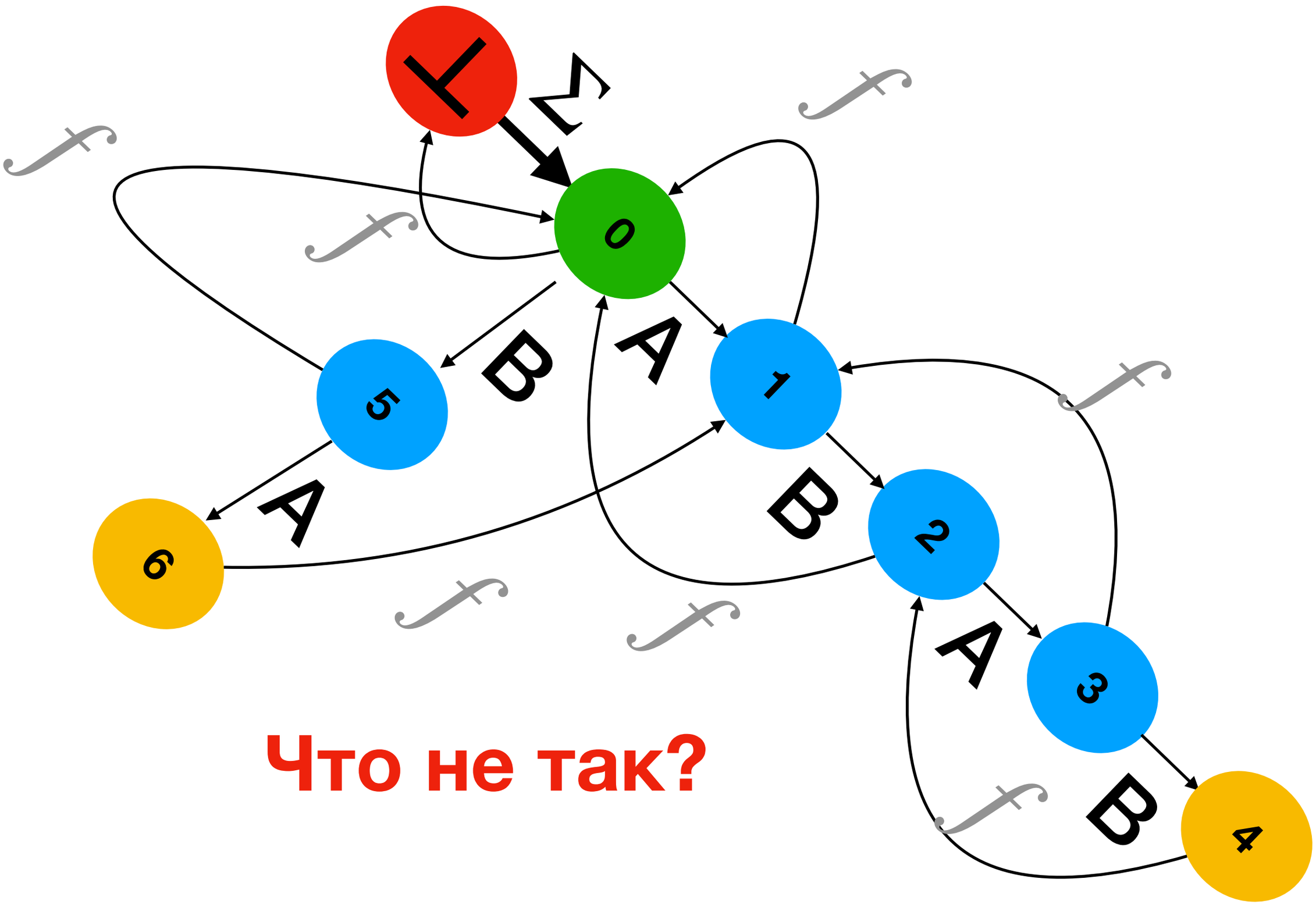
Ахо-Корасик

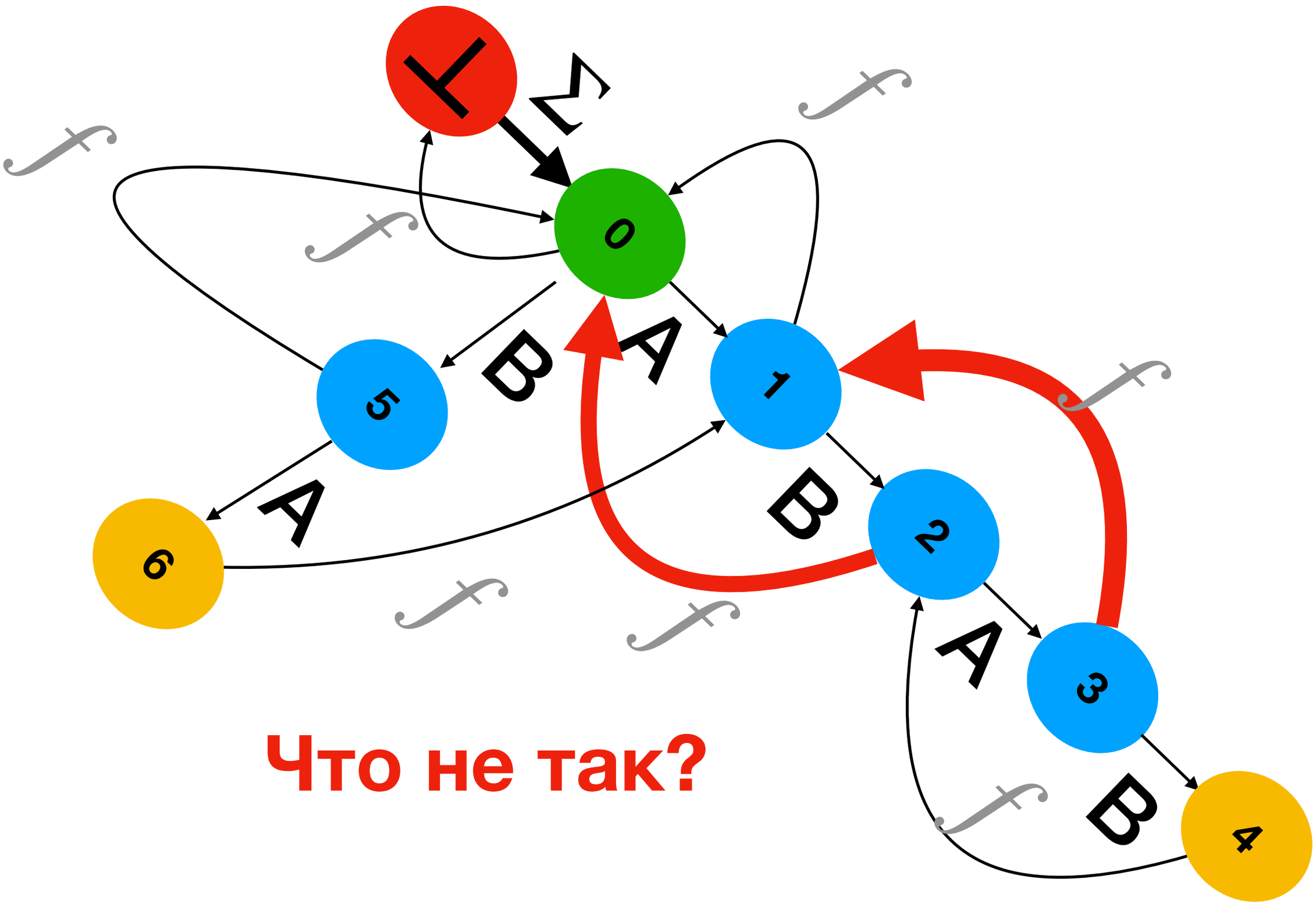
Построим Ахо-Корасик для паттернов BA и ABAB

Ахо-Корасик



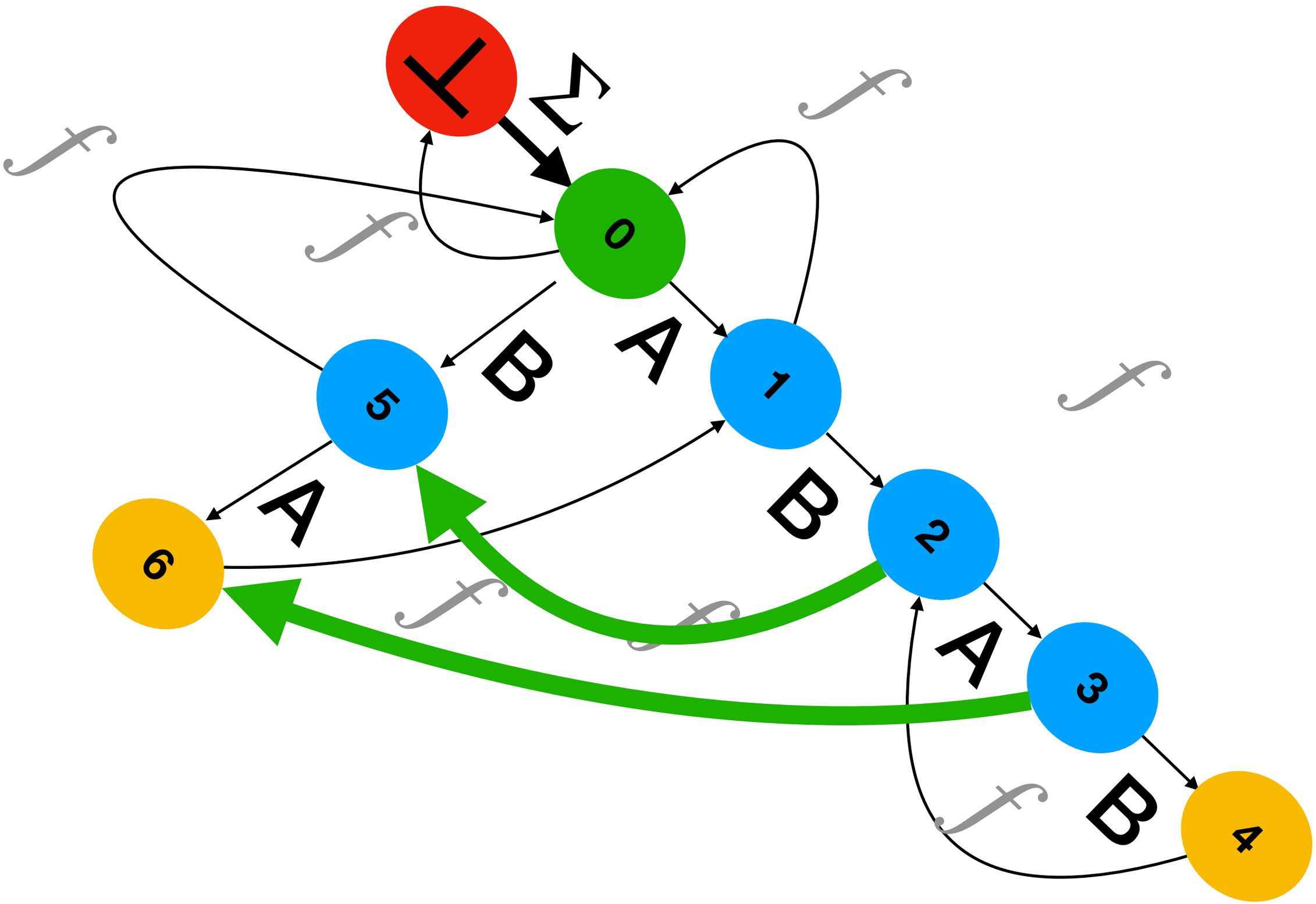
Что не так?

Ахо-Корасик



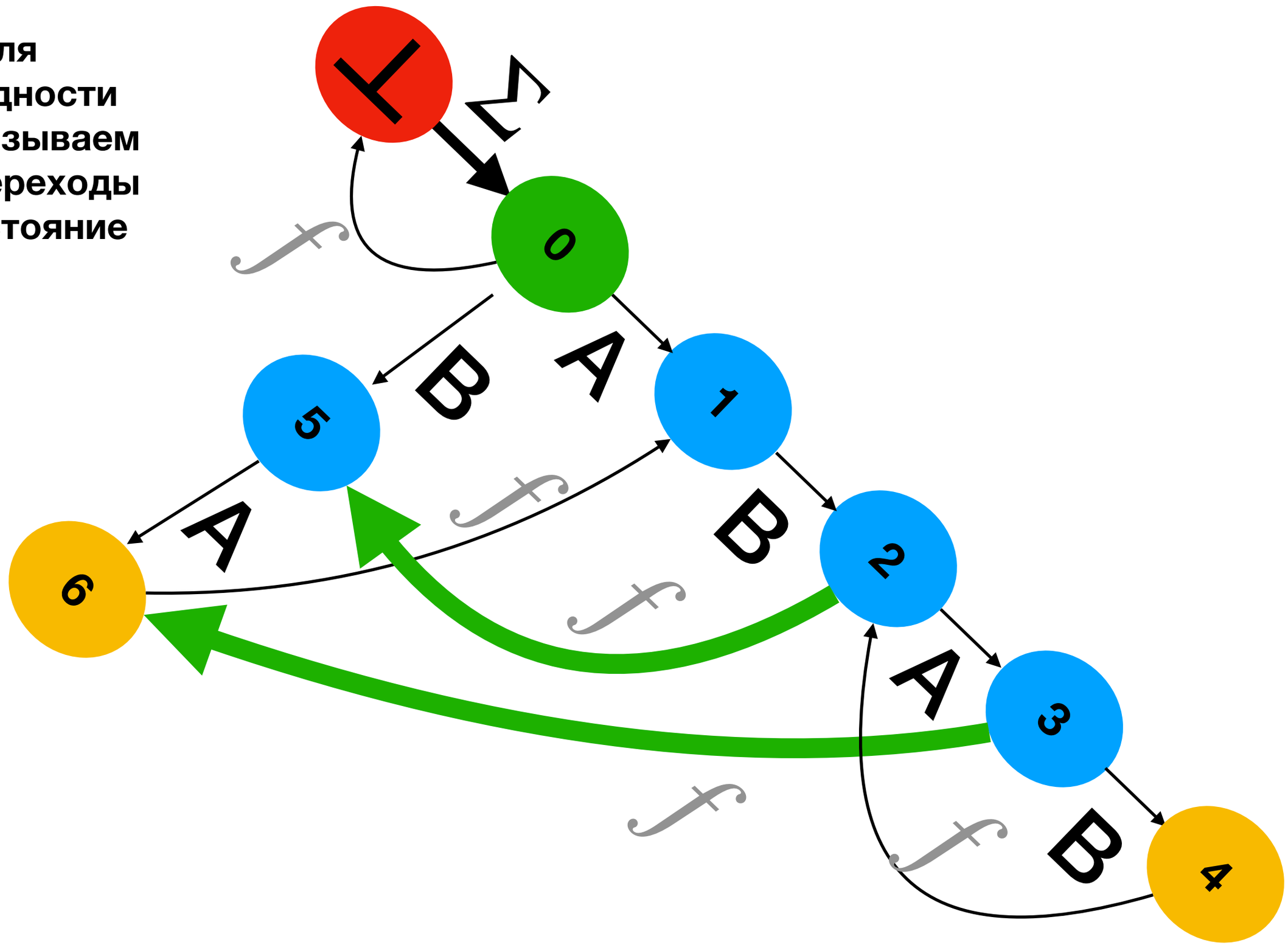
Что не так?

