Фолдинг белка

Задача

Имея белковую последовательность длины N, необходимо уложить (свернуть) ее на двумерную решетку со стороной L так, чтобы суммарная энергия белковой свертки равнялась K.

Решение

Входные данные

Алгоритм принимает на вход следующие величины:

- Целое число L — сторона двумерной решетки.
- Целое число K — суммарная энергия итоговой белковой свертки.
- Белковую последовательность длины N, состоящую из символов H и P. Для удобства представим последовательность как массив булевых переменных $R = \{R_i\},\ i \in [0, N-1]$, где $R_i = True$, если i-ый элемент входной последовательности равен H.
- Координаты двух первых элементов последовательности: (x_0, y_0) и (x_1, y_1).

Переменные

Заведем следующие массивы с булевыми переменными:

- Массивы $X = \{X_{ia}\}$ и $Y = \{Y_{ia}\}$, $i \in [0, N-1],\ a \in [0, L-1]$. Элемент $X_{ia} = True$ тогда и только тогда, когда $x_i = a$. Аналогично для массива Y.
- Массив $neighbours = \{n_{ij}\},\ i \in [0, N-1],\ j \in [0, N-1]$, где $n_{ij} = True$ тогда и только тогда, когда i-ый и j-ый элементы последовательности расположены в соседних узлах решетки.

Определение "соседа"

Так как у нас квадратная решетка, следовательно i и j элементы будут соседями только тогда, когда j-ый элемент находится либо сверху, либо снизу, либо справа, либо слева от i-ого:

$$\begin{aligned}(neighbours_{ij} = 1) \leftrightarrow\ & (x_i = x_j\ \wedge\ y_i = (y_j + 1)) \vee \\ & (x_i = x_j\ \wedge\ y_i = (y_j - 1)) \vee \\ & (x_i = (x_j - 1)\ \wedge\ y_i = y_j) \vee \\ & (x_i = (x_j + 1)\ \wedge\ y_i = y_j)\end{aligned}$$

В условиях переменных X_{ia} и Y_{ia} данная формула будет выглядеть еще сложнее. Поэтому для упрощения перевода ограничений в КНФ, введем несколько новых массивов с булевыми переменными:

- $right = \{right_{ij}\},\ i \in [0, N-1],\ j \in [0, N-1]$, где $right_{ij} = True$ тогда и только тогда, когда элемент последовательности j расположен справа от элемента i.
- Аналогично определяем массивы $left = \{left_{ij}\}$, $up = \{up_{ij}\}$, $down = \{down_{ij}\}$.

Таким образом получаем более простое определение для соседства и ограничения для массивов направлений:

$$neighbours_{ij} \leftrightarrow right_{ij} \vee left_{ij} \vee up_{ij} \vee down_{ij}$$

$$\bigwedge_{i,j} right_{ij} \leftrightarrow (\bigvee_a X_{ia} \wedge X_{j(a+1)}) \wedge (\bigvee_a Y_{ia} \wedge Y_{ja})$$

$$\bigwedge_{i,j} left_{ij} \leftrightarrow (\bigvee_a X_{ia} \wedge X_{j(a-1)}) \wedge (\bigvee_a Y_{ia} \wedge Y_{ja})$$

$$\bigwedge_{i,j} up_{ij} \leftrightarrow (\bigvee_a X_{ia} \wedge X_{ja}) \wedge (\bigvee_a Y_{ia} \wedge Y_{j(a+1)})$$

$$\bigwedge_{i,j} down_{ij} \leftrightarrow (\bigvee_a X_{ia} \wedge X_{ja}) \wedge (\bigvee_a Y_{ia} \wedge Y_{j(a-1)})$$

