Введение

Пептиды – органические молекулы, состоящие из аминокислот (точнее, аминокислотных остатков), “сцепленных” друг с другом. В простейшем случае *линейных пептидов* аминокислоты образуют последовательность или цепь. Такие пептиды можно представлять как строчки над алфавитом, буквы которого соответствуют различным аминокислотам – обычно рассматривается алфавит из 20-и “стандартных” аминокислот.

У каждой аминокислоты есть заранее известная масса. Масса пептида равна сумме масс аминокислот. Поскольку линейные пептиды представляются строчками, можно говорить о префиксах и суффиксах линейного пептида как о других линейных пептидах и, в частности, об их массах.

Для определения строения (порядка аминокислот) неизвестного пептида используется технология *масс-спектрометрии*. Схематично для случая линейных пептидов её можно описать так: берётся большое количество молекул одного и того же пептида, и каждая из них “рвётся” в случайной позиции с образованием двух ионов. После чего, благодаря использованию электромагнитных сил и явления инерции, определяется относительное количество ионов с определенным отношением массы к заряду. На выходе после некоторой немаловажной пост-обработки, которая здесь рассматриваться не будет, получается *эмпирический спектр* пептида – гистограмма распределения относительного количества **префиксов** линейного пептида в зависимости от их **массы**.

Пептиду также можно сопоставить *теоретический спектр*, например, считая, что все префиксы встречаются одинаково часто. Конечно, из-за погрешностей измерений теоретический и эмпирический спектр одного и того же пептида могут не совпадать. В качестве меры “схожести” таких спектров вводится некоторая оценочная функция, например, их скалярное произведение.

Из соображений схожести теоретического и эмпирического спектров, а также, иногда, априорных догадок о строении неизвестного пептида, по результатам масс-спектрометрии можно высказать гипотезу о строении исследуемого образца. Задача оценки надежности идентификации пептида состоит в оценке надежности этой гипотезы. Один из способов формализовать эту задачу – это спросить, с какой вероятностью случайный пептид той же массы будет иметь значение оценочной функции не меньшее, чем предполагаемый. Если считать, что распределение на пептидах равномерное, задача сводится к подсчету количества таких пептидов.

Следует отметить, что в реальных условиях решение этой задачи включает в себя не только непосредственный подсчёт количества подходящих пептидов, но также учёт неидеальности входных данных: например, правильную их дискретизацию. Однако естественно перед работой с этими техническими деталями рассмотреть модельную версию задачи, в которой они не возникают. В этой статье так и сделано.

Мы обсудили основные понятия в случае линейных пептидов. Пептиды могут не быть линейными. Например, они могут быть Y-образные: “общий префикс” аминокислот в какой-то момент “разделяется” на две “ветви”. Такие пептиды и называются разветвлёнными. Данные выше описания достаточно естественным образом перено-

сятся на случай разветвлённых пептидов, но я сразу дам формальные определения для дальнейшей работы.

Постановка задачи

Пусть Σ — алфавит, буквы которого соответствуют аминокислотам. Именем $mass\colon \Sigma^* \to \mathbb{N}_0$ будем обозначать функцию, сопоставляющую аминокислотам их массы и продолженную на множество всех строк.

Определение. Разветвленным пептидом будем называть тройку (b, l, r) строк над алфавитом Σ. Строка b называется общим префиксом разветвленного пептида, а строки l и r – левой и правой ветвью соответственно.

Функция $mass$ естественным образом продолжается на множество разветвлённых пептидов: если $p = (b, l, r)$, то $mass(p) = mass(b) + mass(l) + mass(r)$.