Ахо-Корасик

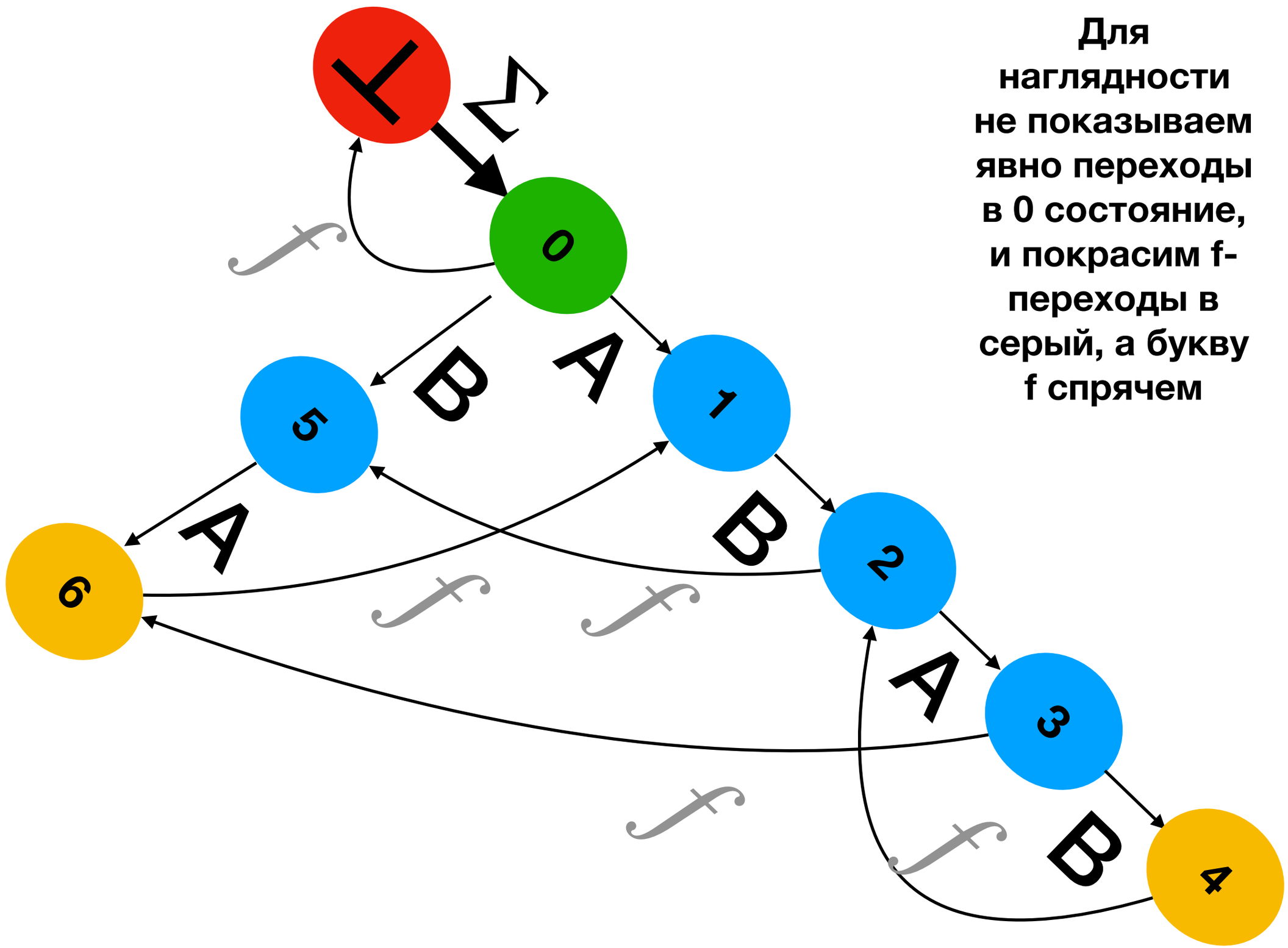


Ахо-Корасик

Для наглядности не показываем явно переходы в 0 состояние

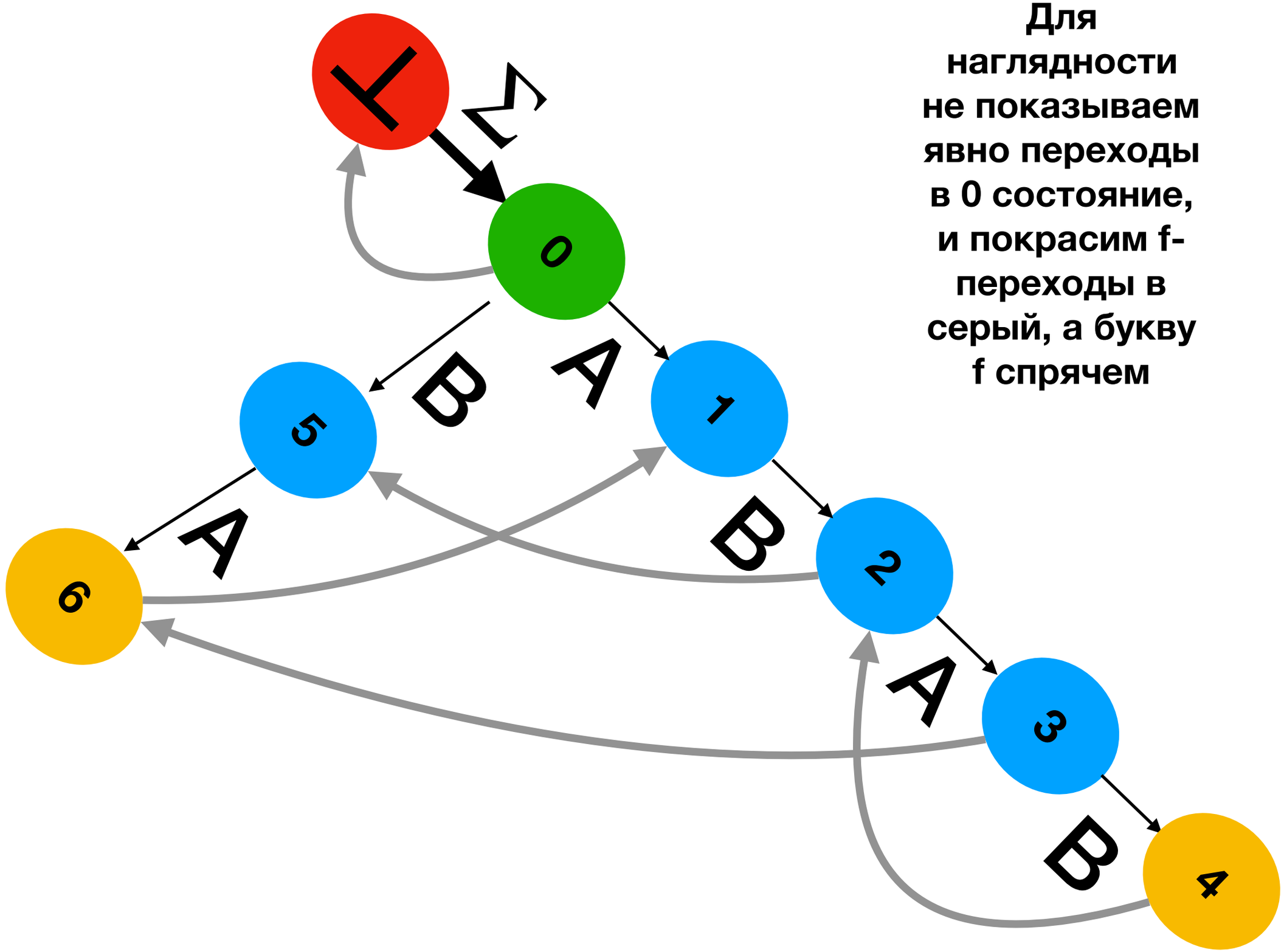


Ахо-Корасик



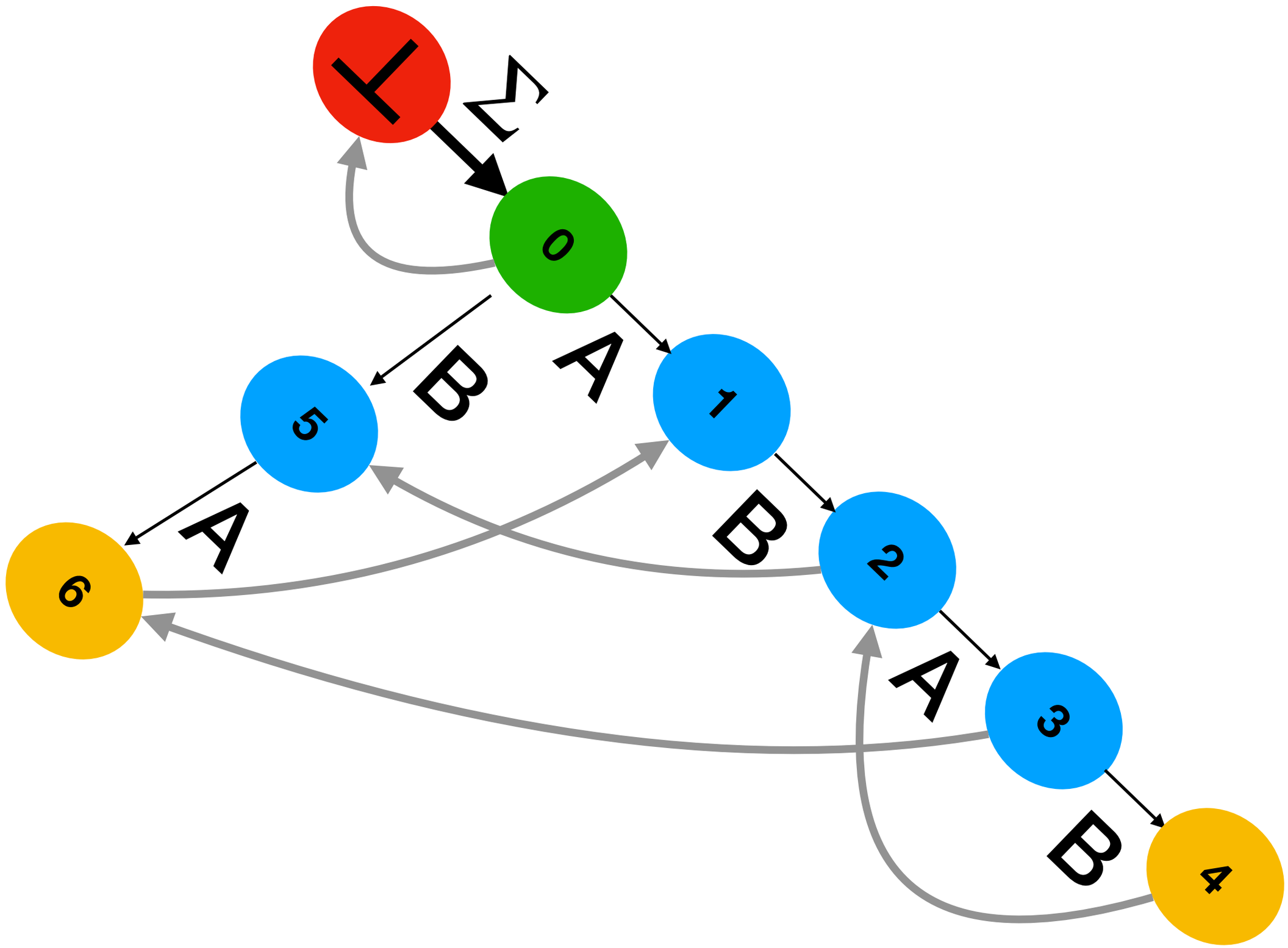
Для наглядности не показываем явно переходы в 0 состояние, и покрасим f-переходы в серый, а букву f спрячем

Ахо-Корасик



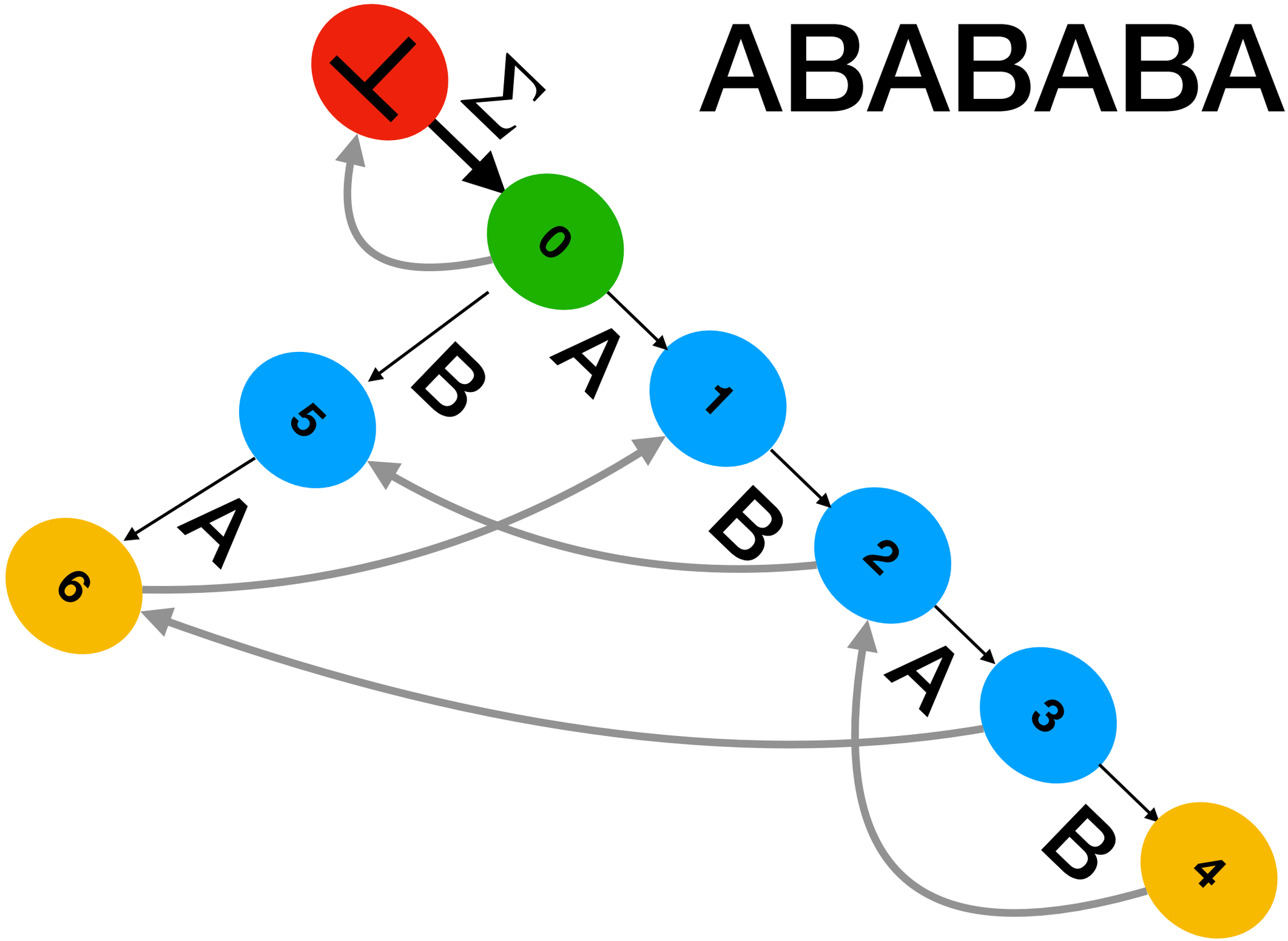
Для наглядности не показываем явно переходы в 0 состояние, и покрасим f-переходы в серый, а букву f спрячем

