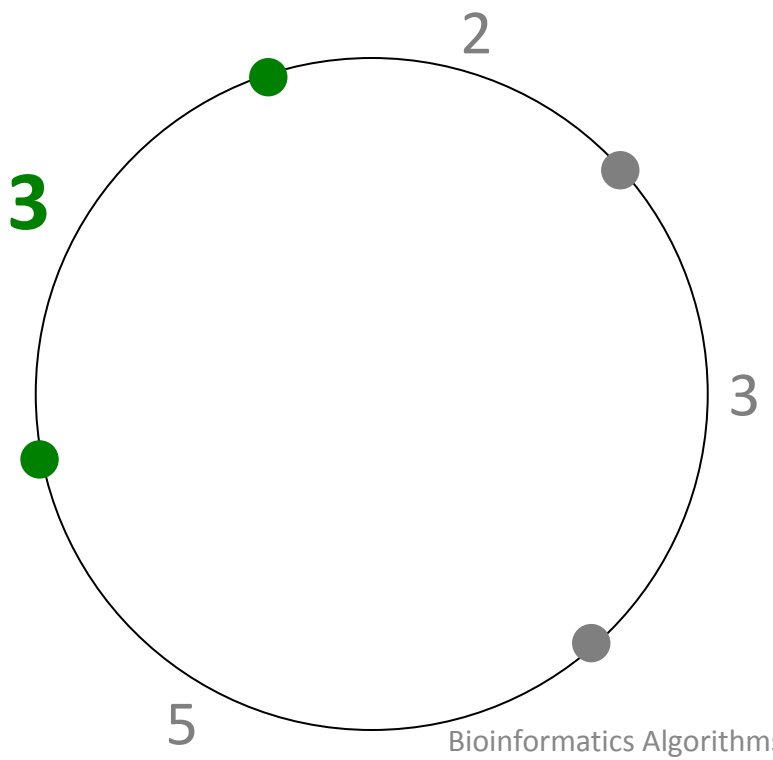
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




2

# Beltway and Turnpike Problems

Given a collection of points on a circle, it is easy to find the pairwise distances between them.

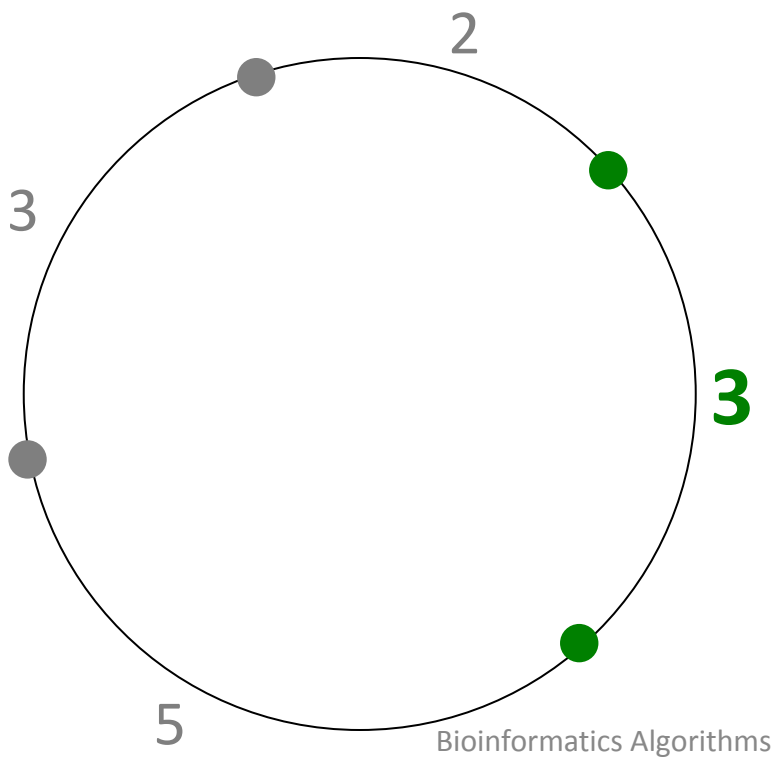



2

3

# Beltway and Turnpike Problems

Given a collection of points on a circle, it is easy to find the pairwise distances between them.



