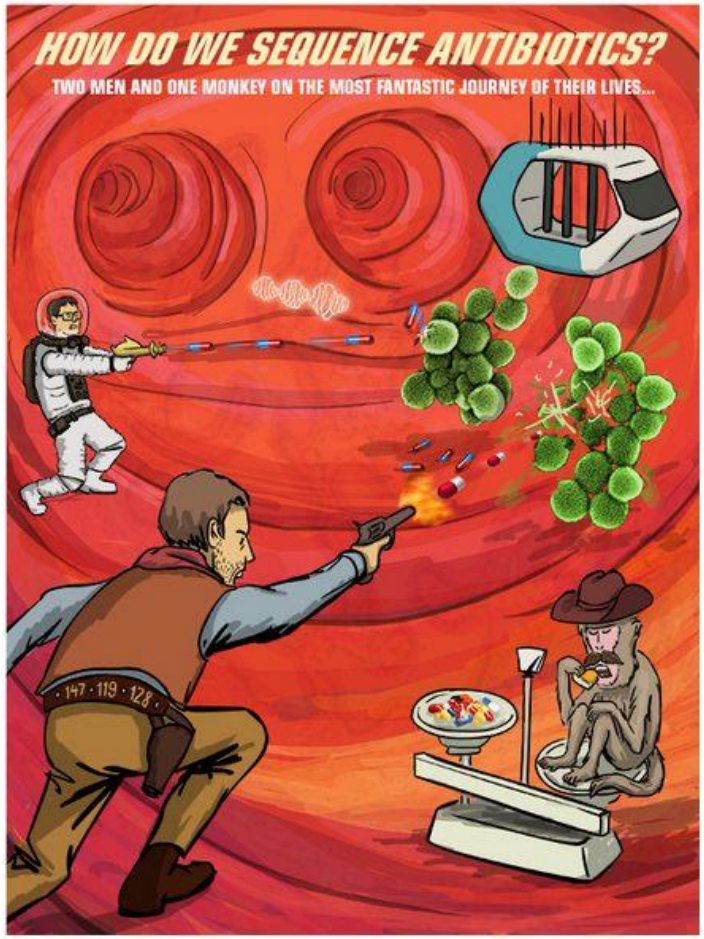
Stepic Full screen



Share Next step › Discussions

Bioinformatics Algorithms

In August 1928, before leaving for vacation, Scottish microbiologist Alexander Fleming stacked his cultures of infection-causing *Staphylococcus* bacteria on a laboratory bench. When he returned to work a few weeks later, Fleming noticed that one culture had been contaminated with *Penicillium* mold, and that the colony of *Staphylococcus* surrounding it had been destroyed! Fleming named the bacteria-killing substance **penicillin**, and he suggested that it could be used to treat bacterial infections in humans.

When Fleming published his discovery in 1929, his article had little immediate impact. Subsequent experiments struggled to isolate the **antibiotic** agent (i.e., the compound that actually killed bacteria) from the fungus. As a result, Fleming eventually concluded that penicillin could not be practically applied to treat bacterial infections and abandoned his antibiotics research.

Searching for new drugs to treat wounded soldiers, the American and British governments intensified their search for antibiotics after the start of World War II; however, the challenge of mass-producing antibiotics remained. In March 1942, half of the total supply of penicillin owned by pharmaceutical giant Merck was used to treat a single infected patient.

Also in 1942, Russian biologists Georgy Gause and Maria Brazhnikova noticed that the *Bacillus brevis* bacterium killed the pathogenic bacterium *Staphylococcus aureus*. In contrast to Fleming's efforts with penicillin, they successfully isolated the antibiotic compound from *Bacillus brevis* and named it *Gramicidin Soviet*. Within a year, this antibiotic was available in Soviet military hospitals.

Meanwhile, US scientists were scouring various food markets for rotten groceries and finally found a moldy cantaloupe in Illinois with the highest concentrations of penicillin. This mundane discovery allowed the United States to produce 2 million doses of penicillin in time for the Allied invasion of Normandy in 1944, thus saving thousands of wounded soldiers' lives.

Gause continued his research into *Gramicidin Soviet* after World War II but failed to elucidate its chemical structure. Taking the torch from Gause, English biochemist Richard Synge studied *Gramicidin Soviet* and a wide array of other antibiotics produced by *Bacillus brevis*. A few years after World War II ended, he demonstrated that they represent short amino acid sequences (i.e., mini-proteins) called **peptides**. Gause received the Stalin Prize in 1946, and Synge won the Nobel Prize in 1952. The former award proved more valuable as it protected Gause from execution during the period of **Lysenkoism**, the bloody Soviet campaign against "bourgeois" geneticists that intensified in the postwar era. See Detour: Gause and Lysenkoism.

The mass-production of antibiotics initiated an evolutionary arms race. Pharmaceutical companies worked to develop new antibiotic drugs, while pathogens acquired resistance against these drugs. Although modern medicine won every battle for six decades, the last ten years have witnessed an alarming rise in antibiotic-resistant bacterial infections that cannot be treated even by the most powerful antibiotics. In particular, the *Staphylococcus aureus* bacterium that Gause had studied in 1942 mutated into a resistant strain known as **Methicillin-resistant *Staphylococcus aureus*** (**MRSA**). MRSA is now the leading cause of death from infections in hospitals; its death rate has even passed that of AIDS in the United States.

With the rise of MRSA at hand, developing new antibiotics represents a central challenge to modern medicine. A difficult problem in antibiotics research is that of **sequencing** newly discovered antibiotics, or determining the order of amino acids making up the antibiotic peptide.

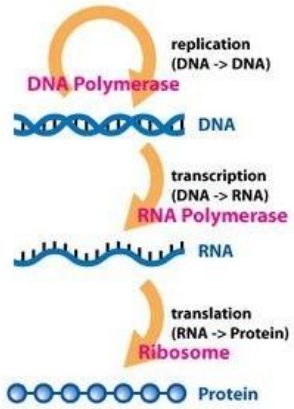
CENTRAL QUESTION: How do we sequence antibiotics?

Stepic Full screen

Let's begin by considering **Tyrocidine B1**, one of many antibiotics produced by *Bacillus brevis*. Tyrocidine B1 is defined by the 10 amino acid-long sequence shown below. Our goal in this section is to figure out how *Bacillus brevis* could have made this antibiotic.

```
Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr
 V   K   L   F   P   W   F   N   Q   Y
```

The **Central Dogma of Molecular Biology** states that "DNA makes RNA makes protein." According to the Central Dogma, a gene from a genome is first **transcribed** into a strand of RNA. You can think of the genome as a large cookbook, in which case the gene and its RNA transcript form a recipe in this cookbook. Then, the RNA transcript is **translated** into an amino acid sequence of a protein.

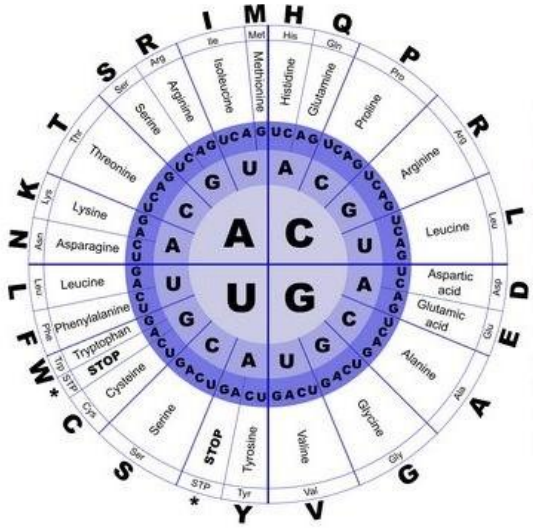


DNA replicates with the help of DNA polymerase, as we saw in the first chapter. DNA also serves as a template for transcription by RNA polymerase into single-stranded RNA, which is then translated by ribosomes into proteins. Courtesy Daniel Horspool.

Share Next step > Discussions

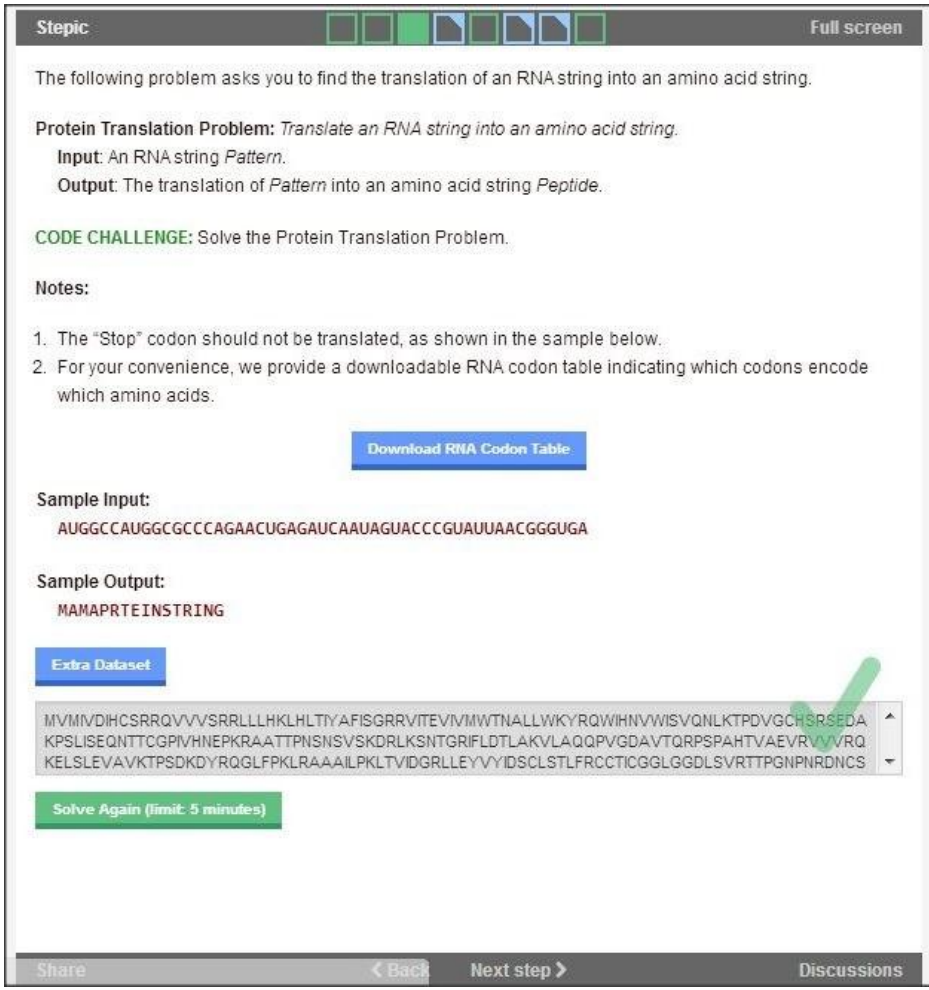
Bioinformatics Algorithms

Much like replication, the chemical machinery underlying transcription and translation is fascinating, but from a computational perspective, both processes are straightforward. Transcription simply transforms a DNA string into an RNA string by replacing all occurrences of `T` with `U`. The resulting strand of RNA is translated into an amino acid sequence via the **genetic code**; this process converts each 3-mer of RNA, called a **codon**, into one of 20 amino acids. As illustrated in the figure below, each of the 64 RNA codons encodes its own amino acid (some codons encode the same amino acid), with the exception of three **stop codons** that do not translate into amino acids and serve to halt translation (see DETOUR: Discovery of Codons). For example, the DNA string `TATACGAAA` transcribes into the RNA string `UAUACGAAA`, which in turn translates into the amino acid string `Tyr-Thr-Lys`.