Определение. Пусть $p = (b, l, r)$ — разветвленный пептид, $mass(p) = M$. Рассмотрим

$$\mathcal{P}_p := \{b_1 \dots b_i \mid 1 \leqslant i \leqslant |b|\} \cup \{bl_1 \dots l_i \mid 1 \leqslant i \leqslant |l|\} \cup \{br_1 \dots r_i \mid 1 \leqslant i \leqslant |r|\}$$

— множество непустых префиксов пептида p. Теоретическим спектром $Spec(p)$ пептида p будем называть строку $s_0 \dots s_M$, где $s_i = \left|\{x \in \mathcal{P}_p \mid mass(x) = i\}\right|$. Отметим, что $s_i \in \{0, 1, 2\}$.

Определение. Счётом пептида p относительно спектра $S \in \mathbb{N}_0^{M+1}$ будем называть число

$$Score_S(p) = \langle Spec(p), S\rangle = \sum_{i=0}^{M} Spec(p)_i S_i$$

— скалярное произведение $Spec(p)$ и S.

Пусть дан некоторый эмпирический спектр $S = s_1 \dots s_M$, где $s_i \in \mathbb{N}_0$. Для удобства будем считать $s_0 = 0$. Наша задача состоит в том, чтобы для каждой массы $0 \leqslant w \leqslant M$ и каждого счёта $0 \leqslant t \leqslant T$ посчитать $ans(w, t)$ — количество разветвлённых пептидов массы w со счётом t относительно данного спектра S. Мы рассмотрим три подхода к её решению. Второй и третий подходы будут использовать идеи из первого, а также иметь меньшую вычислительную сложность.

Линейный случай

Для начала вспомним, как можно решать аналогичную задачу для линейных пептидов, то есть разветвленных пептидов вида $p = (b, \varepsilon, \varepsilon)$. В решении будет использоваться метод динамического программирования. Здесь будет описан алгоритм, параметризованный некоторыми “начальными данными”. Буквально в задаче о линейных пептидах начальные данные имеют определенные задачей значения, но

тот же самый алгоритм с другими начальными данными будет использоваться в дальнейшем.

Определим $\lambda(w, t)$ как количество линейных пептидов массы w со счётом t. Начальными данными линейного случая условимся называть вектор значений $\lambda(0, t)$ по всем t. Заметим, что в задаче о линейных пептидах $\lambda(0, 0) = 1$ и $\lambda(0, t) = 0, t > 0$. Для $w > 0$ $\lambda(w, t)$ можно выразить через значения $\lambda(w', t')$ с $w' < w$ следующим образом:

$$\lambda(w, t) = \sum_{a \in \Sigma} \lambda(w - |a|, t - s_w),$$

где $|a| = mass(a)$ — масса аминокислоты a. Тут подразумевается, что $\lambda(w', t') = 0$ при $w' < 0$ или $t' < 0$.

Действительно, любой линейный пептид массы $w > 0$ представляется единственным образом в виде конкатенации более короткого пептида массы $w' = w - |a|$ и некоторой аминокислоты a. При этом, в силу определения $Score$-функции как скалярного произведения, счёт первого пептида больше счёта второго пептида ровно на значение эмпирического спектра в позиции w, так как в соответствующей позиции теоретического спектра будет стоять 1. Формально нужно было бы ввести понятие (w, t)-подходящих пептидов и рассмотреть их множество $A_{w,t}$ так, чтобы по определению λ получалось $\lambda(w, t) = |A_{w,t}|$, а затем из приведенных выше соображений получить $A_{w,t} = \bigsqcup\limits_{a \in \Sigma} A_{w-|a|,t-s_w} a$, где $Ax = \{wx \mid w \in A\}$, но я не буду этого делать.

Таким образом, перебирая значения w' в возрастающем порядке, значения $\lambda(w, t)$ можно посчитать за время $O(MT|\Sigma|)$. Подробнее о линейном случае рассказано в [1].