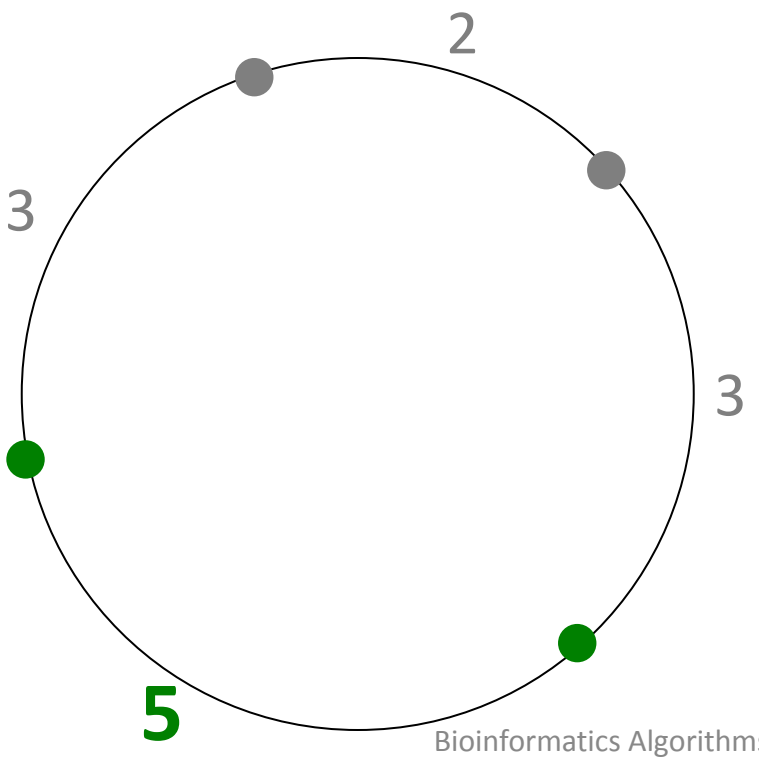

2

3

**3**

# Beltway and Turnpike Problems

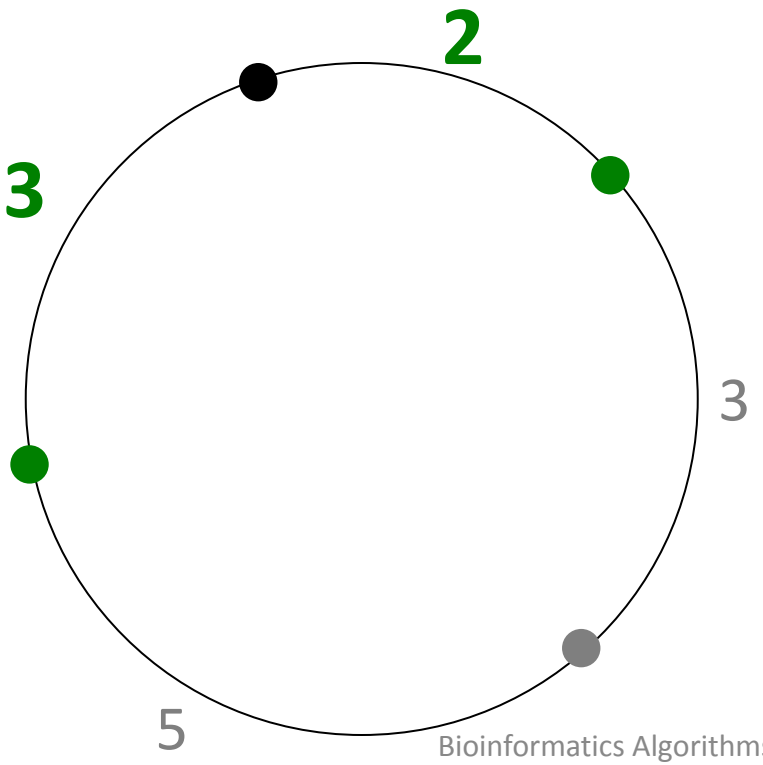
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




2
3
3
**5**

# Beltway and Turnpike Problems

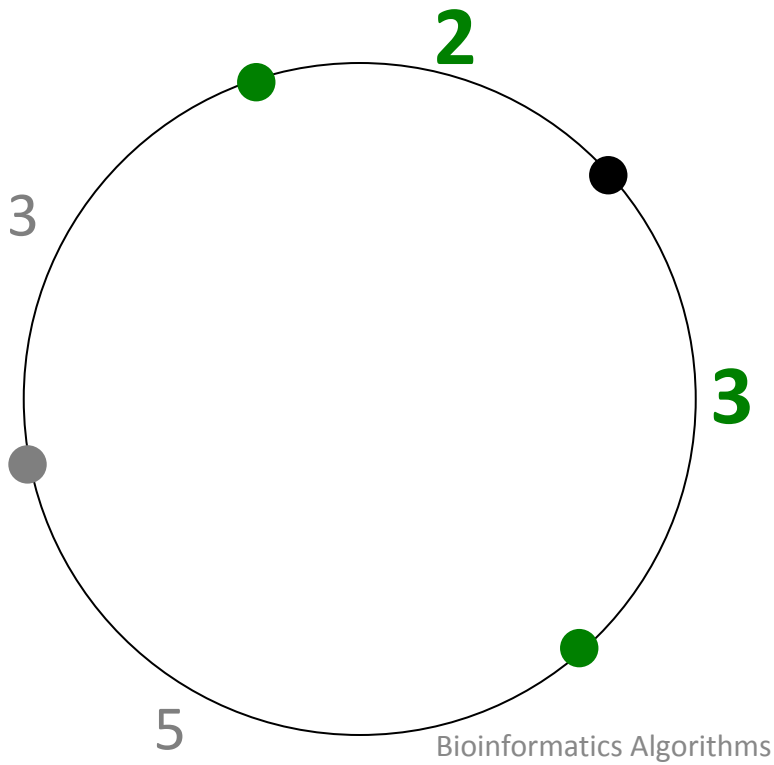
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




2
3
3
5
5

# Beltway and Turnpike Problems

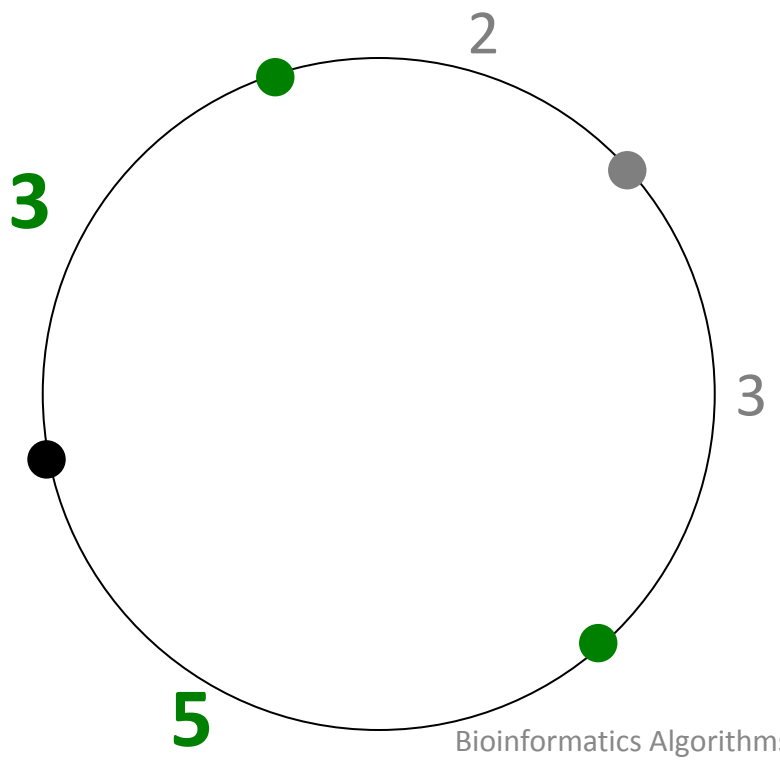
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