The genetic code describes the translation of an RNA 3-mer (codon) into one of 20 different amino acids. The first three circles, moving from the inside out, represent the 1st, 2nd, and 3rd nucleotides of a given codon. The 4th, 5th, and 6th circles define the translated amino acid in three ways: the amino acid's full name, its 3-letter abbreviation, and its single-letter abbreviation. Three of the 64 total RNA codons are stop codons, which halt translation and implicitly add a 21st stop symbol to the amino acid alphabet. Reproduced from Open Clip Art.

Bioinformatics Algorithms

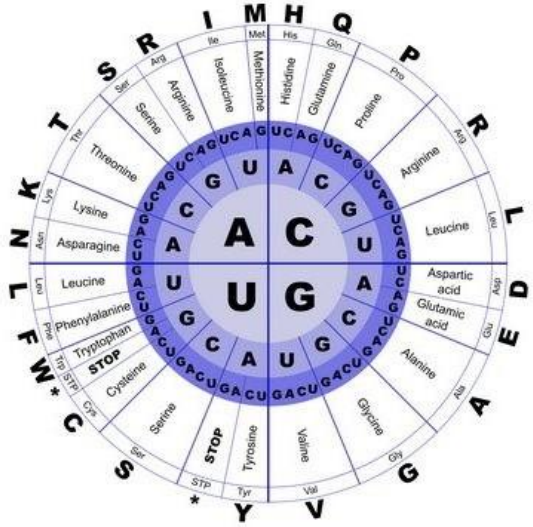


Bioinformatics Algorithms

Stepic Full screen

EXERCISE BREAK: How many DNA strings of length 30 transcribe and translate into Tyrocidine B1?

Recall that the amino acid sequence of Tyrocidine B1 is `Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr`. We also reproduce the genetic code figure below.



Start Quiz

Share Back Submit Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

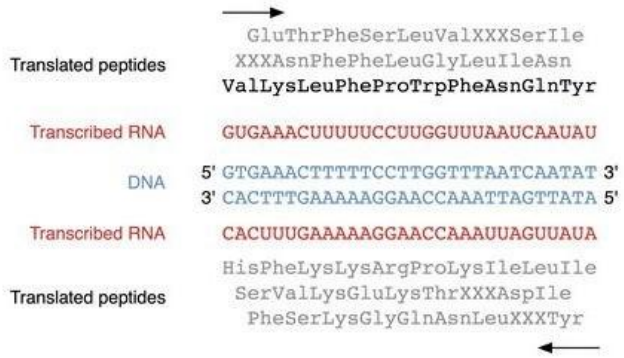
17. The Discovery of Antibiotics

18. How Do Bacteria Make Antibiotics?

19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. The Spectral Convolution Saves the Day
27. Epilogue: From Simulated to Real Spectra

Stepic Full screen

Thousands of different DNA 30-mers could code for Tyrocidine B1, and we would like to know which one appears in the *Bacillus brevis* genome. There are three different ways to divide a DNA string into codons for translation: one starts at position 1, another starts at position 2, and a third starts at position 3. These different ways of dividing a DNA string into codons are called **reading frames**. Since DNA is double-stranded, a genome has six reading frames (three on each strand), as shown in the figure below.



Six different reading frames give six different ways for the same fragment of DNA to be translated (three from each strand). Amino acid strings are read in the 5′ — 3′ direction; the highlighted amino acid string spells out the sequence of Tyrocidine B1. The stop codon is encoded by XXX.

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Stepic  Full screen

We say that a DNA string *Pattern* **encodes** an amino acid string *Peptide* if the RNA string transcribed from either *Pattern* or its reverse complement $\overline{Pattern}$ translates into Peptide.

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).

CODE CHALLENGE: Solve the Peptide Encoding Problem.

Sample Input:

```
ATGGCCATGGCCCCCAGAACTGAGATCAATAGTACCCGTATTAACGGGTGA
MA
```

Sample Output:

```
ATGGCC
GGCCAT
ATGGCC
```

Note: The solution may contain repeated strings if the same string occurs more than once as a substring of *Text* and encodes *Peptide*.

Extra Dataset

Start Quiz (limit: 5 minutes)

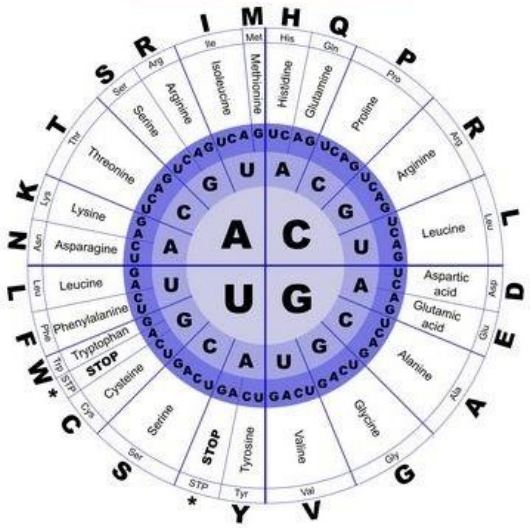
Share ❮ Back Submit Discussions

Bioinformatics Algorithms

Stepic Full screen

EXERCISE BREAK: Solve the Peptide Encoding Problem for *Bacillus brevis* and Tyrocidine B1 (`Val-Lys-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr`). How many starting positions in *Bacillus brevis* encode this peptide? (Genetic code figure reproduced below.)

Download *Bacillus brevis* Genome



Start Quiz

Share ❮ Back Submit Discussions

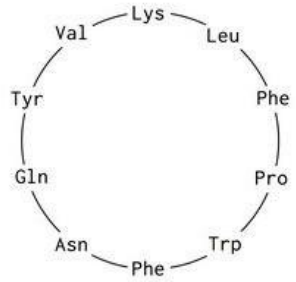
Bioinformatics Algorithms

Stepic Full screen

After solving the Peptide Encoding Problem for Tyrocidine B1, we should be able to find a 30-mer in the *Bacillus brevis* genome encoding Tyrocidine B1, and yet no such 30-mer exists!

STOP and Think: How could a bacterium produce a peptide that is not actually encoded by the bacterium's genome?

Neither Gause nor Synge was aware of it, but tyrocidines and gramicidins are actually **cyclic peptides**; the cyclic representation for Tyrocidine B1 is shown in the figure below.



Thus, Tyrocidine B1 has ten different linear representations, which are shown below. We should therefore run the Peptide Encoding Problem on every one of these sequences to find potential 30-mers coding for Tyrocidine B1.

```
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
```

Yet when we solve the Peptide Encoding Problem for each of the sequences above, we still find no 30-mer in the *Bacillus brevis* genome encoding Tyrocidine B1!

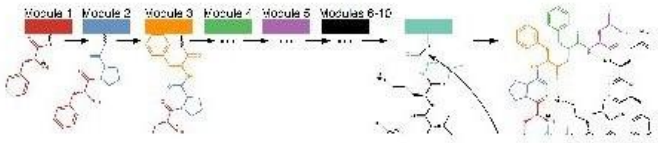
Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Stepic Full screen

Hopefully, you are perplexed, because the Central Dogma of Molecular Biology implies that *all* peptides must be encoded by the genome. Nobel laureate Edward Tatum was also confused, and in 1963, he devised an ingenious experiment. Protein translation is carried out by a molecular machine called a **ribosome**, and so Tatum reasoned that if he inhibited the ribosome, all protein production in *Bacillus brevis* should grind to a halt. To his amazement, all proteins did indeed shut down — except for tyrocidines and gramicidins! His experiment led Tatum to hypothesize that some yet unknown *non-ribosomal* mechanism must assemble these peptides.

In 1969, Fritz Lipmann (another Nobel laureate) demonstrated that tyrocidines and gramicidins are **non-ribosomal peptides** (**NRPs**), synthesized not by the ribosome, but by a giant protein called **NRP synthetase**. This enzyme pieces together antibiotic peptides without any reliance on RNA or the genetic code! We now know that every NRP synthetase assembles peptides by growing them one amino acid at a time, as shown in the figure below.



NRP synthetase is a giant multi-module protein that assembles a cyclic peptide in steps, one amino acid at a time. Each of ten different modules (shown by different colors) adds a single amino acid to the peptide, which in the figure is one of many tyrocidines produced by Bacilus brevis. In a final step, the peptide is circularized.

Share Next step > Discussions

Bioinformatics Algorithms

Stepic Full screen

The reason why many NRPs have pharmaceutical applications is that they have been optimized by eons of evolution as "molecular bullets" that bacteria and fungi use to kill their enemies. If these enemies happen to be pathogens, researchers are eager to borrow these bullets as antibacterial drugs. However, NRPs are not limited to antibiotics: many of them represent anti-tumor agents and immunosuppressors, while others are used by bacteria to communicate with other cells (see Detour: Quorum Sensing).

