Sequencing of Antibiotics

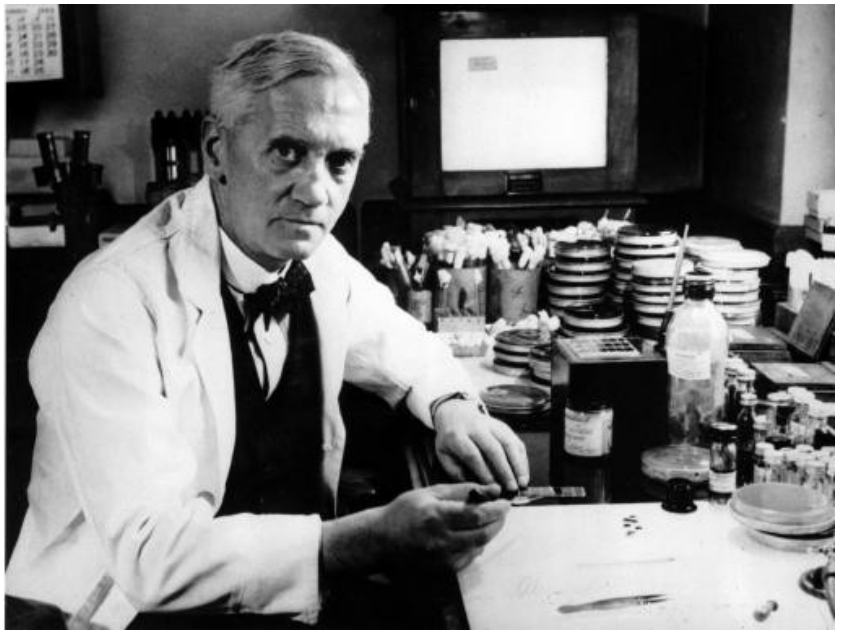
Chapter 2, “How do we sequence antibiotics?”

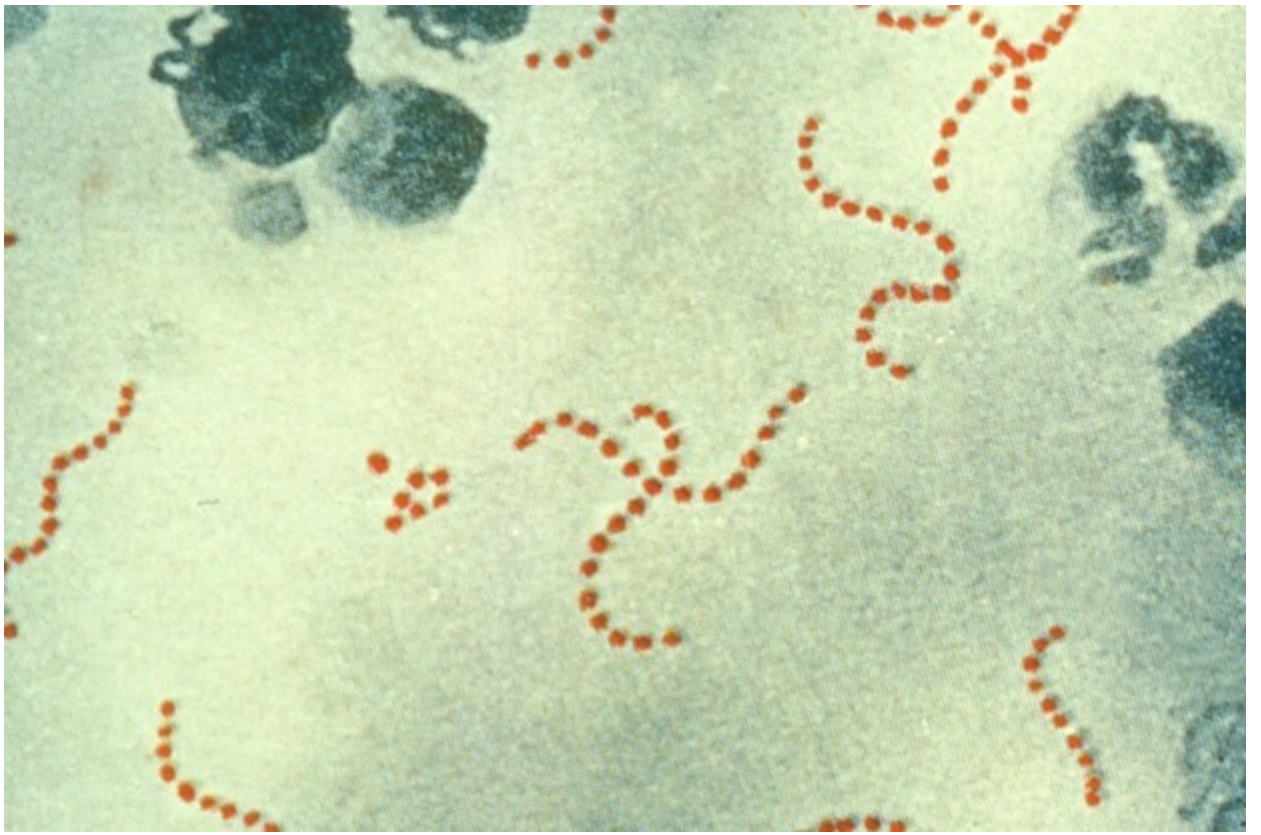
Compaeu, Pevzner, Bioinformatics Algorithms

Discovery of Penicillin (1928)