2
3
3
5
5
**5**

# Beltway and Turnpike Problems

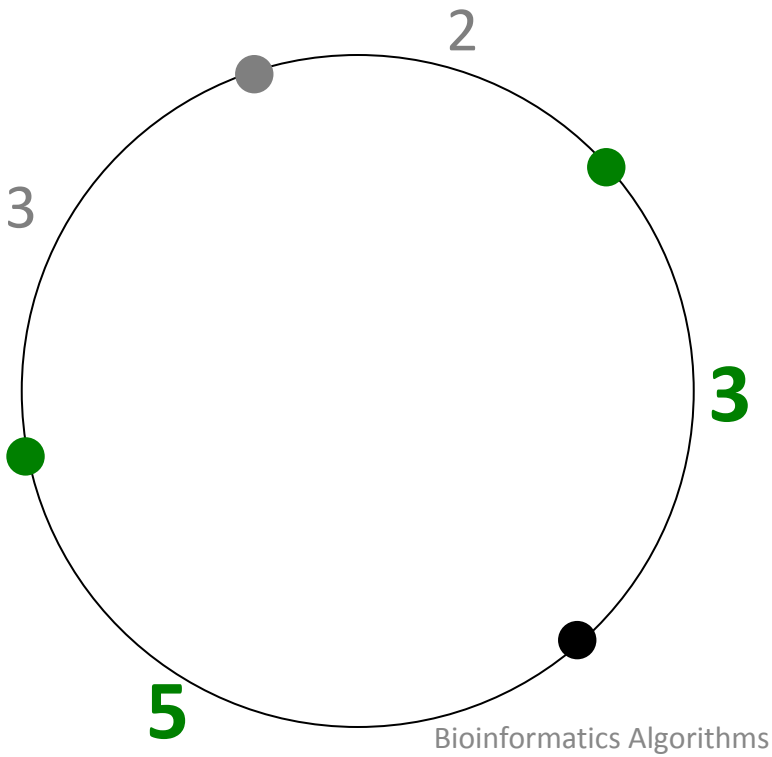
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




2 8
3
3
5
5
5

# Beltway and Turnpike Problems

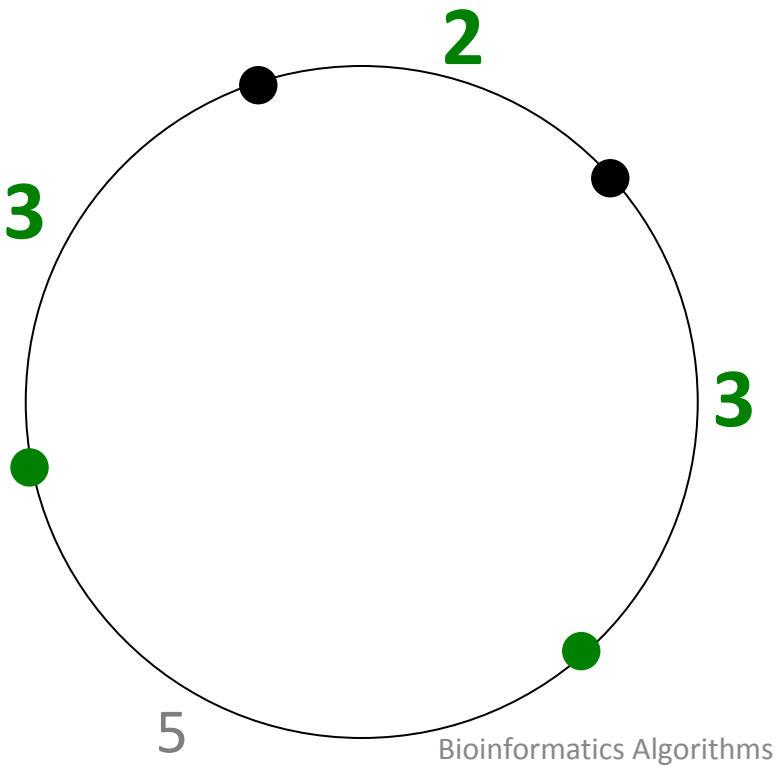
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




2 8
3 **8**
3
5
5
5

# Beltway and Turnpike Problems

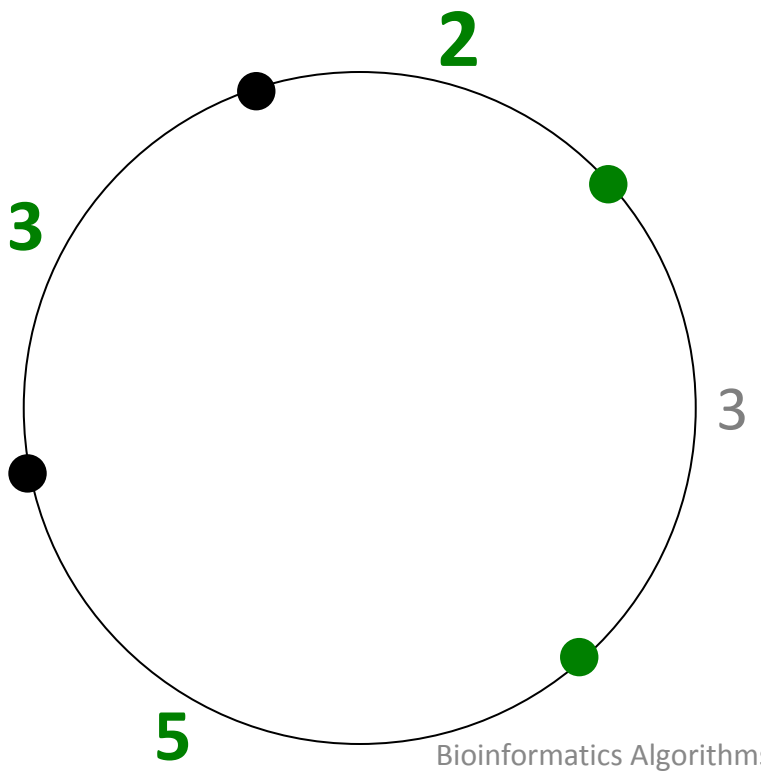
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