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Since NRPs do not adhere to the Central Dogma, any method we use to sequence them cannot rely on genome analysis, which brings us back to where we started. What makes sequencing these peptides even more difficult is that many NRPs (including tyrocidines and gramicidins) are cyclic. Thus, the standard tools for sequencing linear peptides, which we will describe in a follow-up chapter, are not applicable to NRP analysis.

The workhorse of peptide sequencing is the **mass spectrometer**, an expensive molecular scale that shatters molecules into pieces and then weighs the resulting fragments. The mass spectrometer measures the mass of a molecule in **daltons** (**Da**); 1 Da is approximately equal to the mass of a single nuclear particle (i.e., a proton or neutron). We will approximate the mass of a molecule by simply adding the number of protons and neutrons found in the molecule's constituent atoms, which yields the molecule's **integer mass**. For example, the amino acid `Gly`, which has chemical formula C_2H_3ON, has an integer mass of 57, since 2·12 + 3·1 + 1·16 + 1·14 = 57. Yet 1 Da is not exactly equal to the mass of a proton/neutron, and we may need to account for different naturally occurring isotopes of each atom when weighing a molecule (see DETOUR: Molecular Mass). As a result, amino acids typically have non-integer masses (e.g., `Gly` has total mass equal to approximately 57.02 Da); for simplicity, however, we will work with the integer mass table given below.

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

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

Download Table

Tyrocidine B1, which is represented by `VKLFPWFNQY` in the single-letter amino acid alphabet, has total mass 1322 Da (99 + 128 + 113 + 147 + 97 + 186 + 147 + 114 + 128 + 163 = 1322).

Stepic Full screen

The mass spectrometer can break Tyrocidine B1 into two different linear fragments, and it analyzes samples that may contain billions of identical copies of the peptide, with each copy breaking in its own way. One copy may break into `LFP` and `WFNQYVK` (with respective masses 357 and 965), whereas another may break into `PWFN` and `QYVKLF`. Our goal is to use the masses of these and other fragments to sequence the peptide. The collection of all the fragment masses generated by the mass spectrometer is called an **experimental spectrum**.

STOP and Think: How can we use the experimental spectrum to sequence a peptide?

For now, we will assume for simplicity that the mass spectrometer breaks the copies of a cyclic peptide at every possible two bonds, so that the resulting experimental spectrum contains the masses of all possible linear fragments of the peptide, which are called **subpeptides**. For example, the cyclic peptide `NQEL` has 12 subpeptides: `N`, `Q`, `E`, `L`, `NQ`, `QE`, `EL`, `LN`, `NQE`, `QEL`, `ELN`, and `LNQ`. Subpeptides may occur more than once if an amino acid occurs multiple times in the peptide (e.g., `ELEL` also has 12 subpeptides).

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

Sample Input:

```
31315
```

Sample Output:

```
980597910
```

Start Quiz (limit: 5 minutes)

Share Back Submit Discussions

Bioinformatics Algorithms

Stepic Full screen

The **theoretical spectrum** of a cyclic peptide *Peptide*, denoted *Cyclospectrum*(*Peptide*), is the collection of all of the masses of its subpeptides, in addition to the mass 0 and the mass of the entire peptide. Note that the theoretical spectrum may contain duplicate elements, as is the case for NQEL (shown below), where NQ and EL have the same mass.

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

Generating Theoretical Spectrum Problem: *Generate the theoretical spectrum of a cyclic peptide.*
Input: An amino acid string *Peptide*.
Output: *Cyclospectrum*(*Peptide*).

CODE CHALLENGE: Solve the Generating Theoretical Spectrum Problem.

Sample Input:

```
LEQN
```

Sample Output:

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

Extra Dataset

Start Quiz (limit: 5 minutes)

Share Back Submit Discussions

Bioinformatics Algorithms

Stepic Full screen

Generating the theoretical spectrum of a known peptide is easy, but our aim is to solve the reverse problem: we must reconstruct an unknown peptide from its experimental spectrum. We will start by assuming that the biologist is lucky enough to generate an **ideal experimental spectrum**, which is one coinciding with the peptide's theoretical spectrum. For example, consider the theoretical spectrum for Tyrocidine B1 shown below: if an experiment produced this spectrum, how would you reconstruct the amino acid sequence of Tyrocidine B1?

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

Download This Spectrum

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).

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

From now on, we will sometimes work directly with amino acid masses, taking the liberty to represent a peptide by a sequence of integers denoting the peptide's constituent amino acid masses. For example, we represent NQEL as 114-128-129-113 and Tyrocidine B1 (VKLFPWFNQY) as 99-128-113-147-97-186-147-114-128-163. We have therefore replaced an alphabet of 20 amino acids with an alphabet of only 18 integers because two amino acid pairs have the same integer 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

Note that in the general case (when we are not restricted to the amino acid alphabet), the Cyclopeptide Sequencing Problem could have multiple solutions. For example, "peptides" 1-1-3-3 and 1-2-1-4 have the same theoretical spectrum {1, 1, 2, 3, 3, 4, 4, 5, 5, 6, 7, 7}.

STOP and Think: We don't know whether there exist different peptides (in the alphabet of 18 amino acid masses) with identical theoretical spectra — can you find such peptides?

Note: This question may be difficult to answer.

Brute force (also known as **exhaustive search**) is a general problem-solving technique that enumerates all possible solution candidates and then checks whether each candidate satisfies the problem's statement. Such algorithms require little effort to design and are guaranteed to produce a correct solution, but they may take an enormous amount of time, and the number of solution candidates may be too large to check.

Let's design a straightforward brute force algorithm for the Cyclopeptide Sequencing Problem. We denote the total mass of an amino acid string *Peptide* as *Mass*(*Peptide*). In mass spectrometry experiments, whereas the peptide that generated *Spectrum* is unknown, the peptide's mass is typically known and is denoted *ParentMass*(*Spectrum*). Of course, given an ideal experimental spectrum, *Mass*(*Peptide*) is given by the largest mass in the spectrum. The brute force cyclopeptide sequencing algorithm BFCyclopeptideSequencing generates all possible peptides whose mass is equal to *ParentMass*(*Spectrum*) and then checks which of these peptides has theoretical spectra matching *Spectrum*.

```
BFCyclopeptideSequencing(Spectrum)
  for every peptide with Mass(Peptide) equal to ParentMass(Spectrum)
    if Spectrum = Cyclospectrum(Peptide)
      output Peptide
```

There should be no question that BFCyclopeptideSequencing will solve the Cyclopeptide Sequencing Problem. However, we should be concerned about its running time: how many peptides are there having mass equal to *ParentMass*(*Spectrum*)?

Stepic  Full screen

Counting Peptides with Given Mass Problem: *Compute the number of peptides of given total mass.*

Input: An integer *m*.

Output: The number of linear peptides having integer mass *m*.

CODE CHALLENGE: Solve the Counting Peptides with Given Mass Problem. Recall that we assume that peptides are formed from the following 18 amino acid masses:

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

Sample Input:

```
1024
```

Sample Output:

```
14712706211
```

Suggestion: If you have difficulty solving this problem or getting the runtime down, please return to it after learning more about dynamic programming algorithms in a later chapter.

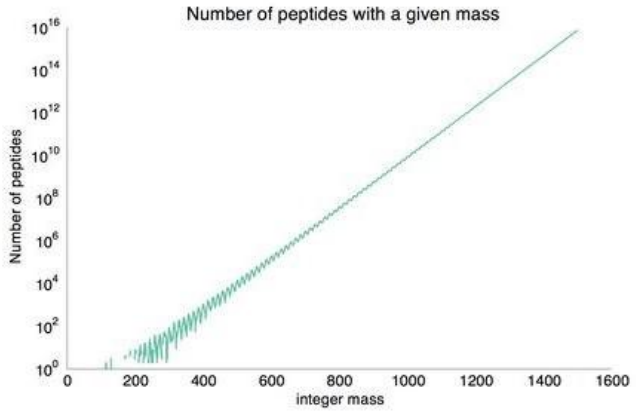
Start Quiz (limit: 5 minutes)

Share ‹ Back Submit Discussions

Bioinformatics Algorithms

Stepic Full screen

Solving the Counting Peptides with Given Mass Problem for mass equal to 1322 reveals that *trillions* of peptides have the same integer mass as Tyrocidine B1 (see the figure below). Therefore, BFCyclopeptideSequencing is completely impractical, and we will not even bother asking you to implement it.



The number of peptides with given mass grows exponentially.

EXERCISE BREAK: This figure suggests that for large m, the number of peptides with given integer mass m can be approximated as $k \cdot C^m$, where k and C are constants. Find C. (Give your answer as a decimal; the allowable error is 0.002).

Start Quiz

Share ❮ Back Submit Discussions

Bioinformatics Algorithms

Just because the algorithm from the previous section failed miserably does not mean that all brute force approaches are automatically doomed. Can we design a faster brute force algorithm based on a different idea?

Instead of checking all peptides with a given mass, our new approach to solving the Cyclopeptide Sequencing Problem works by slowly building candidate *linear* peptides whose theoretical spectra are "consistent" with the experimental spectrum. Given an experimental spectrum *Spectrum*, we will form a collection *List* of candidate linear peptides initially consisting of the **0-peptide**, which is just an empty string "" having mass 0. At the next step, we will expand *List* to contain all linear peptides of length 1. We continue this process, creating 18 new peptides of length $k + 1$ for each amino acid string *Peptide* of length k in *List* by appending every possible amino acid mass to the end of *Peptide*.

To prevent the number of candidate peptides from increasing exponentially, every time we expand *List*, we trim it by keeping only those linear peptides that remain consistent with the experimental spectrum. We then check to see if any of these new linear peptides, when circularized, provides a solution to the Cyclopeptide Sequencing Problem.

STOP and Think: What should it mean for a peptide to be consistent with the experimental spectrum?

More generally, brute force algorithms that enumerate all candidate solutions but discard large subsets of hopeless candidates by using various consistency conditions are known as **branch-and-bound algorithms**. Each such algorithm consists of a **branching step** to increase the number of candidate solutions, followed by a **bounding step**. In our branch-and-bound algorithm for the Cyclopeptide Sequencing Problem, the branching step will extend each candidate peptide of length k into 18 peptides of length $k + 1$, and the bounding step will remove inconsistent peptides from consideration.