Ахо-Корасик

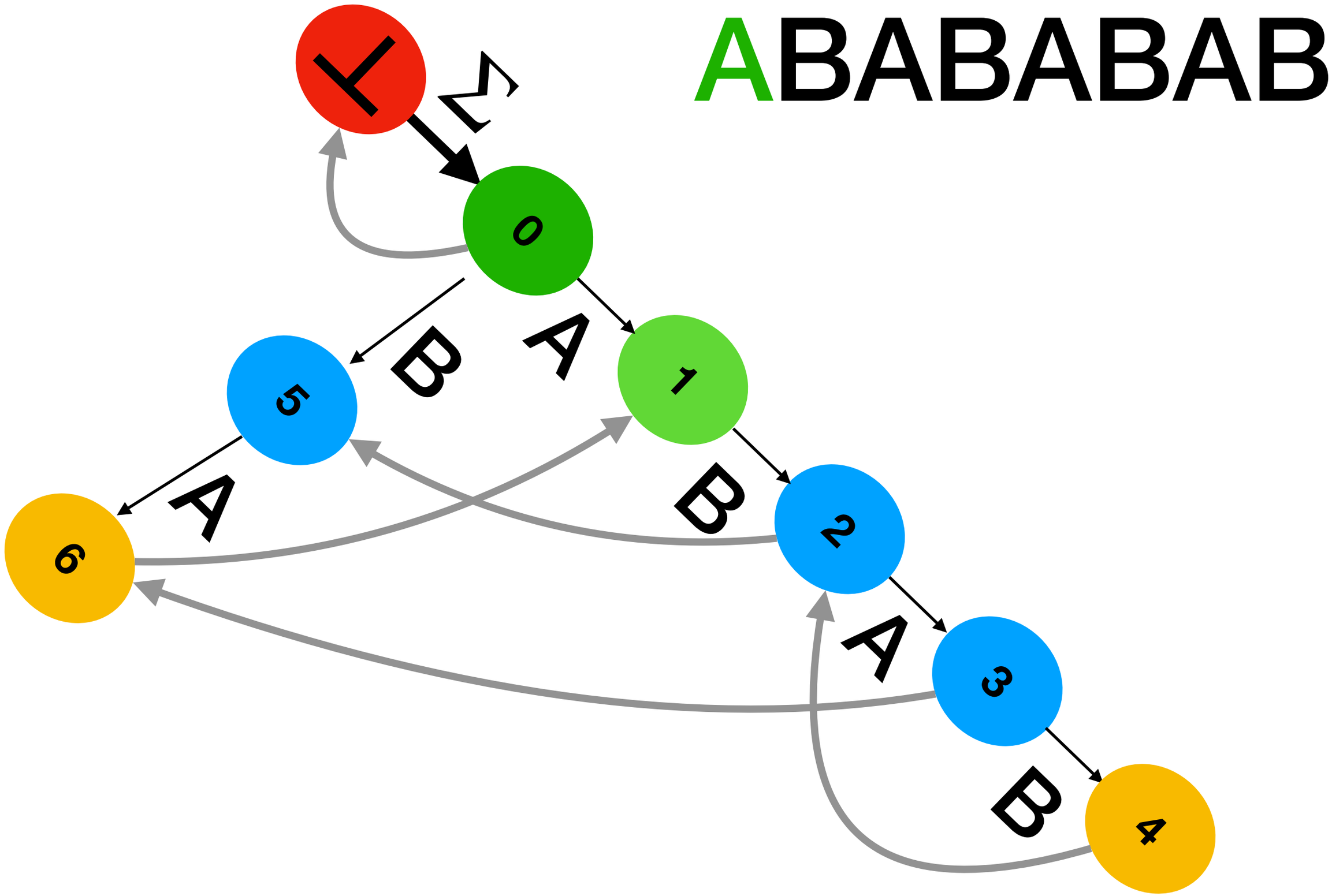


Попробуем поискать в строке

ABABABAB

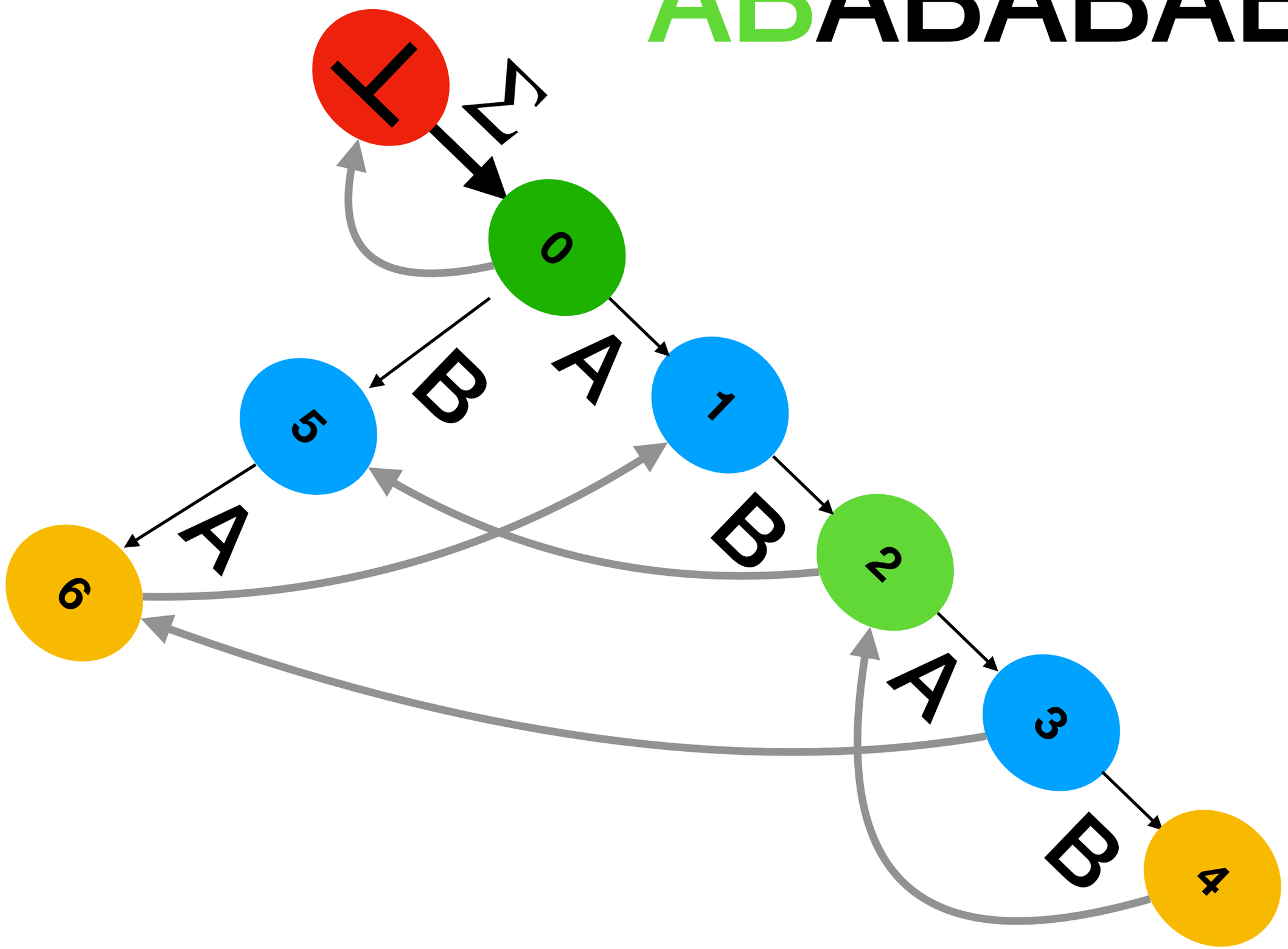


Попробуем поискать в строке



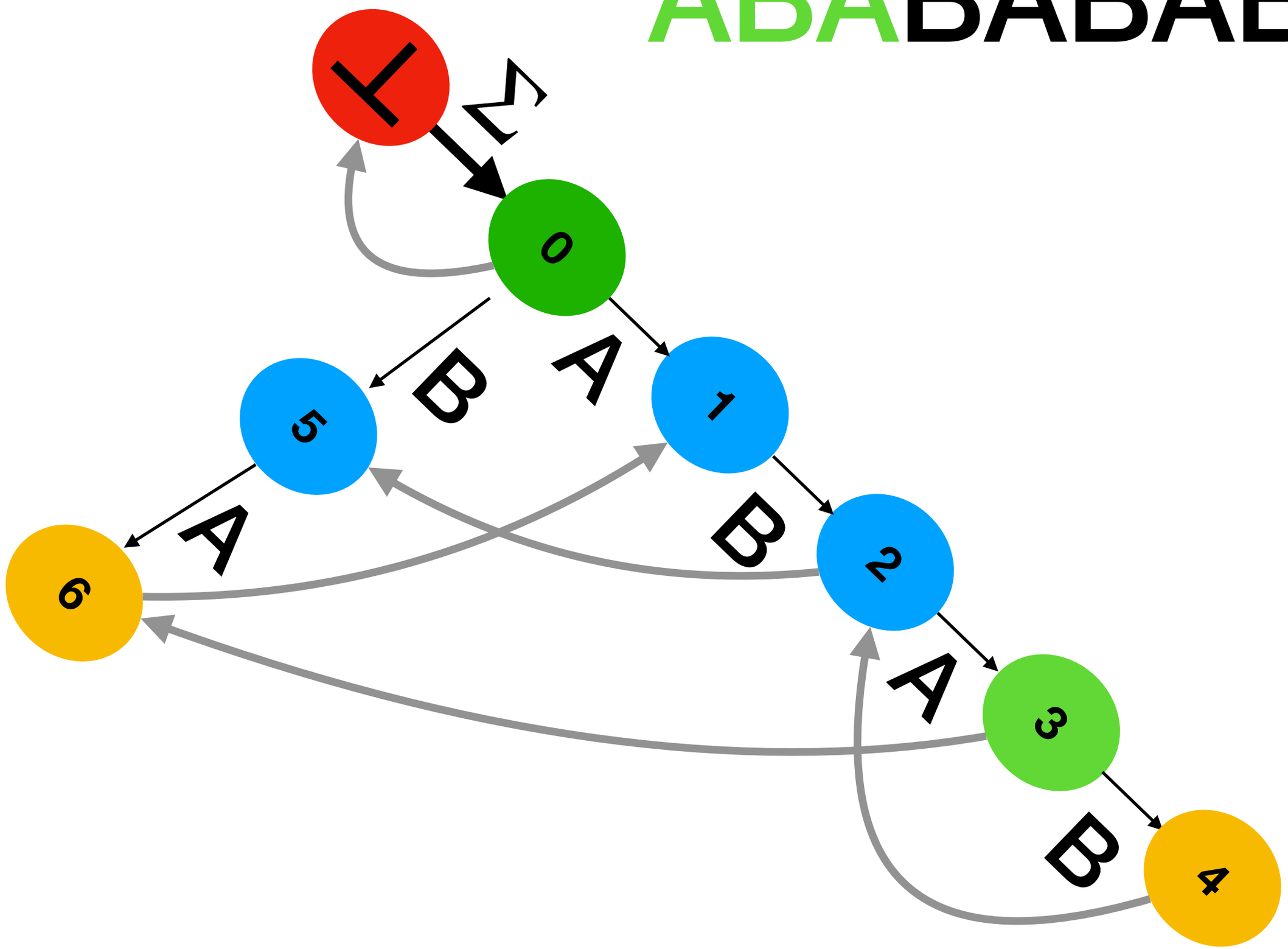
Попробуем поискать в строке

ABABABAB



Попробуем поискать в строке

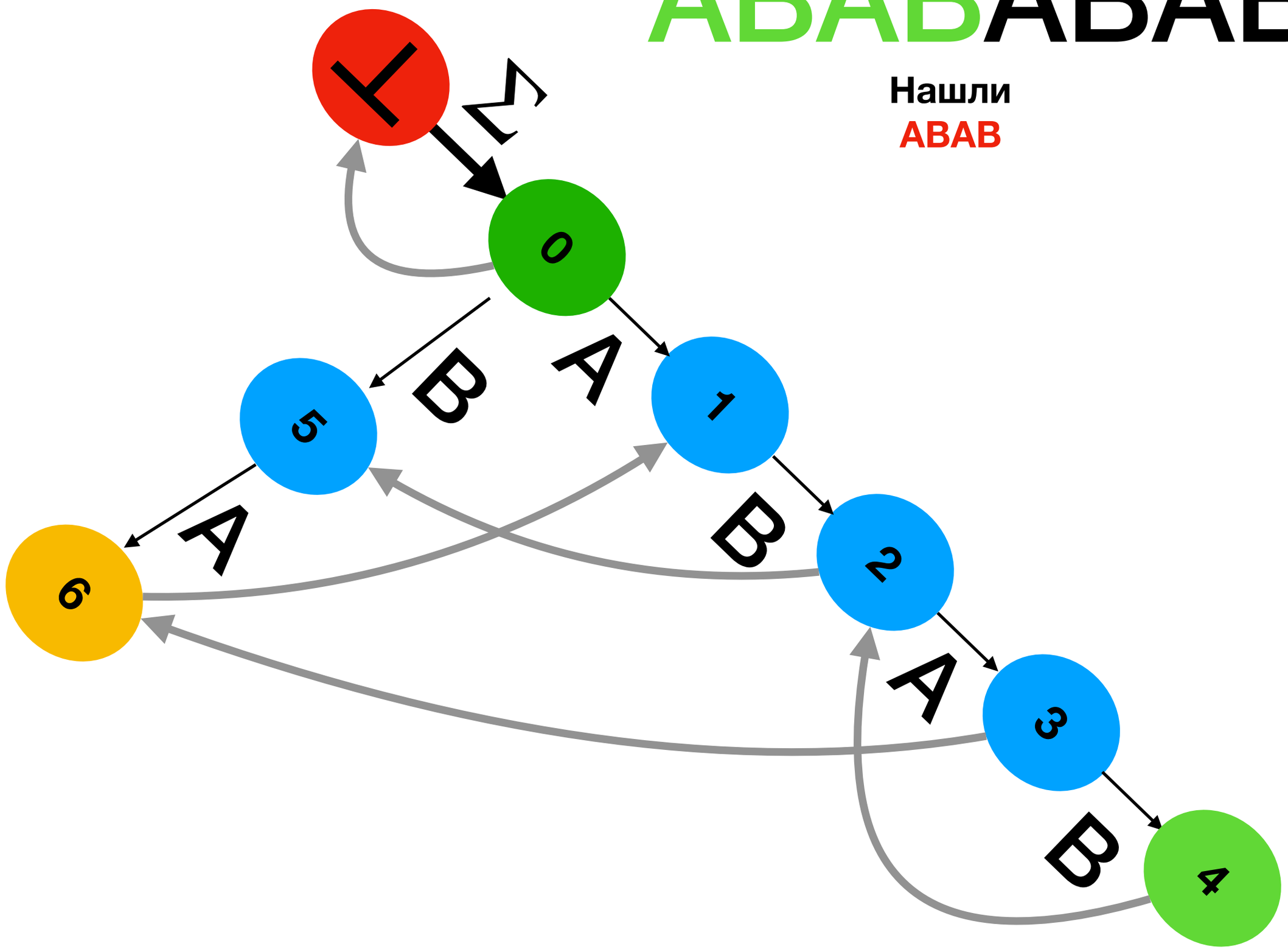
ABABABAB



Попробуем поискать в строке

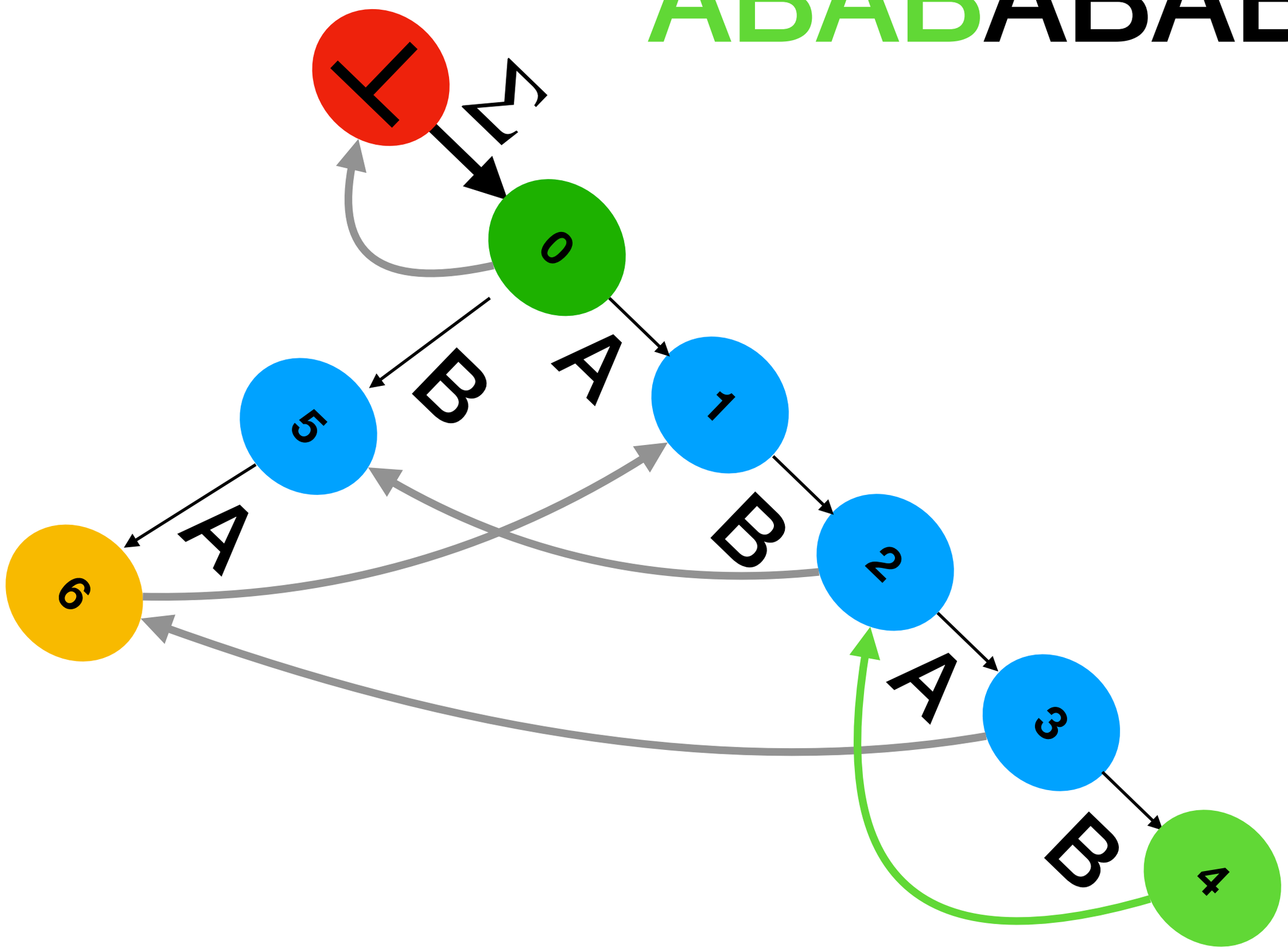
ABABABAB

Нашли
ABAB



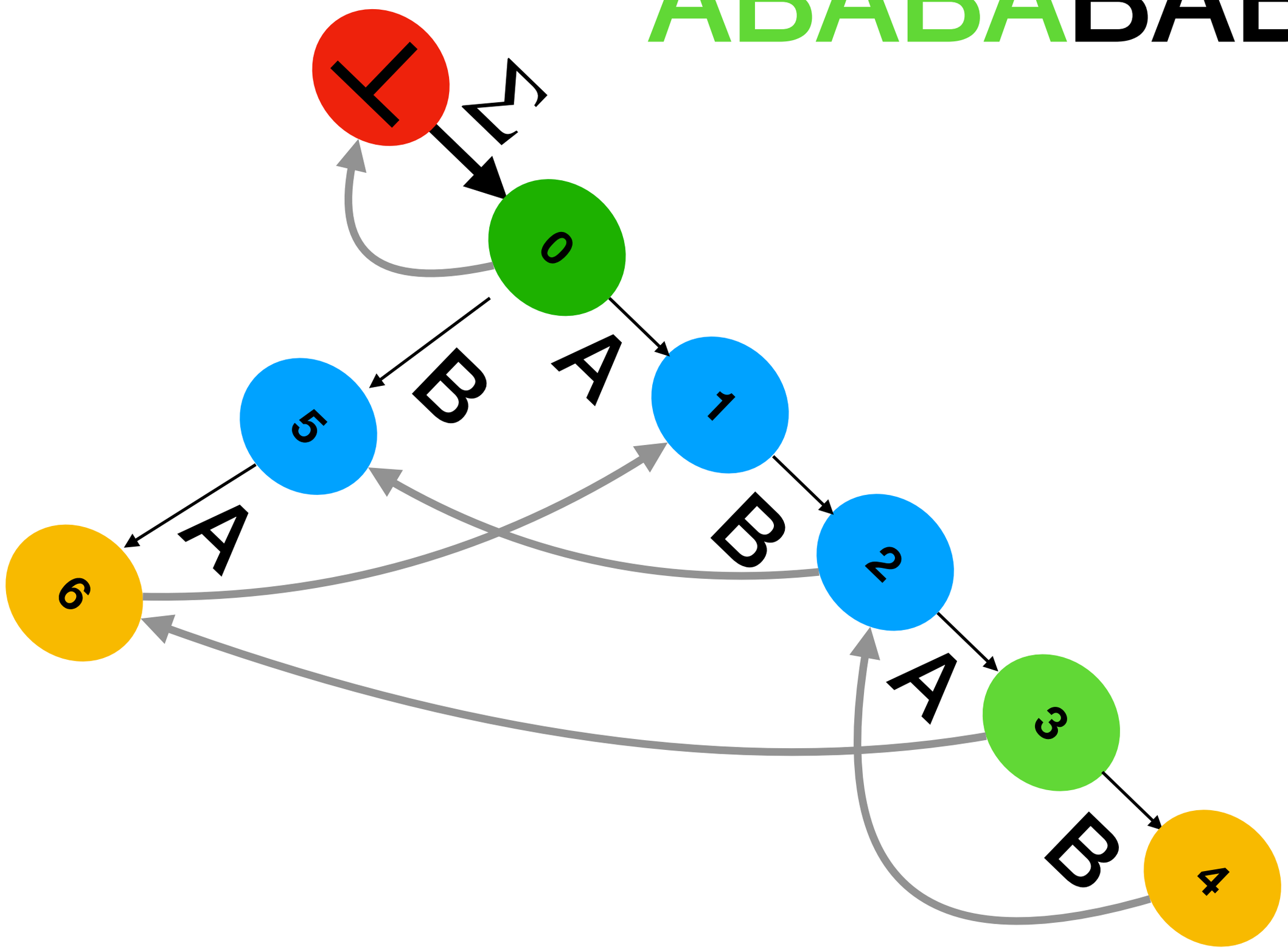
Попробуем поискать в строке

ABABABAB

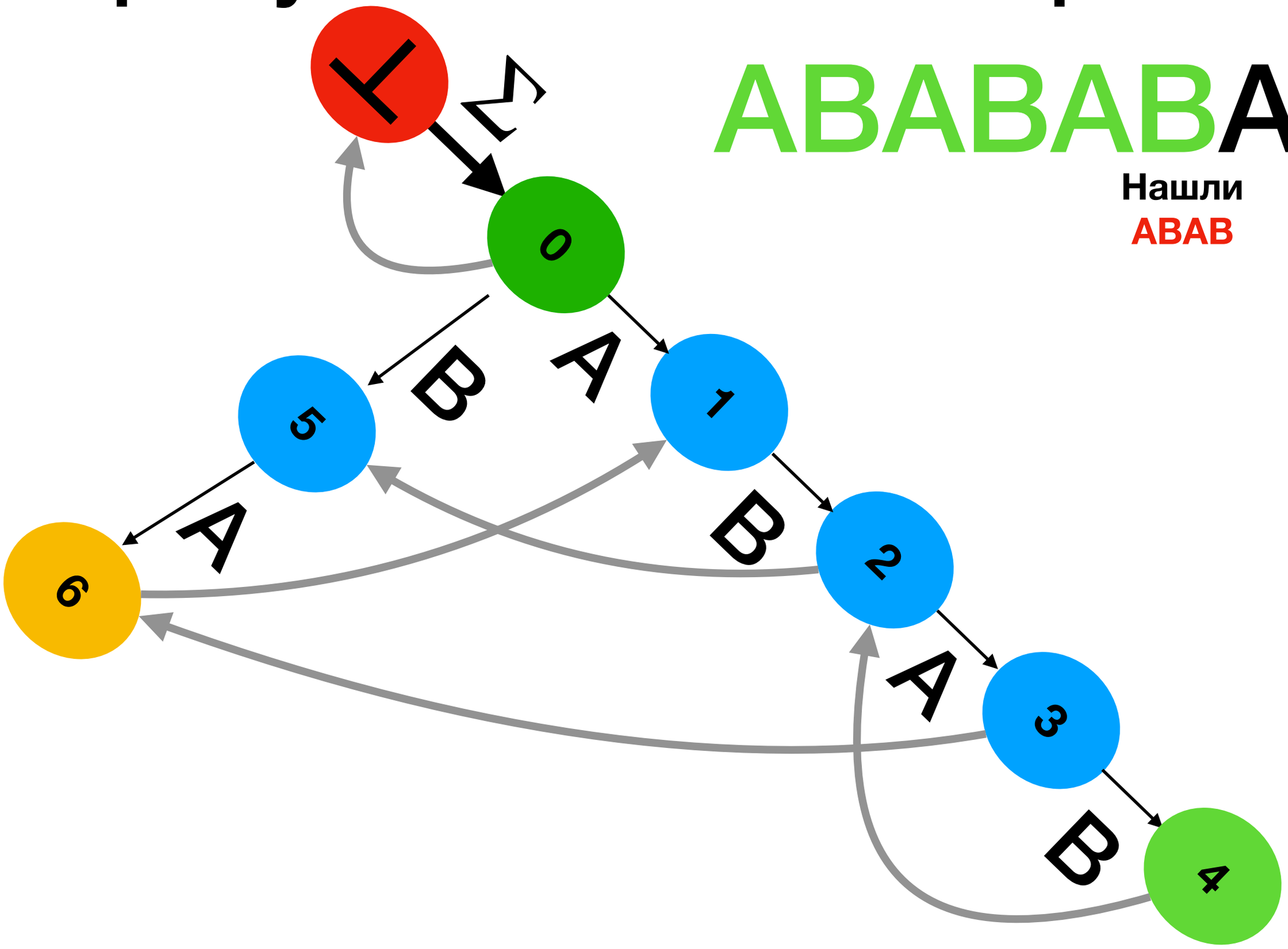


Попробуем поискать в строке

ABABABAB



Попробуем поискать в строке



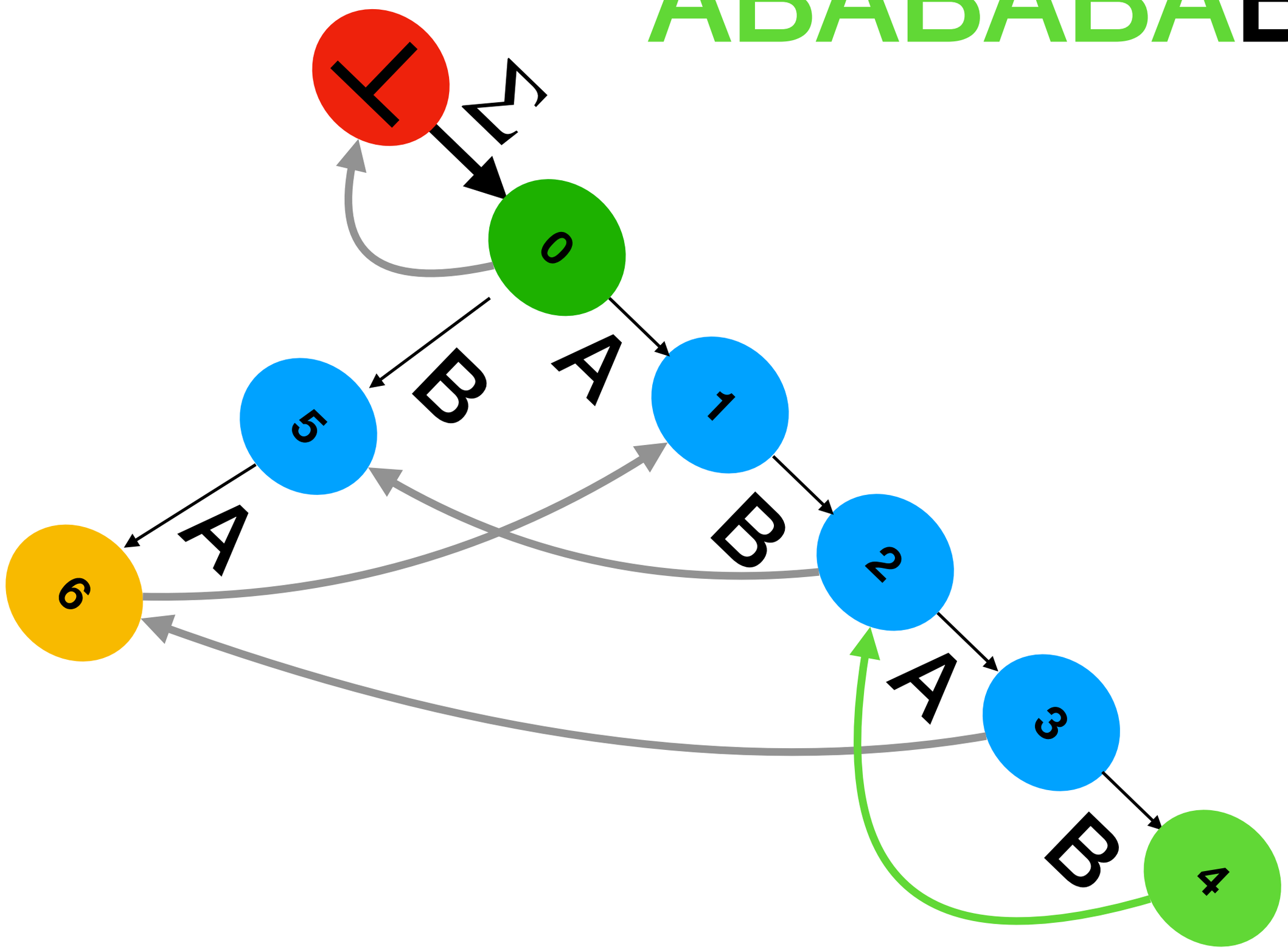
ABABABAB

Нашли

ABAB

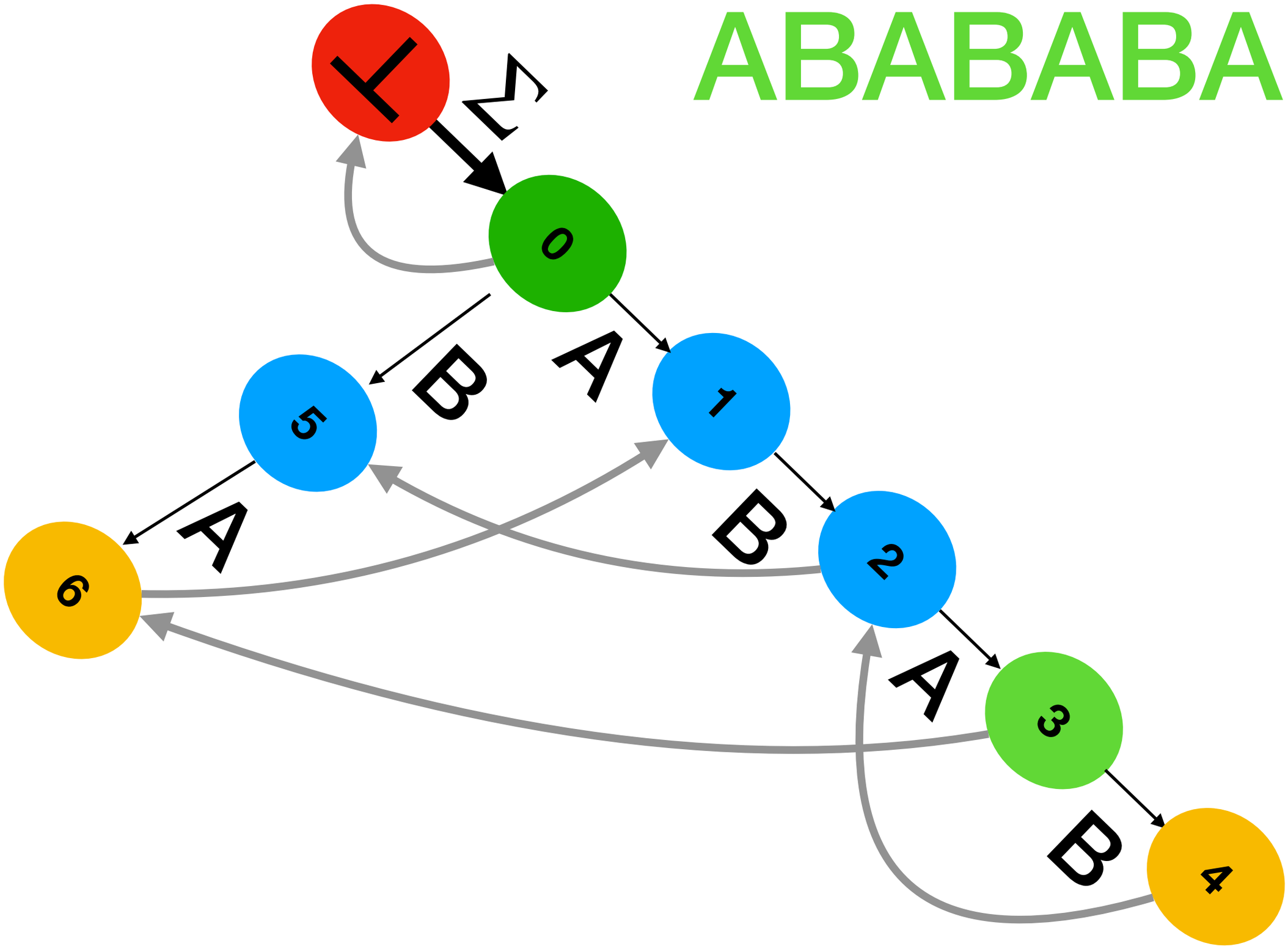
Попробуем поискать в строке

ABABABAB

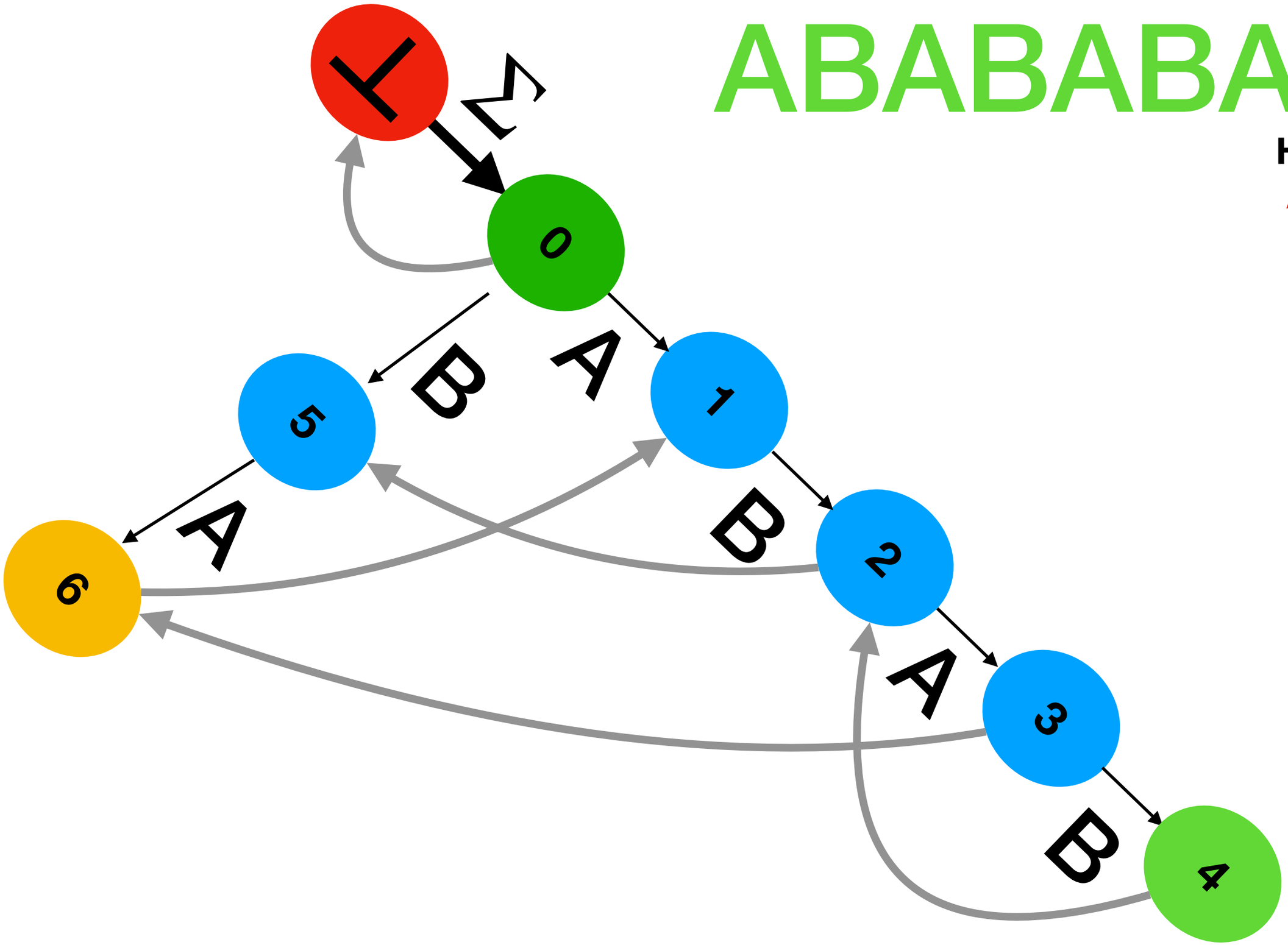