| | |
|---|---|
| 2 | 8 |
| 3 | 8 |
| 3 | **8** |
| 5 | |
| 5 | |
| 5 | |

# Beltway and Turnpike Problems

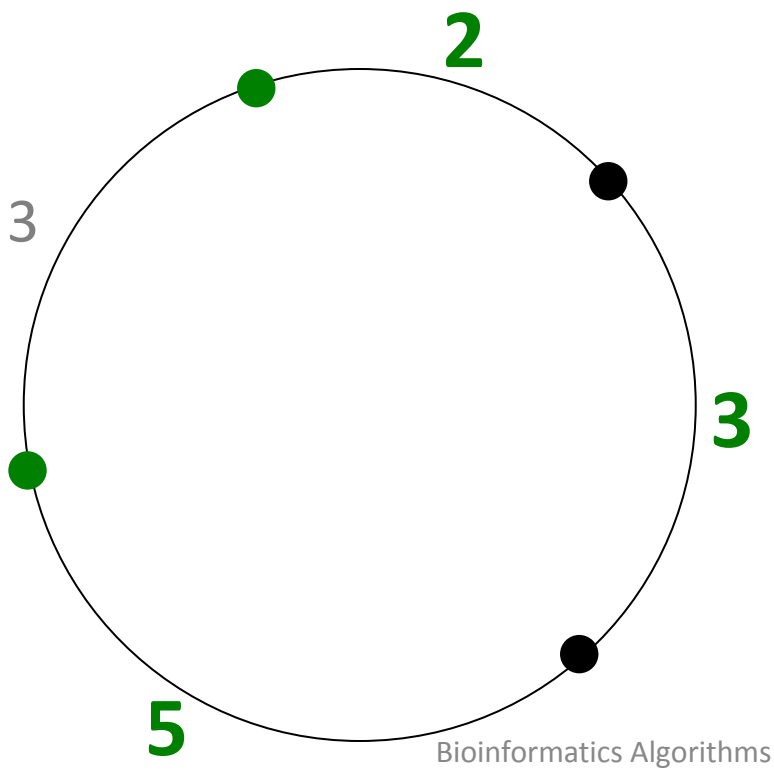
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




| | |
|---|---|
| 2 | 8 |
| 3 | 8 |
| 3 | 8 |
| 5 | **10** |
| 5 | |
| 5 | |

# Beltway and Turnpike Problems

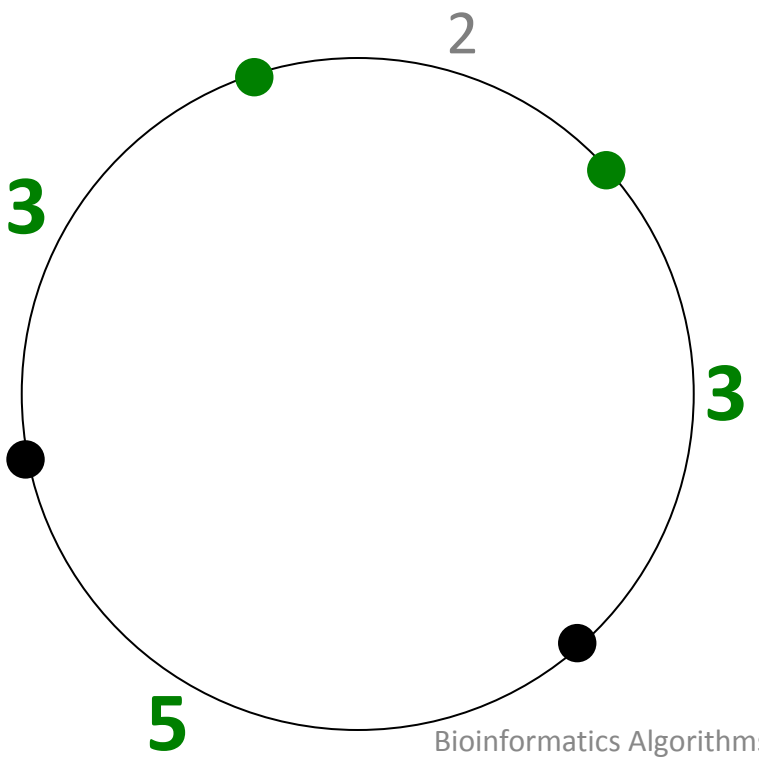
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




