Лаб. 2 часть 2

Секвенирование антибиотиков

Формат сдачи: Jupyter Notebook или любой другой Notebook с краткими ответами на вопросы от Question-а и кодом для задания на слайде 16.

Лично преподавателю, показать ответы и запустить работающий код.

Материалы:

Компо, Певзнер «Алгоритмы в биоинформатике» с. 204 (202) – 235 (233)

Секвенирование антибиотиков-циклопептидов

Компо, Певзнер «Алгоритмы в биоинформатике» С. 204 (202) – глава 4

Идея

Многие антибиотики являются циклическими нерибосомными пептидами, что не позволяет получать их последовательность ДНК (соответственно аминокислот также) напрямую из генома.

А эта информация нам жизненно необходима, что же делать?

Решение

Не паниковать! И использовать масс-спектрометр

Компо, Певзнер «Алгоритмы в биоинформатике» С. 211 (209)

Задача секвенирования циклопептидов

Пока для простоты будем считать, что масс-спектрометр разрывает копии циклического пептида по любым возможным связям, так что результирующий экспериментальный спектр содержит массы всех возможных линейных фрагментов пептида, называемых **субпептидами**. Например, циклический пептид NQEL имеет 12 возможных субпептидов: N, Q, E, L, NQ, QE, EL, LN, NQE, QEL, ELN и LNQ. Мы также предполагаем, что субпептиды могут встречаться более одного раза, если аминокислота встречается в пептиде несколько раз (например, ELEL также имеет 12 субпептидов: E, L, E, L, EL, LE, EL, LE, ELE, LEL, ELE и LEL).

Упражнение.

Расправьте спину, оторвитесь от монитора, закройте глаза и представьте что-то, что делает вас счастливыми.

А теперь как вы думаете зовут 2-х котиков (female and male) Анны Дмитриевны?)

Теоретический спектр циклического пептида *Peptide*, обозначаемого *Cyclospectrum*(*Peptide*), представляет собой совокупность всех масс его субпептидов в дополнение к массе 0 и массе всего пептида, причем массы упорядочены от наименьшей к наибольшей. Будем считать, что теоретический спектр может содержать повторяющиеся элементы, как в случае NQEL (показан ниже), где **NQ** и **EL** имеют одинаковую массу.

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

Таблица молекулярных масс аминокислот

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

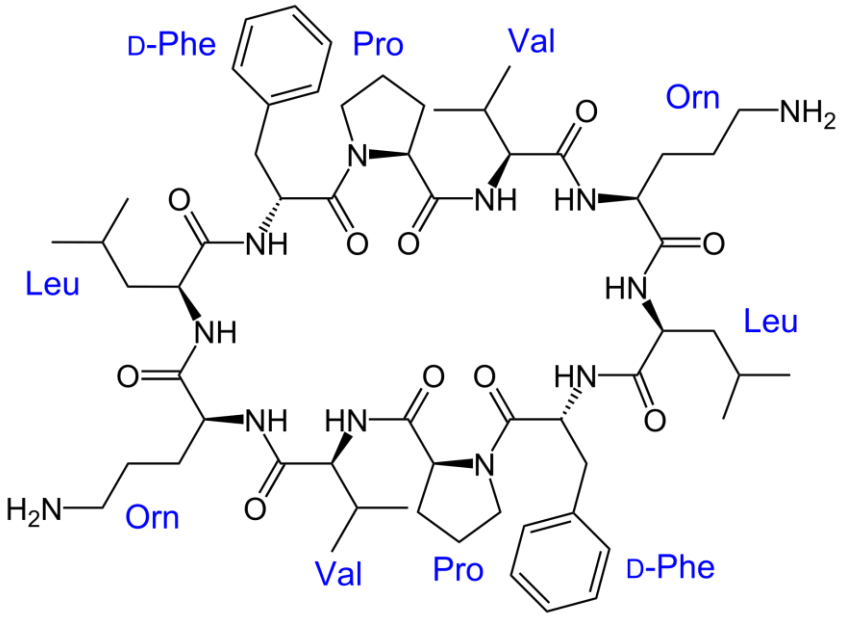
Алфавит аминокислот:

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

Прим. касательно практики
При построении теоретического спектра, если массы каких-либо субпептидов совпадают,
мы их повторяем (т.е. уникальность каждой массы обеспечивать не нужно).

Уже знакомый нам **Грамицидин С**
И его последовательность аминокислот:

Val-Gly-Ala-Leu-Ala-Val-Val-Val-Trp-Trp-Trp-Leu



1) Представьте строкой целочисленных аминокислотных масс

2) Какая общая масса Грамицидина С?

В предыдущей части лабораторной мы работали
с **теоретическим** спектром циклопептида.
На практике же (в ходе эксперимента) получают **экспериментальный** спектр.
И чаще всего они не совпадают.