Попробуем поискать в строке

ABABABAB

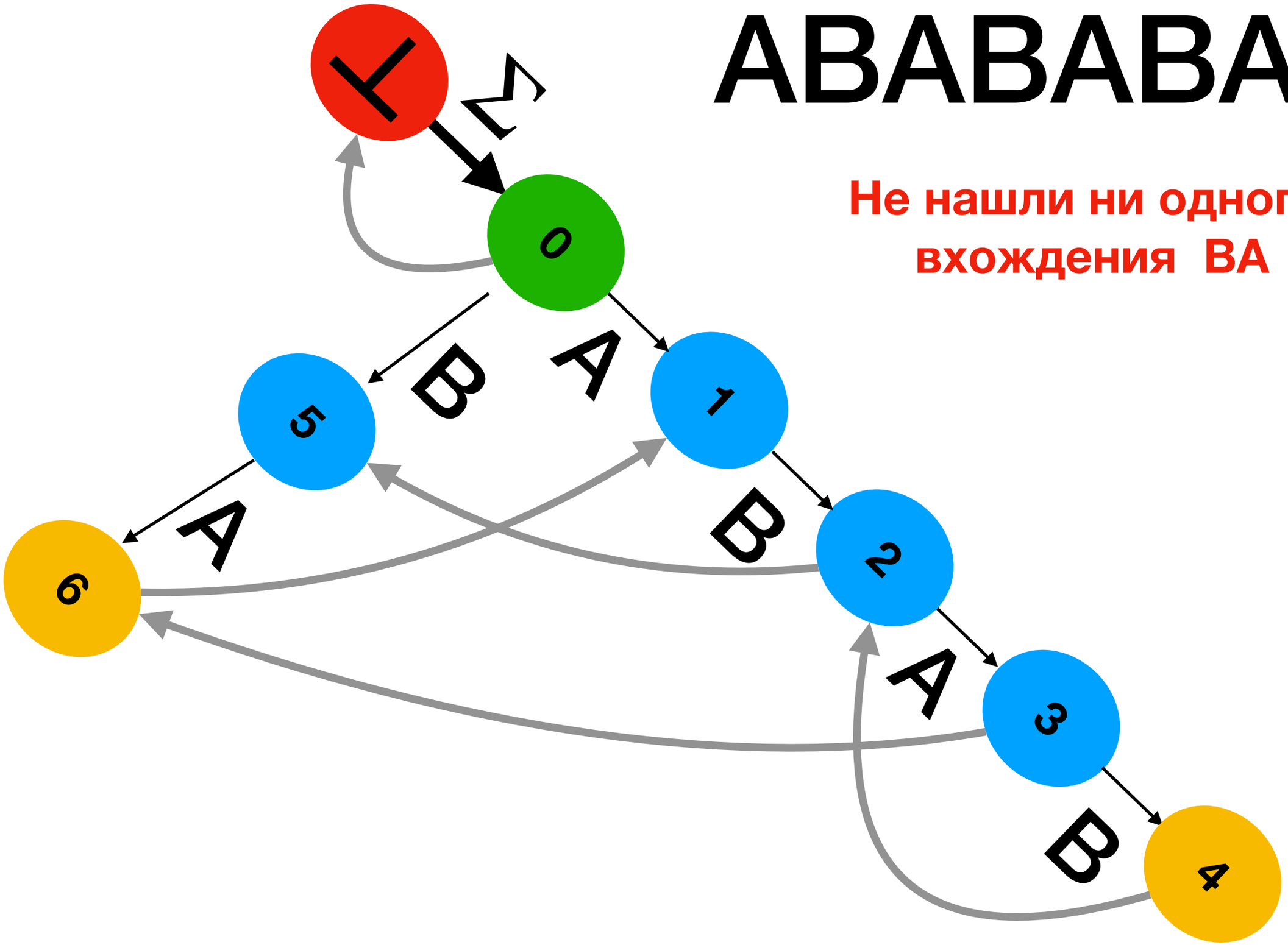


Попробуем поискать в строке



Попробуем поискать в строке

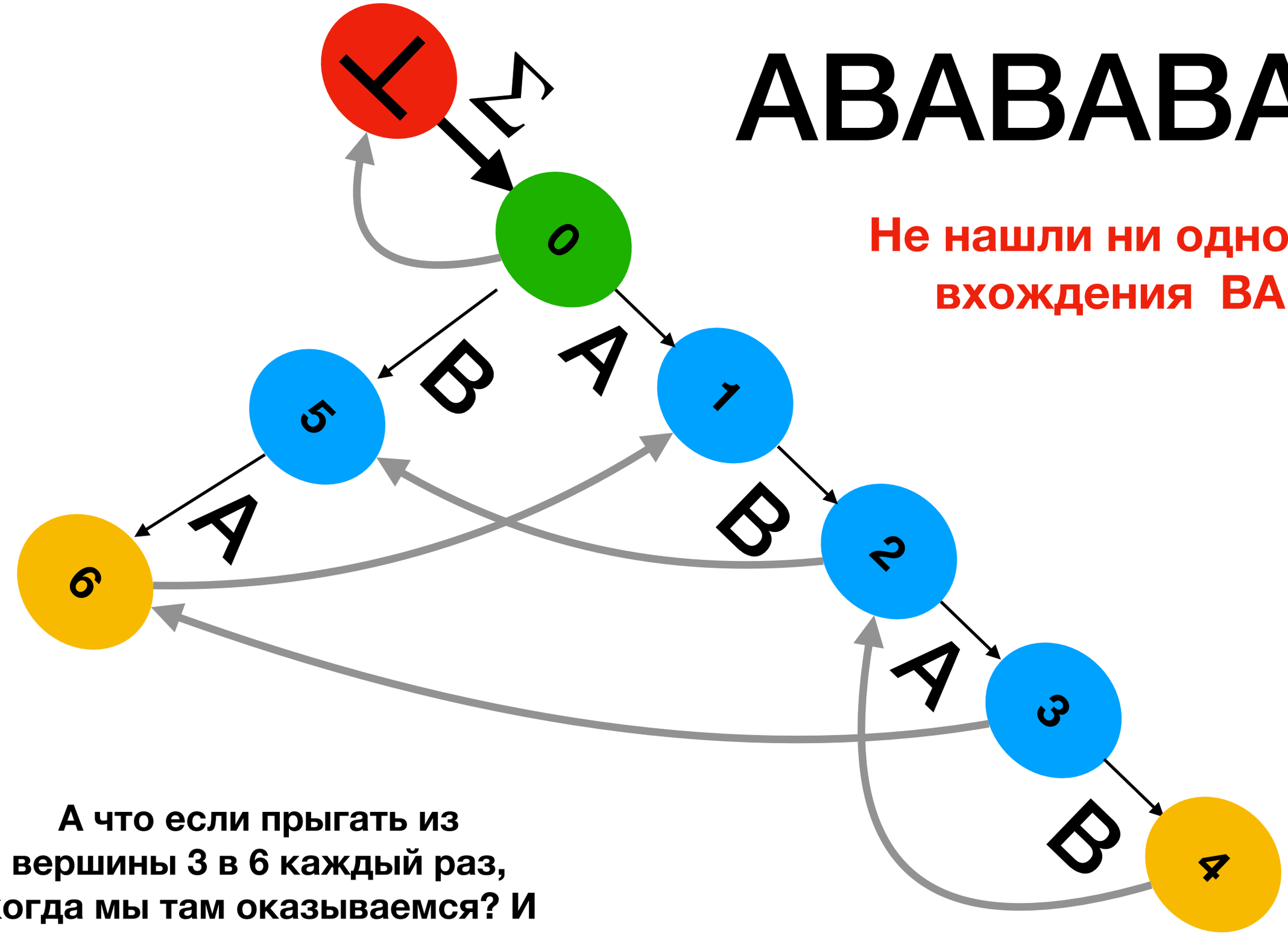
ABABABAB



Не нашли ни одного вхождения BA

Попробуем поискать в строке

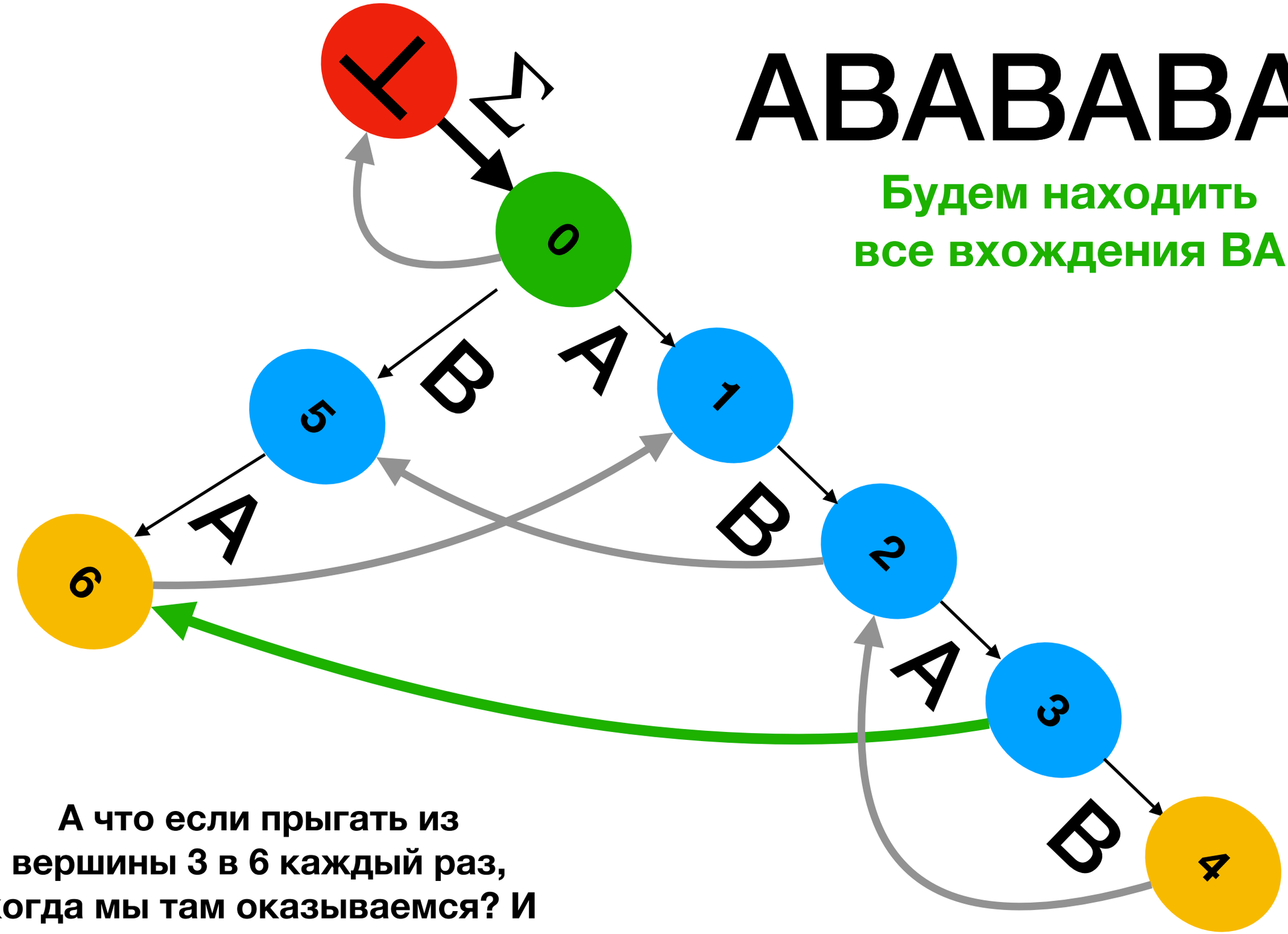
ABABABAB



Не нашли ни одного вхождения BA

А что если прыгать из вершины 3 в 6 каждый раз, когда мы там оказываемся? И идти по двум ветвям?

Попробуем поискать в строке

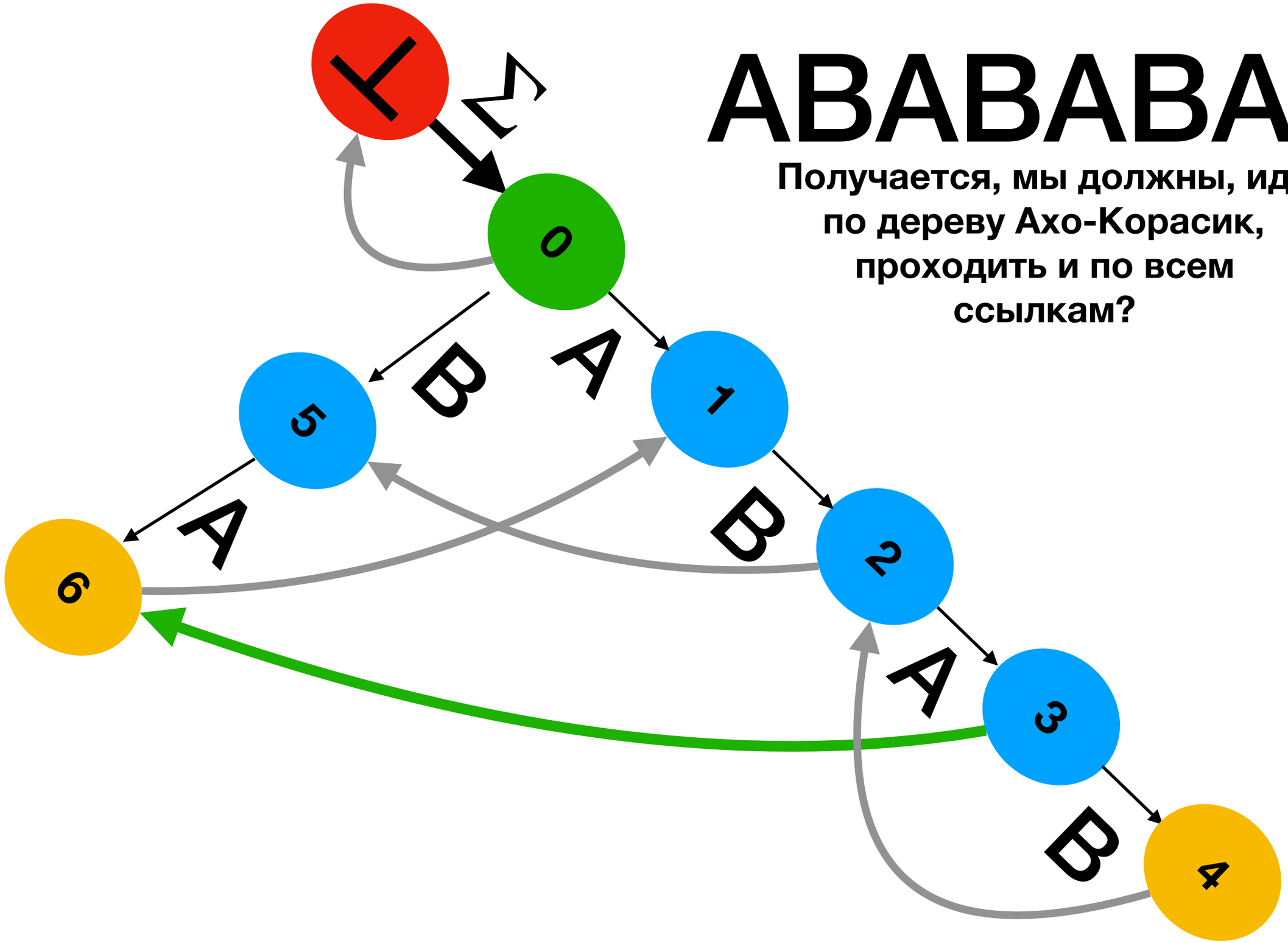


ABABABAB

Будем находить все вхождения BA

А что если прыгать из вершины 3 в 6 каждый раз, когда мы там оказываемся? И идти по двум ветвям?

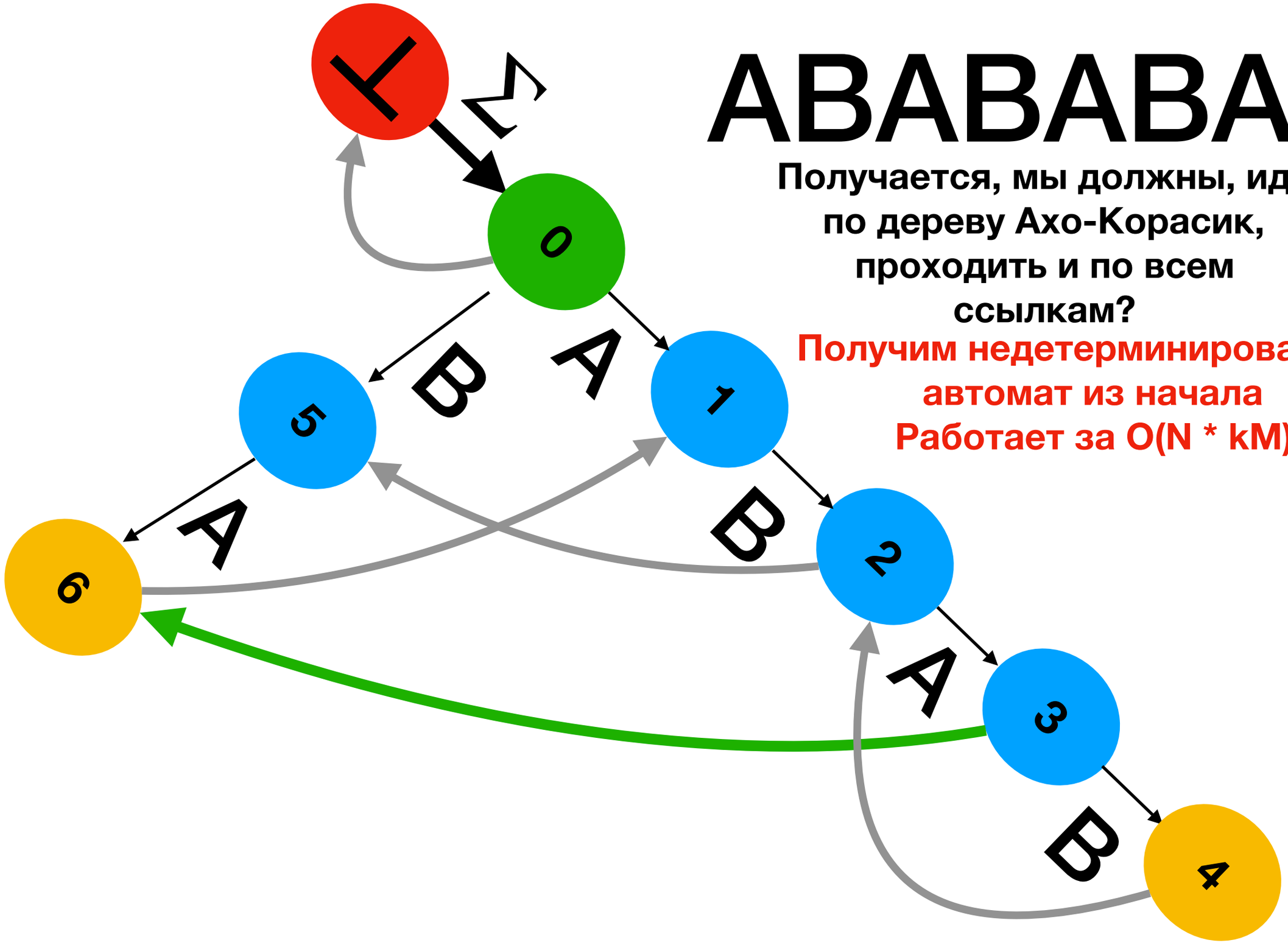
Попробуем поискать в строке



ABABABAB

Получается, мы должны, идя по дереву Ахо-Корасик, проходить и по всем ссылкам?