Начальные ограничения

- Сперва следует задать начальные координаты:

$$X_{0x_0} \wedge Y_{0y_0} \wedge X_{1x_1} \wedge Y_{1y_1}$$

- Для данного i существует ровно один $X_{ia} = 1$ (аналогично для Y):

$$\bigwedge_i \bigvee_a X_{ia}$$

$$\bigwedge_i \bigwedge_a \bigwedge_{b!=a} \overline{(X_{ia} \wedge X_{ib})}$$

$$\bigwedge_i \bigvee_a Y_{ia}$$

$$\bigwedge_i \bigwedge_a \bigwedge_{b!=a} \overline{(Y_{ia} \wedge Y_{ib})}$$

- Никакие два различных элемента не совпадают по координатам:

$$\bigwedge_i \bigwedge_{j!=i} \bigwedge_a \bigwedge_b (X_{ia} \wedge X_{ja}) \rightarrow \overline{Y_{ib} \wedge Y_{jb}}$$

- Последовательность связна:

$$\bigwedge_{i=0}^{N-2} neighbours_{i(i+1)}$$

- Определение соседей для всех пар элементов:

$$\bigwedge_{i,j} neighbours_{ij} \leftrightarrow right_{ij} \vee left_{ij} \vee up_{ij} \vee down_{ij}$$

- Избегаем "ходьбы" в нескольких направлениях сразу:

$$\bigwedge_{i,j} \overline{(right_{ij} \wedge left_{ij})}$$

$$\bigwedge_{i,j} \overline{(right_{ij} \wedge up_{ij})}$$

$$\bigwedge_{i,j} \overline{(right_{ij} \land down_{ij})}$$

$$\bigwedge_{i,j} \overline{(left_{ij} \land up_{ij})}$$

$$\bigwedge_{i,j} \overline{(left_{ij} \land down_{ij})}$$

$$\bigwedge_{i,j} \overline{(down_{ij} \land up_{ij})}$$

- Не позволяем белку "ходить" через границы решетки:

$$\bigwedge_{i!=j} \overline{(X_{i(L-1)} \land right_{ij})}$$

$$\bigwedge_{i!=j} \overline{(X_{i0} \land left_{ij})}$$

$$\bigwedge_{i!=j} \overline{(Y_{i0} \land down_{ij})}$$

$$\bigwedge_{i!=j} \overline{(Y_{i(L-1)} \land up_{ij})}$$

- "Свяжем" массивы направлений соседства с массивами X и Y:

$$\bigwedge_{i,j} \bigwedge_{a=0}^{L-2} (right_{ij} \land X_{ia}) \to X_{j(a+1)}$$

$$\bigwedge_{i,j} \bigwedge_{a=1}^{L-1} (left_{ij} \land X_{ia}) \to X_{j(a-1)}$$

$$\bigwedge_{i,j} \bigwedge_{a=0}^{L-2} (up_{ij} \land Y_{ia}) \to Y_{j(a+1)}$$

$$\bigwedge_{i,j} \bigwedge_{a=1}^{L-1} (down_{ij} \land Y_{ia}) \to Y_{j(a-1)}$$

Упрощение ограничений для переменных направлений

Рассмотрим ограничение для массива $right$:

$$\bigwedge_{i,j} right_{ij} \leftrightarrow (\bigvee_{a} X_{ia} \land X_{j(a+1)}) \land (\bigvee_{a} Y_{ia} \land Y_{ja})$$

Данное ограничение невозможно выразить формулой полиноминальной длины в КНФ. Чтобы решить эту проблему, выразим $right$ через дополнительные переменные, как в случае с массивом $neighbours$.

Введем следующие дополнительные массивы:

- $MatchX = \{MatchX_{ij}\},\ i \in [0, N-1],\ j \in [0, N-1]$, где $MatchX_{ij} = True$, если $\exists a : X_{ia} \land X_{ja} = True$.
- $MatchX^+ = \{MatchX^+_{ij}\},\ i \in [0, N-1],\ j \in [0, N-1]$, где $MatchX^+_{ij} = True$, если $\exists a : X_{ia} \land X_{j(a+1)} = True$.