Первый подход

Случай пустого общего префикса

Рассмотрим случай пептидов вида $p = (\varepsilon, l, r)$, то есть разветвленных пептидов с пустым общим префиксом. Определим $\chi(w_1, w_2, t)$ как количество пептидов $p = (\varepsilon, l, r)$ таких, что $mass(l) = w_1, mass(r) = w_2$ и $Score(p) = t$. Начальными данными случая пептидов с пустым общим префиксом будем называть значения $\chi(0, 0, t)$ по всем t. Опять же, в текущей задаче $\chi(0, 0, 0) = 1$ и $\chi(0, 0, t) = 0$ при $t > 0$. Алгоритм вычисления всех значений χ следующий. Сперва вычислим $\chi(w_1, 0, t)$ по рекуррентной формуле:

$$\chi(w_1, 0, t) = \sum_{a \in \Sigma} \chi(w_1 - |a|, 0, t - s_{w_1}),$$

то есть полностью аналогично линейному случаю. Затем для каждого фиксированного w_1 вычислим недостающие значения $\chi(w_1, w_2, t)$ с $w_2 > 0$:

$$\chi(w_1, w_2, t) = \sum_{a \in \Sigma} \chi(w_1, w_2 - |a|, t - s_{w_2}),$$

то есть по существу аналогично линейному случаю с начальными данными $\lambda(0, t) = \chi(w_1, 0, t)$. Аналогичным линейному случаю образом легко проверить, что эти рекур-

рентные формулы корректны. Время работы этого алгоритма есть $O(M^2T|\Sigma|)$.

Дважды посчитанные пептиды

С практической точки зрения мы не хотим различать пептиды вида $p = (\varepsilon, l, r)$ и $p' = (\varepsilon, r, l)$. Условимся называть такие пептиды зеркальными по отношению друг к другу. Заметим, что если $mass(l) = mass(r) = w$ и $Score(p) = t$, и к тому же $p \neq p'$, то $Score(p') = t$ и в предложенном алгоритме такие пары пептидов были посчитаны дважды при вычислении $\chi(w, w, t)$. Пусть $\sigma(w, t)$ — количество пептидов вида $p = (\varepsilon, l, l)$, которые называются симметричными, со счётом t и массой $2w$. Тогда правильные значения $\chi(w, w, t)$ можно получить как:

$$\chi(w, w, t) \leftarrow \frac{\chi(w, w, t) - \sigma(w, t)}{2} + \sigma(w, t)$$

Таким образом, нам достаточно вычислить значения $\sigma(w, t)$. Делается это аналогично линейному случаю: $\sigma(0, 0) = \chi(0, 0, 0) = 1$, $\sigma(0, t) = \chi(0, 0, t) = 0$ при $t > 0$ и

$$\sigma(w, t) = \sum_{a \in \Sigma} \sigma(w - |a|, t - 2s_w),$$

поскольку при переходе от $p = (\varepsilon, l, l)$ к $p' = (\varepsilon, la, la)$ в $Spec(p)$ на позиции $w = mass(la)$ появляется пик интенсивности 2 — по единице от каждой из ветвей. Время вычисления $\sigma(w, t)$ есть $O(MT|\Sigma|)$, то есть в M раз меньше, чем время вычисления $\chi(w_1, w_2, t)$. Эта же разница сохранится и при переходе к общему случаю разветвленных пептидов с необязательно пустым префиксом.

Далее мы не будем подробно рассматривать детали, касающиеся учета пар зеркальных пептидов в общем случае. Все подробности можно посмотреть в коде реализации предложенных алгоритмов.

Вычисление ответа для случая пептидов с пустым общим префиксом

Имея значения $\chi(w_1, w_2, t)$ несложно вычислить $ans'(w, t)$ — количество пептидов с пустым общим префиксом массы w со счётом t. А именно:

$$ans'(w, t) = \sum_{\substack{w_1, w_2 \geqslant 0 \\ w_1 + w_2 = w \\ w_1 \leqslant w_2}} \chi(w_1, w_2, t)$$

Условие $w_1 \leqslant w_2$ появляется, поскольку мы не различаем зеркальные пептиды в том числе с различными массами левой и правой ветвей. Несложно убедиться, что $\chi(w_1, w_2, t) = \chi(w_2, w_1, t)$ при $w_1 \neq w_2$, поскольку отображение $(\varepsilon, l, r) \mapsto (\varepsilon, r, l)$ задает биекцию между соответствующими множествами. Заметим также, что если мы не хотим учитывать пептиды, у которых одна или обе ветви пустые, нам достаточно заменить условие $w_1, w_2 \geqslant 0$ на $w_1, w_2 \geqslant 1$ — это будет важно в дальнейшем.