Попробуем поискать в строке



ABABABAB

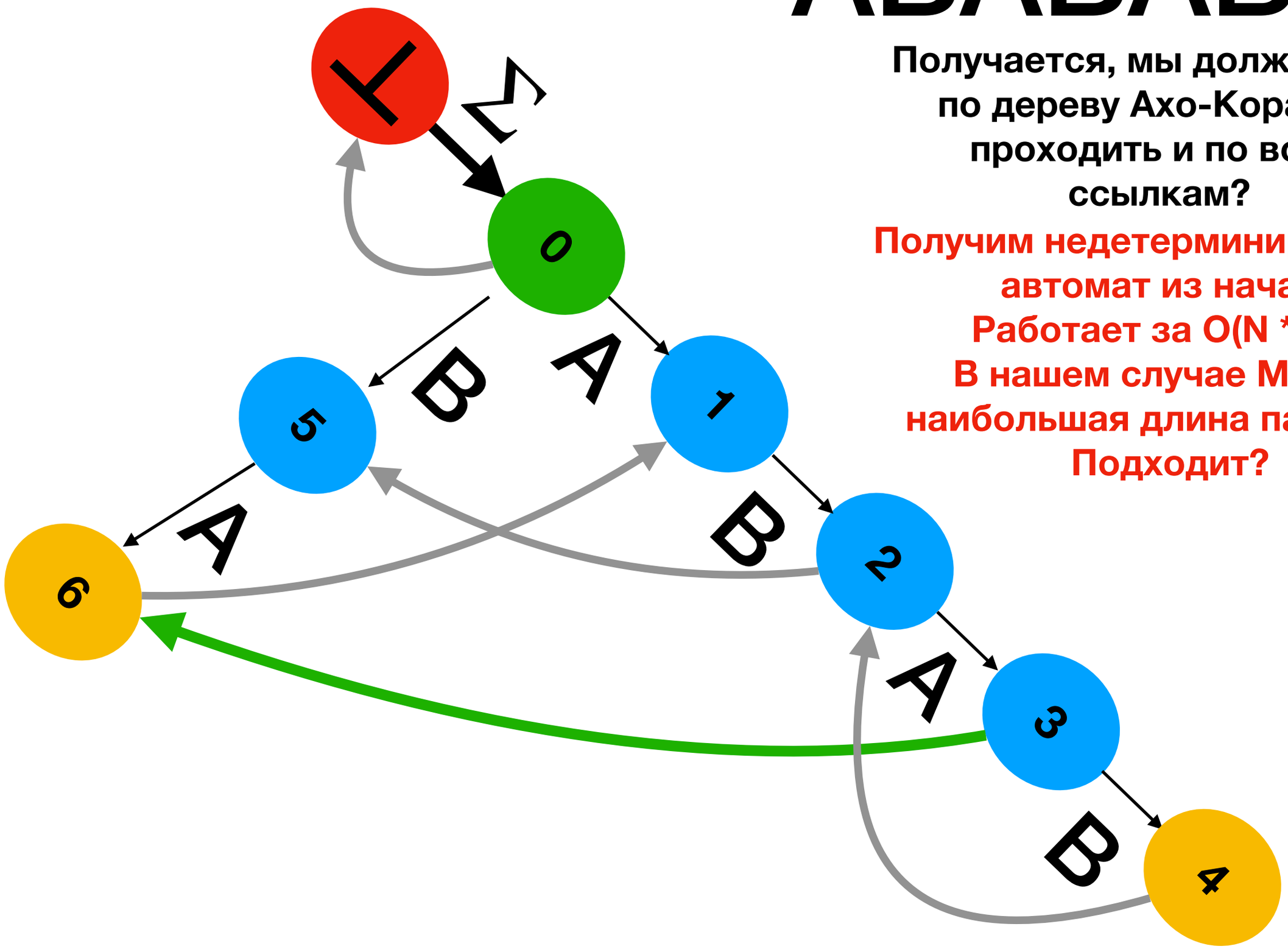
Получается, мы должны, идя по дереву Ахо-Корасик, проходить и по всем ссылкам?

Получим недетерминированный автомат из начала

Работает за O(N * kM)

Попробуем поискать в строке

АВАВАВАВ

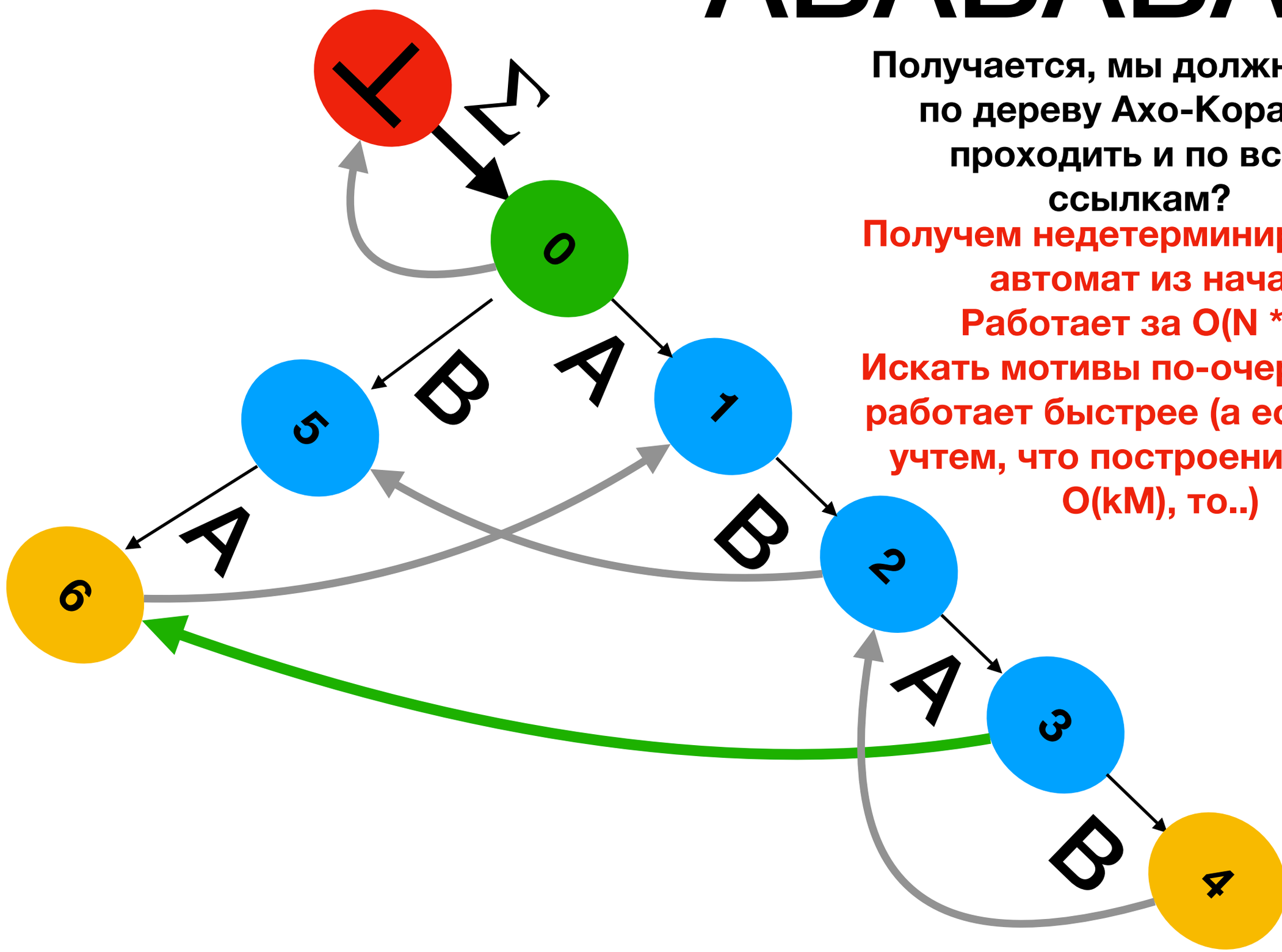


Получается, мы должны, идя по дереву Ахо-Корасик, проходить и по всем ссылкам?

Получим недетерминированный автомат из начала
Работает за O(N * kM)
В нашем случае M - это наибольшая длина паттерна.
Подходит?

Попробуем поискать в строке

АВАВАВАВ



Получается, мы должны, идя по дереву Ахо-Корасик, проходить и по всем ссылкам?

Получем недетерминированный автомат из начала

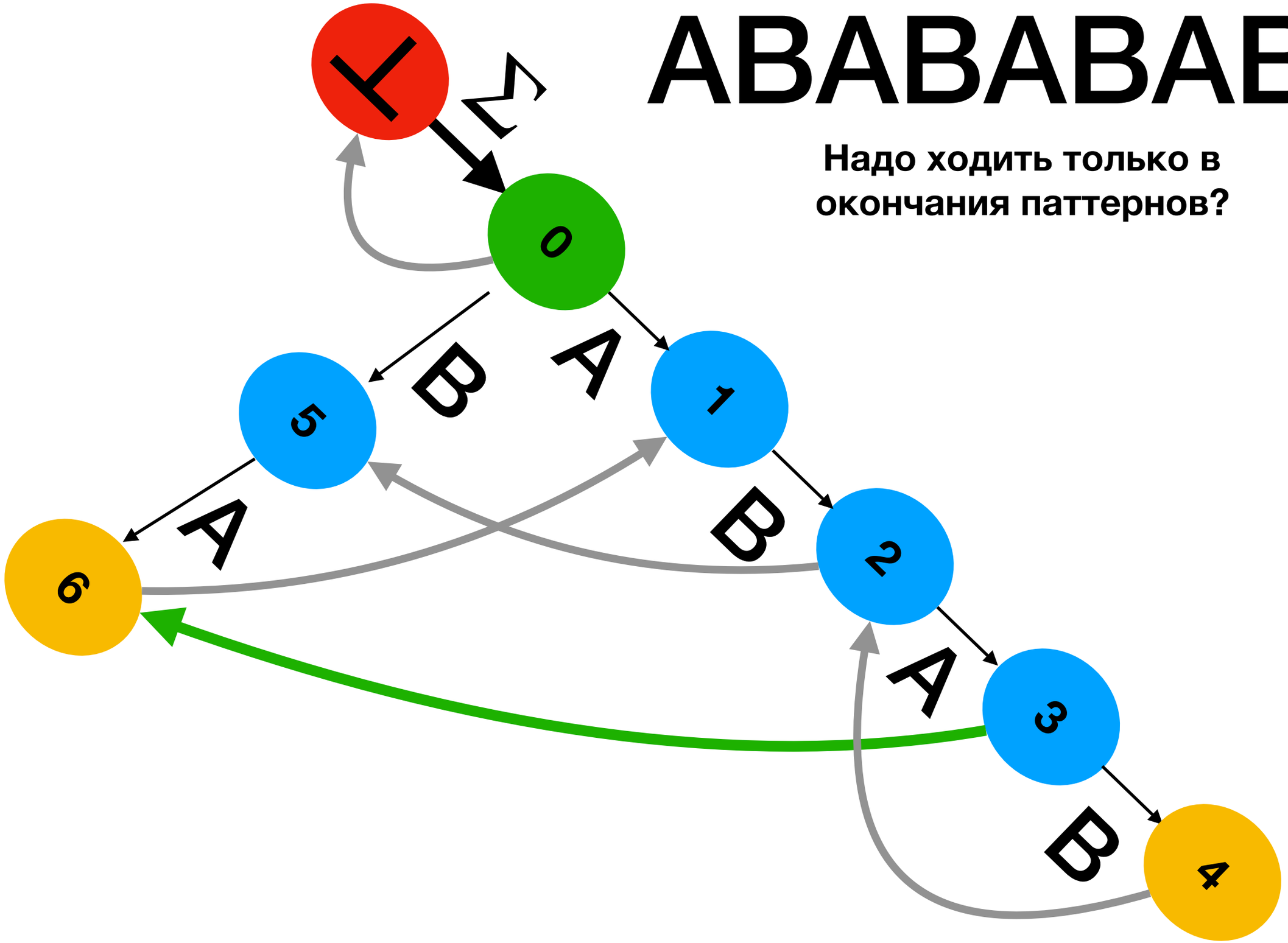
Работает за O(N * kM)

Искать мотивы по-очереди КМП работает быстрее (а если еще и учтем, что построение бора - O(kM), то..)

Попробуем поискать в строке

ABABABAB

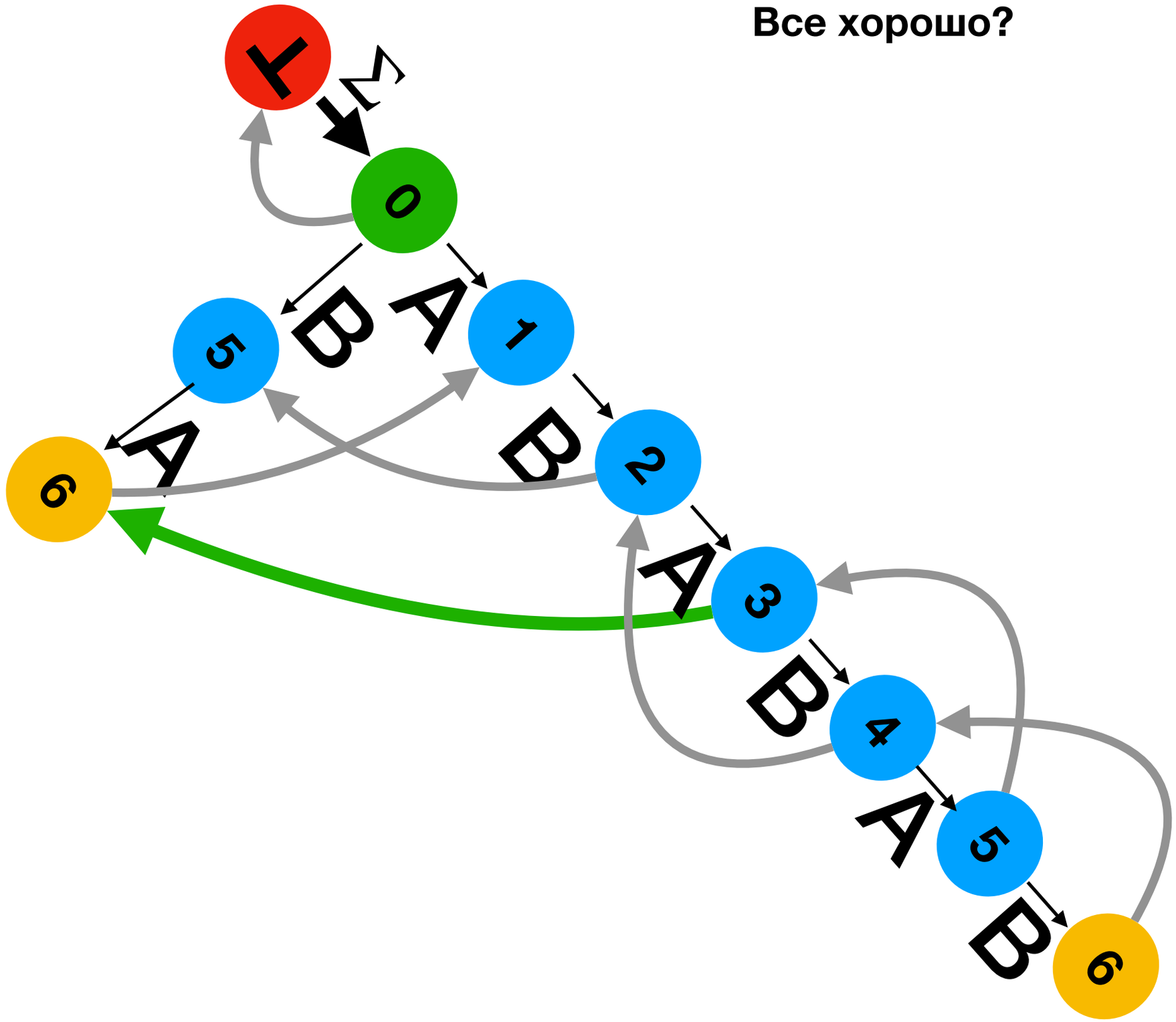
Надо ходить только в окончания паттернов?



Ахо-Корасик

Построим Ахо-Корасик для паттернов BA и ABABAB

Все хорошо?

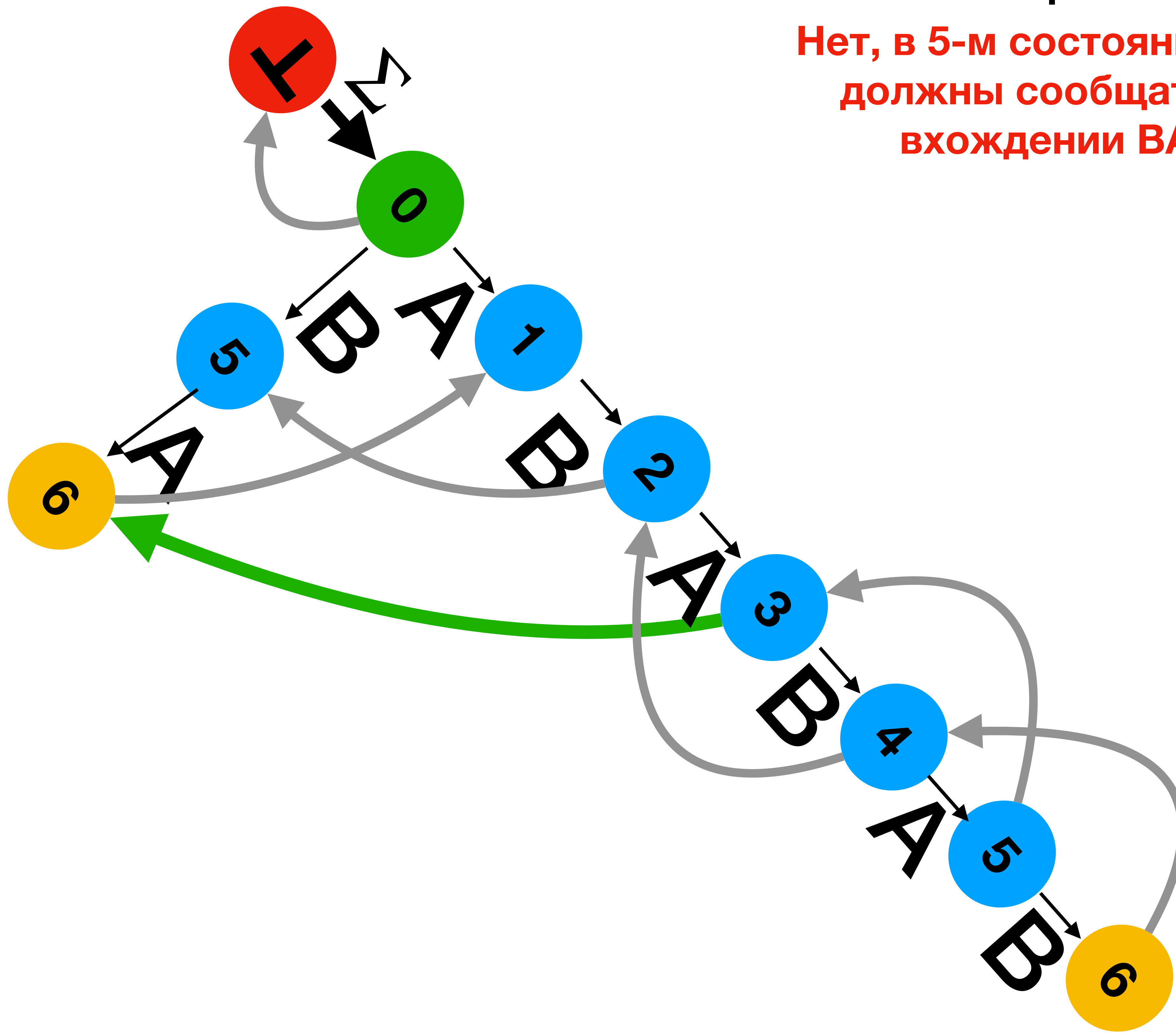


Ахо-Корасик

Построим Ахо-Корасик для паттернов BA и ABABAB

Все хорошо?