- $MatchX^- = \{MatchX^-_{ij}\},\ i \in [0, N-1],\ j \in [0, N-1]$, где $MatchX^-_{ij} = True$, если $\exists a : X_{ia} \wedge X_{j(a-1)} = True$.
- Также заведем аналогичные массивы $MatchY,\ MatchY^+,\ MatchY^-$

Так как ограниение все еще не выражается формулой полиноминальной длины в КНФ, введем еще одну группу дополнительных массивов для проверки соответствия между x_i и x_j:

- $MatchXA = \{MatchXA_{ija}\},\ i \in [0, N-1],\ j \in [0, N-1],\ a \in [0, L-1]$, где $MatchXA_{ija} = True$, если $X_{ia} \wedge X_{ja} = True$.
- $MatchXA^+ = \{MatchXA^+_{ija}\},\ i \in [0, N-1],\ j \in [0, N-1],\ a \in [0, L-2]$, где $MatchXA^+_{ija} = True$, если $X_{ia} \wedge X_{j(a+1)} = True$.
- $MatchXA^- = \{MatchXA^-_{ija}\},\ i \in [0, N-1],\ j \in [0, N-1],\ a \in [1, L-1]$, где $MatchXA^-_{ija} = True$, если $X_{ia} \wedge X_{j(a-1)} = True$.
- Также заведем аналогичные массивы $MatchYA,\ MatchYA^+,\ MatchYA^-$.

Теперь мы можем выразить ограничение для переменных направлений через новые переменные:

- Для направлений:

$$\bigwedge_{i,j} right_{ij} \leftrightarrow (MatchY_{ij} \wedge MatchX^+_{ij})$$
$$\bigwedge_{i,j} left_{ij} \leftrightarrow (MatchY_{ij} \wedge MatchX^-_{ij})$$
$$\bigwedge_{i,j} up_{ij} \leftrightarrow (MatchY^+_{ij} \wedge MatchX_{ij})$$
$$\bigwedge_{i,j} down_{ij} \leftrightarrow (MatchY^-_{ij} \wedge MatchX_{ij})$$

- Для первой группы дополнительных переменных по X (по Y выражается аналогично):

$$\bigwedge_{i,j} MatchX_{ij} \leftrightarrow \bigvee_{a=0}^{L-1} MatchXA_{ija}$$
$$\bigwedge_{i,j} MatchX^+_{ij} \leftrightarrow \bigvee_{a=0}^{L-2} MatchXA^+_{ija}$$
$$\bigwedge_{i,j} MatchX^-_{ij} \leftrightarrow \bigvee_{a=1}^{L-1} MatchXA^-_{ija}$$

- Для второй группы дополнительных переменных по X (по Y выражается аналогично):

$$\bigwedge_{i,j} \bigwedge_{a=0}^{L-1} MatchXA_{ija} \leftrightarrow X_{ia} \wedge X_{ja}$$
$$\bigwedge_{i,j} \bigwedge_{a=0}^{L-2} MatchXA^+_{ija} \leftrightarrow X_{ia} \wedge X_{j(a+1)}$$
$$\bigwedge_{i,j} \bigwedge_{a=1}^{L-1} MatchXA^-_{ija} \leftrightarrow X_{ia} \wedge X_{j(a-1)}$$

Взяв ограничения из этого пункта и начальные ограничения, мы получаем формулу, которая позволяет уложить белковую последовательность на решетку. Остается только придумать ограничения для получения свертки с нужной энергией.

Идея решения

Существует два способа решить задачу получения нужной энергии свертки:

- Пусть Z — это множество всех возможных HH-пар аминокислот в последовательности, исключая последовательные пары. Тогда необходимо, чтобы существовало сочетание длины K из множества Z, все пары в котором являются соседями:

$$\bigvee_{t=1}^{C_Z^K} \bigwedge_{(i,j)\in O_t} (R_i \wedge R_j \wedge neighbours[i][j]),\ \text{где } O_t = \{(i_t, j_t)_{k_t}\},\ k_t \in [1, t]$$

  Ограничение достаточно простое, но зависимость числа дизъюнктов в формуле от $|Z|$ является экспотенциальной из-за перебора по сочетаниям, поэтому данное решение не подходит.