Note that the spectrum of a *linear* peptide does not contain as many masses as the spectrum of a *cyclic* peptide. For instance, the theoretical spectrum of the cyclic peptide NQEL contains 14 masses (corresponding to "", N, Q, E, L, LN, NQ, QE, EL, ELN, LNQ, NQE, QEL, and NQEL); however, the theoretical spectrum of the *linear* peptide NQEL, shown below, does not contain masses corresponding to LN, LNQ, or ELN, since these subpeptides "wrap around" the end of the linear peptide.

0	113	114	128	129	242	242	257	370	371	484
""	L	N	Q	E	NQ	EL	QE	QEL	NQE	NQEL

EXERCISE BREAK: How many subpeptides does a linear peptide of given length n have? (Include the 0-peptide and the entire peptide.)

Start Quiz (limit: 5 minutes)

Bioinformatics Algorithms

Given an experimental spectrum *Spectrum* of a cyclic peptide, a linear peptide is **consistent** with *Spectrum* if every mass in its theoretical spectrum is contained in *Spectrum*. If a mass appears more than once in the spectrum of the linear peptide, then it must appear *at least* that many times in *Spectrum* in order for the linear peptide to be consistent with *Spectrum*.

The key to our new algorithm is that every linear subpeptide of a cyclic peptide *Peptide* is automatically consistent with *Cyclospectrum*(*Peptide*). Thus, to solve the Cyclopeptide Sequencing Problem for *Spectrum*, we can safely ban all peptides that are *inconsistent* with *Spectrum* from the growing *List*, which powers the bounding step that we described above.

What about the branching step? Given the current collection of linear peptides *List*, define *Expand*(*List*) as a new collection containing all possible extensions of peptides in *List* by a single amino acid mass. We can now provide the pseudocode for the branch-and-bound algorithm, called CYCLOPEPTIDESEQUENCING.

```
CYCLOPEPTIDESEQUENCING(Spectrum)
  List ← {0-peptide}
  while List is nonempty
    List ← Expand(List)
    for each peptide Peptide in List
      if Cyclospectrum(Peptide) = Spectrum
        output Peptide
        remove Peptide from List
      else if Peptide is not consistent with Spectrum
        remove Peptide from List
```

Stepic  Full screen

CODE CHALLENGE: Implement CyclopeptideSequencing (pseudocode reproduced below).

Note: After the failure of the first brute-force algorithm we considered, you may be hesitant to implement this algorithm for fear that its runtime will be prohibitive. The potential problem with CyclopeptideSequencing is that it may generate incorrect *k*-mers at intermediate stages (i.e., *k*-mers that are not subpeptides of a correct solution). You may wish to wait to implement CyclopeptideSequencing until after the next section, where we will analyze this algorithm.

```
CyclopeptideSequencing(Spectrum)
  List ← {0-peptide}
  while List is nonempty
    List ← Expand(List)
    for each peptide Peptide in List
      if Cyclospectrum(Peptide) = Spectrum
        output Peptide
        remove Peptide from List
      else if Peptide is not consistent with Spectrum
        remove Peptide from List
```

Sample Input:

```
0 113 128 186 241 299 314 427
```

Sample Output:

```
186-128-113 186-113-128 128-186-113 128-113-186 113-186-128 113-128-186
```

Extra Dataset

Start Quiz (limit: 5 minutes)

Share ❮ Back Submit Discussions

Bioinformatics Algorithms

Stepic Full screen

To see that CyclopeptideSequencing may generate incorrect *k*-mers at intermediate stages, let's walk through this algorithm for the following *Spectrum*:

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

Download This Spectrum

CyclopeptideSequencing first expands *List* into the set of all 1-mers consistent with *Spectrum*:

97	99	101	103
P	V	T	C

The algorithm next appends each of the 18 amino acid masses to each of the 1-mers above. The resulting *List* containing 4 · 18 = 72 peptides of length 2 is then trimmed to keep only the 10 peptides that are consistent with *Spectrum*:

97-99	97-101	97-103	99-97	99-101
PV	PT	PC	VP	VT
99-103	101-97	101-99	103-97	103-99
VC	TP	TV	CP	CV

Share Next step > Discussions

Bioinformatics Algorithms

After expansion and trimming in the next iteration, *List* contains 15 consistent 3-mers:

97-99-103	97-99-101	97-101-97	97-101-99	97-103-99
PVC	PVT	PTP	PTV	PCV
99-97-103	99-97-101	99-101-97	99-103-97	101-97-99
VPC	VPT	VTP	VCP	TPV
101-97-103	101-99-97	103-97-101	103-97-99	103-99-97
TPC	TVP	CPT	CPV	CVP

With one more iteration, *List* contains 10 consistent 4-mers. Observe that the six 3-mers highlighted in red above failed to expand into any 4-mers below, and so we now know that CYCLOPEPTIDESEQUENCING may generate some incorrect *k*-mers.

97-99-103-97	97-101-97-99	97-101-97-103	97-103-99-97	99-97-101-97
PVCP	PTPV	PTPC	PCVP	VPTP
99-103-97-101	101-97-99-103	101-97-103-99	103-97-101-97	103-99-97-101
VCPT	TPVC	TPCV	CPTP	CVPT

In the final iteration, we generate 10 consistent 5-mers:

97-99-103-97-101	97-101-97-99-103	97-101-97-103-99	97-103-99-97-101	99-97-101-97-103
PVCPT	PTPVC	PTPCV	PCVPT	VPTPC
99-103-97-101-97	101-97-99-103-97	101-97-103-99-97	103-97-101-97-99	103-99-97-101-97
VCPTP	TPVCP	TPCVP	CPTPV	CVPTP

All these linear peptides correspond to the same cyclic peptide PVCPT, thus solving the Cyclopeptide Sequencing Problem. You may wish to complete the Code Challenge from the previous lesson (step 4) if you have not done so already, after which you can verify that CYCLOPEPTIDESEQUENCING quickly reconstructs Tyrocidine B1.

Although it is difficult to imagine a worst-case in which CYCLOPEPTIDESEQUENCING takes a long time to run, no one has been able to guarantee that this algorithm will not generate a huge number of incorrect k-mers. Computer scientists classify an algorithm as **polynomial** if its running time can be bounded by a polynomial in the length of the input data. On the other hand, an algorithm is **exponential** if its runtime on some datasets is exponential in the length of the input data. BFCYCLOPEPTIDESEQUENCING is exponential, and although CYCLOPEPTIDESEQUENCING is much faster in practice, this algorithm has not been proven to be polynomial. Thus, from the perspective of an "Algorithms 101" course focusing on theoretical computer science, the practical CYCLOPEPTIDESEQUENCING algorithm is just as inefficient as the first brute force algorithm we considered, since neither algorithm's running time can be bounded by a polynomial. See the Open Problems section to learn more about the algorithmic challenges related to this problem.

Stepic Full screen

Although CyclopeptideSequencing successfully reconstructed Tyrocidine B1, this algorithm only works in the case of an ideal experimental spectrum, when the experimental spectrum of a peptide coincides exactly with its theoretical spectrum. This inflexibility of CyclopeptideSequencing presents a practical barrier, since mass spectrometers generate spectra that are far from ideal — they are characterized by having both **false masses** and **missing masses**. A false mass is present in the experimental spectrum but absent from the theoretical spectrum; a missing mass is present in the theoretical spectrum but absent from the experimental spectrum.

For example, compare the following theoretical and experimental spectrum of the cyclic peptide NQEL, which displays three **missing** and two **false** masses.

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

What is particularly worrisome about this example is that the mass of the amino acid E (129) is missing, and the mass of the amino acid V (99) is false; as a result, the first step of CyclopeptideSequencing would establish {V, L, N, Q} as the amino acid composition of our candidate peptides, which is incorrect. In fact, *any* false or missing mass will cause CyclopeptideSequencing to throw out the correct peptide, because its theoretical spectrum differs from the experimental spectrum.

STOP and Think: How would you reformulate the Cyclopeptide Sequencing Problem to handle the case of experimental spectra with errors?

Share Next step > Discussions

Bioinformatics Algorithms

To generalize the Cyclopeptide Sequencing Problem to handle "noisy" spectra having false and missing masses, we need to relax the requirement that a candidate peptide's theoretical spectrum must match the experimental spectrum *exactly*, and instead incorporate a **scoring function** that will select the peptide whose theoretical spectrum matches the given experimental spectrum *the most closely*. Given a cyclic peptide *Peptide* and a spectrum *Spectrum*, we define *Score*(*Peptide*, *Spectrum*) as the number of masses shared between *Cyclospectrum*(*Peptide*) and *Spectrum*. Recalling our example above: if

Spectrum = {0, 99, 113, 114, 128, 227, 257, 299, 355, 356, 370, 371, 484}

then *Score*(NQEL, *Spectrum*) = 11. We now redefine the Cyclopeptide Sequencing Problem for the case of noisy spectra:

To generalize the Cyclopeptide Sequencing Problem to handle "noisy" spectra having false and missing masses, we need to relax the requirement that a candidate peptide's theoretical spectrum must match the experimental spectrum *exactly*, and instead incorporate a **scoring function** that will select the peptide whose theoretical spectrum matches the given experimental spectrum *the most closely*. Given a cyclic peptide *Peptide* and a spectrum *Spectrum*, we define *Score*(*Peptide*, *Spectrum*) as the number of masses shared between *Cyclospectrum*(*Peptide*) and *Spectrum*. Recalling our example above: if

Spectrum = {0, 99, 113, 114, 128, 227, 257, 299, 355, 356, 370, 371, 484}

then *Score*(NQEL, *Spectrum*) = 11. We now redefine the Cyclopeptide Sequencing Problem for the case of noisy spectra:

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*).

Our goal is to adapt the CYCLOPEPTIDESEQUENCING algorithm to find a peptide with maximum score. Remember that this algorithm had a bounding step in which all candidate peptides having inconsistent spectra were thrown out; we need to somehow revise this step to include more candidate peptides, while

Bioinformatics Algorithms

To limit the number of candidate peptides under consideration, we will replace *List* with a *Leaderboard*, which holds the *N* highest scoring candidate peptides for further extension. At each step, we will expand all candidate peptides found in *Leaderboard*, then eliminate those peptides whose newly calculated scores are not high enough to keep them on the *Leaderboard*. This idea is similar to the notion of a "cut" in a golf tournament; after the cut, only the top *N* golfers are allowed to play in the next round, since they are the only players who have a reasonable chance of winning.