Иногда в экспериментальном спектре появляются
ложные массы (зеленые), а каких-то не хватает (синие).

Теоретический:	0		113	114	128	**129**	227	**242**	**242**	257		355	356	370	371	484
Экспериментальный:	0	**99**	113	114	128		227			257	**299**	355	356	370	371	484

3) Как из теоретического спектра получить массу пептида?

Стоит также заметить, что

В общем случае (не ограничиваясь алфавитом аминокислот) одному теоретическому спектру может соответствовать несколько пептидов.

Пример: «пептиды» 1-1-3-3 и 1-2-1-4 имеют спектр [0 1 1 2 3 3 4 4 5 5 6 7 7 8]

Верно ли это для алфавита аминокислот? Как вы думаете?
(вопрос не обязательный к ответу, но идеи приветствуются)

Решаемая основная задача

По экспериментальному спектру восстановить последовательность аминокислот циклопептида.

Но в данной лабораторной мы решим упрощенную задачу

1. Предположим, что наш экспериментальный спектр **идеален**, т.е. совпадает с теоретическим.

Задача

Задача секвенирования циклопептида: *при заданном идеальном спектре найти циклический пептид, теоретический спектр которого соответствует экспериментальному спектру.*

Input: набор (возможно, повторяющихся) целых чисел *Spectrum*, соответствующих экспериментальному спектру.

Output: аминокислотная строка *Peptide* такая, что *Cyclospectrum*(*Peptide*) = *Spectrum* (если такая строка существует).

Решение. Полный перебор (brute force algorithm)

1) По спектру получить массу пептида
2) Сгенерировать все пептиды с такой массой. А сколько их (см.график)?
3) Для каждого пептида построить теоретический спектр и сравнить с заданным (экспериментальным).

Я ждал 14 лет… Могу еще подождать!

Гарри, ты ли это?

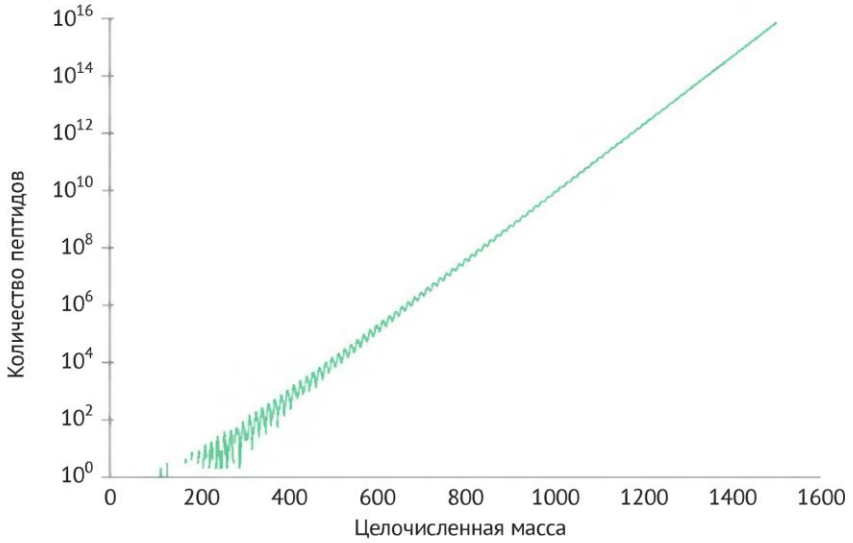
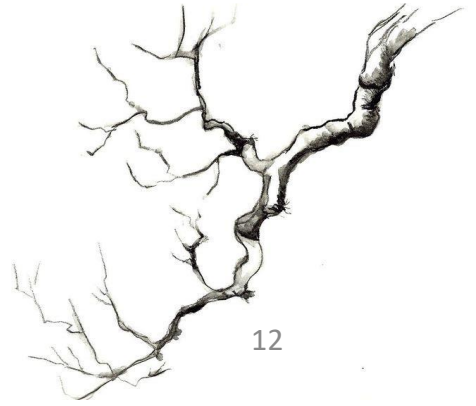


Рис. 4.8 Количество пептидов с заданной целочисленной массой растет экспоненциально

Улучшенное решение. Добавим ветвей и границ

Вместо проверки всех циклических пептидов с заданной массой наш новый метод решения проблемы секвенирования циклопептидов будет «выращивать» линейные пептиды-кандидаты, чьи теоретические спектры «согласуются» с экспериментальным спектром.

Для экспериментального спектра *Spectrum* циклического пептида линейный пептид **согласуется** с *Spectrum*, если каждая масса в его теоретическом спектре содержится в *Spectrum*. Если масса появляется в теоретическом спектре линейного пептида более одного раза, то она должна появляться как минимум столько раз в *Spectrum*, чтобы линейный пептид соответствовал *Spectrum*.