Нет, в 5-м состоянии мы должны сообщать о вхождении BA

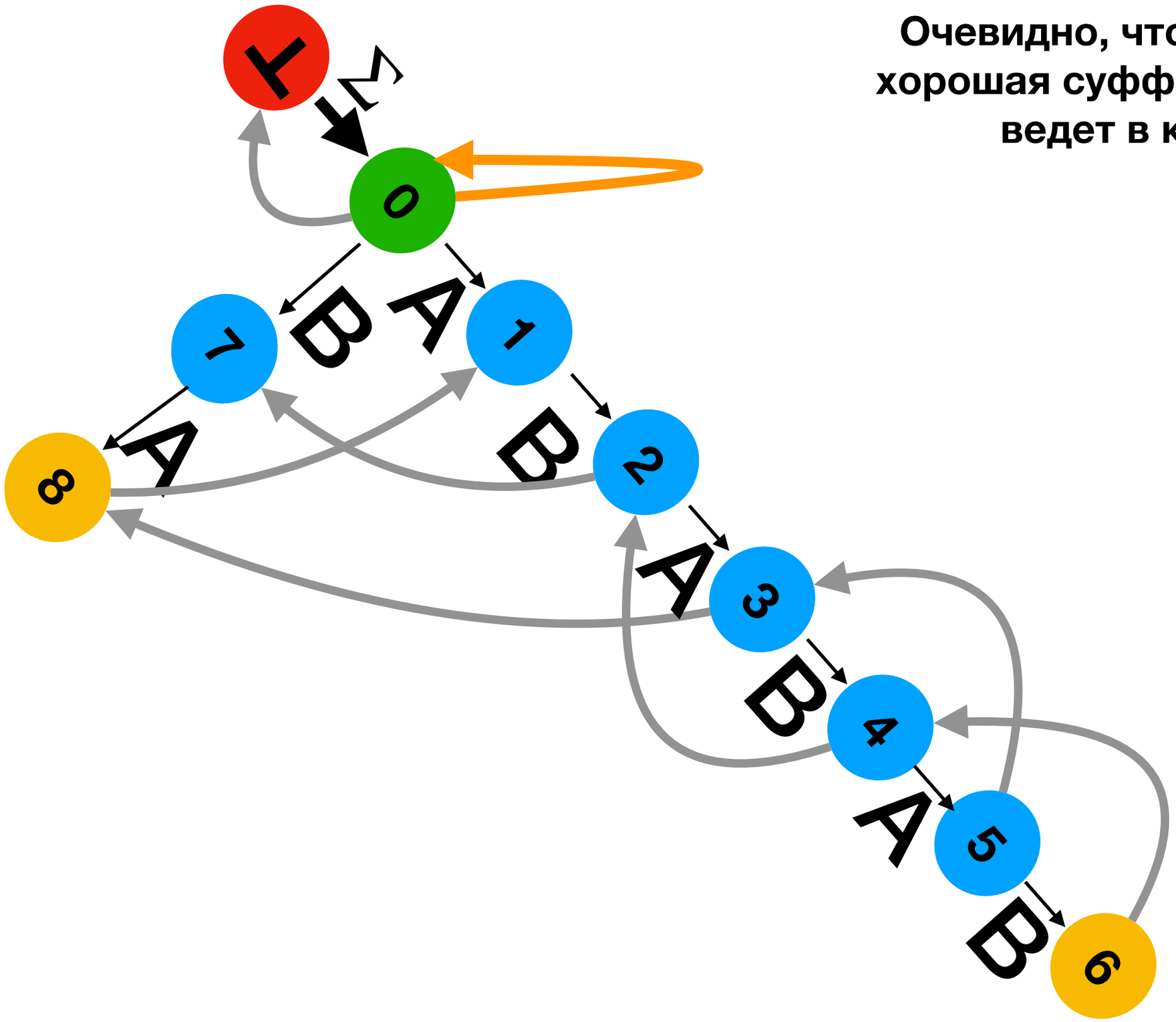


Хорошая суффиксная ссылка

Хорошей (сжатой, функцией выхода) суффиксной ссылкой данной вершины будем называть ссылку ведущую в наибольший собственный суффикс соответствующей вершине строки, который также является паттерном. Если такого суффикса нет, то хорошая суффиксная ссылка ведет в корень

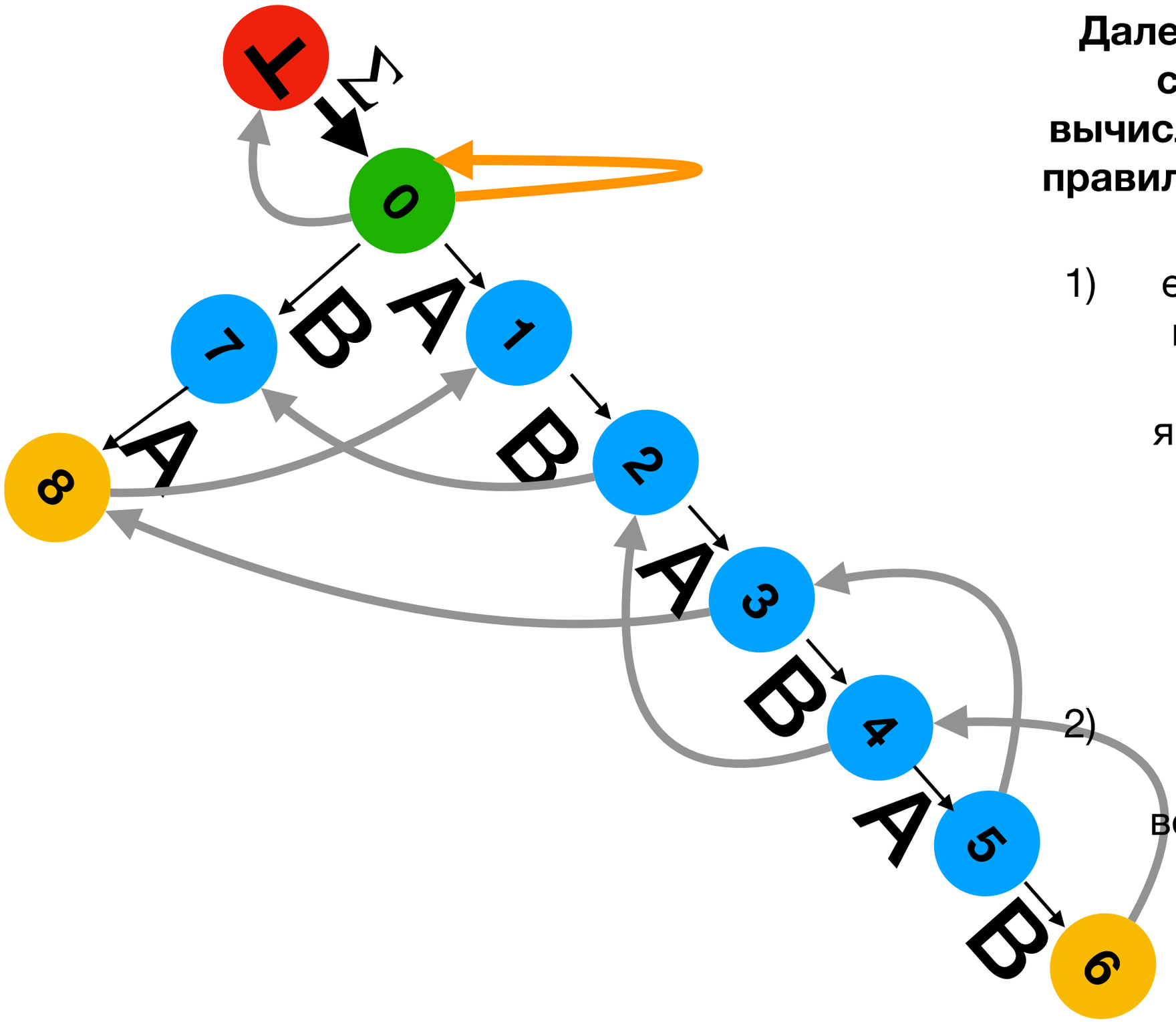
Например, если у нас есть паттерны ABCD и BC, то из вершины, которой соответствует строка ABC должна вести хорошая суффиксная ссылка в вершину, соответствующую строке BC

Подсчитаем суффиксные ссылки



Очевидно, что для корня хорошая суффикса ссылка ведет в корень

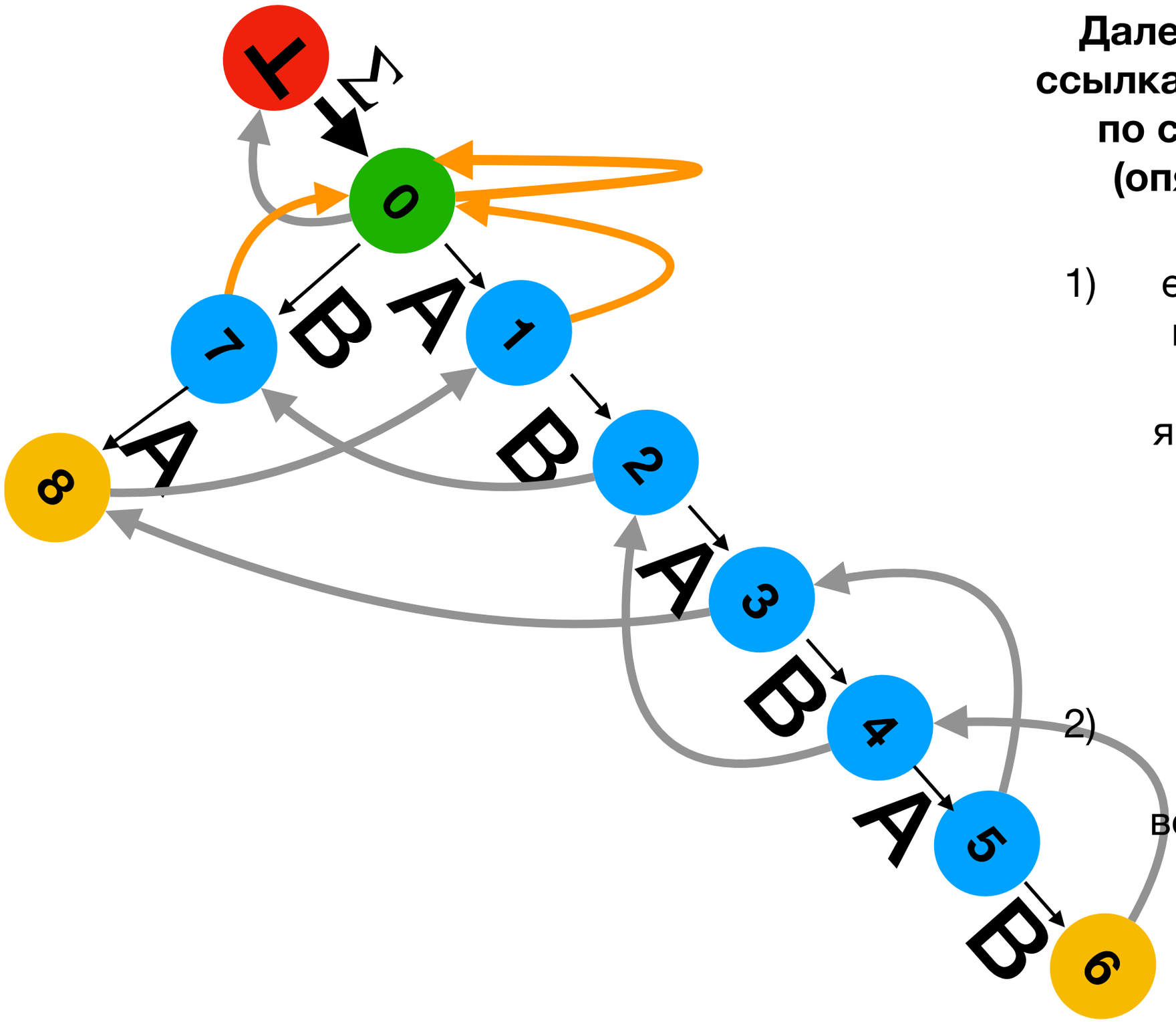
Подсчитаем суффиксные ссылки



Далее хорошая суффикса ссылка вершины v вычисляется по следующему правилу (опять идем от корня к концам):

1) если суффиксная ссылка вершины ***v*** указывает на вершину ***u***, которая является концом паттерна или корнем, то хорошей суффиксной ссылкой вершины ***v*** будет суффиксная ссылка вершины ***v***
иначе - хорошей суффиксной ссылкой вершины ***v*** будет хорошая суффиксная ссылка вершины ***u***

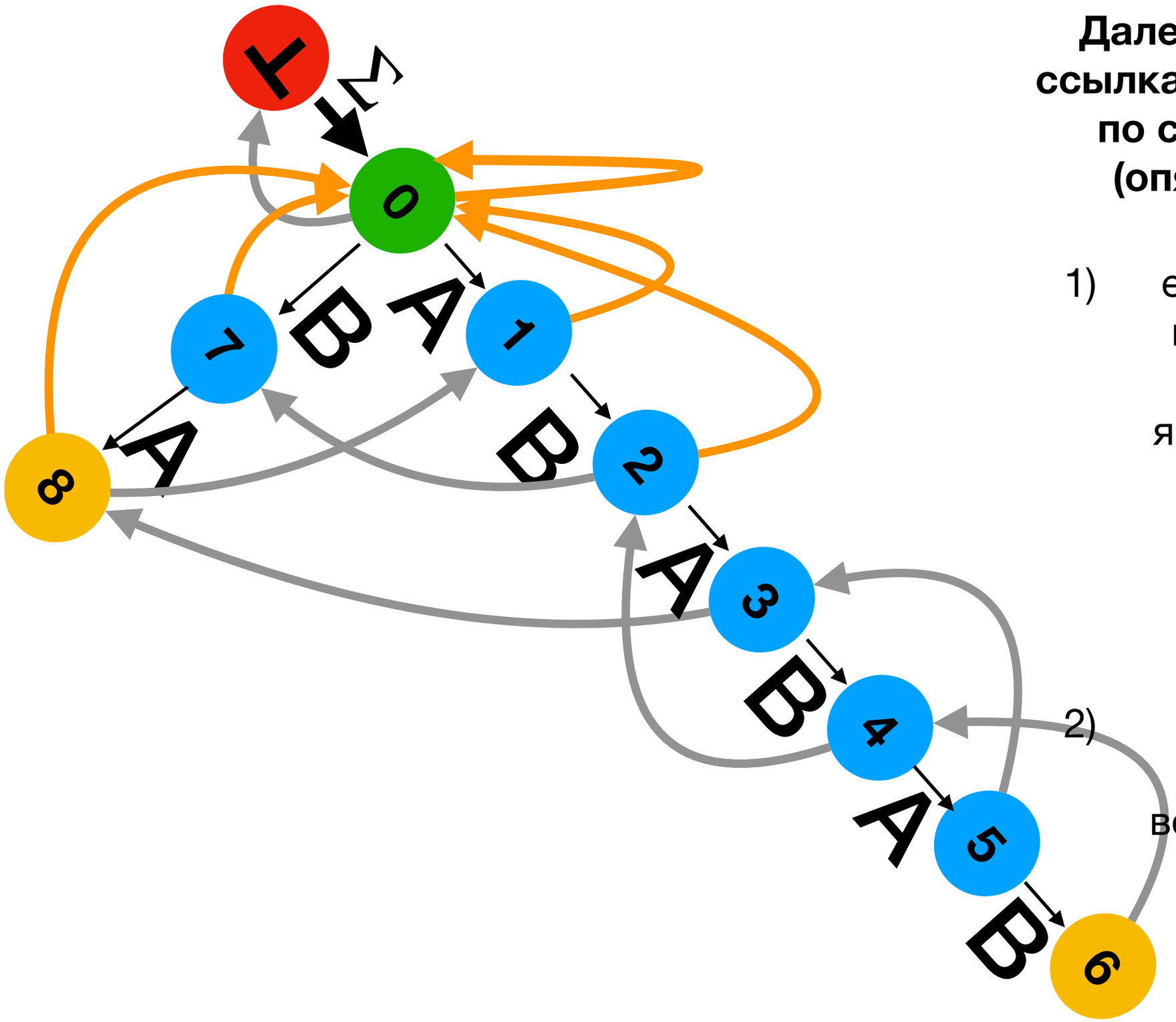
Подсчитаем суффиксные ссылки



Далее хорошая суффикса ссылка вершины вычисляется по следующему правилу (опять идем от корня к концам):

1) если суффиксная ссылка вершины ***v*** указывает на вершину ***u***, которая является концом паттерна или корнем, то хорошей суффиксной ссылкой вершины ***v*** будет суффиксная ссылка вершины ***v***
иначе - хорошей суффиксной ссылкой вершины ***v*** будет хорошая суффиксная ссылка вершины ***u***

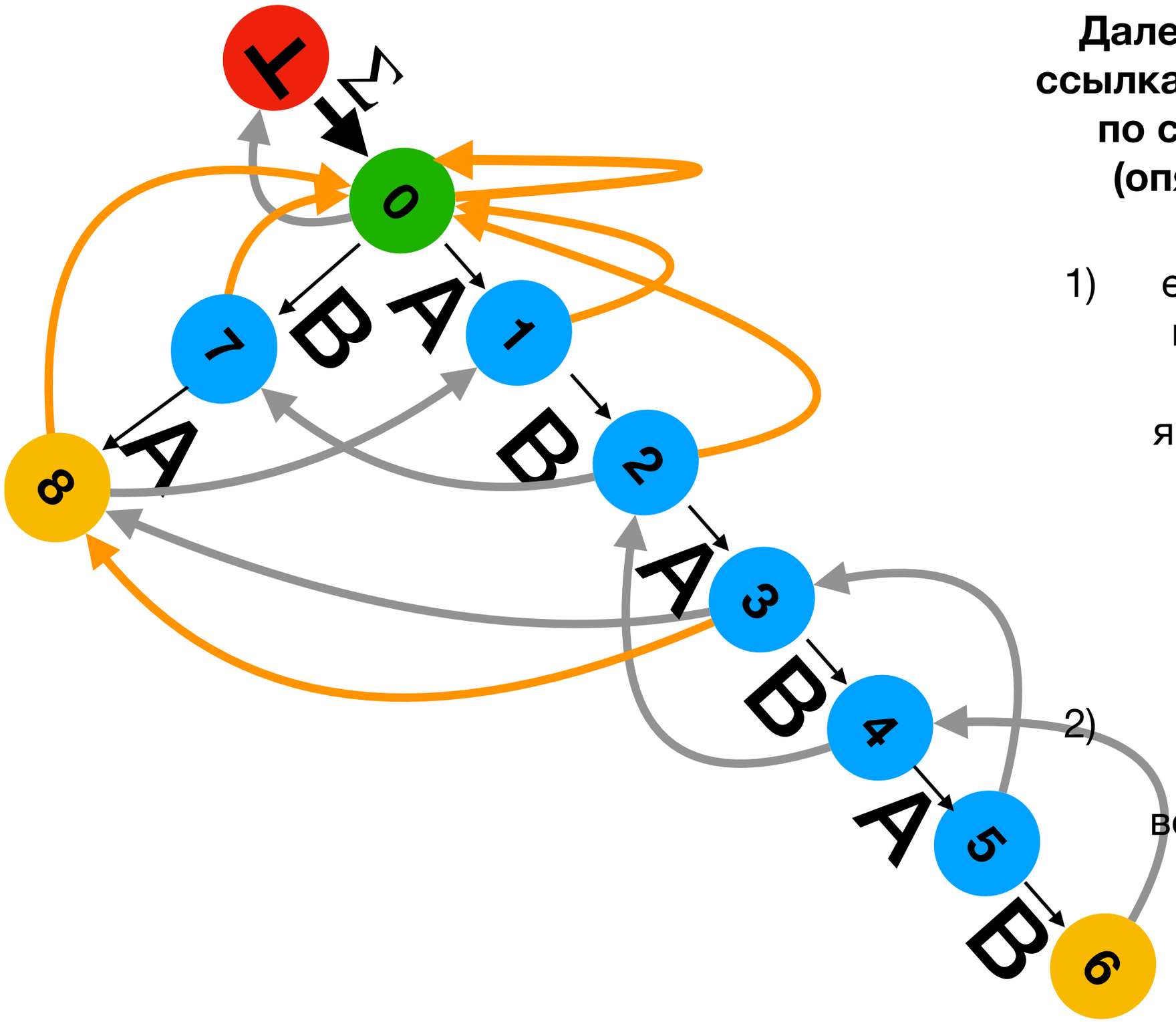
Подсчитаем суффиксные ссылки



Далее хорошая суффикса ссылка вершины вычисляется по следующему правилу (опять идем от корня к концам):

1) если суффиксная ссылка вершины ***v*** указывает на вершину ***u***, которая является концом паттерна или корнем, то хорошей суффиксной ссылкой вершины ***v*** будет суффиксная ссылка вершины ***v***
иначе - хорошей суффиксной ссылкой вершины ***v*** будет хорошая суффиксная ссылка вершины ***u***

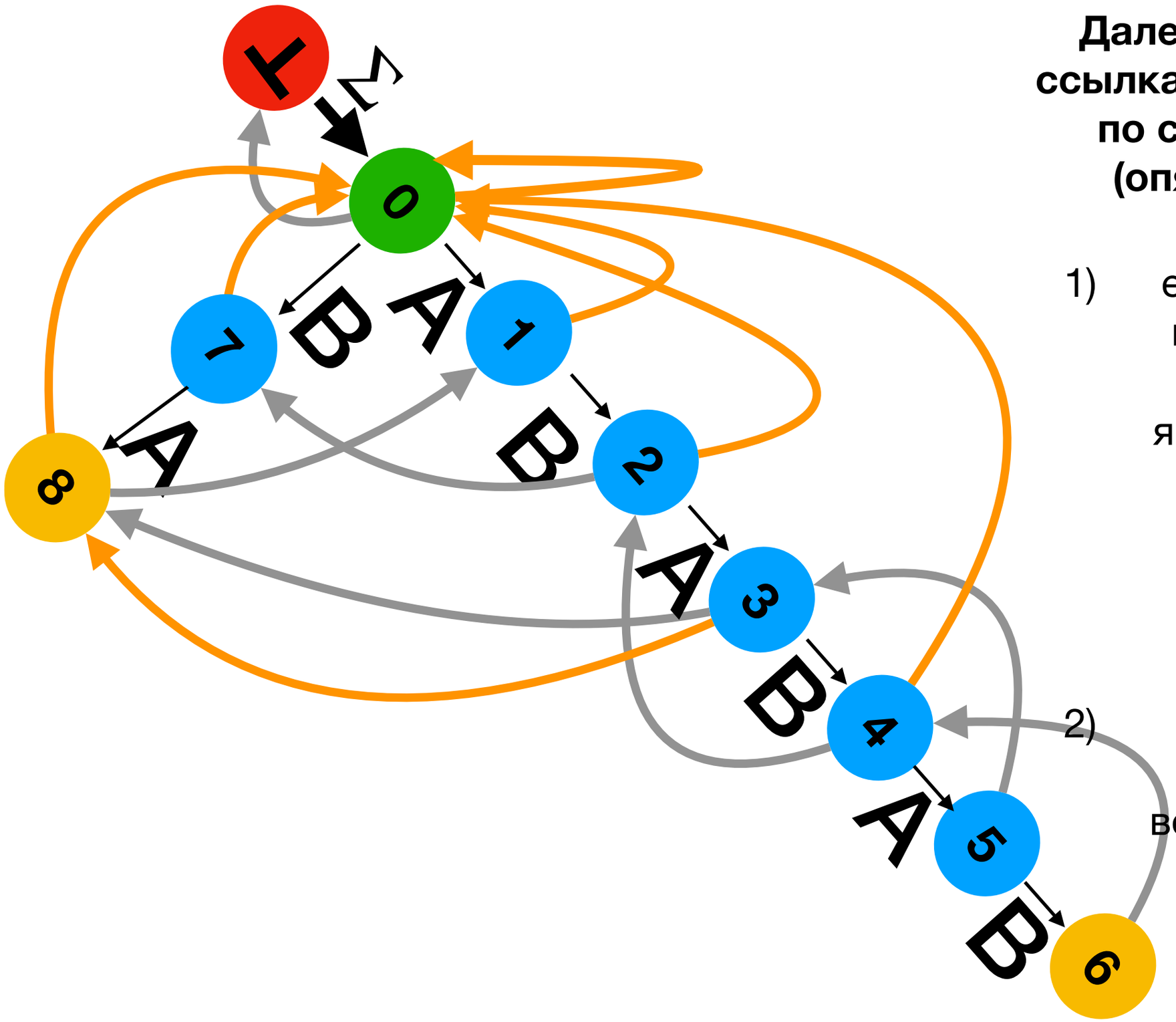
Подсчитаем суффиксные ссылки



Далее хорошая суффикса ссылка вершины вычисляется по следующему правилу (опять идем от корня к концам):