| | |
|---|---|
| 2 | 8 |
| 3 | 8 |
| 3 | 8 |
| 5 | 10 |
| 5 | **10** |
| 5 | |

# Beltway and Turnpike Problems

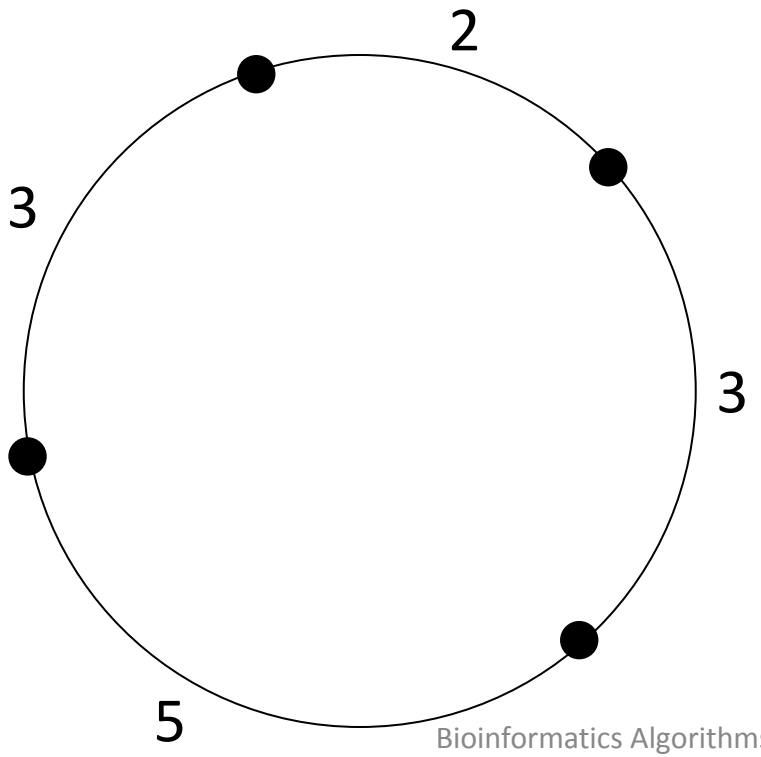
Given a collection of points on a circle, it is easy to find the pairwise distances between them.




| | |
|---|---|
| 2 | 8 |
| 3 | 8 |
| 3 | 8 |
| 5 | 10 |
| 5 | 10 |
| 5 | **11** |

# Beltway and Turnpike Problems

Given a collection of points on a circle, it is easy to find the pairwise distances between them.

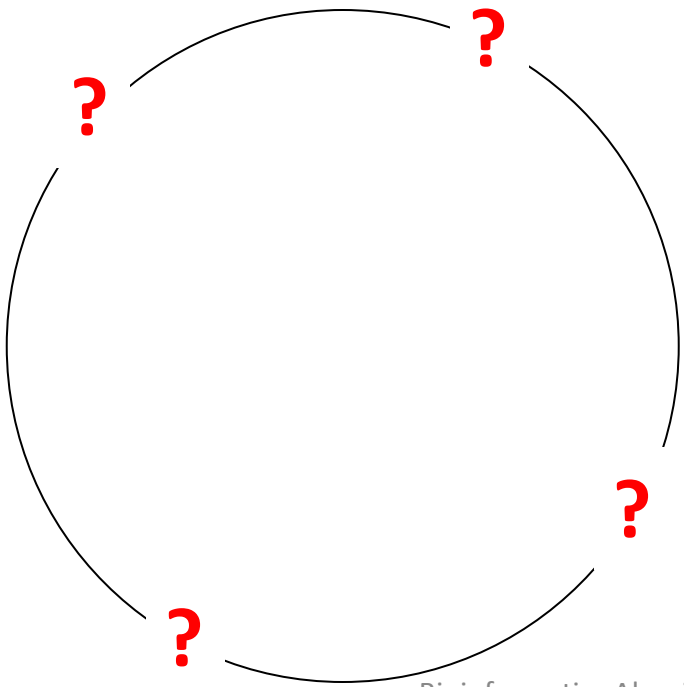



EASY

| | |
|---|---|
| 2 | 8 |
| 3 | 8 |
| 3 | 8 |
| 5 | 10 |
| 5 | 10 |
| 5 | 11 |

# Beltway and Turnpike Problems

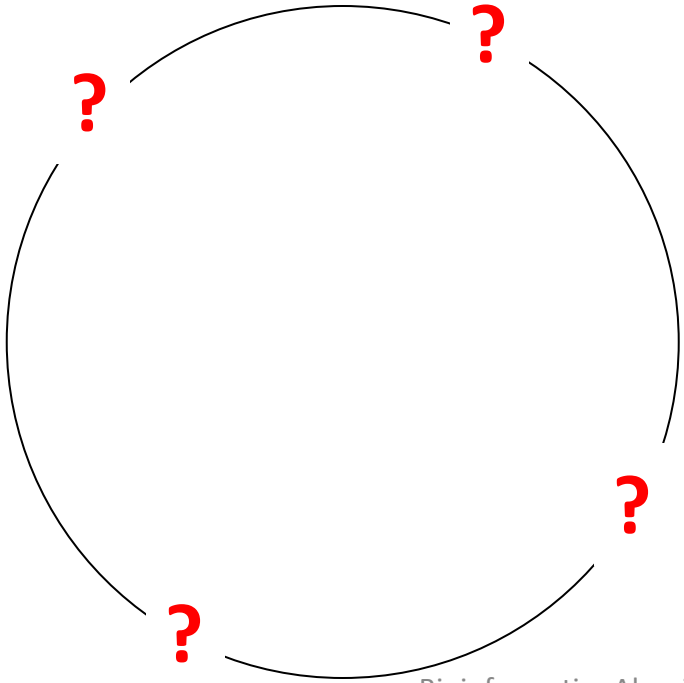
What if we are given the pairwise distances and want to reconstruct the points?




| | |
|---|---|
| 2 | 8 |
| 3 | 8 |
| 3 | 8 |
| 5 | 10 |
| 5 | 10 |
| 5 | 11 |

# Beltway and Turnpike Problems

This is a harder problem, known as the **Beltway Problem** (think: cities on a circular road).