- Fights syphilis, infections caused by *staphylococci* and *streptococci*

Alexander Fleming

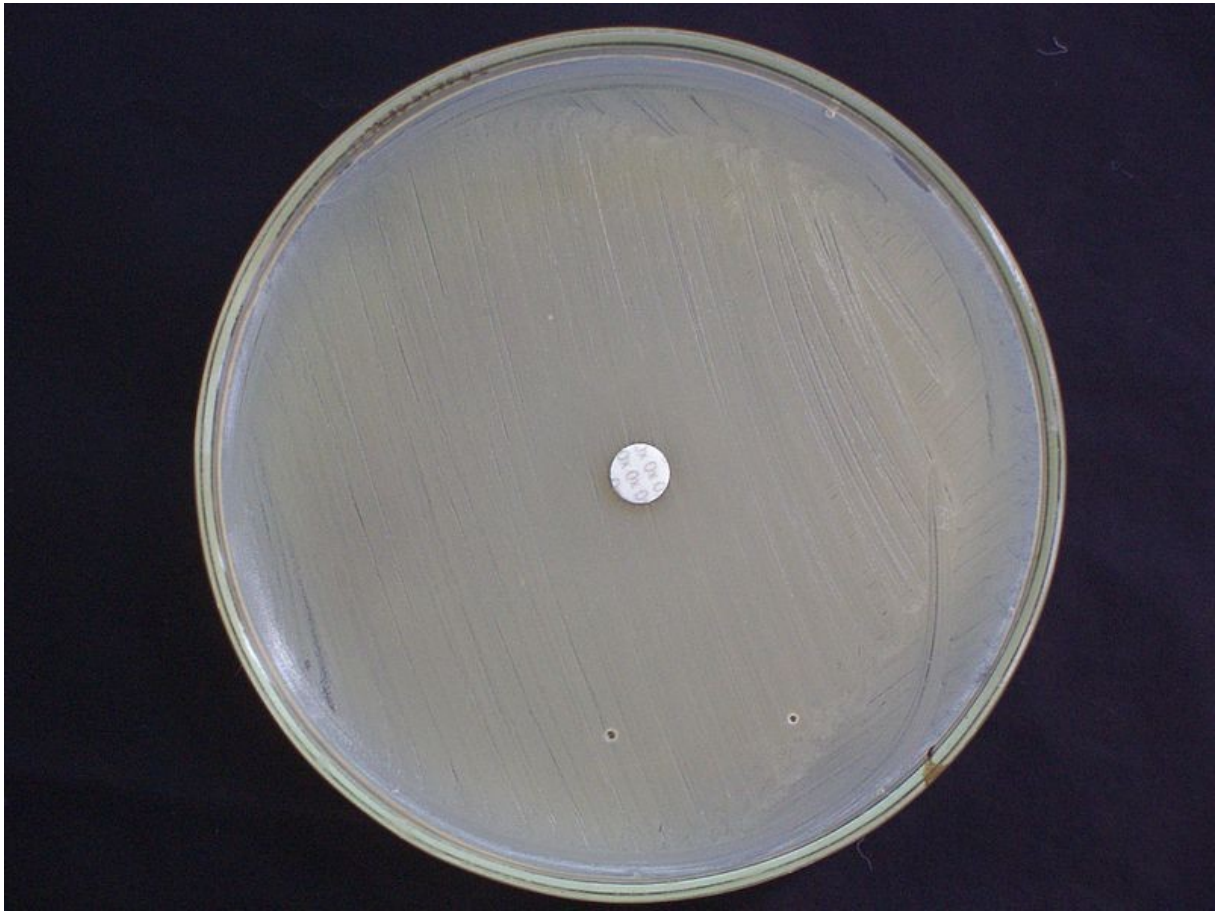




Streptococcus pyogenes

Antibiotic Resistant Bacteria

- Methicillin-resistant Staphylococcus aureus (MRSA)
  - Has developed resistance to penicillins through natural selection



Antibiotics

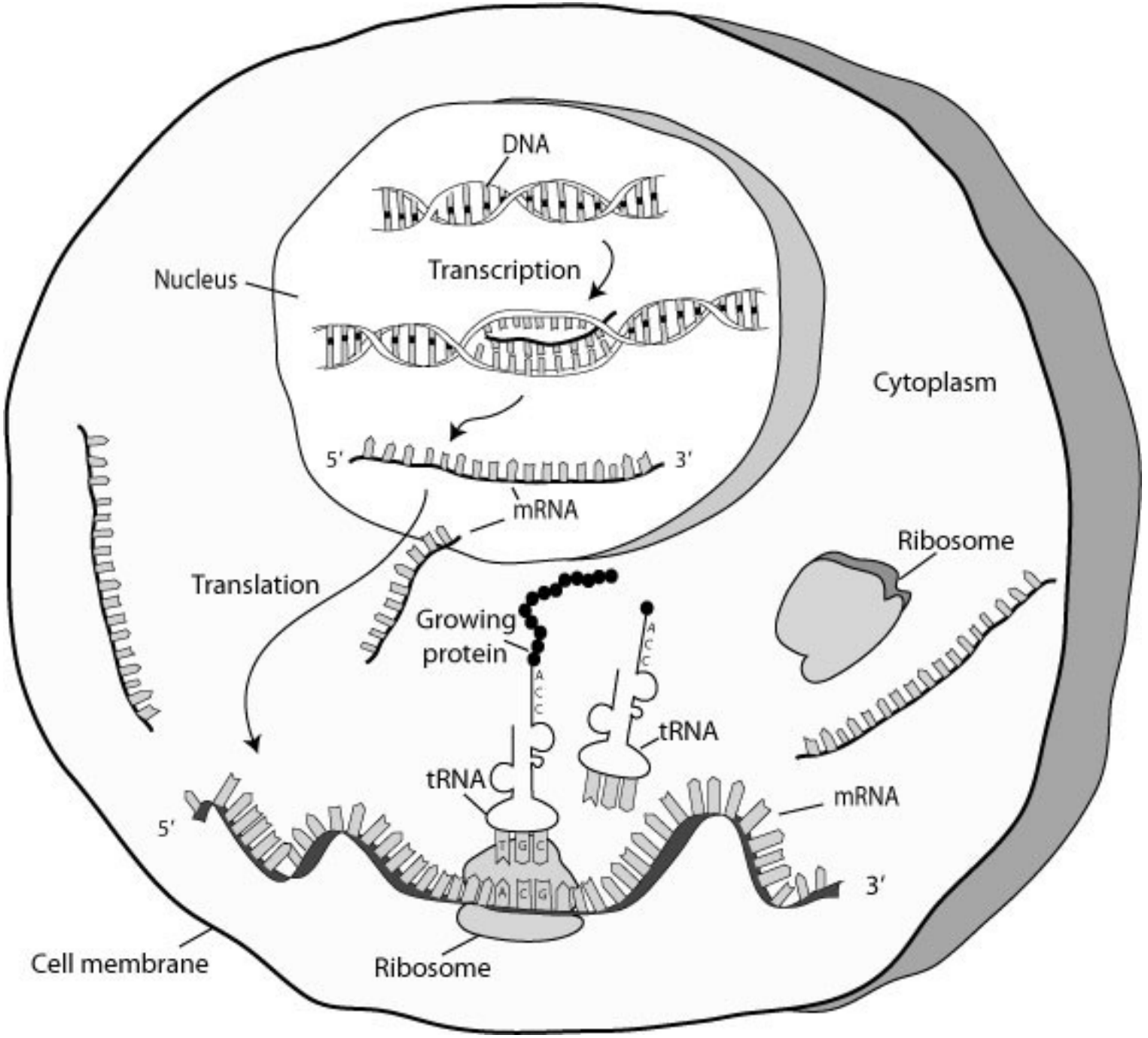
- Substances that kill bacteria
  - Penicilin mould (fungi), other bacteria
- Eg. Tyrocidine B1 produced by *Bacillus brevis*
- Tyrocidine B1 is a “mini-protein” - a short string of amino acids (peptide)