1) если суффиксная ссылка вершины ***v*** указывает на вершину ***u***, которая является концом паттерна или корнем, то хорошей суффиксной ссылкой вершины ***v*** будет суффиксная ссылка вершины ***v***
иначе - хорошей суффиксной ссылкой вершины ***v*** будет хорошая суффиксная ссылка вершины ***u***

Подсчитаем суффиксные ссылки

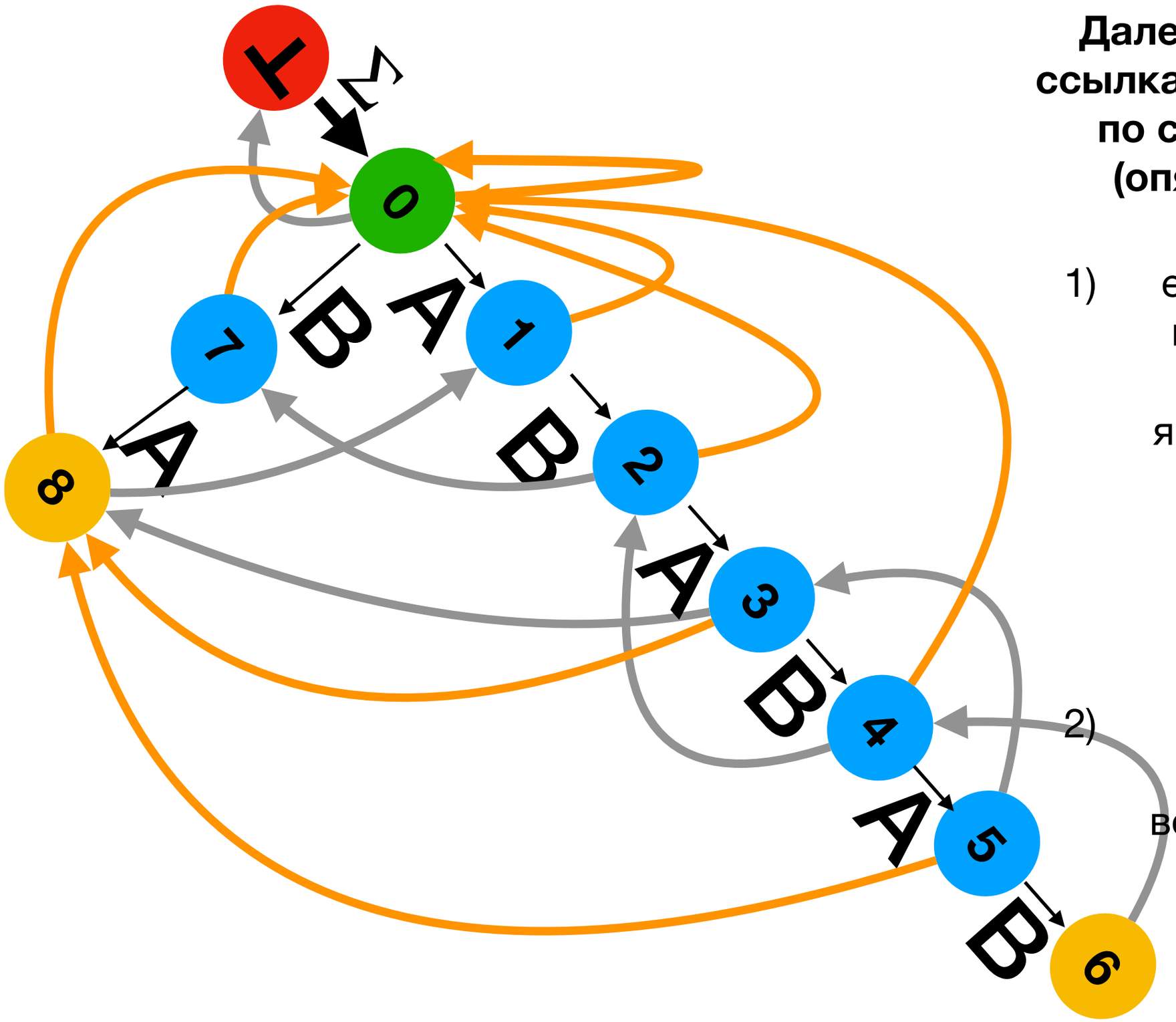


Далее хорошая суффикса ссылка вершины вычисляется по следующему правилу (опять идем от корня к концам):

1) если суффиксная ссылка вершины ***v*** указывает на вершину ***u***, которая является концом паттерна или корнем, то хорошей суффиксной ссылкой вершины ***v*** будет суффиксная ссылка вершины ***v***

2) иначе - хорошей суффиксной ссылкой вершины ***v*** будет хорошая суффиксная ссылка вершины ***u***

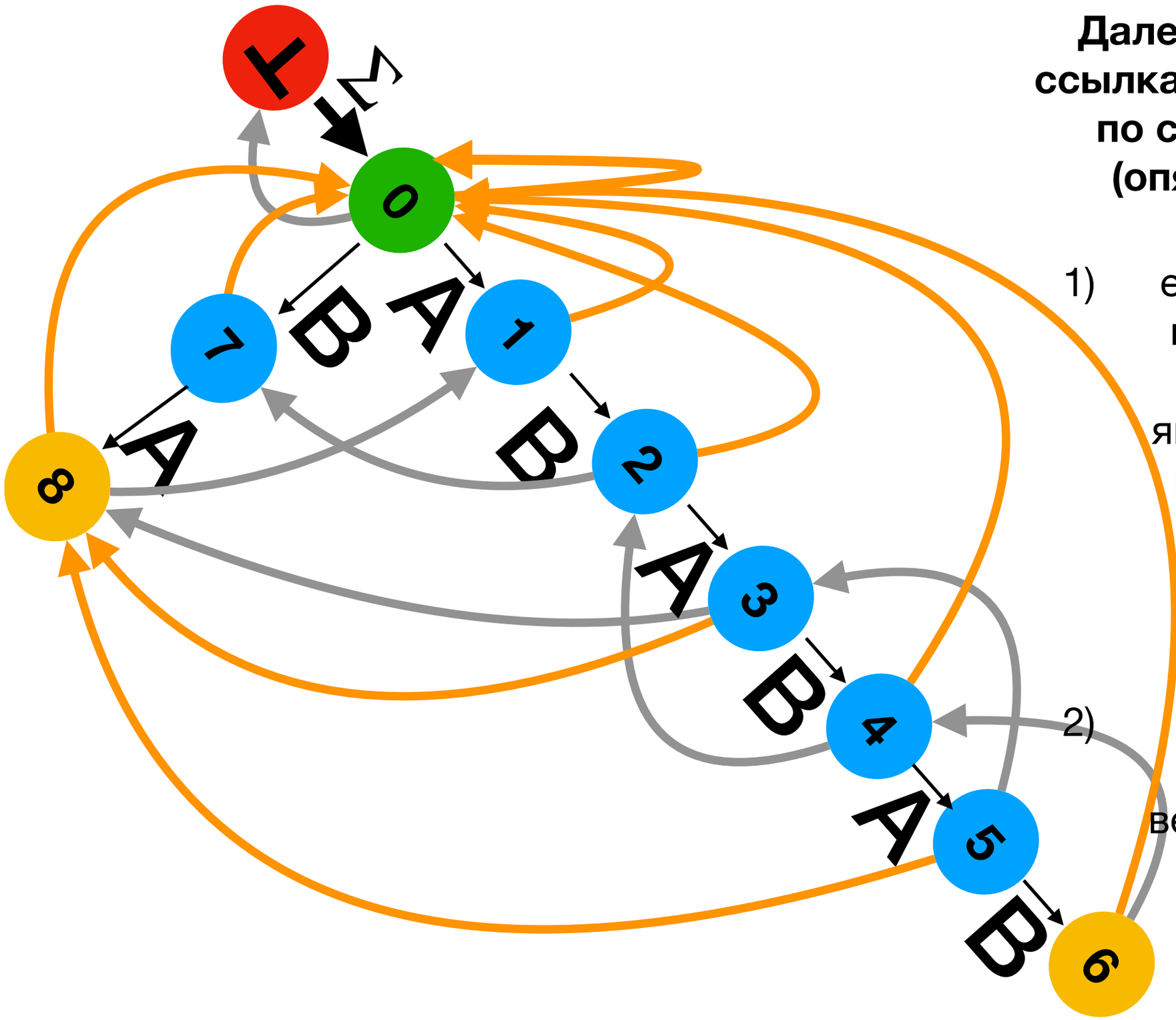
Подсчитаем суффиксные ссылки



Далее хорошая суффикса ссылка вершины вычисляется по следующему правилу (опять идем от корня к концам):

1) если суффиксная ссылка вершины ***v*** указывает на вершину ***u***, которая является концом паттерна или корнем, то хорошей суффиксной ссылкой вершины ***v*** будет суффиксная ссылка вершины ***v***
иначе - хорошей суффиксной ссылкой вершины ***v*** будет хорошая суффиксная ссылка вершины ***u***

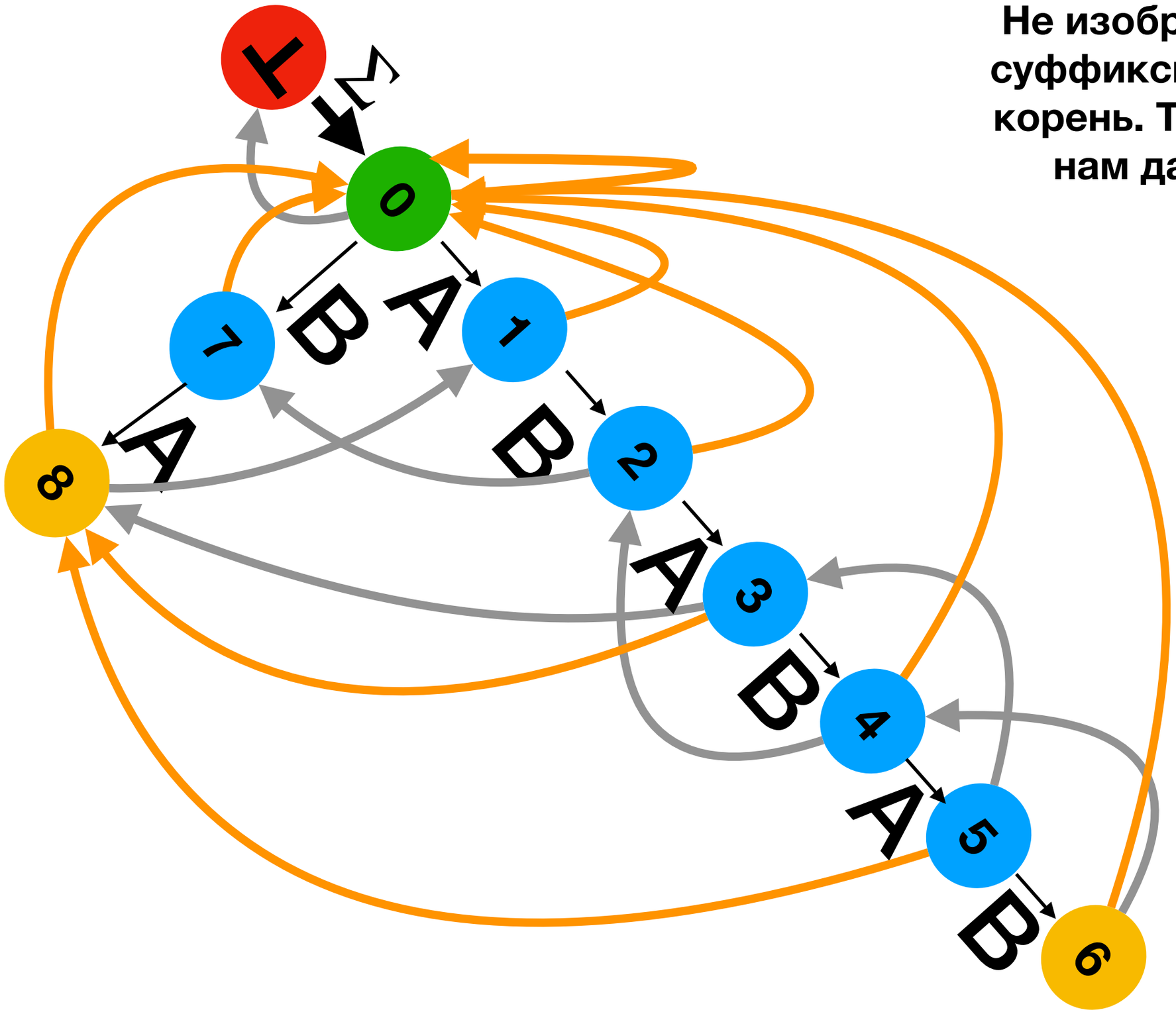
Подсчитаем суффиксные ссылки



Далее хорошая суффикса ссылка вершины вычисляется по следующему правилу (опять идем от корня к концам):

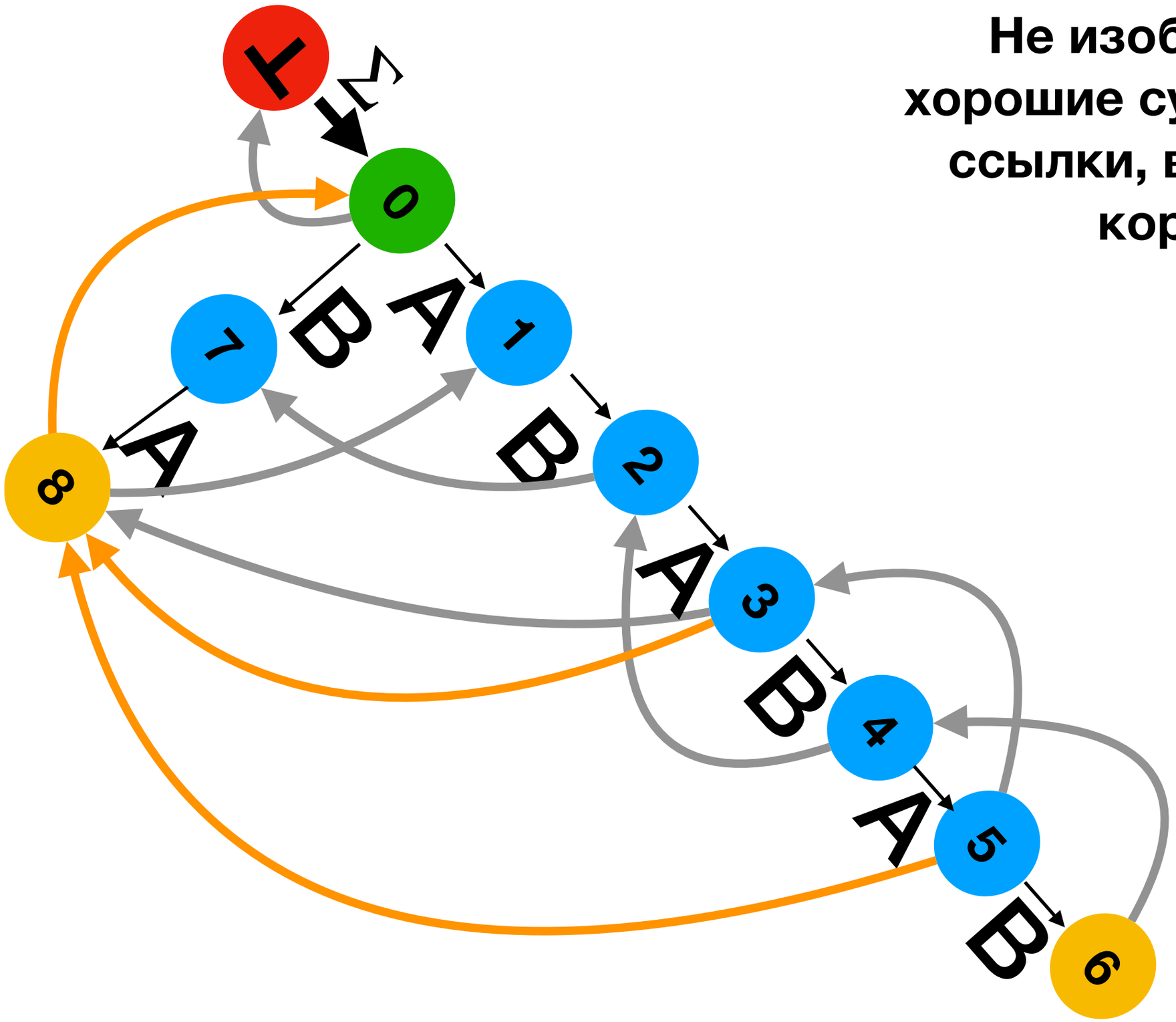
1) если суффиксная ссылка вершины ***v*** указывает на вершину ***u***, которая является концом паттерна или корнем, то хорошей суффиксной ссылкой вершины ***v*** будет суффиксная ссылка вершины ***v***
иначе - хорошей суффиксной ссылкой вершины ***v*** будет хорошая суффиксная ссылка вершины ***u***

Подсчитаем суффиксные ссылки



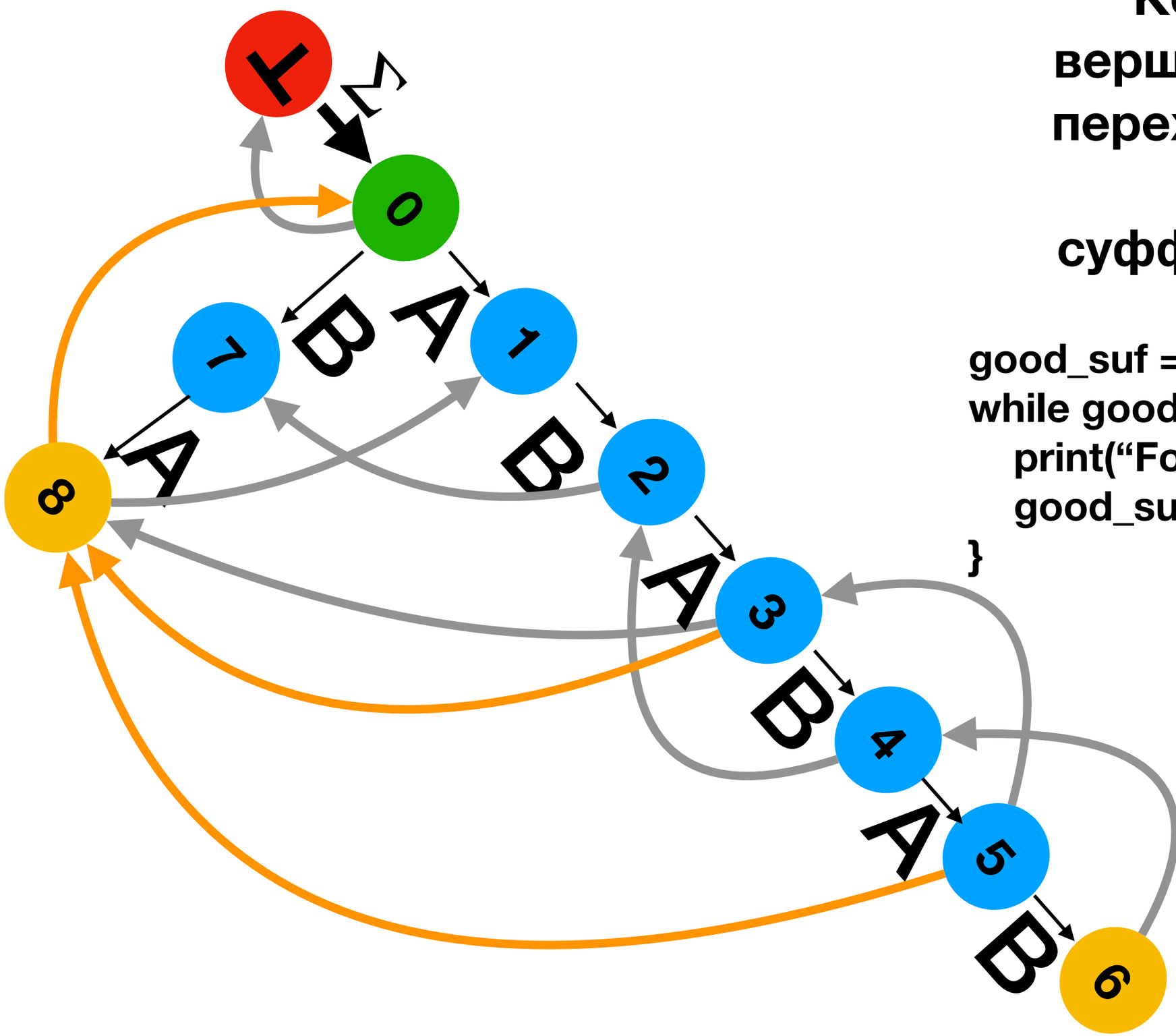
Не изображаем все хорошие суффиксы ссылки, ведущие в корень. Только одну, которая нам далее понадобится