Переход к общему случаю

Посчитаем теперь $\eta(w_0, w_1, w_2, t)$ — количество разветвленных пептидов $p = (b, l, r)$ с $mass(b) = w_0$, $mass(l) = w_1$, $mass(r) = w_2$ и счётом t. Здесь мы будем считать, что $w_1, w_2 > 0$, то есть что $l, r \neq \varepsilon$. В противном случае возник бы неприятный эффект: например, (aa, bc, ε) и (aab, c, ε) считались бы различными разветвленными пептидами, хотя по существу это один и тот же (линейный) пептид. Ясно, что

$$ans(w, t) = \sum_{\substack{w_0 \geqslant 0,\ w_1, w_2 \geqslant 1 \\ w_0 + w_1 + w_2 = w \\ w_1 \leqslant w_2}} \eta(w_0, w_1, w_2, t)$$

Итак, сначала мы считаем $\lambda(w_0, t)$ так же, как в линейном случае для данного спектра S. Затем для каждого фиксированного w_0 мы делаем следующее: сперва мы сдвигаем спектр S на w_0 позиций влево, то есть определяем спектр S' длины $M - w_0$ с $s'_i = s_{w_0+i}$. Затем считаем $\chi_{w_0}(w_1, w_2, t)$ как в случае пептидов с пустым общим префиксом для спектра S', но с нестандартными начальными данными: полагаем $\chi_{w_0}(0, 0, t) = \lambda(w_0, t)$. После чего присваиваем $\eta(w_0, w_1, w_2, t) = \chi_{w_0}(w_1, w_2, t)$.

Действительно, количество пептидов вида $(b, \varepsilon, \varepsilon)$ с $mass(b) = w_0$ и счётом t это и есть $\lambda(w_0, t)$, а рекуррентные соотношения для количеств пептидов вида (b, l, r) будут такими же, как и для (ε, l, r), но со сдвинутым на $w_0 = mass(b)$ спектром.

Время работы получившегося алгоритма составляет $O(M^3 T|\Sigma|)$. Заметим также, что не обязательно хранить значения $\eta(w_0, w_1, w_2, t)$ для всех w_0. Напротив, вычислив $\chi_{w_0}(w_1, w_2, t)$, можно сразу обновить текущие значения $ans(w, t)$, добавив в соответствующую сумму очередные слагаемые.

Альтернативный переход к общему случаю

Рассмотрим также немного отличный от предыдущего переход к общему случаю. Он будет иметь худшее время работы, но его идея будет полезна во втором подходе к решению задачи. Итак, давайте снова посчитаем $\lambda(w_0, t)$ как в линейном случае и для каждого w_0 посчитаем $\chi'_{w_0}(w_1, w_2, t)$ как в случае пептидов с пустым общим префиксом для спектра S', но уже со стандартными начальными данными. Тогда

$$\eta(w_0, w_1, w_2, t) = \sum_{s=0}^{t} \lambda(w_0, s) \cdot \chi'_{w_0}(w_1, w_2, t - s)$$

— свёртка λ и χ'_{w_0} по переменной t. Её можно посчитать с помощью быстрого преобразования Фурье за время $O(Tlog(T))$. Таким образом, итоговое время работы получается равным $O(M^3 T(|\Sigma| + log(T)))$

Второй подход

Вернемся снова к решению задачи для пептидов с пустым общим префиксом. Посчитаем сначала $\lambda(w_1, t)$ — линейный случай для данного спектра. Тогда количество

пептидов с пустым общим префиксом, массой w и счётом t $x(w,t)$ выражается через $\lambda(w_1,t)$:

$$x(w,t) = \sum_{\substack{0 \leqslant w_1 \leqslant w \\ 0 \leqslant s \leqslant t}} \lambda(w_1,s) \cdot \lambda(w-w_1, t-s)$$

Это двумерная свертка $\lambda(w_1,t)$ с собой же, которую можно вычислить с помощью двумерного FFT за время $O(MTlog(MT))$.