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

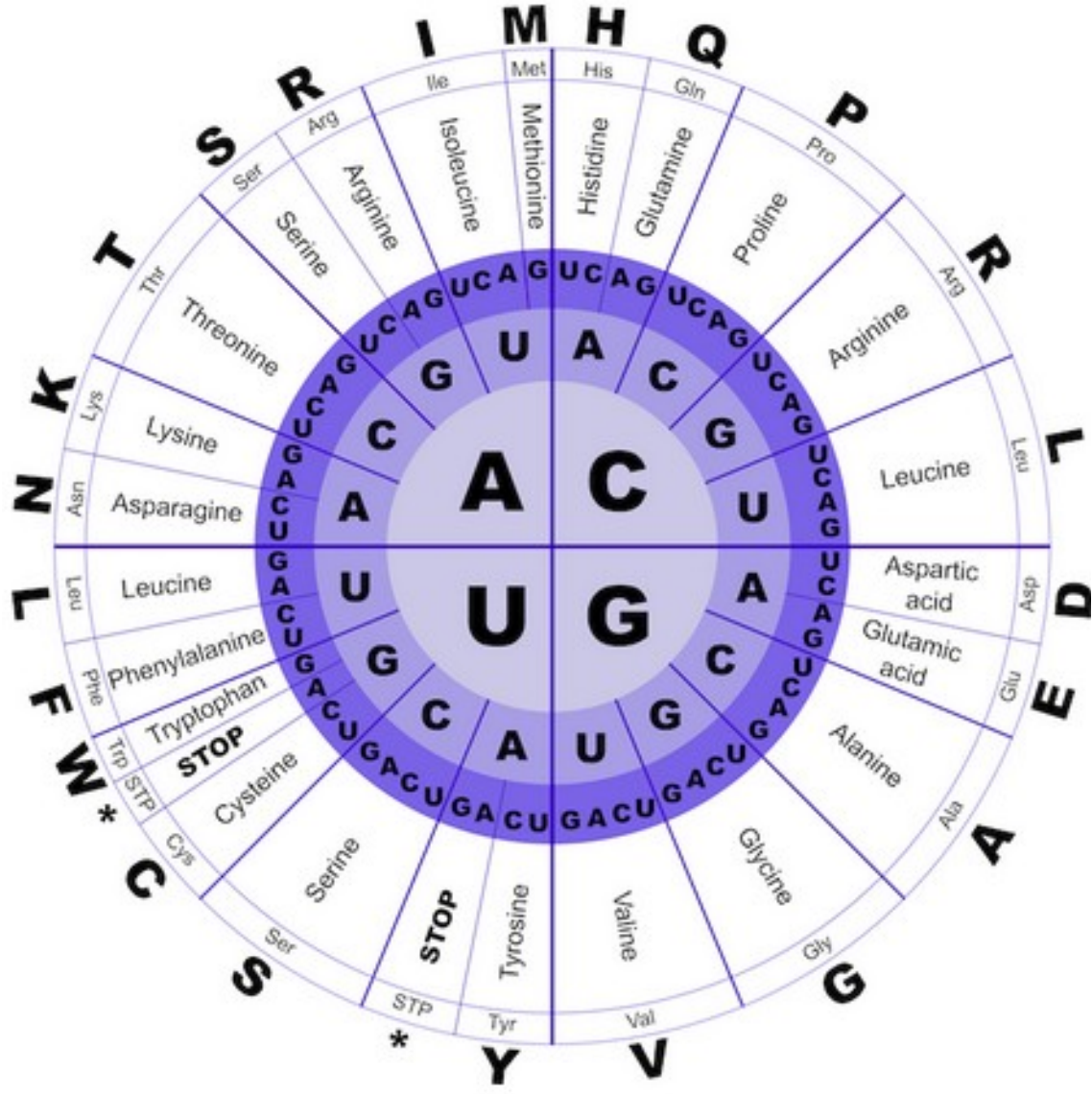
How does *Bacillus brevis* produce Tyrocidine B1?

The Central Dogma of Biology

The Central Dogma of Molecular Biology



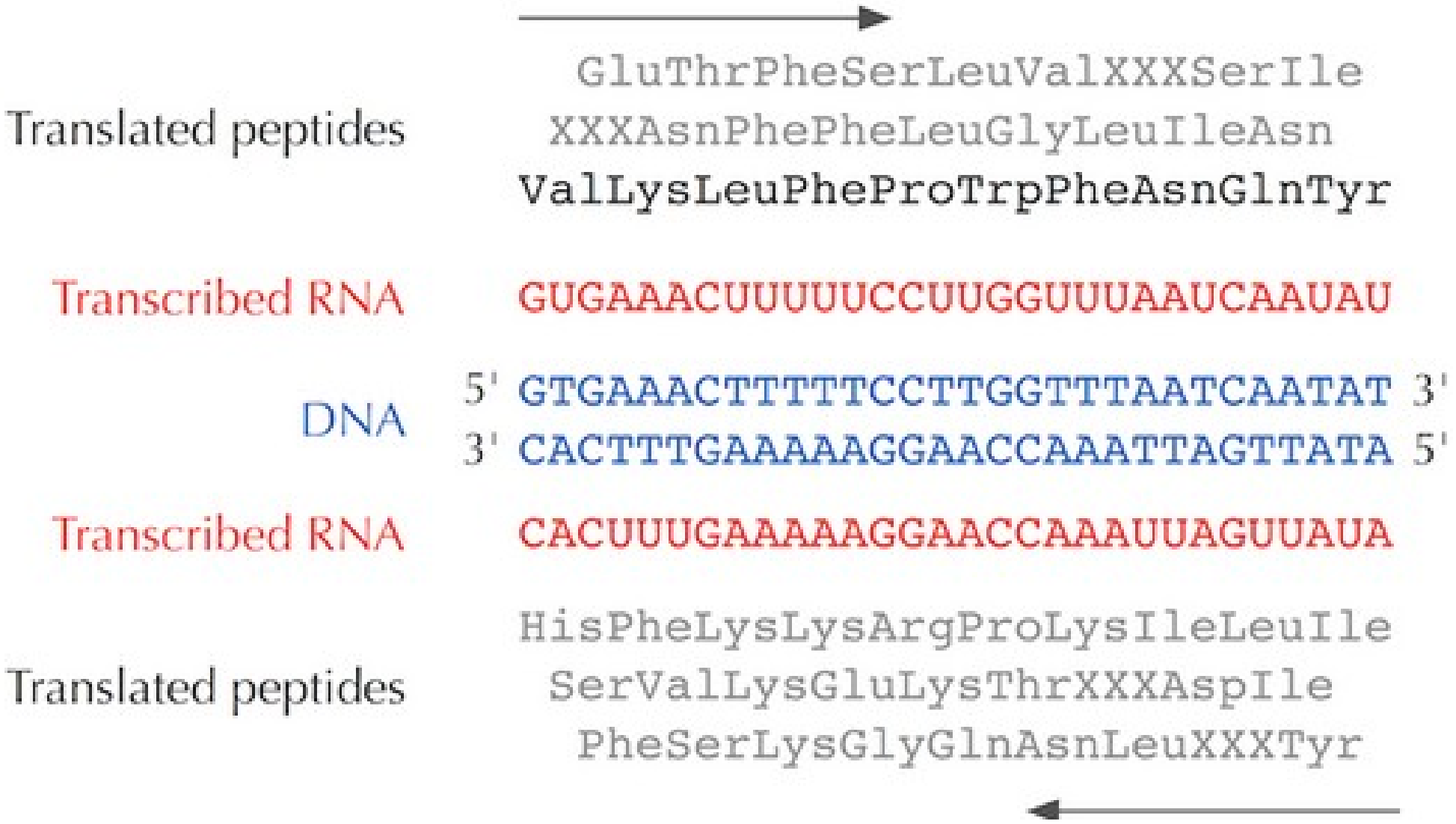
Protein → DNA



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

In how many ways can Tyrocidine B1 be encoded?