To be fair, a cut should include anyone who is tied with the *N*th-place competitor. Thus, *Leaderboard* should be trimmed down to the "*N* highest-scoring peptides including ties", which may include more than *N* peptides. Given a list of peptides *Leaderboard*, a spectrum *Spectrum*, and an integer *N*, *Cut*(*Leaderboard*, *Spectrum*, *N*) returns the top *N* highest-scoring peptides in *Leaderboard* (including ties) with respect to *Spectrum*. We now introduce LEADERBOARDCYCLOPEPTIDESEQUENCING:

```
LEADERBOARDCYCLOPEPTIDESEQUENCING(Spectrum, N)
  Leaderboard ← {0-peptide}
  LeaderPeptide ← 0-peptide
  while Leaderboard is non-empty
    Leaderboard ← Expand(Leaderboard)
    for each Peptide in Leaderboard
      if Mass(Peptide) = ParentMass(Spectrum)
        if Score(Peptide, Spectrum) > Score(LeaderPeptide, Spectrum)
          LeaderPeptide ← Peptide
      else if Mass(Peptide) > ParentMass(Spectrum)
        remove Peptide from Leaderboard
    Leaderboard ← Cut(Leaderboard, Spectrum, N)
  output LeaderPeptide
```

STOP and Think: How should you select the leaderboard number *N* for practical applications of LEADERBOARDCYCLOPEPTIDESEQUENCING?

Stepic Full screen

CODE CHALLENGE: Implement LeaderboardCyclopeptideSequencing.

Input: Integer *N* and a collection of integers *Spectrum*.

Output: *LeaderPeptide* after running LeaderboardCyclopeptideSequencing(*Spectrum*, *N*).

Sample Input:

```
10
0 71 113 129 147 200 218 260 313 331 347 389 460
```

Sample Output:

```
113-147-71-129
```

Extra Dataset

Note: Multiple solutions may exist. You may return any one.

Start Quiz (limit: 5 minutes)

Share Back Submit Discussions

Bioinformatics Algorithms

Stepic  Full screen

We point out that because the highest-scoring peptide may be cut early on, this algorithm is not guaranteed to correctly solve the Cyclopeptide Sequencing Problem. It is instead a **heuristic**, or a technique that trades precision for speed. Whenever we develop such a method, we must ask ourselves: *how accurate is this heuristic?* Consider the simulated spectrum $Spectrum_{10}$ of Tyrocidine B1 shown below, which has approximately 10% missing/false masses. Note that the blue masses are not actually in the spectrum, but we show them so that it is clear which masses are missing.

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			

Download This Spectrum

Applying LeaderboardCyclopeptideSequencing to this spectrum (with N = 1000) results in the correct cyclic peptide VKLFPWFNQY, which has a score of 86. The next closest contender is the similar peptide YKLFPWFNQV, which has a score of 76.

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Stepic Full screen

So far, LEADERBOARDCYCLOPEPTIDESEQUENCING has worked well, but as the number of errors increases, so does the likelihood that this algorithm will return the incorrect peptide. Let's see how this algorithm performs on a noisier simulated spectrum; below, we show $Spectrum_{25}$ for Tyrocidine B1, which has 25% missing/false 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							

Download This Spectrum

EXERCISE BREAK: Run LEADERBOARDCYCLOPEPTIDESEQUENCING on $Spectrum_{25}$ with N = 1000. You should find 14 linear peptides of maximum score 83 (corresponding to 6 different cyclic peptides). What are they? (Return your peptides in integer format, with each peptide separated by a single space, e.g., `113-147-71-129 71-147-129-113`.)

Note: More peptides of score 83 can be found if you use a larger value of N; please use N = 1000.

Start Quiz (limit: 5 minutes)

Share Back Submit Discussions

Bioinformatics Algorithms

Stepic Full screen

When run on $Spectrum_{25}$, LeaderboardCyclopeptideSequencing (with N = 1000) identifies VKLFPADFNQY (score: 83) as a highest-scoring cyclic peptide instead of the correct peptide VKLFPWFNQY (score: 82). These two peptides are similar, owing to the fact that the combined mass of A (71) and D (115) is equal to the mass of W (186).

STOP and Think: How could we have eliminated the incorrect peptide VKLFPADFNQY from consideration for $Spectrum_{25}$?

Notice that although the correct and incorrect peptides are similar, their amino acid compositions differ. If we could figure out the amino acid composition of Tyrocidine B1 from its spectrum alone and run LeaderboardCyclopeptideSequencing on this smaller alphabet (rather than on the alphabet of all amino acids), then we could eliminate the incorrect peptide VKLFPADFNQY from consideration.

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Stepic Full screen

When we apply LeaderboardCyclopeptideSequencing for the extended alphabet to $Spectrum_{10}$, one of the highest-scoring peptides is VKLFPWFNQ-98-65, in contrast to the correct peptide VKLFPWFNQY (*Mass*(Y) = 98 + 65 = 163). Apparently, non-standard amino acids successfully competed with standard amino acids for the limited number of positions on the leaderboard, resulting in VKLFPWFNQ-98-65 winning over VKLFPWFNQY. Since LeaderboardCyclopeptideSequencing fails to identify the correct peptide even with only 10% false and missing masses, our stated aim from the previous section is now even more important. We must determine the amino acid composition of an unknown peptide from its spectrum so that we may run LeaderboardCyclopeptideSequencing on this smaller alphabet of amino acids.

STOP and Think: Do you have any ideas for how we can determine which amino acids are present in an unknown peptide using only an experimental spectrum?

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Stepic Full screen

Until now, we have assumed that 20 amino acids form the building blocks of proteins; these building blocks are called **proteinogenic amino acids**. There are actually two additional proteinogenic amino acids, called **selenocysteine** and **pyrrolysine**, which are incorporated into proteins by special biosynthetic mechanisms (see Detour: Selenocysteine and Pyrrolysine). Yet in addition to the 22 proteinogenic amino acids, NRPs contain **non-proteinogenic amino acids**, which increases the number of possible building blocks for antibiotic peptides from 20 to over 100.

Enlarging the amino acid alphabet spells trouble for our current approach to cyclopeptide sequencing. Indeed, the correct peptide now must "compete" with many more incorrect ones for a place on the leaderboard, increasing the chance that the correct peptide will be cut along the way.

For example, although Tyrocidine B1 contains only proteinogenic amino acids, its closest relative, Tyrocidine B (`Val-Orn-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr`) contains a non-proteinogenic amino acid called **ornithine** (`Orn`). Because so many non-proteinogenic amino acids exist, bioinformaticians often assume that any integer from 57 to 200 may represent the mass of an amino acid; the "lightest" amino acid, `Gly`, has mass 57 Da, and most amino acids have masses smaller than 200 Da.

EXERCISE BREAK: Applying LeaderboardCyclopeptideSequencing on the extended amino acid alphabet to $Spectrum_{10}$ with $N = 1000$ returns 38 different linear peptides of maximum score. What are they? (Return your answer in integer format separated by a space, e.g., `113-147-71-129 199-200-61`.)

Note: This exercise is difficult because it may take a long time to run if your solution is inefficient.

Start Quiz (limit: 5 minutes)

Share Submit Discussions

Bioinformatics Algorithms

Stepic Full screen

When we apply LeaderboardCyclopeptideSequencing for the extended alphabet to $Spectrum_{10}$, one of the highest-scoring peptides is VKLFPWFNQ-98-65, in contrast to the correct peptide VKLFPWFNQY (*Mass*(Y) = 98 + 65 = 163). Apparently, non-standard amino acids successfully competed with standard amino acids for the limited number of positions on the leaderboard, resulting in VKLFPWFNQ-98-65 winning over VKLFPWFNQY. Since LeaderboardCyclopeptideSequencing fails to identify the correct peptide even with only 10% false and missing masses, our stated aim from the previous section is now even more important. We must determine the amino acid composition of an unknown peptide from its spectrum so that we may run LeaderboardCyclopeptideSequencing on this smaller alphabet of amino acids.

STOP and Think: Do you have any ideas for how we can determine which amino acids are present in an unknown peptide using only an experimental spectrum?

Share < Back Next step > Discussions

Bioinformatics Algorithms

Stepic Full screen

At the end of the last section, you may have guessed that the best way to determine the amino acid composition of a peptide from its experimental spectrum would be to take the smallest masses present in the spectrum (between 57 and 200 Da). However, if only a single amino acid mass is missing, then this approach will not be able to reconstruct the peptide's amino acid composition.

Let's take a different approach. Say that our experimental spectrum contains the masses of subpeptides NQE and NQ. If we subtract these two masses, then we will obtain the mass E for free, even if it was not present in the experimental spectrum! If the underlying peptide is NQEL, then we can also find the mass of E by subtracting the masses of QE and Q or NQEL and LNQ.

Following this example, we define the **convolution** of a spectrum by taking all positive differences of masses in the spectrum. The tables on the next steps show the convolutions of the theoretical and simulated spectra of NQEL that we encountered before (spectra reproduced below).

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

Share Next step ❯ Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. **The Spectral Convolution Saves the**

Stepic Full screen

	"" 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
0													
113	113												
114	114	1											
128	128	15	14										
129	129	16	15	1									
227	227	114	113	99	98								
242	242	129	128	114	113	15							
242	242	129	128	114	113	15							
257	257	144	143	129	128	30	15	15					
355	355	242	241	227	226	128	113	113	98				
356	356	243	242	228	227	129	114	114	99	1			
370	370	257	256	242	241	143	128	128	113	15	14		
371	371	258	257	243	242	144	129	129	114	16	15	1	
484	484	371	370	356	355	257	242	242	227	129	128	114	113

Spectral convolution for the theoretical spectrum of NQEL. The most frequent elements in the convolution between 57 and 200 are (multiplicities in parentheses): 113 (8), 114 (8), 128 (8), 129 (8).