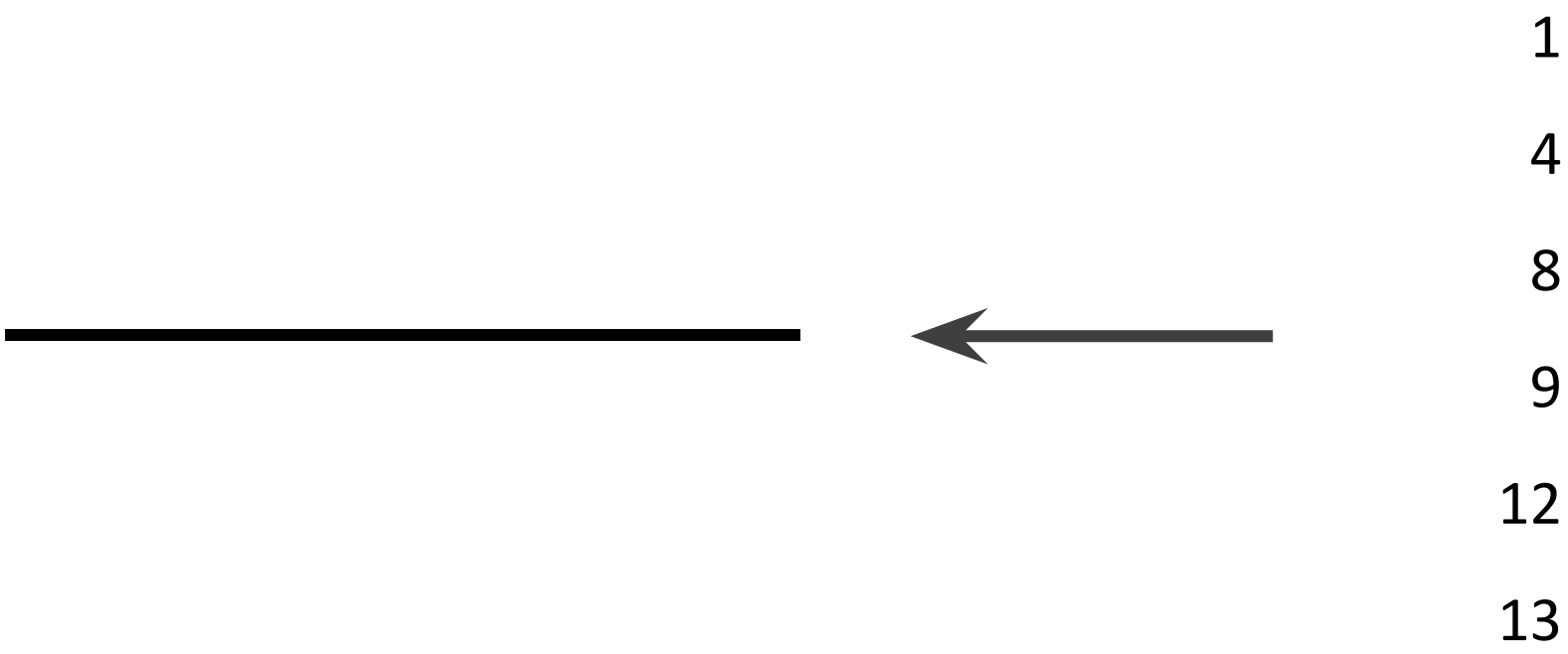



HARD

| | |
|---|---|
| 2 | 8 |
| 3 | 8 |
| 3 | 8 |
| 5 | 10 |
| 5 | 10 |
| 5 | 11 |

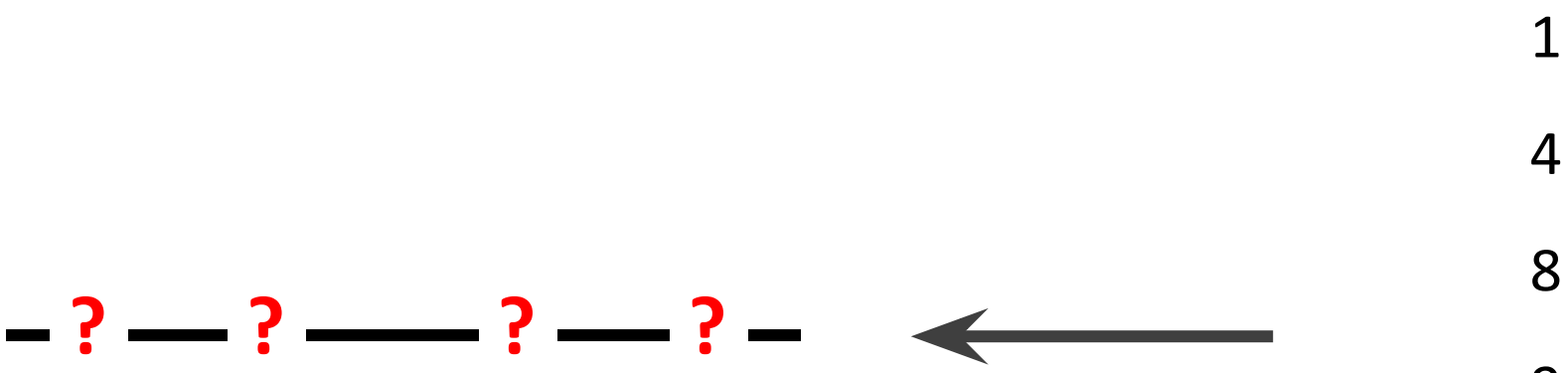
# Beltway and Turnpike Problems

If the points are on a *line segment* instead of a circle, we have the **Turnpike Problem**.



# Beltway and Turnpike Problems

If the points are on a *line segment* instead of a circle, we have the **Turnpike Problem**.



# Beltway and Turnpike Problems

No one has ever found a polynomial algorithm for either the Beltway or Turnpike Problem.

However, the Turnpike Problem does have a **pseudo-polynomial** algorithm, which is polynomial in the length of the *segment*.

# Beltway and Turnpike Problems

No one has ever found a polynomial algorithm for either the Beltway or Turnpike Problem.