Example Solution



Peptide Encoding Problem

Find substrings of a genome encoding a given amino acid sequence.

Input: A DNA string *Text* and an amino acid string *Peptide.*

Output: All substrings of *Text* encoding *Peptide* (if any such substrings exist).

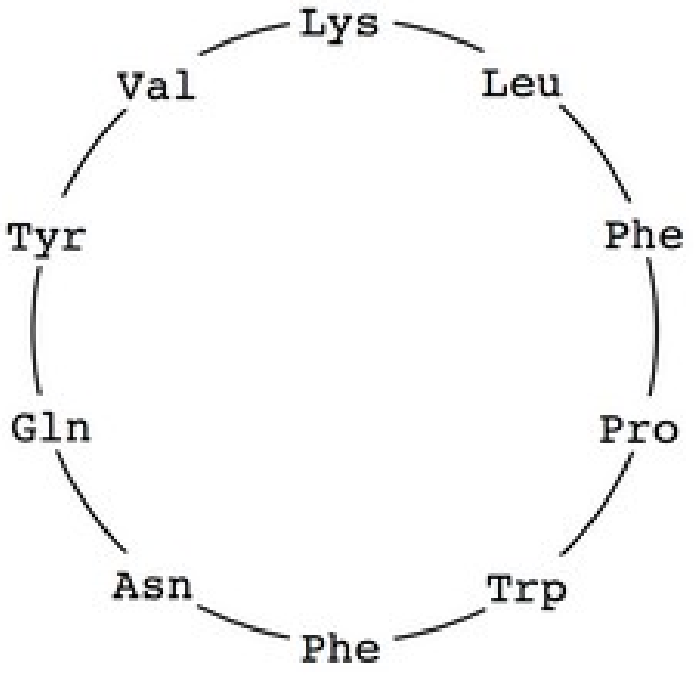
Peptide Encoding Problem

Find substrings of a genome encoding a given amino acid sequence.

Input: A DNA string *Text* and an amino acid string *Peptide.*

Output: All substrings of *Text* encoding *Peptide* (if any such substrings exist).

Structure of Tyrocidine B1



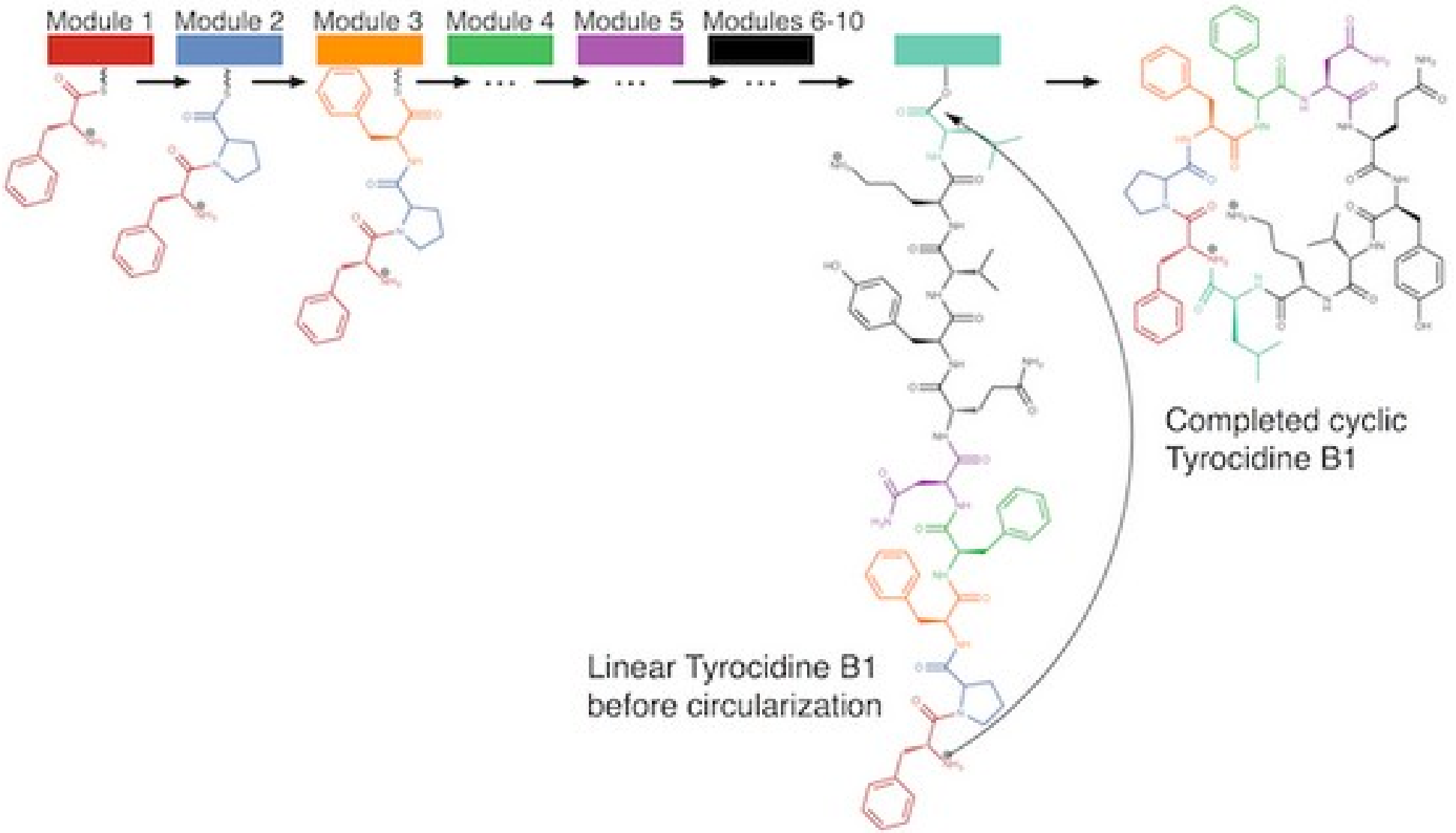
1 Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr
2 Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr-Val
3 Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr-Val-Lys
4 Phe-Pro-Trp-Phe-Asn-Gln-Tyr-Val-Lys-Leu
5 Pro-Trp-Phe-Asn-Gln-Tyr-Val-Lys-Leu-Phe
6 Trp-Phe-Asn-Gln-Tyr-Val-Lys-Leu-Phe-Pro
7 Phe-Asn-Gln-Tyr-Val-Lys-Leu-Phe-Pro-Trp
8 Asn-Gln-Tyr-Val-Lys-Leu-Phe-Pro-Trp-Phe
9 Gln-Tyr-Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn
10 Tyr-Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln

Tyrocidine B1 is a cyclic peptide (left), and so it has ten different linear representations (right).