As predicted, some of the values in these tables appear more frequently than others. For example, **113** (the mass of L) appears eight times in the table above; we say that **113** has **multiplicity** 8. Six occurrences of **113** in the table above correspond to subpeptide pairs differing in an L: L and "; LN and N; EL and E; LNQ and NQ; QEL and QE; NQEL and NQE.

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Stepic

Full screen

	"" 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 for the simulated spectrum of NQEL. The most frequent elements in the convolution between 57 and 200 are (multiplicities in parentheses): 113 (4), 114 (4), 128 (4), 99 (3), 129 (3).

Interestingly, 129 (the mass of E) pops up three times in the above convolution of the simulated spectrum, even though 129 was missing from the spectrum itself.

We now should feel confident about using the most frequently appearing integers in the convolution as a guess for the amino acid composition of an unknown peptide. In our simulated spectrum for NQEL, the most frequent elements of the convolution in the range from 57 to 200 are:

113 (4), 114 (4), 128 (4), 99 (3), 129 (3)

Note that these most frequent elements capture all four amino acids in NQEL.

Share

< Back

Next step >

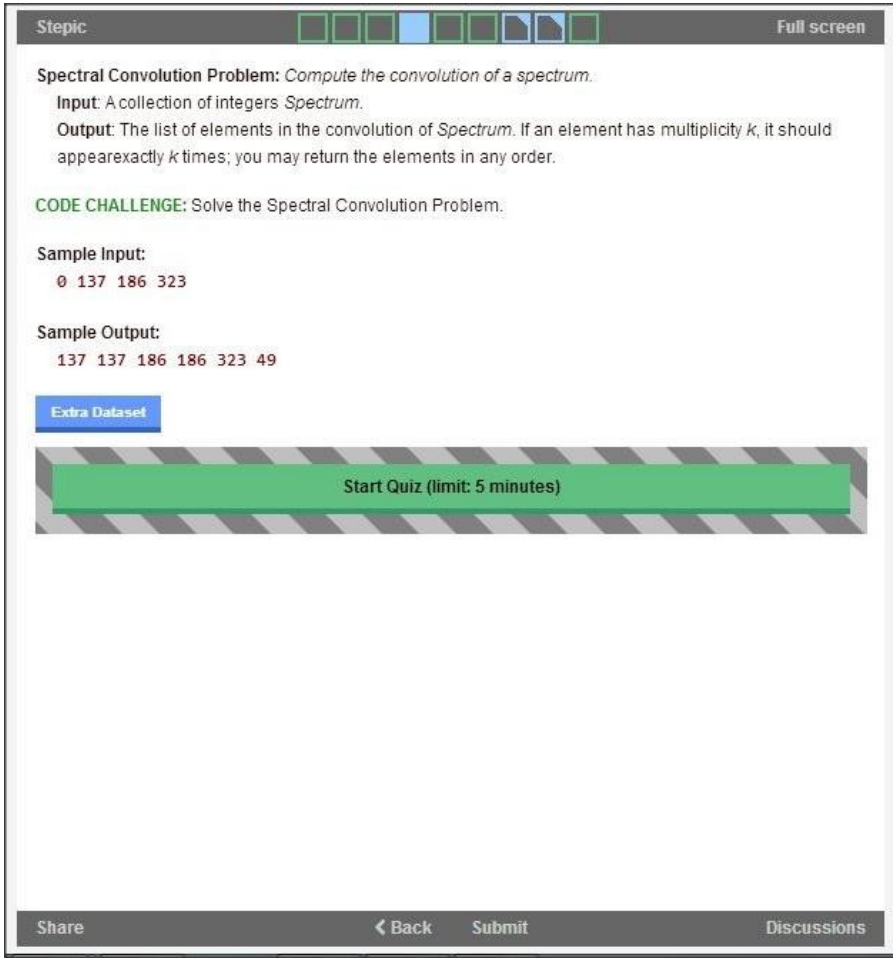
Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. **The Spectral Convolution Saves the**



Bioinformatics Algorithms

Stepic

Full screen

Recall that LeaderboardCyclopeptideSequencing failed to reconstruct Tyrocidine B1 from $Spectrum_{10}$ when using the extended alphabet of amino acids. The twenty most frequent elements of its spectral convolution in the range from 57 to 200 are (with multiplicities in parentheses):

147 (35)	128 (31)	97 (28)	113 (28)	114 (26)
186 (23)	57 (21)	163 (21)	99 (18)	145 (18)
130 (16)	154 (16)	129 (15)	73 (14)	146 (14)
187 (14)	98 (13)	148 (13)	164 (13)	170 (13)

This list captures all eight different amino acid masses making up Tyrocydine B1, which are colored green in the list above. Furthermore, all eight masses appear in the ten most frequent elements from the spectral convolution! The figure on the next step shows the convolution of $Spectrum_{10}$.

Share

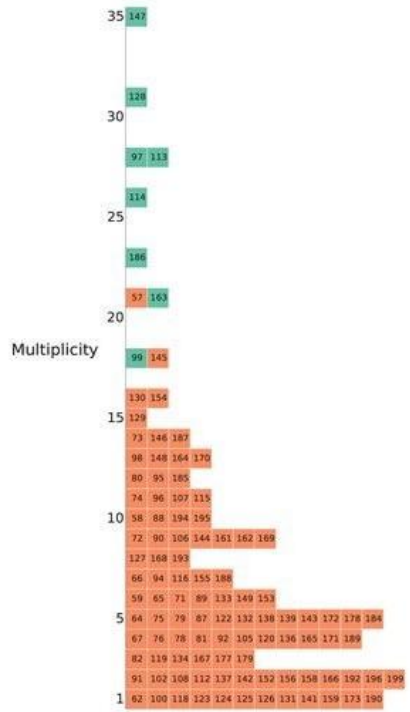
< Back

Next step >

Discussions

Bioinformatics Algorithms

Stepic Full screen



The spectral convolution of the experimental spectrum $Spectrum_{10}$ for Tyrocidine B1. For each element of the spectral convolution (shown as a number within a box), its y-coordinate represents the number of times that the element appears in the spectral convolution. Green boxes represent masses of amino acids from Tyrocidine B1.

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids

26. The Spectral Convolution Saves the

Stepic Full screen

We now have the outline for an algorithm. Given an experimental spectrum, we first compute the convolution of an experimental spectrum. We then select the *M* most frequent elements between 57 and 200 to form a putative alphabet of amino acid masses; in order to be fair, we should include the top *M* elements of the convolution "with ties". Finally, we run the algorithm LeaderboardCyclopeptideSequencing, where the amino acid masses are restricted to this alphabet. We call this algorithm ConvolutionCyclopeptideSequencing.

CODE CHALLENGE: Implement ConvolutionCyclopeptideSequencing.

Input: An integer *M*, an integer *N*, and a collection of (possibly repeated) integers *Spectrum*.

Output: A cyclic peptide *LeaderPeptide* with amino acids taken only from the top *M* elements (and ties) of the convolution of *Spectrum* that fall between 57 and 200, and where the size of *Leaderboard* is restricted to the top *N* (and ties).

Sample Input:

```
20
60
57 57 71 99 129 137 170 186 194 208 228 265 285 299 307 323 356 364 394 422 493
```

Sample Output:

```
99-71-137-57-72-57
```

Extra Dataset

Start Quiz (limit: 5 minutes)

Share Back Submit Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. **The Spectral Convolution Saves the**

Stepic Full screen

EXERCISE BREAK: Run ConvolutionCyclopeptideSequencing on $Spectrum_{25}$ (reproduced below) with N = 1000 and M = 20. Identify the 23 highest-scoring linear peptides. (Return the peptides in integer format separated by a single space, e.g., `123-57-200-143 199-143-121-60`)

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							

Start Quiz (limit: 5 minutes)

Share ❮ Back Submit Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. **The Spectral Convolution Saves the**

Stepic Full screen

ConvolutionCyclopeptideSequencing (with N = 1000 and M = 20) now correctly reconstructs Tyrocidine B1 from $Spectrum_{10}$. The true test of this algorithm is whether it will work on a noisier spectrum. Recall that our previous algorithm failed to identify the correct peptide for $Spectrum_{25}$. In contrast, the improved ConvolutionCyclopeptideSequencing (with N = 1000 and M = 20) now correctly identifies Tyrocidine B1 from this spectrum!

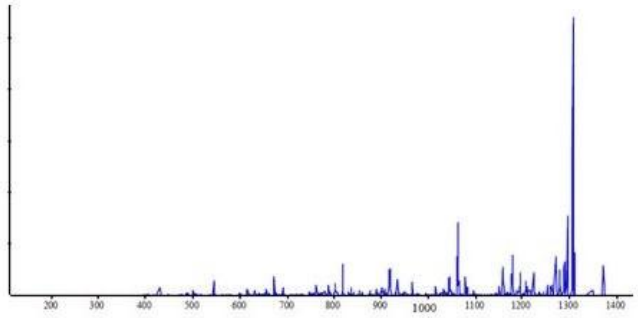
Share ‹ Back Next step › Discussions

Bioinformatics Algorithms

Stepic Full screen

In this chapter, we have sheltered you from the gruesome realities of mass spectrometry by providing simulated spectra that are relatively easy to sequence (even those with false and missing masses). We committed a sin of omission by loosely describing the mass spectrometer as a "scale" and assuming that this complex machine simply weighs tiny peptide fragments one at a time. In truth, the mass spectrometer first converts subpeptides into ions (i.e., charged particles). Ionization of particles helps the mass spectrometer sort the ions by using an electromagnetic field; ions are separated not by their mass, but rather according to their **mass/charge ratio**. If fragment ion NQY (integer mass: 114 + 128 + 163 = 405) has charge +1, then it contains one additional proton, resulting in a total integer mass of 406 and a mass/charge ratio of 406/1 = 406. To be more precise, the *monoisotopic mass* of NQY is approximately 114.043 + 128.058 + 163.063 = 405.164, and the mass of a proton is 1.007 Da, which makes the mass charge/ratio more closely equal to (405.164 + 1.007)/1 = 406.171.