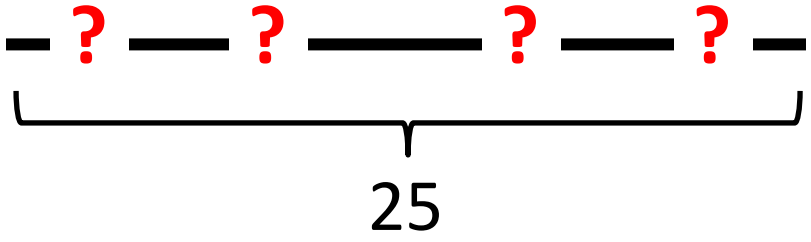
However, the Turnpike Problem does have a **pseudo-polynomial** algorithm, which is polynomial in the length of the *segment*.

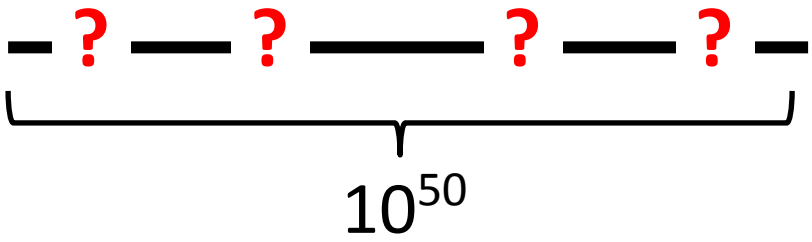
– ? — ? ——— ? — ? –  – ? — ? ——— ? — ? –

# Beltway and Turnpike Problems

No one has ever found a polynomial algorithm for either the Beltway or Turnpike Problem.

However, the Turnpike Problem does have a **pseudo-polynomial** algorithm, which is polynomial in the length of the *segment*.

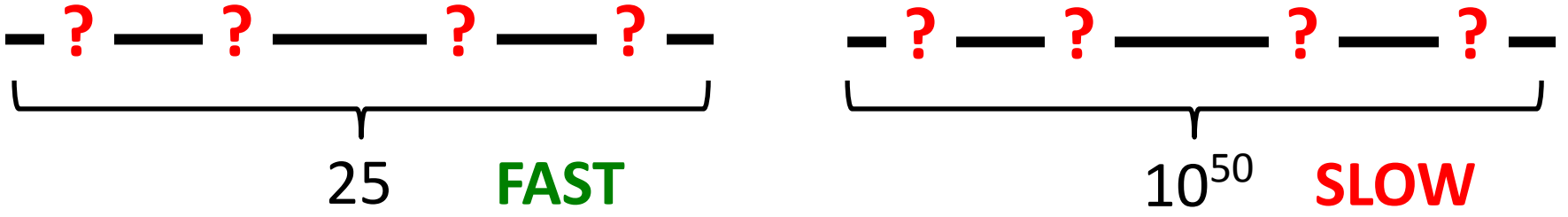


# Beltway and Turnpike Problems

No one has ever found a polynomial algorithm for either the Beltway or Turnpike Problem.

However, the Turnpike Problem does have a **pseudo-polynomial** algorithm, which is polynomial in the length of the *segment*.



# Beltway and Turnpike Problems

**Question:** Can you find a pseudo-polynomial algorithm for the Beltway Problem?

# Toward a Computational Problem

**Theoretical spectrum**: mass of *every possible* subpeptide, plus 0 and the mass of the peptide.

**Peptide**

**NQEL**

**Spectrum**

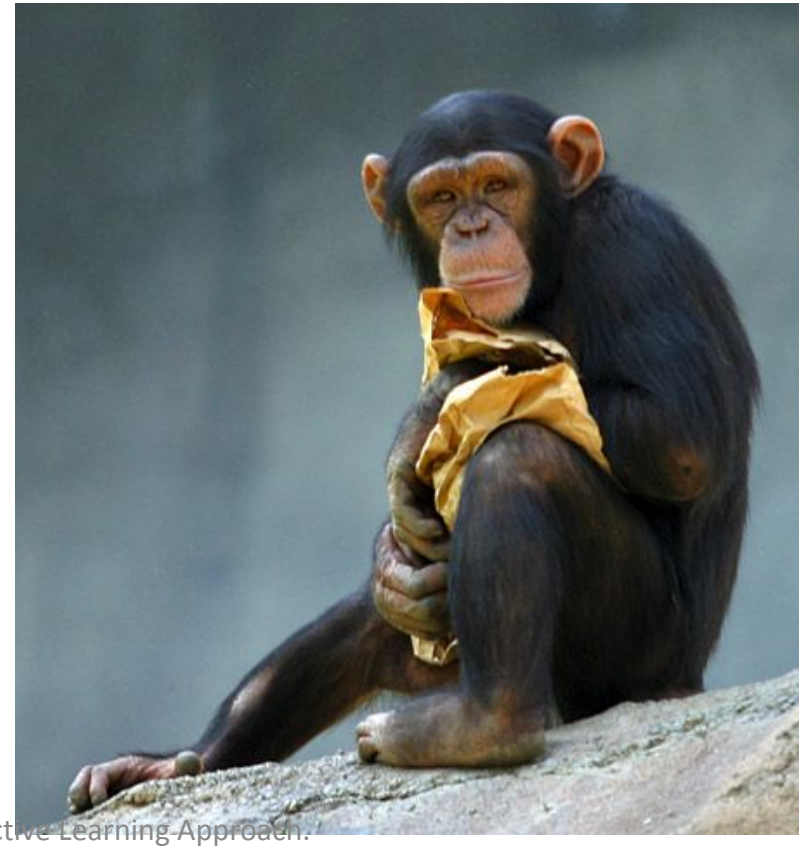
| Subpeptide | Mass |
|---|---|
| L | **113** |
| N | **114** |
| Q | **128** |
| E | **129** |
| LN | **227** |
| NQ | **242** |
| EL | **242** |
| QE | **257** |
| LNQ | **355** |
| ELN | **356** |
| QEL | **370** |
| NQE | **371** |
| NQEL | **484** |
| “” | **0** |

# Sequencing Cyclic Peptides in Primates

**θ-defensin:** cyclic peptide discovered in macaques (1999); has strong anti-HIV activity.

# Sequencing Cyclic Peptides in Primates

Humans and chimps don't make θ-defensin!