DNA sequences needed to encode Tyrocidine B1 is not present in *Bacillus brevis* genome*!*

Non Ribosomal Peptides (NRP)

- Edward Tatum (1963) and Fritz Lipmann (1969)



NRP

- Molecular bullets used by bacteria and fungi
- Produce antibiotics, anti-tumour agents, immunosuppressors, and inter-bacteria communication

The central dogma of Biology is not applicable to NRP

The Mass Spectrometer

Dalton

- Mass of a proton or a neutron
- Mass of a molecule is the sum of its protons/neutrons
- Mass of Glycine (C_2H_3ON) ≈ 12 · 2 + 1 · 3 + 16 · 1 + 14 · 1 ≈ **57 Da**
- Actual Mass: 57.02 Da

Integer Mass Table

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

Mass of Tyrocidine B1 = 1322

Integer Mass Table

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

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

The Mass Spectrometer

NQEL
x 10^6

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

The collection of all the fragment masses generated by the mass spectrometer is called an **experimental spectrum**.

Theoretical Spectrum

The **theoretical spectrum** of a cyclic peptide *Peptide*, denoted *Cyclospectrum*(*Peptide*), is the collection of all of the masses of its subpeptides

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

Includes 0 and the mass of the entire peptide.
Includes possible duplicates.

Theoretical Spectrum Problem

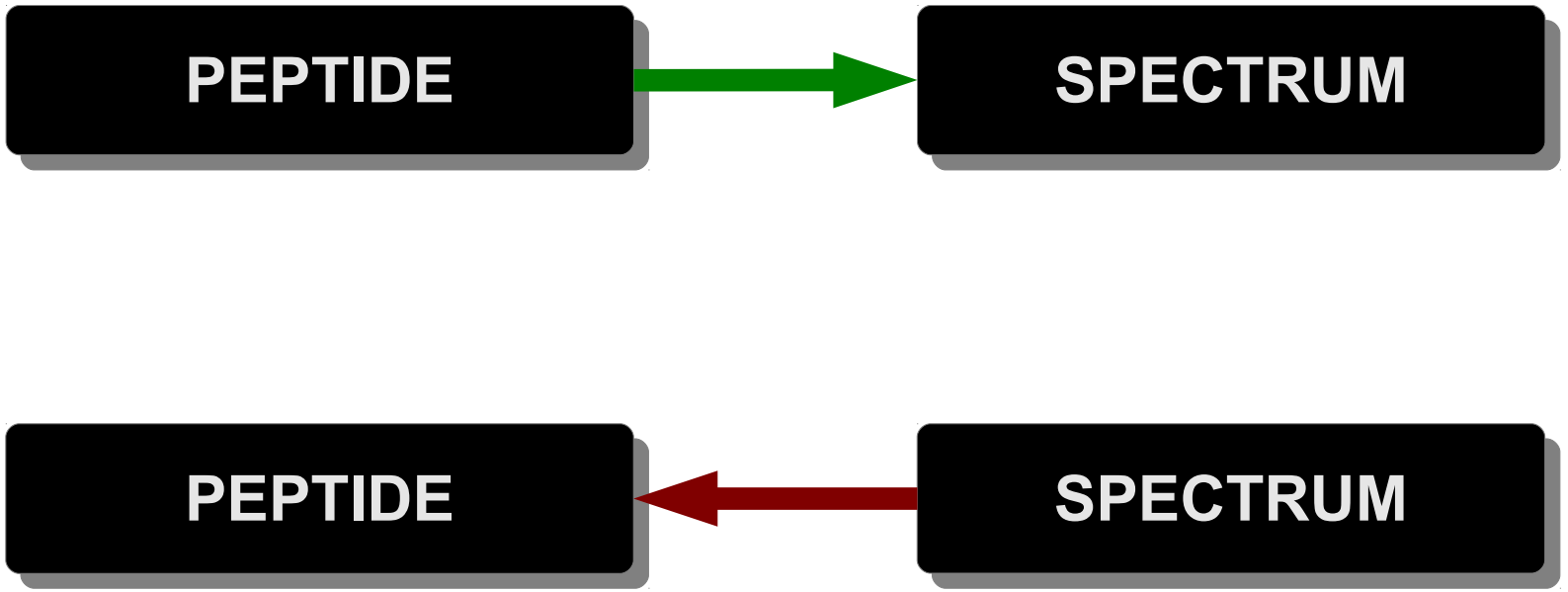
Generate the theoretical spectrum of a cyclic peptide.

Input: An amino acid string *Peptide*.

Output: *Cyclospectrum*(*Peptide*).

How many subpeptides does a cyclic peptide of length *n* have?

The Computational Problem



Cyclopeptide Sequencing Problem

Given an ideal experimental spectrum, find a cyclic peptide whose theoretical spectrum matches the experimental spectrum.

Input: A collection of (possibly repeated) integers *Spectrum* corresponding to an ideal experimental spectrum.

Output: An amino acid string *Peptide* such that *Cyclospectrum*(*Peptide*) = *Spectrum* (if such a string exists).

Brute Force Cyclopeptide Sequencing

- The mass of the entire peptide is usually known

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

Brute Force Algorithm: “Try all” candidate solutions of a given kind.

There are > 10^{13} candidates!

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

Integer Mass Table

G	A	S	P	V	T	C	I/L	N
57	71	87	97	99	101	103	113	114

D	K/Q	E	M	N	F	R	Y	W
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

Integer Mass Table

G	A	S	P	V	T	C	I/L	N
57	71	87	97	99	101	103	113	114

D	K/Q	E	M	N	F	R	Y	W
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

Integer Mass Table

G	A	S	P	V	T	C	I/L	N
57	71	87	97	99	101	103	113	114

D	K/Q	E	M	N	F	R	Y	W
115	128	129	131	137	147	156	163	186

Start building the tree with P, V, T, C as candidate solutions.

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

Start building the tree with P, V, T, C as candidate solutions.

Enumerate all possible 2-mers starting from the 4 candidates

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

PV is consistent with the Spectrum

Mass(P) = 97

Mass(V) = 99

Mass(PV) = 196

Save consistent 2-mers, repeat ...

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

The final cyclic peptide: PTPVC

From Theoretical to Noisy Spectra

Experimental spectra often produce errors

Consider the following spectra with NQEL includes **Missing** and **False** **Masses**

Theoretical →

0		113	114	128	129	227	242	242	257		355	356	370	371	484
0	99	113	114	128		227			257	299	355	356	370	371	484

Experimental →

Score a peptide on how many masses its spectrum shares with the experimental spectrum

Cyclopeptide Sequencing Problem (for noisy spectra)

Find a cyclic peptide having maximum score against an experimental spectrum.

Input: A collection of integers *Spectrum*.

Output: A cyclic peptide *Peptide* maximizing *Score*(*Peptide*, *Spectrum*) over all peptides *Peptide* such that *Mass*(*Peptide*) is equal to *ParentMass*(*Spectrum*).

