Определение 1. Пусть задано множество строк $A = \alpha_1, \alpha_2, ..., \alpha_n$ над конечным алфавитом. Требуется найти такую строку ω, что все строки из множества A являются подстроками ω и длина ω минимальна. Назовем эту задачу "задачей о надстроке"или SSP.

Определение 2. $OUT(v)$

Определение 3. задача ограниченного направленного Гамильтонова пути.

Определение 4. $overlap(s_1, s_2)$ Это длина наибольшего x, такого что $s_1 = ax$ и $s_2 = xb$ для некоторых строк a и b. Проще говоря это длина максимально возможного перекрытия двух строк.

Теорема 3. SSP с бесконечным алфавитом является NP-полной.

Доказательство: Пусть $G = (V, E)$ это граф из задачи ограниченного напревленного Гамильтонова пути, где V это множество целых чисел от 1 до n(1 это начальная вершина и n это конечная вершина) и $|E| = m$. Мы построим строки для G над алфавитом $\Sigma = V \cup B \cup S$, где $B = \{\overline{v}|v \in V - \{n\}\}$ множество "баръерных символов а S - множество специальных символов. Барьерные символы являются локальными для каждой вершины, когда как специальные являются глобальными символами для всего графа G. Для каждой вершины $v \in V - \{v\}$ мы сопоставим множество A_v содержащее $2OUT(v)$ строк. Пусть $R_v = \{\omega_0, ..., \omega_{OUT(v)-1}\}$ это множество вершин соединенных с v. Тогда, $A_v = \{\overline{v}\omega_i\overline{v}|\omega_i \in R_v\} \cup \{\omega_i\overline{v}\omega_{i\oplus 1|\omega_i \in R_v}\}$, где $\oplus$ обозначает сложение по модулю $OUT(v)$. Для каждой вершины $v \in V - \{1, n\}$ создаем множество из одного элемента, содержащее строку $v\#\overline{v}$ называемую соединителем или коннектором. Обозначим S множество, которое содержит терминальные строки $T = \{\%\#\overline{1}, n\#\$\}$. Пусть S это объединение $A_j, 1 \leq j < n$; $C_j, 1 \leq i < n$ и T. Утверждается, что G имеет направленный Гамильтонов путь в том и только в том случае, если S имеет надстроку длины $2m + 3n$. Предположим, что в G есть направленный Гамильтонов путь. Пусть v, ω_i ребро из этого пути. Для начала построим надстроку длины $2OUT(v) + 2$ для A_v вида $\overline{v}\omega_i\overline{v}\omega_{i\oplus 1}\overline{v}...\overline{v}\omega_i$, называемую ω_i-стандартной надстрокой для A_v. Эта надстрока сформирована "схлапыванием"перекрытий строк A_v в порядке $\overline{v}\omega_i\overline{v}, \omega_i\overline{v}\omega_{i+\oplus 1}, \overline{v}\omega_{i+\oplus 1}\overline{v}, ..., \overline{v}\omega_{i\oplus OUT(v)}\overline{v}, \omega_{i\oplus OUT(v)}\overline{v}\omega_i$, где каждая последующая пара имеет перекрытие длины 2. Отметим, что множество ω_i-стандартных надстрок для A_v переходят трук в друга в соответствии с циклическими перестановками целых числе от 0 до $OUT(v)-1$ Пусть $u_1, u_2, ..., u_n$ обозначает направленный Гамильтонов путь где $u_1 = 1, u_n = n$ и обозначим стандартную u_j надстроку для A_{u_i} как $STD(\overline{u_i}, u_j)$. Мы можем построить надстроку для для S как схлапывание стандартных надстрок и строк из S в конкретном порядке: $\%\#\overline{1}, STD(\overline{1}, u_2), u_2\#\overline{u_2}, STD(\overline{u_2}, u_3), u_3\#\overline{u_3}, ...\overline{u_{n-1}}\#\overline{u_{n-1}}, STD(\overline{n_{n-1}}, n), n\#\$$ Надстрока имеет длину $\sum_{i=1}^{n-1}(2OUT(i) + 2) + (n-2) + 4 = 2m + 3n$.

Чтобы доказать обратное утверждение, мы покажем, что $2m+3n$ это нижняя граница размера надстроки S и затем покажем, что эта нижняя граница может быть достигнута только в случае если надстрока кодирует Гамильтонов путь. Всего мы имеем $2m + n$ строк, в сумме их длина $3(2m + n)$. Наибольшее "сжатие"дает порядок в котором каждая строка кроме первой имеет перекрытие длины 2 с обеих сторон. Этот порядок должен дать надстроку длины $3(2m + n) - 2(2m + n - 1) = 2m + n + 2$. Однако, n - 2 коннектора могут иметь перекрытие только длины с обеих сторон, т.к. ни одна строка не начинается и не заканчивается с символа #. К тому же терминальные строки могут перекрываться максимум на 1 символ только с одной строны. Соблюдая эти условия, мы имеем нижнюю границу на длину надстроки в $(2m + n + 2) + 2(n - 2) + 2 = 2m + 3n$ для S. ОТтметим, что она начинается с $\%\#\overline{1}$ и заканчивается $n\#\$$ Рассмотрим два вхождения # в такую надстроку. Обозначим за x то, что находится между этими двумя знаками #. Первый символ из x должен быть барьерным, а последний не барьерным, поскольку они являются подстрокой соединителя. Если в x нет соеденителей, то тогда все подстроки кроме первой и последней должны иметь перекрытие 2 с обеих сторон. Первая строка должна быть $\overline{v}u_j\overline{v}$, следующая $u_j\overline{v}u_{j\oplus 1}$ и так далее. Более того, все строки в A_v кроме двух последних должны иметь перекрытие длины 2 с обеих сторон, так каждая последующая строка должна быть "добита"уникальной строкой, которая перекрывается с ней на 2 символа. Таким образом все строки в A_v должны появляться в конкретном порядке, и если x содержит одну строку из A_v, то он обязан содержать их все. Таким образом, x - это стандартная надстрока для A_v Применяя рассуждения выше ко всем вхождениям пар # мы получаем $n - 1$ различную стандартную строку. Мы можем восстановить Гамильтонов Путь смотря на символы следующие за каждым вхождением #, причем символы с чертой и без черты каждого соединителя отвечают одной и той же верщине в G. Отметим, что по построению $\%\#\overline{1}$ и $b\#\$$, мы получаем путь из 1 в n

4 приближенный алгоритм

Построим граф $G = (V, E)$, где $V = 1..n, E = \{(i, j, overlap(s_i, s_j))|i, j = 1..n, i \neq j\}$ - последнее множество, это множество троек (начальная вершина, конечная вершина, вес). Затем для данного графа G найдем покрытие циклами минимального суммарного веса. Это и будет 4-приближенный алгоритм для задачи.

Вычислим жадное назначение для данного графа G. Будем хранить его в массиве из n чисел, обозначим его за A.
объявим все ребра незачеркнутыми
повторять пока остаются незачеркнутые ребра.
1)Выберем ребро (i, j) максимального веса среди незачеркнутых
2)Зачеркнем все рабра выходящие из i и входящие в j
3)A[i] = j