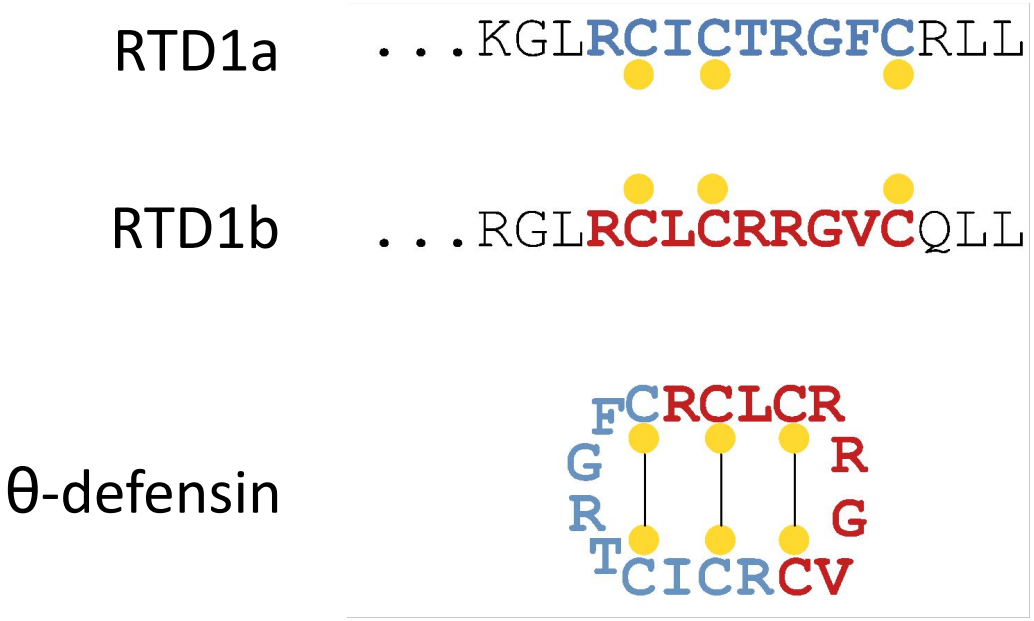




Courtesy: Aaron Logan

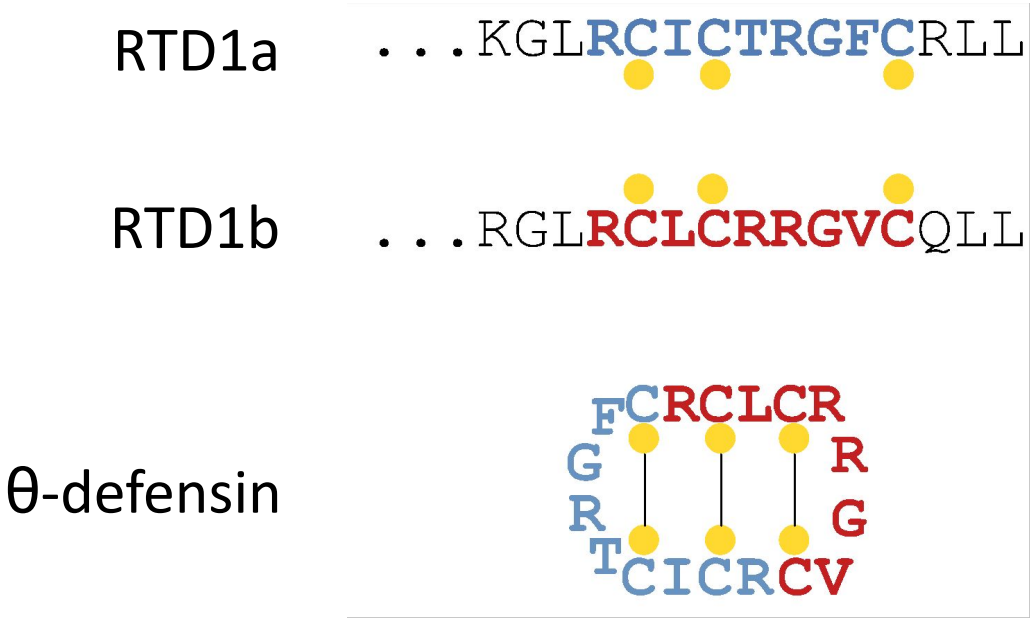
# Sequencing Cyclic Peptides in Primates

θ-defensin is formed from two proteins encoded by the RTD1a and RTD1b genes, which we lack…



# Sequencing Cyclic Peptides in Primates

θ-defensin is formed from two proteins encoded by the RTD1a and RTD1b genes, which we lack…




…but humans do have very similar genes!

# Sequencing Cyclic Peptides in Primates

A mutation occurred in the human-chimp ancestor, creating a premature stop codon.

# Sequencing Cyclic Peptides in Primates

A mutation occurred in the human-chimp ancestor, creating a premature stop codon.

Can we get θ-defensin back?

# Sequencing Cyclic Peptides in Primates

A mutation occurred in the human-chimp ancestor, creating a premature stop codon.

Can we get θ-defensin back? **YES!**

# Sequencing Cyclic Peptides in Primates

A mutation occurred in the human-chimp ancestor, creating a premature stop codon.

Can we get θ-defensin back? **YES!**

We still have the "cut-and-paste" enzymes needed to create θ-defensin. *But why?*

# Sequencing Cyclic Peptides in Primates

If the enzymes needed for θ-defensin aren't used, they would erode into "pseudogenes"…

# Sequencing Cyclic Peptides in Primates

If the enzymes needed for θ-defensin aren't used, they would erode into "pseudogenes"…

**…so why do we have these enzymes?**

# Sequencing Cyclic Peptides in Primates

**Current paradigm**: humans don't produce cyclic peptides.

But maybe, like antibiotics, they've been there all along, waiting to be discovered…

**Question:** Do humans produce cyclic peptides?