Cyclopeptide Sequencing Problem (for noisy spectra)

Theoretical →

0		113	114	128	129	227	242	242	257		355	356	370	371	484
0	99	113	114	128		227			257	299	355	356	370	371	484

Experimental →

Score a peptide on how many masses its spectrum shares with the experimental spectrum

Instead of checking for consistency with the spectrum, Peptides are bound based on their scores. (Top N with matches are retained)

LeaderBoard Cyclic Peptide Sequencing

1. Add *0-peptide* to the LeaderBoard
2. Extend the LeaderBoard with the 18 amino acids
3. Cut low scoring peptides from the LeaderBoard (keep top N with ties)
4. Update *LeaderPeptide* from the LeaderBoard if its *mass* = *parent mass*
5. Eliminate peptides from LeaderBoard if mass > parent mass
6. Iterate (2 → 5) till LeaderBoard is empty
7. Return *LeaderPeptide*

Testing on a Tyrocidine B1 Spectrum

The algorithm is a heuristic. It sacrifices precision for speed.

Spectrum$_{10}$: 10% false and 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	1039	1060	1061	1062	1078	1080	1081	1095
1136	1159	1175	1175	1194	1194	1208	1209	1223	1225	1322			

Testing on a Tyrocidine B1 Spectrum

The algorithm is a heuristic. It sacrifices precision for speed.

Spectrum$_{10}$: 10% false and 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		1039	1060	1061	1062	1078	1080	1081	1095
1136	1159	1175	1175	1194	1194	1208	1209	1223	1225	1322			

Highest scoringpeptide: VKLFPWFNQY

A Noisier Tyrocidine B1 Spectrum

Spectrum$_{25}$: 25% false and 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 and 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 scoringpeptide: VKLFPADFNQY

The 2nd highest scoringpeptide: VKLFPWFNQY.
mass A + D = mass W = 186

From 18 to 100+ Amino Acids

- NRPs contain non-standard amino acids because they are free from the Central Dogma
- Any integer between 57 – 200 is a valid mass.

Tyrocidine B

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

Back to the Noisy Spectra

Spectrum$_{10}$: 10% false and 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		1039	1060	1061	1062	1078	1080	1081	1095
1136	1159	1175	1175	1194	1194	1208	1209	1223	1225	1322			

Highest scoringpeptide: VKLFPWFN-98-65

Restricting the Amino Acid Alphabet

Goal: Reduce the number of amino acids to be considered

The **NQEL** experimental spectrum:

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

Mass (**E**) = **129**. It is not present in the spectrum. But, ...

Restricting the Amino Acid Alphabet

Goal: Reduce the number of amino acids to be considered

The **NQEL** experimental spectrum:

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

Mass (**E**) = **129**. It is not present in the spectrum. But, ...

Mass **(QE)** – Mass **(Q) = 257** – **128** = **129**.

Restricting the Amino Acid Alphabet

Goal: Reduce the number of amino acids to be considered

The **NQEL** experimental spectrum:

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

Mass (**E**) = **129**. It is not present in the spectrum. But, ...

Mass **(**QE**)** – Mass (Q) **= 257** – **128** = **129**.
Mass (**ELN**) – Mass (**LN**) = **356 – 227** = **129.**

Restricting the Amino Acid Alphabet

Goal: Reduce the number of amino acids to be considered

The **NQEL** experimental spectrum:

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

Mass (**E**) = **129**. It is not present in the spectrum. But, ...

Mass **(**QE**)** – Mass (Q) **= 257** – **128** = **129**.
Mass (ELN) – Mass (LN) = **356 – 227** = **129.**
Mass (NQEL) – Mass (LNQ) = **484 – 355 = 129.**

Spectral Convolution

Positive difference between every pair of masses in the spectrum

0		113	114	128	129	227	242	242	257		355	356	370	371	484
0	99	113	114	128		227			257	299	355	356	370	371	484

Spectral Convolution

	“” 0	false 99	L 113	N 114	Q 128	LN 227	QE 257	false 299	LNQ 355	ELN 356	QEL 370	NQE 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

Spectral Convolution

Most occuring masses in the range 57 – 200 in the spectral convolution:

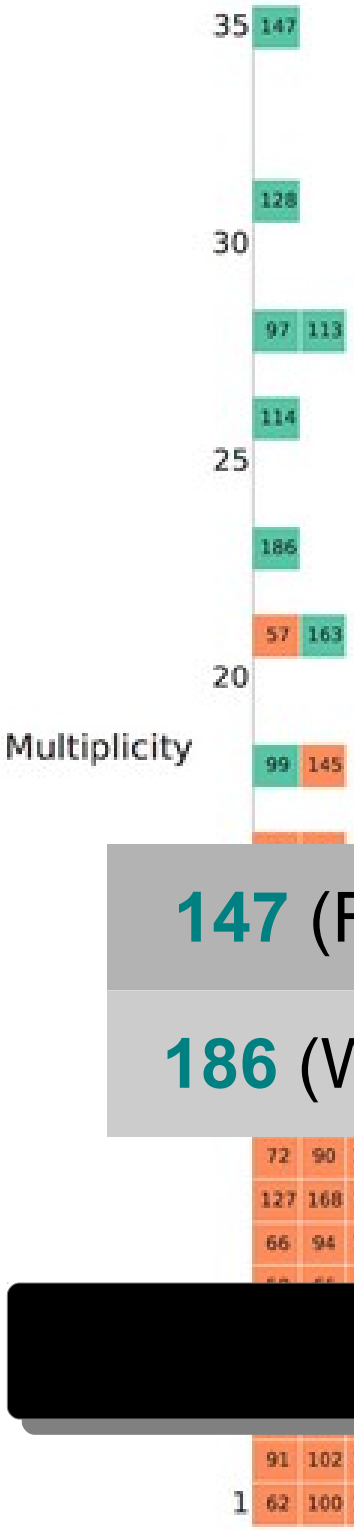
99 (V), 113 (L), 114 (N), 128 (Q), 129 (E).

4 out of the 5 amino acids (L, N, Q, E) occur in the peptide

ConvolutionCycloPeptideSequencing

1. Form Spectral Convolution of the spectrum
2. Take the most M frequent elements in the convolution (between 57 and 200)
3. Run *LeaderBoardCycloPeptideSequencing* forming peptides only on these *M* integers.