Идея - Линейные пептиды-кандидаты

Растим дерево, где от каждого узла идет 18 новых (размер алфавита аминокислот).
Для каждого нового проверяем
1) соотносится ли спектр построенного линейного пептида с заданным спектром циклопептида
Если нет, то обрезаем узел.

Если да, то

2) совпадает ли масса с массой циклопептида
Если да, то закольцовываем линейный пептид, строим теоретический спектр и проверяем на соответствие заданному.

Ключом к нашему новому алгоритму является то, что каждый линейный субпептид циклического пептида *Peptide* согласуется с *Cyclospectrum*(*Peptide*).

Таким образом, чтобы решить задачу секвенирования циклопептидов для *Spectrum*, мы можем безопасно запретить все пептиды, несовместимые со *Spectrum*, из растущего набора *Peptide*, что усиливает шаг ограничения, описанный нами выше.

Упражнение. Сколько субпептидов имеет линейный пептид заданной длины n? (Включите пустой пептид и весь пептид целиком.)

Input: целое число *n*.

Output: количество субпептидов линейного пептида длины *n*.

Как насчет шага ветвления? Имея текущую коллекцию линейных пептидов *Peptides*, определите *Expand(Peptides)* как новый набор, содержащий все возможные расширения пептидов в Peptides на одну массу аминокислоты. Теперь мы можем предоставить псевдокод для алгоритма ветвей и границ, называемого **CyclopeptideSequencing**.

```
CyclopeptideSequencing(Spectrum)
    CandidatePeptides ← набор, содержащий только пустой пептид FinalPeptides ←
    пустой набор строк
        while CandidatePeptides не пустой
            CandidatePeptides ← Expand(CandidatePeptides)
            for каждого пептида Peptide в CandidatePeptides
                if Mass(Peptide) = ParentMass(Spectrum)
                    if Cyclospectrum(Peptide) = Spectrum и Peptide не присутствует
                    в FinalPeptides
                        добавить Peptide к FinalPeptides
                    удалить Peptide из CandidatePeptides
                else if Peptide не согласуется со Spectrum
                    удалить Peptide из CandidatePeptides
    return FinalPeptides
```

Задание

Реализовать алгоритм Ветвей и границ для
построения аминокислотной последовательности циклопептида
по заданному идеальному экспериментальному спектру.

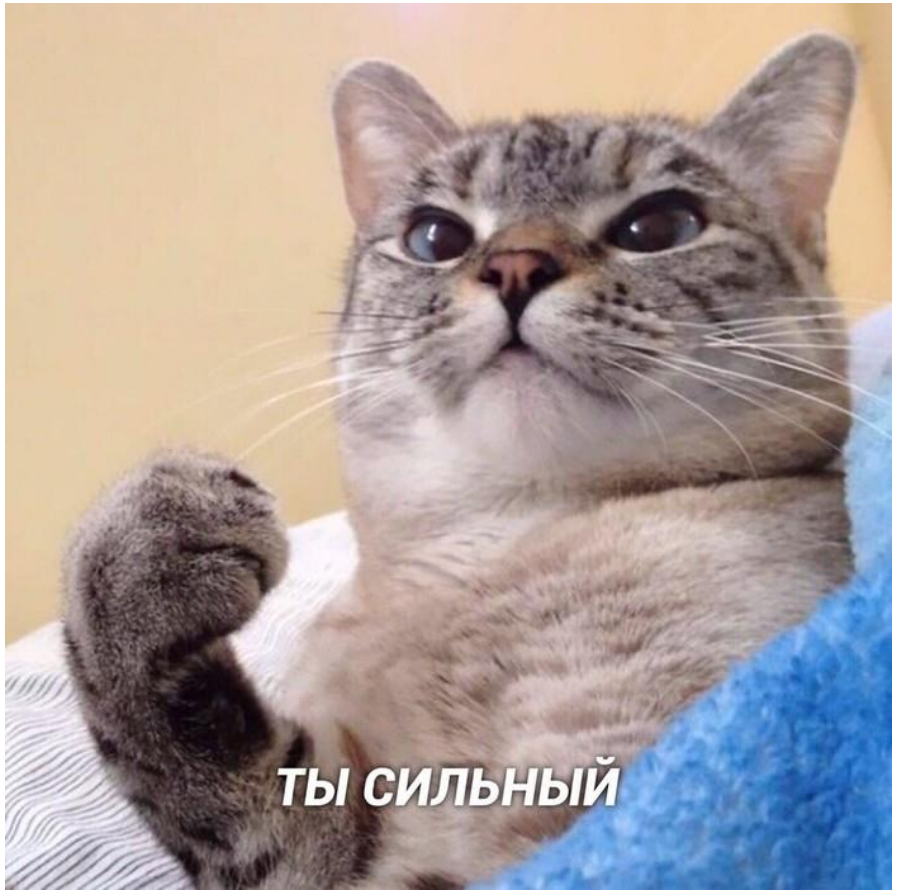
Входные данные

За идеальный экспериментальный спектр взять теоретический спектр следующего пептида:
WFNQYVK.