В общем случае $ans(w,t)$ можно посчитать аналогичным предыдущему подходу образом: для каждого фиксированного w_0 посчитать $x_{w_0}(w,t)$ в точности так же, как в случае пустого общего префикса, но для сдвинутого спектра S', а затем свернуть $x_{w_0}(w,t)$ с $\lambda(w_0,t)$ по переменной t с помощью одномерного FFT, где $\lambda(w_0,t)$ — линейный случай, посчитанный уже для исходного спектра S. Итоговое время работы получается равным $O(M^2Tlog(MT))$.

Третий подход

Третий подход является развитием первого. Основная идея заключается в том, чтобы избавиться от лишней размерности w_0 в $\eta(w_0,w_1,w_2,t)$. Для этого мы определяем $\pi(w_1,w_2,t)$ как количество пептидов $p=(b,l,r)$ таких, что $l \neq \varepsilon$, $mass(b)+mass(l) = w_1$, $mass(r) = w_2$ и $Score(p) = t$, а также вспомогательную величину $\gamma(w_1,w_2,t)$, равную количеству линейных пептидов $(b,\varepsilon,\varepsilon)$ массы w_2 со счётом t, таких что $\exists m \in \{0,\dots,|b|\} : mass(b_1 \dots b_m) = w_1$, то есть таких, что некоторый их префикс имеет массу ровно w_1.

Сперва поймём, как вычислить $\gamma(w_1,w_2,t)$. Посчитаем $\lambda(w_1,t)$ — линейный случай для данного спектра S. Затем для каждого фиксированного w_1 посчитаем $\lambda_{w_1}(\delta,t)$ — линейный случай для сдвинутого на w_1 влево спектра S' с нестандартными начальными данными: полагаем $\lambda_{w_1}(0,t) = \lambda(w_1,t)$. После этого можно присвоить $\gamma(w_1, w_1+\delta, t) = \lambda_{w_1}(\delta,t)$.

Теперь поймём, как вычислить $\pi(w_1,w_2,t)$. Это можно сделать по следующему рекуррентному соотношению:

$$\pi(w_1,w_2,t) = \sum_{a \in \Sigma} \pi(w_1-|a|, w_2, t-s_{w_1}) + \gamma(w_1-|a|, w_1-|a|+w_2, t-s_{w_1})$$

Действительно, для любого пептида $p=(b,la,r)$ с параметрами w_1,w_2,t имеет место следующая альтернатива:

1. $l \neq \varepsilon$ и p получен приписыванием a к левой ветви пептида (b,l,r) с параметрами $w_1-|a|, w_2, t-s_{w_1}$, за что отвечает первое слагаемое.

2. $l = \varepsilon$ и p получен началом ветвления в позиции, соответствующей массе

   $$mass(b) = mass(bl) = mass(bla) - |a| = w_1 - |a|$$

   линейного пептида $(br,\varepsilon,\varepsilon)$ массы $mass(br) = w_1 - |a| + w_2$, имеющего префикс массы ровно $w_1-|a|$. За этот случай отвечает второе слагаемое.

Причем эти случаи не пересекаются и полностью исчерпывают возможности получения пептида с параметрами w_1, w_2, t. Итоговый ответ считается как:

$$ans(w,t) = \sum_{\substack{w_1,w_2 \geqslant 1 \\ w_1+w_2=w}} \pi(w_1, w_2, t)$$

Время работы этого алгоритма есть $O(M^2T|\Sigma|)$.

Список литературы

[1] Sangtae Kim, Nitin Gupta, and Pavel A. Pevzner. *Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases*, 2008.