Tyrocidine B1 Spectrum$_{10}$



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

Winning peptide: VKLFPWFNQY

Tyrocidine B1 Spectrum$_{25}$

Winning peptide: VKLFPWFNQY

Truth about Spectra

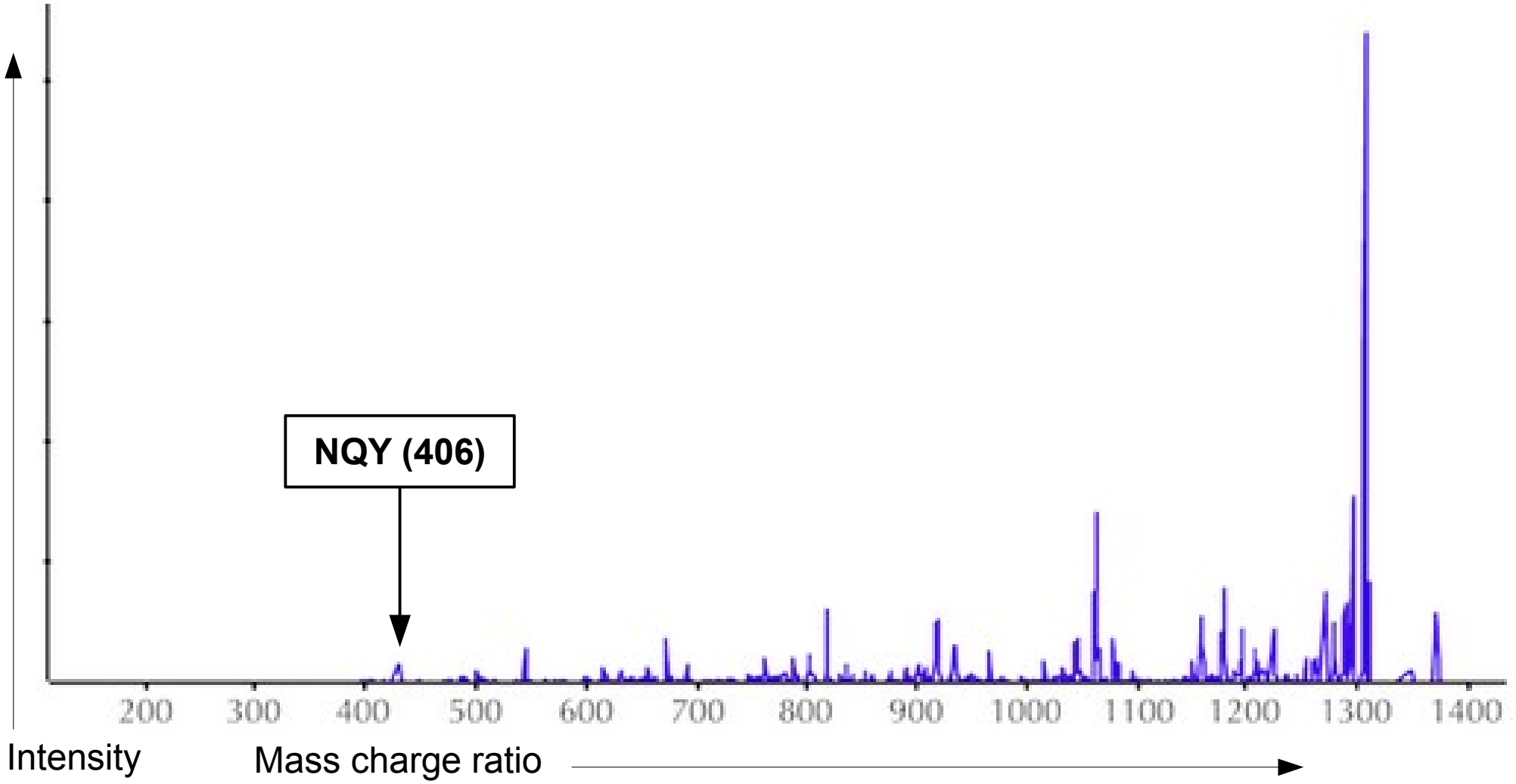
- Spectrum_{25} is much less noisy than the spectra obtained in practice
- The mass spectrometer doesn't “weigh” peptide fragments!

Truth about Spectra

- Ionize the peptide fragments
- Sort fragments using an electromagnetic field
- Measure **mass/charge ratio** of each fragment
- Determine **intensity** (# of ions) at each mass/charge ratio
- If fragment ion NQY (114 + 128 + 163 = 405) has charge +1, then it contains one additional proton, resulting in a total integer mass of 406
  - mass/charge ratio = 406/1 = 406.

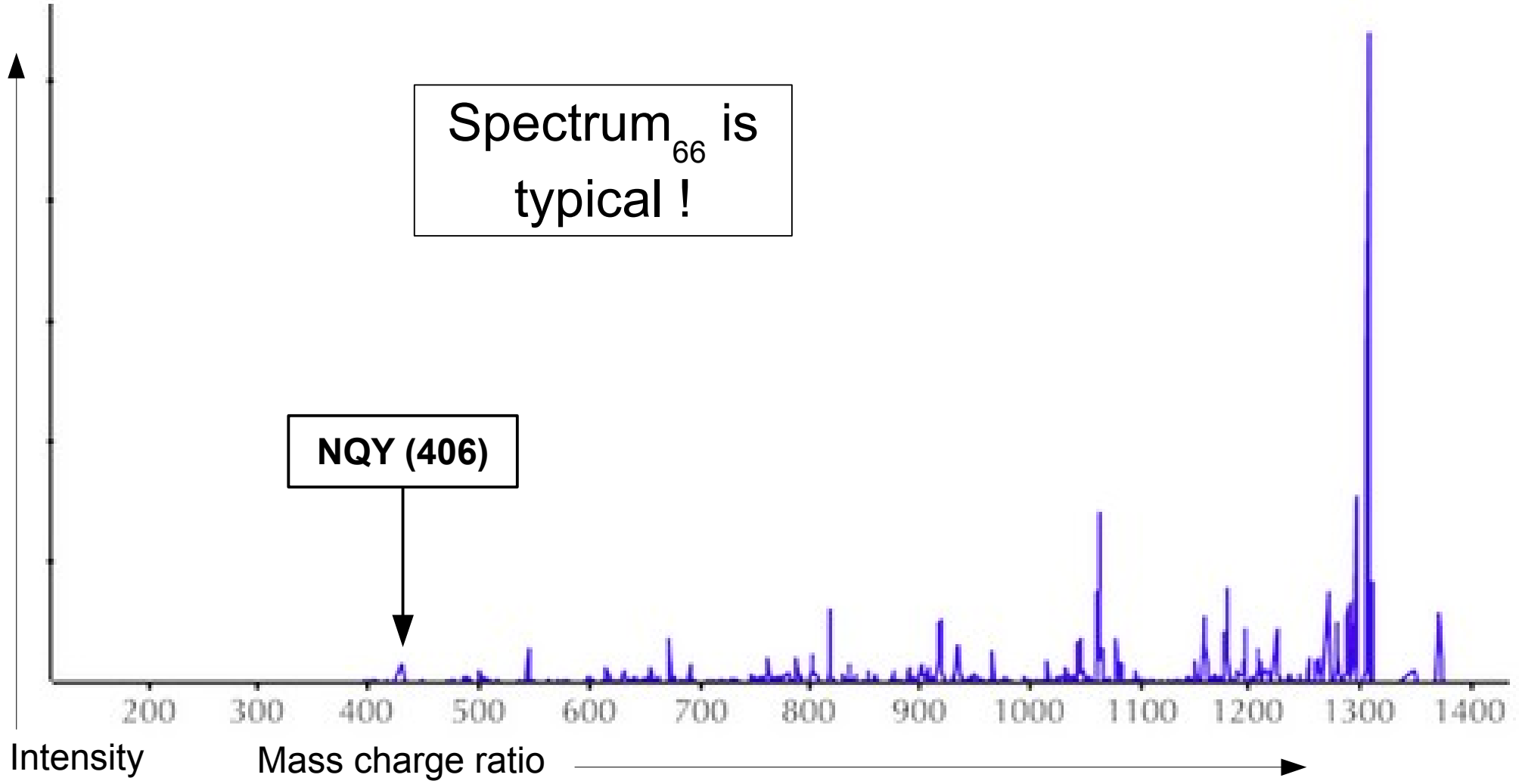
A Real Tyrocidine B1 Spectrum

Spectrum: Graph of Intensity vs. Mass/charge ratio



A Real Tyrocidine B1 Spectrum

Challenge: Reconstruct a peptide from real spectrum.