Т.е. решается как будто задача декодирования (сначала получите спектр, а затем по спектру пептид).

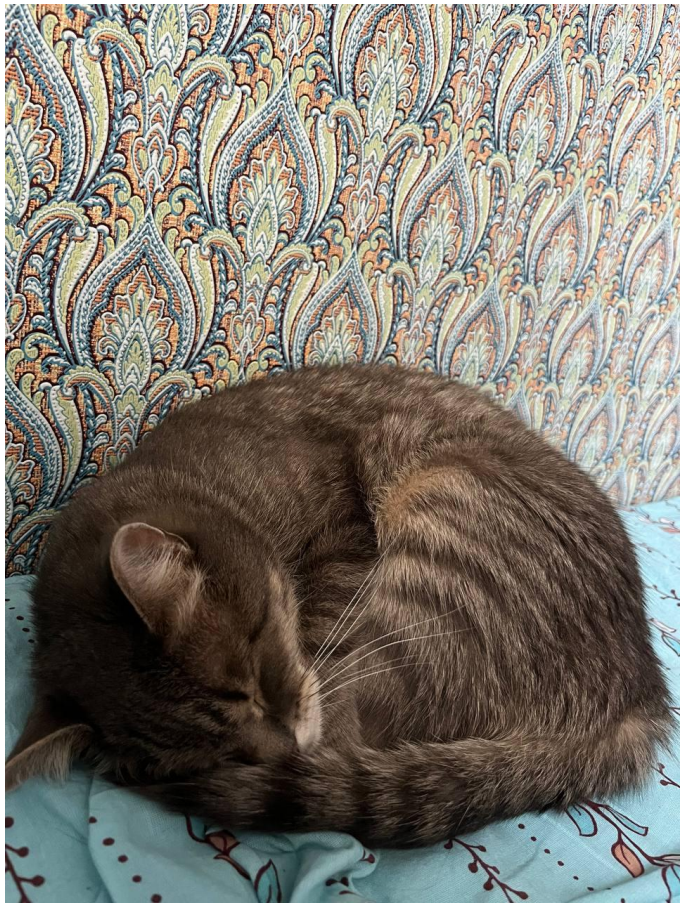
Что вы можете сказать про сложность построенного алгоритма?

Рыся и Мальчик верят в тебя!

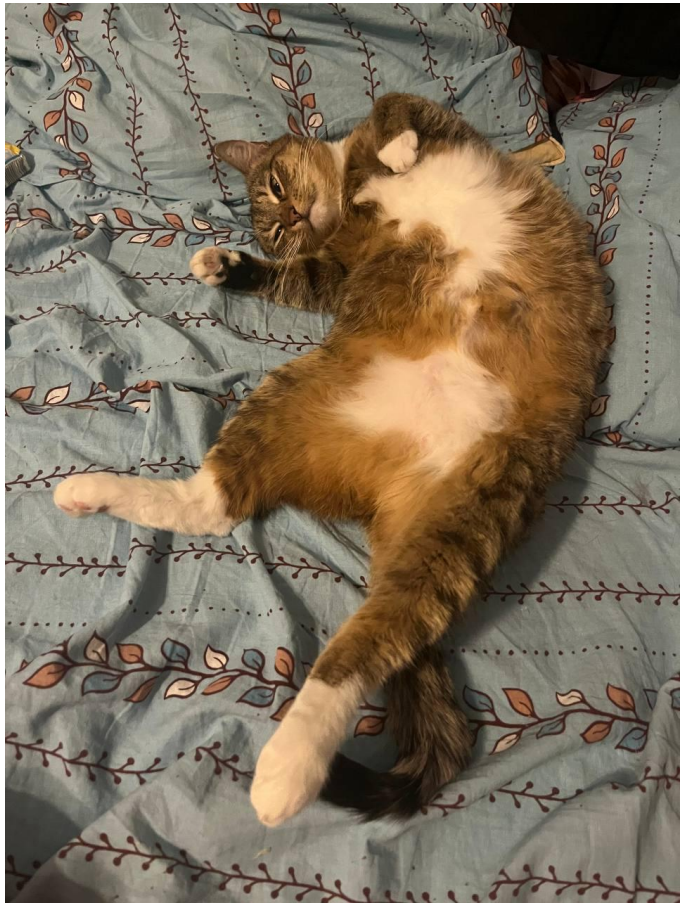


← рандомный котя, я его первый раз вижу (:

Помни: Не сиди без дела, а лежи



Мальчик



Рыся