- Введем два дополнительных массива булевых переменных: $H^1 = \{H^1_{ia}\}$ и $H^2 = \{H^2_{ia}\}$, где $i \in [0, K-1],\ a \in [0, N-1]$. Пусть $H^1_{ia} = True$, если в HH-паре с номером i первый элемент имеет порядковый номер a, а $H^2_{ia} = True$, если в HH-паре с номером i второй элемент имеет порядковый номер a. Тогда с помощью некоторого количества ограничений на эти массивы, мы сможем получить свертку с нужным количеством суммарной энергии.

Ограничения решения

Введем следующие ограничения:

- Сперва отметём все P-аминокислоты, так как они не формируют HH-пары:

$$\bigwedge_{i,a} \overline{R_a} \wedge \overline{H^1_{ia}}$$

$$\bigwedge_{i,a} \overline{R_a} \wedge \overline{H^2_{ia}}$$

- Для данного i существует ровно один $H^1_{ia} = 1$ (аналогично для H^2):

$$\bigwedge_{i} \bigvee_{a} H^1_{ia}$$

$$\bigwedge_{i} \bigwedge_{a} \bigwedge_{b!=a} \overline{(H^1_{ia} \wedge H^1_{ib})}$$

$$\bigwedge_{i} \bigvee_{a} H^2_{ia}$$

$$\bigwedge_{i} \bigwedge_{a} \bigwedge_{b!=a} \overline{(H^2_{ia} \wedge H^2_{ib})}$$

- Потребуем, чтобы все пары были различны:

$$\bigwedge_{i} \bigwedge_{j!=i} \bigwedge_{a} \bigwedge_{b} (H^1_{ia} \wedge H^1_{ja}) \rightarrow \overline{H^2_{ib} \wedge H^2_{jb}}$$

- Потребуем, чтобы никакие две пары не являлись перестановками друг друга:

$$\bigwedge_{i} \bigwedge_{j!=i} \bigwedge_{a} \bigwedge_{b} (H^1_{ia} \wedge H^2_{ib}) \rightarrow \overline{H^1_{jb} \wedge H^2_{ja}}$$

- Пары из последовательных элементов и из одного и того же элемента не учитываются в решении:

$$\bigwedge_i \overline{(H^1_{i0} \wedge H^2_{i0})} \wedge \overline{(H^1_{i1} \wedge H^2_{i0})} \wedge \overline{(H^1_{i(N-1)} \wedge H^2_{i(N-1)})} \wedge \overline{(H^1_{i(N-2)} \wedge H^2_{i(N-1)})}$$

$$\bigwedge_i \bigwedge_{a=1}^{N-2} H^2_{ia} \to \overline{(H^1_{ia} \vee H^1_{i(a+1)} \vee H^1_{i(a-1)})}$$

- Если $H^1_{ia} = True \Rightarrow \exists b : H^2_{ib} = True$ (и наоборот):

$$\bigwedge_{i,a} H^1_{ia} \to \bigvee_b H^2_{ib}$$

$$\bigwedge_{i,a} H^2_{ia} \to \bigvee_b H^1_{ib}$$

- "Свяжем" начальные ограничения с массивами H^1 и H^2:

$$\bigwedge_{i,a,b} (H^1_{ia} \wedge H^2_{ib}) \to (neighbours_{ab} \wedge neighbours_{ba})$$

Данные ограничения вместе с начальными и пространственными ограничениями позволяют решить задачу. Остается только преобразовать все формулы в КНФ, что явяляется тривиальной задачей для получившихся ограничений.