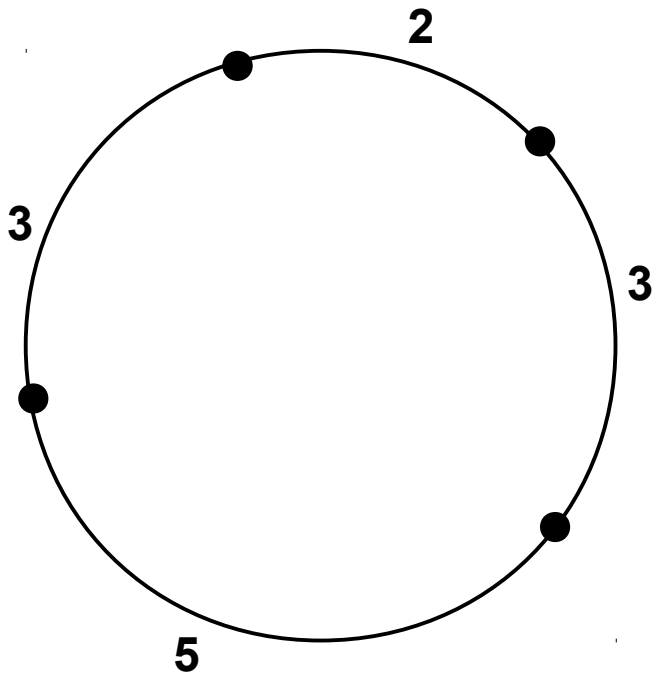


Open Problems

- Beltway and Turnpike problems
- Sequencing cyclic peptides in primates

Beltway and Turnpike Problems

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

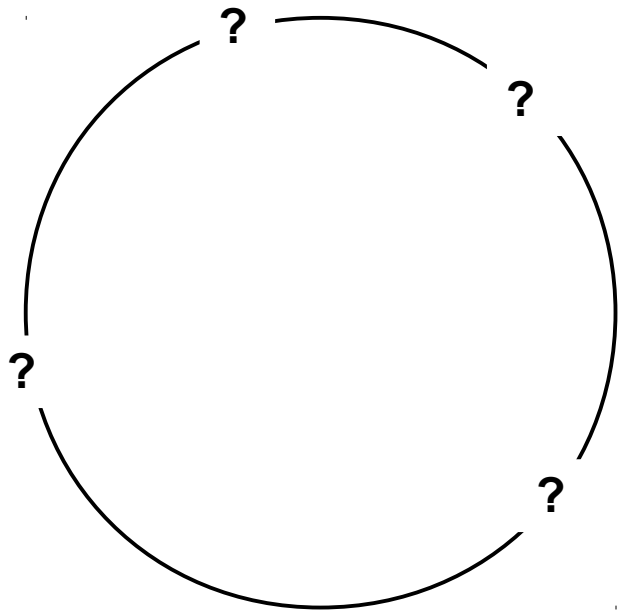


EASY →

2, 3, 5, 3
5, 8, 8, 5
10, 11, 10

Beltway and Turnpike Problems

- Given pairwise distances, reconstruct the points.



HARD

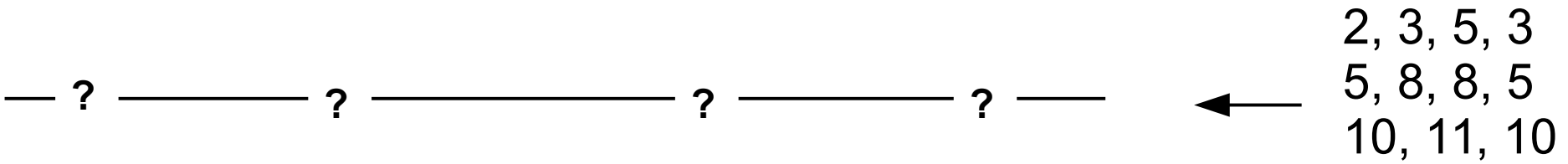
2, 3, 5, 3
5, 8, 8, 5
10, 11, 10

Beltway Problem

No polynomial solution has been found, yet.

Beltway and Turnpike Problems

- The points are on a line segment
- **Pseudo-polynomial** solution exists
  - Poly in the length of the segment



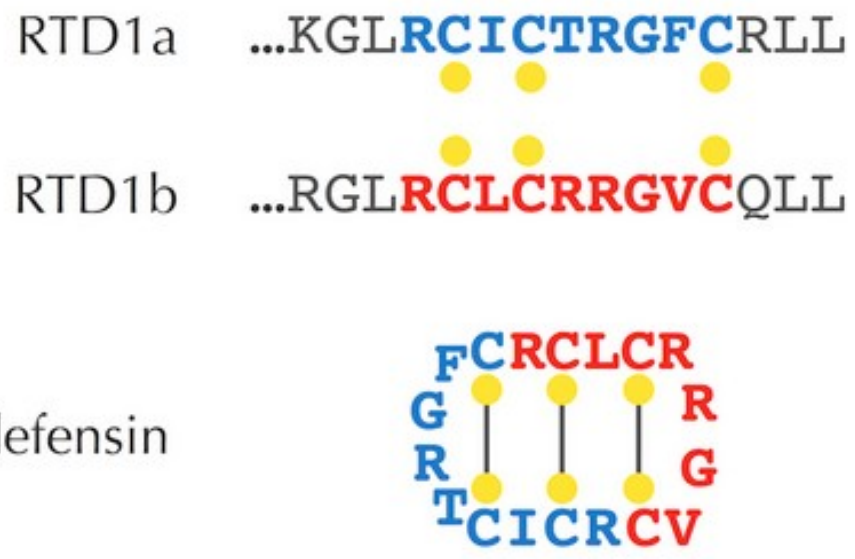
Turnpike Problem

Sequencing Cyclic Peptides in Primates

- **θ-defensin**: cyclic peptide discovered in macaques in 1999
- Has strong anti-HIV activity

Sequencing Cyclic Peptides in Primates

- Humans and chimps do not produce θ-defensin
- 2 proteins encoded by RTD1a and RTD1b genes
  - Humans do not have these genes



Humans have very similar genes in their bone marrow!

Sequencing Cyclic Peptides in Primates

- A mutation occured in the human, chip ancestor resulting a premature stop codon in the genes
- Humans have the cut-and-paste enzymes needed to create θ-defensin
  - Why?
- Current paradigm: Humans do not produce cyclic peptides.