The mass spectrometer outputs a collection of **peaks**, which are shown below for a real Tyrocidine B1 spectrum. Each peak's *x*-coordinate represents its mass/charge ratio, and its height represents the **intensity** (i.e., relative abundance) of particles having that mass/charge ratio. For example, in the experimental spectrum of Tyrocidine B1 shown below, you will find a small (almost invisible) peak with a mass/charge ratio of 406.3,which corresponds to the fragment ion NQY having mass/charge ratio 406.171, with an error of approximately 0.13 Da.



Share Next step > Discussions

Bioinformatics Algorithms

Stepic Full screen

As you can imagine, we must navigate a few practical barriers in order to analyze real spectra. First, the charge of each peak is unknown, often forcing researchers to try all possible charges from 1 to some parameter *maxCharge*, where the particular choice of *maxCharge* depends on the fragmentation technology used. This procedure generates *maxCharge* masses for each peak, so that the larger the value of *maxCharge*, the more false masses in the spectrum.

Second, the real spectrum on the previous step has nearly 1000 peaks, most of which are **false peaks**, meaning that their mass/charge ratio does not correspond to any subpeptide's mass/charge ratio (for any charge value). Fortunately, false peaks typically have low intensities, necessitating a pre-processing step that removes low-intensity peaks before applying an algorithm. Below is the list of the 95 mass/charge ratios for peaks that "survived" this preprocessing step for the Tyrocidine B1 spectrum. Their intensities may nevertheless vary by 2-3 orders of magnitude; for example, the intensity of the peak having mass/charge ratio 372.2 is 300 times smaller than the intensity of the peak with mass/charge ratio 1306.5.

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

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Stepic Full screen

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

Only 31 of these 95 mass/charge ratios (shown in bold above and reproduced below) can be matched to subpeptides of Tyrocidine B1 (with *maxCharge* = 1 and maximum allowable mass discrepancy of 0.3 Da):

Mass	Subpeptide	Mass	Subpeptide	Mass	Subpeptide
406.2	NQY	431.2	FPW	448.2	WFN
486.2	KLFP	488.2	VKLF	505.2	NQYV
544.2	LFPW	545.2	PWFN	632.3	QYVKL
672.3	KLFPW	673.3	PWFNQ	691.3	LFPWF
692.3	FPWFN	746.3	NQYVKL	771.3	VKLFPW
819.4	KLFPWF	836.4	PWFNQY	876.4	QYVKLFP
918.4	VKLFPWF	933.4	LFPWFNQ	934.4	YVKLFPW
935.4	PWFNQYV	966.4	WFNQYVK	1061.5	KLFPWFNQ
1063.5	PWFNQYVK	1079.5	WFNQYVKL	1096.5	LFPWFNQY
1176.5	NQYVKLFPW	1195.6	LFPWFNQYV	1210.6	FPWFNQYVK
1224.6	KLFPWFNQY				

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. The Spectral Convolution Saves the Day

Stepic Full screen

You can now see that sequencing Tyrocidine B1 from a real spectrum, for which two-thirds of all masses are false, presents a much more difficult problem than sequencing this peptide from the simulated $Spectrum_{25}$. In the following challenge problem, you will need to further develop the methods we studied in this chapter to analyze a real spectrum.

FINAL CHALLENGE: Tyrocidine B1 is just one of many known NRPs produced by *Bacillus brevis*. A single bacterial species may produce dozens of different antibiotics, and even after 70 years of research, there are likely undiscovered antibiotics produced by *Bacillus brevis*. Try to sequence the tyrocidine corresponding to the real experimental spectrum below. Since the fragmentation technology used for generating the spectrum tends to produce ions with charge +1, you can safely assume that all charges are +1. Return the peptide as a collection of space-separated integer masses.

371.5	375.4	390.4	392.2	409.0	420.2	427.2	443.3	446.4	461.3
471.4	477.4	491.3	505.3	506.4	519.2	536.1	546.5	553.3	562.3
588.2	600.3	616.2	617.4	618.3	633.4	634.4	636.2	651.5	652.4
702.5	703.4	712.5	718.3	721.0	730.3	749.4	762.6	763.4	764.4
779.6	780.4	781.4	782.4	797.3	862.4	876.4	877.4	878.6	879.4
893.4	894.4	895.4	896.5	927.4	944.4	975.5	976.5	977.4	979.4
1005.5	1007.5	1022.5	1023.7	1024.5	1039.5	1040.3	1042.5	1043.4	1057.5
1119.6	1120.6	1137.6	1138.6	1139.5	1156.5	1157.6	1168.6	1171.6	1185.4
1220.6	1222.5	1223.6	1239.6	1240.6	1250.5	1256.5	1266.5	1267.5	1268.6

Download Spectrum

Start Quiz

Share Back Submit Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. The Spectral Convolution Saves the Day

Stepic Full screen

Beltway and Turnpike Problems

Research Director: Hosein Mohimani

In the case of the alphabet of arbitrary integers, the Cyclopeptide Sequencing Problem corresponds to a famous computer science problem known as the **Beltway Problem**. The Beltway Problem asks you to find a set of points on a circle such that the distances between all pairs of points (where distance is measured around the circle) match a given collection of integers.

The Beltway Problem's analogue in the case when the points lie along a line segment instead of on a circle is called the **Turnpike Problem**. The terms "beltway" and "turnpike" arise from an analogy with exits on circular and linear roads, respectively. In the case of n points on a circle and line, the inputs for the Beltway and Turnpike Problems consist of $n(n-1)+2$ and $n(n-1)/2+2$ distances, respectively (these formulas include the distance 0 as well as the length of the entire segment).

Various attempts to find polynomial algorithms for the Beltway and Turnpike Problems (or to prove that they are NP-hard) have failed. However, there is a **pseudo-polynomial algorithm** for the Turnpike Problem (see Detour: Pseudo-polynomial Algorithm for the Turnpike Problem). In contrast to a truly polynomial algorithm, which can be bounded by a polynomial in the length of the input, a pseudo-polynomial algorithm for the Turnpike Problem is polynomial in the total length of the line segment. For example, if n points are separated by huge distances exceeding 2^{100}, then a polynomial algorithm would still be fast, whereas a pseudo-polynomial algorithm would be impossibly slow. Note that although the distances themselves will be huge, each distance can be stored using only 100 bits, implying that the length of the input is small even for such huge distances.

Pseudo-polynomial algorithms are useful in practice because practical instances of the problems typically do not include huge numbers. Interestingly, although a pseudo- polynomial algorithm exists for the Turnpike Problem, such an algorithm for the seemingly similar Beltway Problem remains undiscovered. Can you develop such an algorithm, or perhaps prove that one does not exist?

Share Next step > Discussions

Bioinformatics Algorithms

Sequencing Standard Subpeptides

Research Director: Hosein Mohimani

Although most NRPs indeed contain non-proteinogenic amino acids, most amino acids in a typical NRP are proteinogenic. For example, the only non-proteinogenic amino acid appearing in tyrocidines is ornithine (Orn), which occurs in at most one position of the peptide. Existing algorithms do not account for the fact that individual NRPs often have only one or two non-proteinogenic amino acids and attempt to solve the Cyclopeptide Sequencing Problem for the alphabet of all integers from 57 to 200.

Given a cyclic peptide, we define its **standard subpeptide** as the longest subpeptide consisting entirely of standard amino acids. For example, the standard subpeptide of Tyrocidine B (`Val-Orn-Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr`) is `Leu-Phe-Pro-Trp-Phe-Asn-Gln-Tyr`. Since finding the complete reconstruction of an NRP in the presence of non-proteinogenic amino acids poses a difficult computational problem, can you reconstruct the standard subpeptide of a cyclic NRP?

Bioinformatics Algorithms

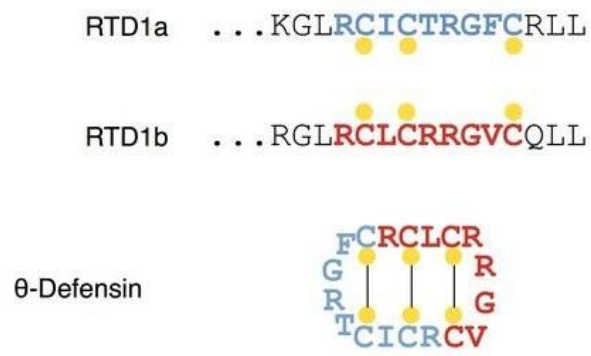
Stepic Full screen

Sequencing Cyclic Peptides in Primates

Research Director: Hosein Mohimani

Bacteria and fungi do not have a monopoly on producing cyclic peptides — animals and plants make them too (albeit through a completely different mechanism). The first cyclic peptide found in animals (called **θ-defensin**) was discovered in 1999 in macaques. θ-defensin prevents viruses from entering cells and has strong anti-HIV activity. The question of how primates make θ-defensin remains a mystery.

Needless to say, there is no 54-mer in the macaque genome encoding the 18 amino acid-long θ-defensin. Instead, this cyclic peptide is formed by concatenating two 9 amino acid-long peptides excised from two different proteins called **RTD1a** and **RTD1b**, as shown in the figure below. It remains unclear which enzymes do this elaborate cutting and pasting.



The 18 amino acid-long θ-defensin peptide is formed by cutting two 9 amino acid-long peptides RCICTRGFC and RCLCRRGVC from the RTD1a and RTD1b proteins, concatenating them, and then circularizing the resulting peptide (along with introducing three disulfide bridges that form bonds across the peptide).

Share ❮ Back Next step ❯ Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. The Spectral Convolution Saves the Day

Interestingly, macaques and baboons produce θ-defensin, whereas humans and chimpanzees do not. This discrepancy makes us wonder whether a mutation occurred in the human-chimpanzee ancestor a few million years ago that resulted in the loss of this very useful peptide. Interestingly, genes very similar to RTD1a and RTD1b do exist in humans, but a codon in one of these genes mutated into a stop codon, thus shortening the encoded protein. Since this stop codon is located before the 9 amino acid-long peptide contributing to θ-defensin, humans do not produce this peptide and thus cannot produce θ-defensin.

In a remarkable experiment, Venkataraman et al., 2009 demonstrated that humans could get θ-defensin back! Certain drugs can force the ribosome to ignore stop codons and continue translating RNA, even after encountering a stop codon. The researchers demonstrated that after treatment with such a drug, human cells began producing the human version of θ-defensin. The surprising conclusion of this experiment is that although humans and chimpanzees lost θ-defensin millions of years ago, we still possess the mysterious enzymes required to cut and paste its constituent peptides.

Some biologists believe that since the enzymes making θ-defensin still work in humans, they must be needed for something else. If these enzymes did not provide some selective advantage, then over time, mutations would cause their genes to become **pseudogenes**, or non-functional remnants of previously working genes. The most natural explanation for why these enzymes are still functional is that humans produce still undiscovered cyclic peptides, and that the enzymes needed for θ-defensin are also used to "cut-and- paste" other (still undiscovered) cyclic peptides. The hypothesis that we may possess undiscovered cyclic peptides is not as improbable as you might think because biologists still lack robust algorithms for cyclopeptide discovery from the billions of spectra generated in hundreds of labs analyzing the human proteome.

By default, researchers assume that all spectra ever acquired in human proteome studies originated from linear peptides. Could they be wrong? Can you devise a fast cyclopeptide-sequencing algorithm that would be able to analyze publicly available spectra and hopefully discover human cyclopeptides?

Bioinformatics Algorithms

Stepic Full screen

The term **Lysenkoism** refers to the political control over genetics in the USSR by Trofim Lysenko that began in the late 1920s and lasted for three decades until the death of Stalin. Lysenkoism was built on theories of inheritance by acquired characteristics, which ran counter to Mendelian laws.

In 1928, Lysenko, a previously unknown agronomist, claimed to have developed a new agricultural technique that had tripled wheat crop yield. During Stalin's rule, Soviet propaganda focused on inspirational stories of workers and peasants, and it portrayed Lysenko as a genius, even though he had manufactured his experimental data. Empowered by his sudden hero status, Lysenko denounced genetics and started promoting his own "scientific" views. He called geneticists "fly lovers and people haters" and claimed that they were trying to undermine the onward march of Soviet agriculture.

Trofim Lysenko

Share Next step > Discussions

Bioinformatics Algorithms

Gause found himself among the few Soviet biologists who were not afraid of publicly denouncing Lysenko. By 1935, Lysenko announced that by opposing his theories, geneticists were directly opposing the teachings of Marxism. Stalin, who was in the audience, was the first to applaud, calling out, "Bravo, Comrade Lysenko!" This event gave Lysenko free reign to slander any geneticists who spoke out against him; many of Lysenkoism's opponents were imprisoned or even executed.

Lysenko (left) speaking in the Kremlin in 1935, with Joseph Stalin (right) in the audience.

After World War □, Lysenko did not forget Gause's criticism: Lysenko's supporters demanded that Gause be expelled from the Russian Academy of Sciences. Lysenkoists made various attempts to invite Gause to denounce genetics and accept their pseudoscience. Gause was probably the only Soviet biologist at that time who could simply ignore such "invitations", the only other contemporary opponents of Lysenkoism being top Soviet nuclear physicists. However, Stalin left Gause and the physicists alone; in Stalin's mind, the development of antibiotics and the atomic bomb were too important. In 1949, when the director of the Russian secret police (Lavrentiy Beria) told Stalin of the dissident scientists, Stalin responded, "Make sure that our scientists have everything needed to do their job", adding, "there will always be time to execute them [later]".

Bioinformatics Algorithms

Stepic Full screen

In 1961, Sydney Brenner and Francis Crick established the rule of "one codon, one amino acid" during protein translation. They observed that deleting a single nucleotide or two consecutive nucleotides in a gene dramatically altered the protein product. Paradoxically, deleting *three* consecutive nucleotides resulted in only minor changes in the protein. For example, the phrase

THE · SLY · FOX · AND · THE · SHY · DOG

turns into gibberish after deleting one letter:

THE · SYF · OXA · NDT · HES · HYD · OG

or after deleting two letters:

THE · SFO · XAN · DTH · ESH · YDO · G

but it makes sense after deleting three letters:

THE · FOX · AND · THE · SHY · DOG

In 1964, Charles Yanofsky demonstrated that a gene and the protein that it produces are **collinear**, meaning that the first codon codes for the first amino acid in the protein, the second codon codes for the second amino acid, etc. For the next thirteen years, biologists believed that a protein was encoded by a long string of *contiguous* nucleotide triplets. However, the discovery of **split human genes** in 1977 proved otherwise and necessitated the computational problem of predicting the locations of genes using only the genomic sequence.

Share Next step › Discussions

Bioinformatics Algorithms

Stepic Full screen

The traditional view that bacteria act as loners and have few interactions with the rest of their colony has been challenged by the discovery of a communication method called **quorum sensing**. This finding has shown that bacteria are capable of coordinated activity when migrating to a better nutrient supply or adopting a **biofilm** formation for defense within hostile environments. The "language" used in quorum sensing is often based on the exchange of peptides (as well as other molecules) called **bacterial pheromones**. The nature of communications between bacteria can be amicable or adversarial.

When a single bacterium releases pheromones into its environment, their concentration is often too low to be detected; however, once the population density increases, pheromone concentrations reach a threshold level that allows the bacteria to activate certain genes in response.

For example, *Burkholderia cepacia* is a pathogen affecting individuals with **cystic fibrosis**. Most patients colonized with *B. cepacia* are coinfected with *Pseudomonas aeruginosa*. The correlation of the two strains in these patients led biologists to hypothesize that interspecies communication with *P. aeruginosa* may help *B. cepacia* enhance its own pathogenicity. Indeed, the addition of *P. aeruginosa* to clones of *B. cepacia* results in a significant increase in the synthesis of proteases (i.e., enzymes needed to break down proteins), suggesting the presence of quorum sensing — *B. cepacia* may profit from pheromones made by a different species in order to improve its own chances of survival.

Share Next step > Discussions

Bioinformatics Algorithms

Stepic Full screen

The **dalton** (abbreviated **Da**) is the unit used for measuring atomic masses on a molecular scale. One dalton is equivalent to one twelfth of the mass of carbon-12 and has a value of approximately $1.66 \cdot 10^{-27}$ kg. The **monoisotopic mass** of a molecule is equal to the sum of the masses of the atoms in that molecule, using the mass of the most abundant **isotope** for each element. See the table below.

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

Share Next step > Discussions

Bioinformatics Algorithms

Stepic Full screen

Selenocysteine is a proteinogenic amino acid that exists in all kingdoms of life as a building block of a special class of proteins called selenoproteins. Unlike other amino acids, selenocysteine is not directly encoded in the genetic code. Instead, it is encoded in a special way by a UGA codon, which is normally a stop codon through a mechanism known as **translational recoding**.

Pyrrolysine is a proteinogenic amino acid that exists in some archaea and methane-producing bacteria. In organisms incorporating pyrrolysine, this amino acid is encoded by UAG, which also normally acts as a stop codon.

Share Next step > Discussions

Bioinformatics Algorithms

Stepic Full screen

If $A = (a_1 = 0, a_2, \ldots, a_n)$ is a set of n points on a line segment in increasing order ($a_1 < a_2 < \cdots < a_n$), then ΔA denotes the collection of all pairwise differences between points in A. For example, if $A = (0, 2, 4, 7, 10)$, then

$$\Delta A = (-10, -8, -7, -6, -5, -4, -3, -3, -2, -2, 0, 0, 0, 0, 0, 2, 2, 3, 3, 4, 5, 6, 7, 8, 10)$$

The turnpike problem asks us to reconstruct A from ΔA.

Turnpike Problem: *Given all pairwise distances between points on a line segment, reconstruct the positions of those points.*

Input: A collection of integers L.

Output: A set A such that $\Delta A = L$.

CODE CHALLENGE: Solve the Turnpike Problem.

Sample Input:

```
-10 -8 -7 -6 -5 -4 -3 -3 -2 -2 0 0 0 0 0 2 2 3 3 4 5 6 7 8 10
```

Sample Output:

```
0 2 4 7 10
```

Start Quiz (limit: 5 minutes)

Share Submit Discussions

Bioinformatics Algorithms

Chapter 1: Where Does DNA Replication Begin?: ...

Chapter 2: How Do We Sequence Antibiotics?:

17. The Discovery of Antibiotics
18. How Do Bacteria Make Antibiotics?
19. Dodging the Central Dogma
20. Sequencing Antibiotics by Shattering Them into Pieces
21. A Brute Force Algorithm for Cyclopeptide Sequencing
22. A Faster Algorithm for Cyclopeptide Sequencing
23. How Fast is This Algorithm?
24. Adapting Cyclopeptide Sequencing for Spectra with Errors
25. From 20 to More than 100 Amino Acids
26. The Spectral Convolution Saves the Day

We will now outline an approach to solving the Turnpike Problem that is polynomial in the length of the line segment. Given a collection of integers $A = (a_1 < a_2 < \cdots < a_n)$, the **generating function** of A is the polynomial

$$A(x) = \sum_{i-1}^{n} x^{a_i}$$

For example, if $A = (0, 2, 4, 7, 10)$, then

$$A(x) = x^0 + x^2 + x^4 + x^7 + x^{10}$$

$$\Delta A(x) = x^{-10} + x^{-8} + x^{-7} + x^{-6} + x^{-5} + x^{-4} + 2x^{-3} + 2x^{-2} + 5x^0 + 2x^2 + 2x^3 + x^4 + x^5 + x^6 + x^7 + x^8 + x^{10}$$

You can verify that the generating function for $\Delta A(x)$ is equal to $A(x) \cdot A(x^{-1})$. Thus, the Turnpike Problem reduces to a problem about polynomial factorization. Just as an integer can be broken down into its prime factors, a polynomial with integer coefficients can be factored into "prime" polynomials having integer coefficients. If we can factor $\Delta A(x)$ and determine which prime factors correspond to $A(x)$ and which prime factors correspond to $A(x^{-1})$, then we will know $A(x)$ and therefore A. Rosenblatt and Seymour, 1982 described such a method to represent $A(x)$ as $A(x) \cdot A(x^{-1})$. Since a polynomial can be factored in time polynomial in its maximum exponent, $\Delta A(x)$ can be factored in time polynomial in the total length of the line segment, which yields the desired pseudo-polynomial algorithm for the Turnpike Problem.

STOP and Think: Can the generating function approach be modified to address the case when there are errors in the pairwise differences?

Bioinformatics Algorithms