Подсчитаем суффиксные ссылки



Не изображаем хорошие суффиксные ссылки, ведущие в корень

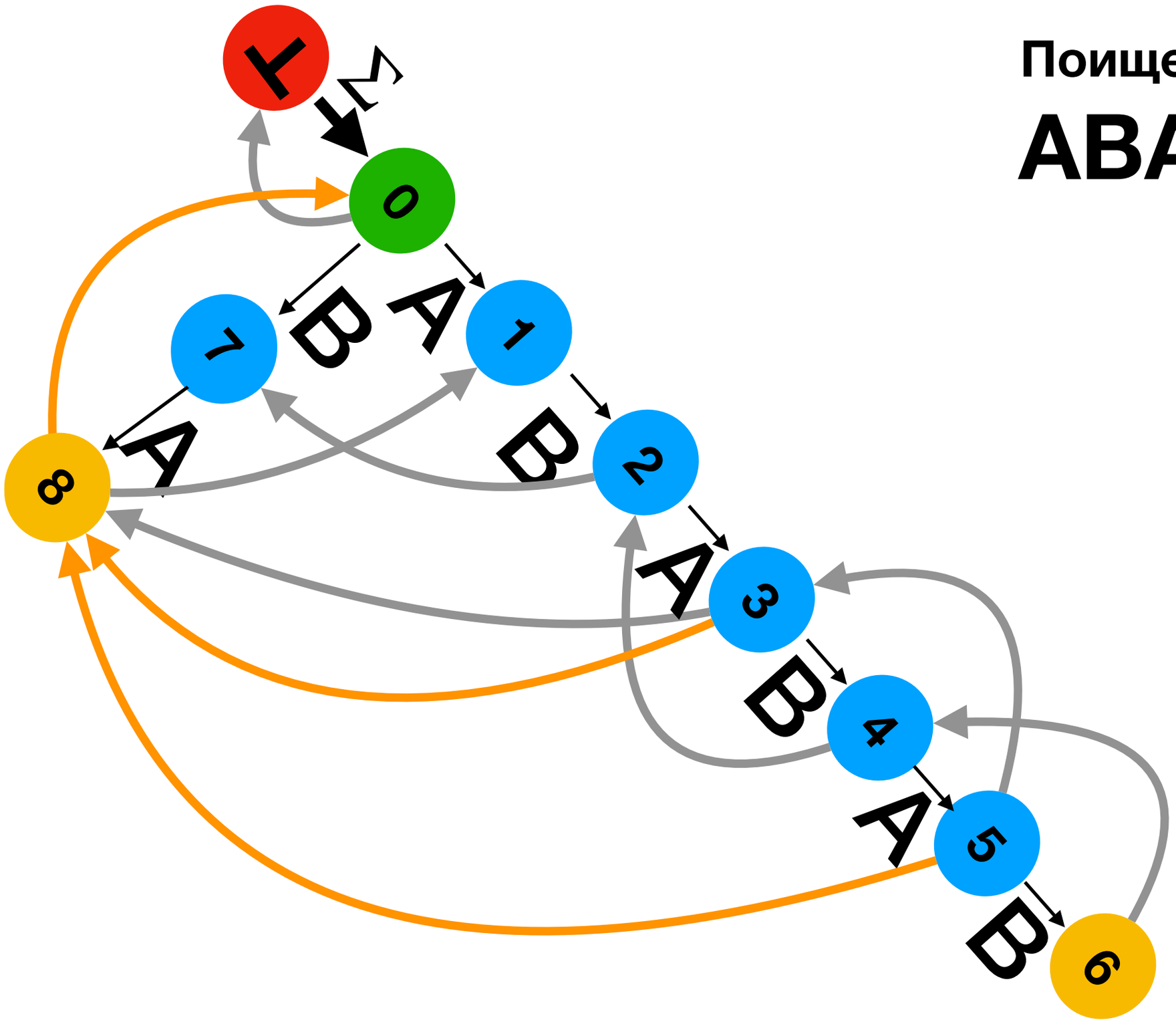
Подсчитаем суффиксные ссылки



Когда заходим в вершину по основному переходу, проходим по ее хорошим суффиксным ссылкам до корня

```
good_suf = cur_node.good_suf
while good_suf != root {
    print("Found {}", good_suf)
    good_suf = good_suf. good_suf
}
```

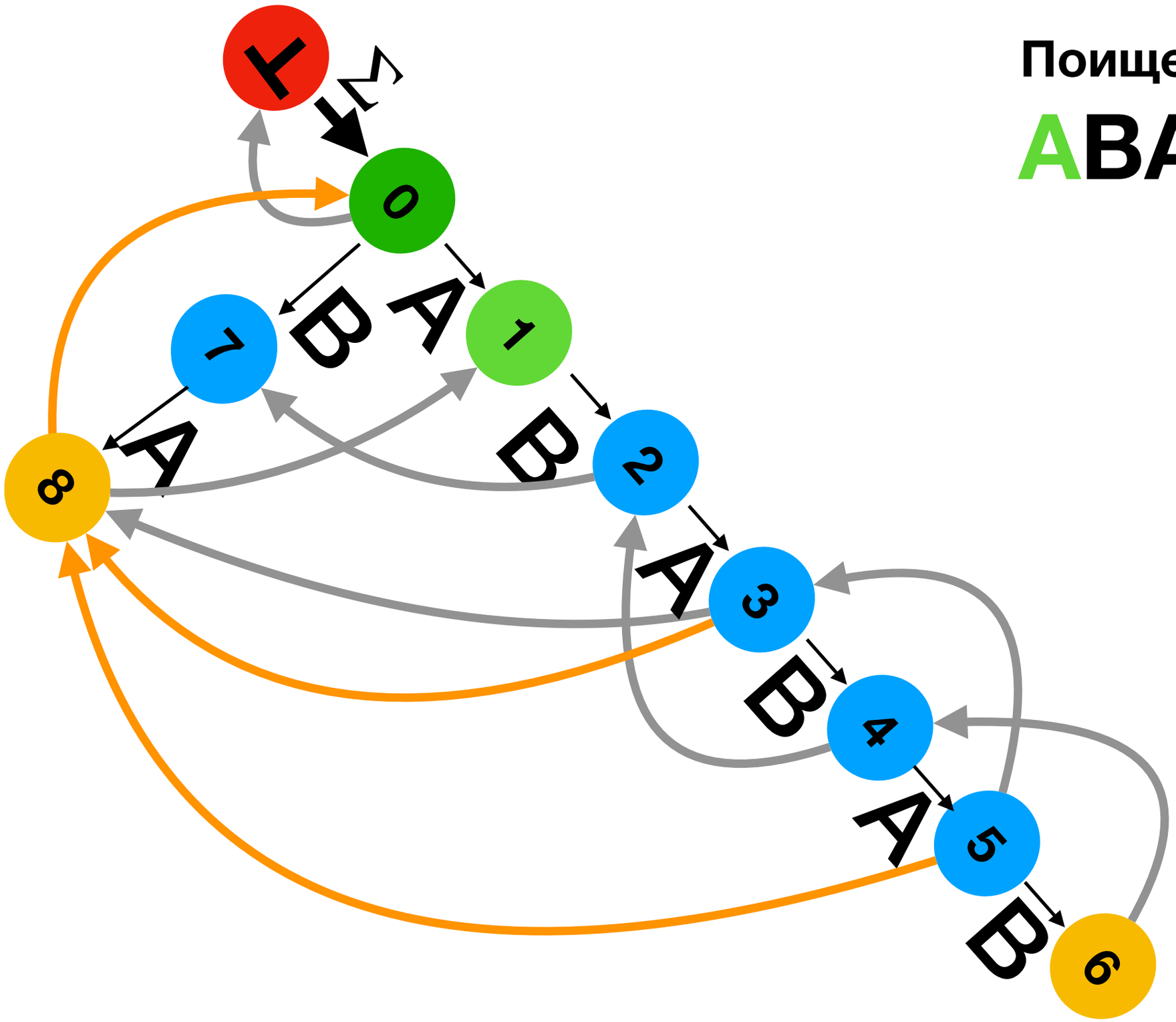
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

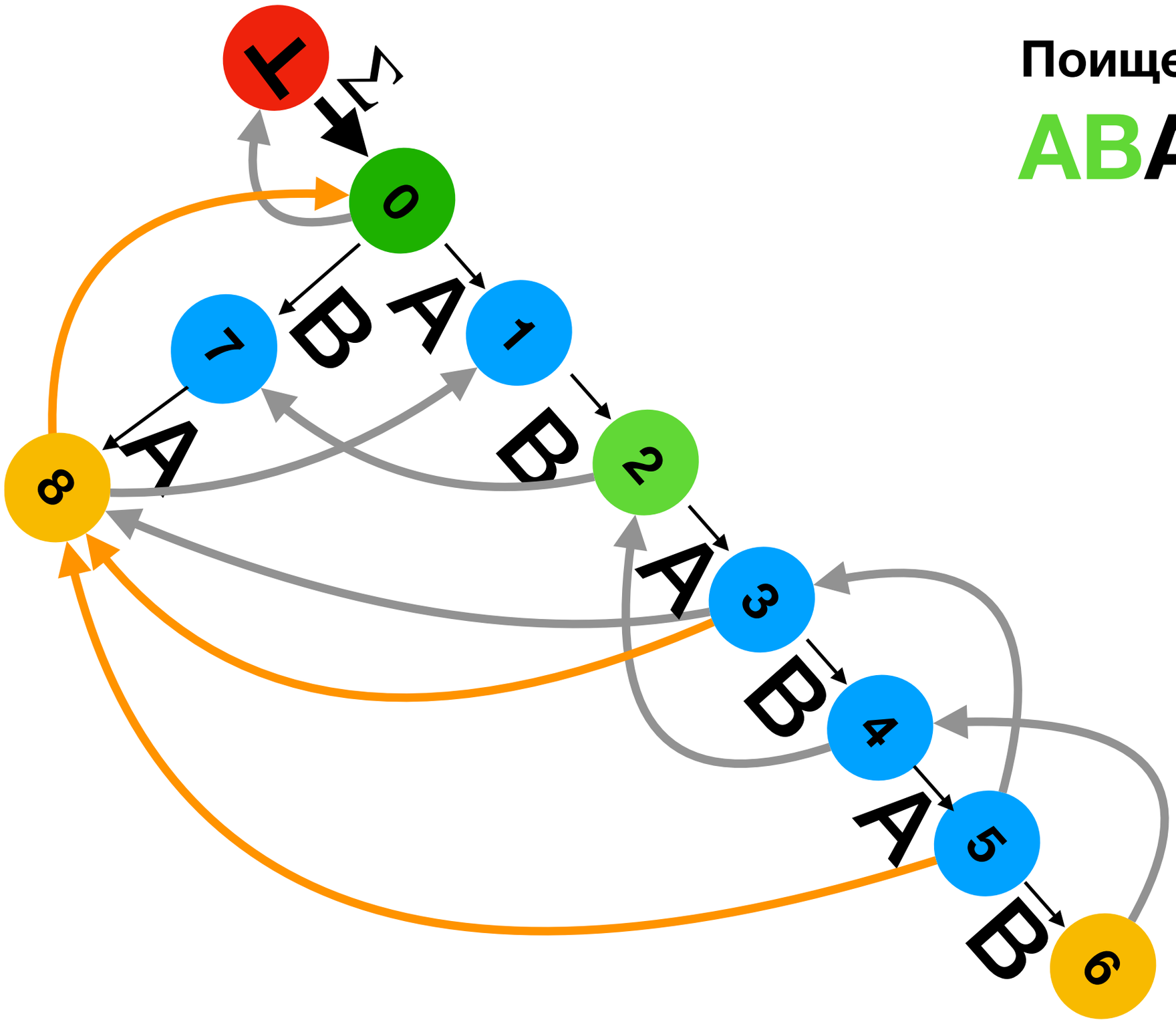
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

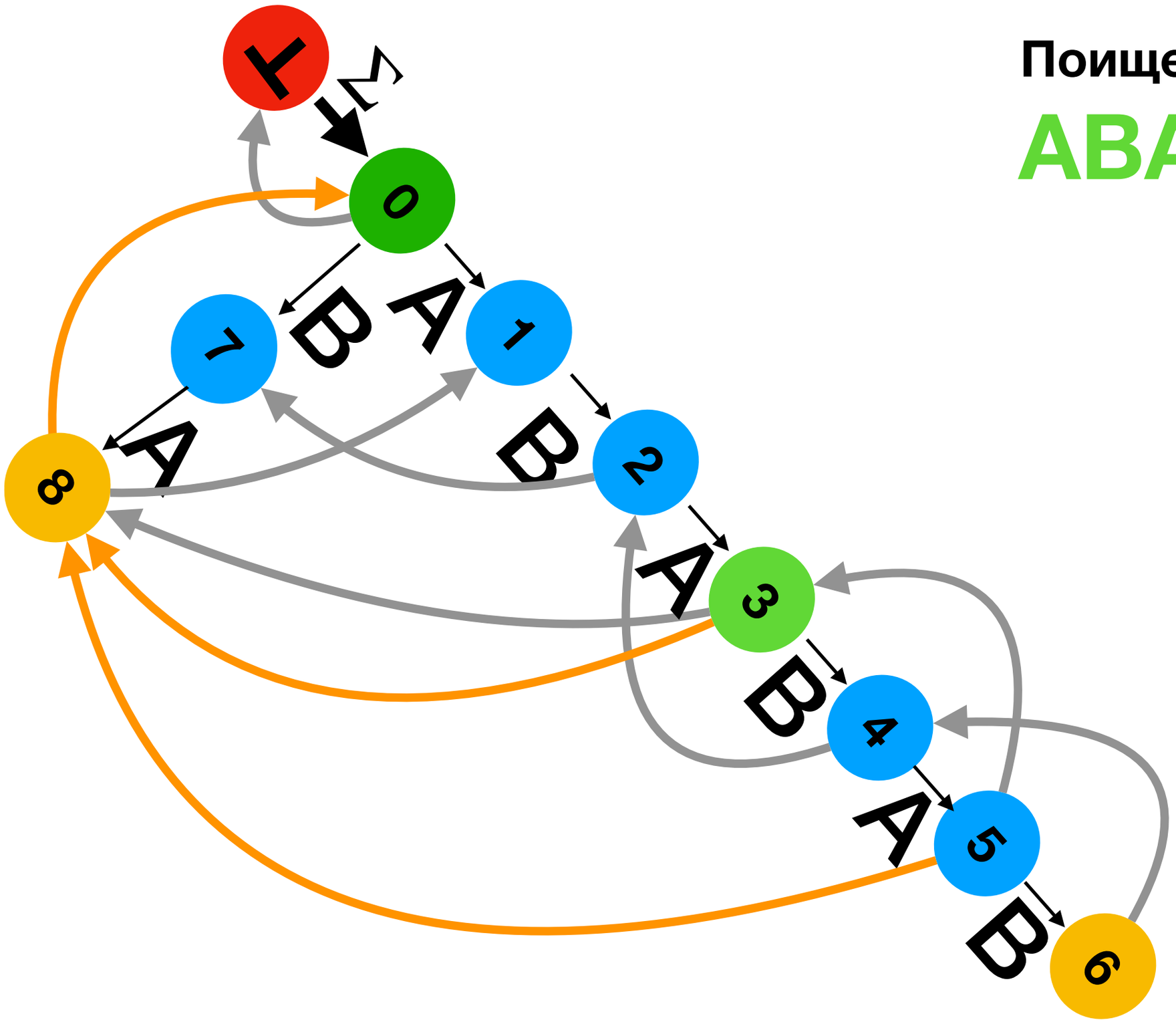
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

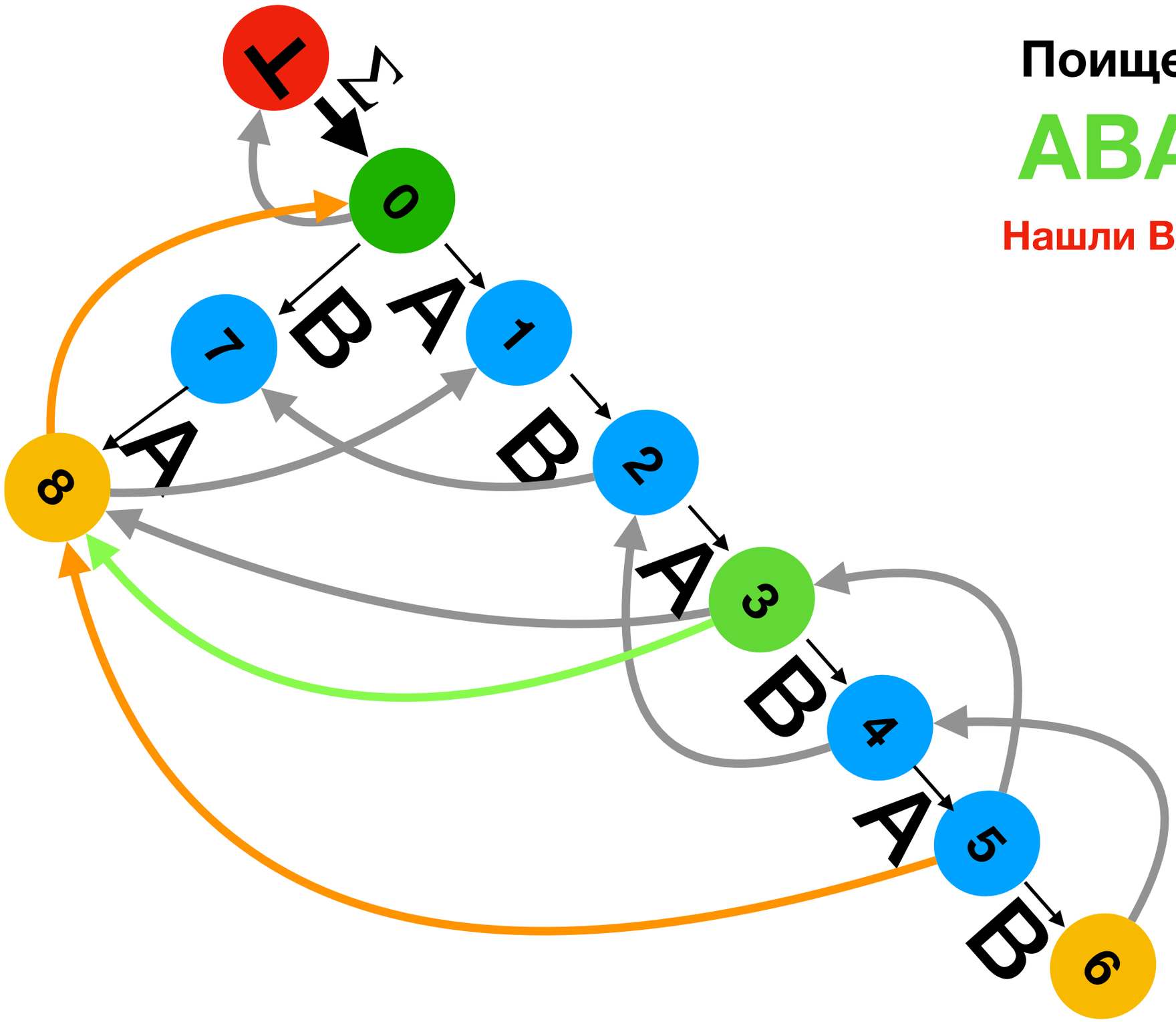
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

Подсчитаем суффиксные ссылки

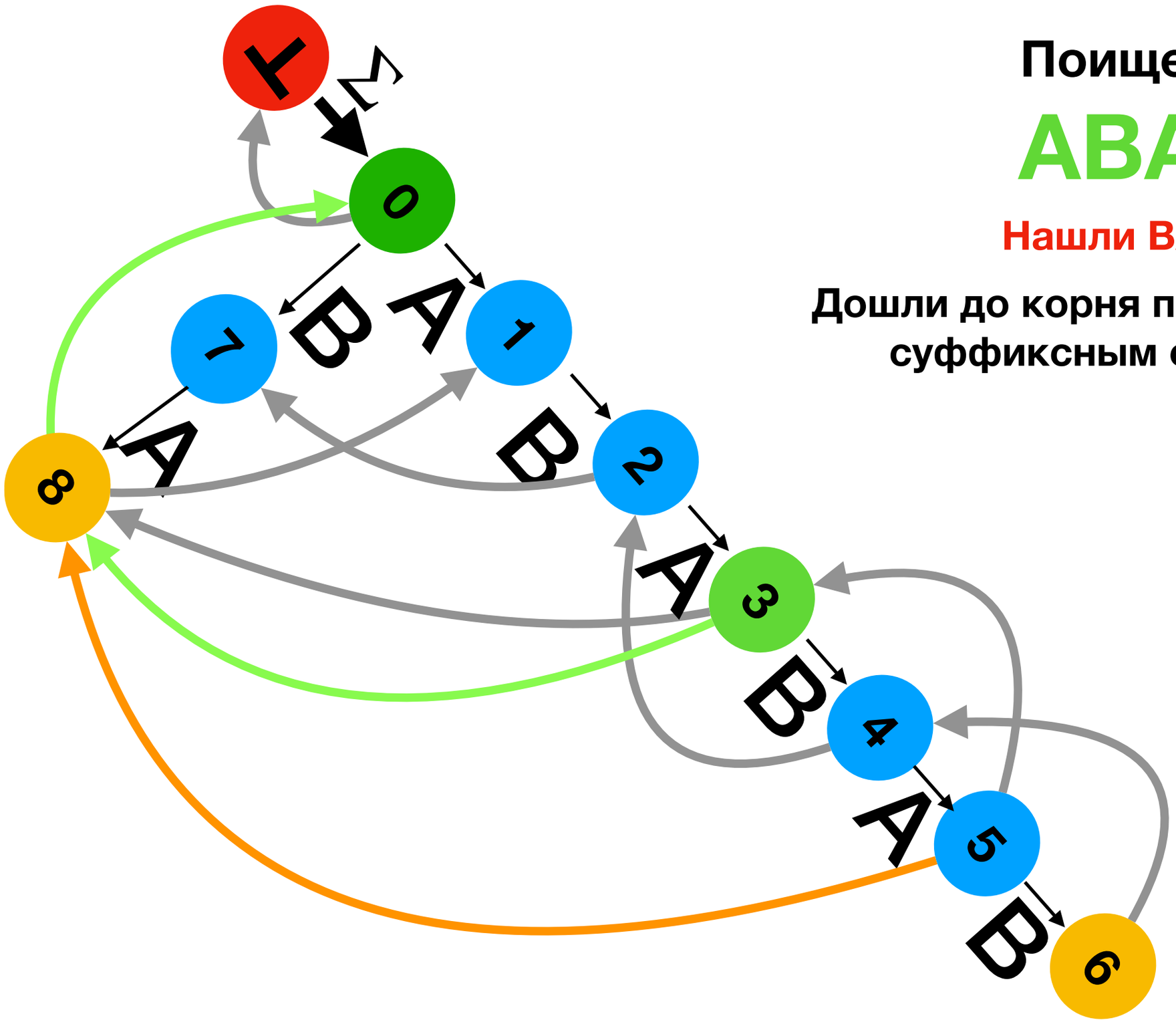


Поищем в строке

ABABABB

Нашли BA

Подсчитаем суффиксные ссылки



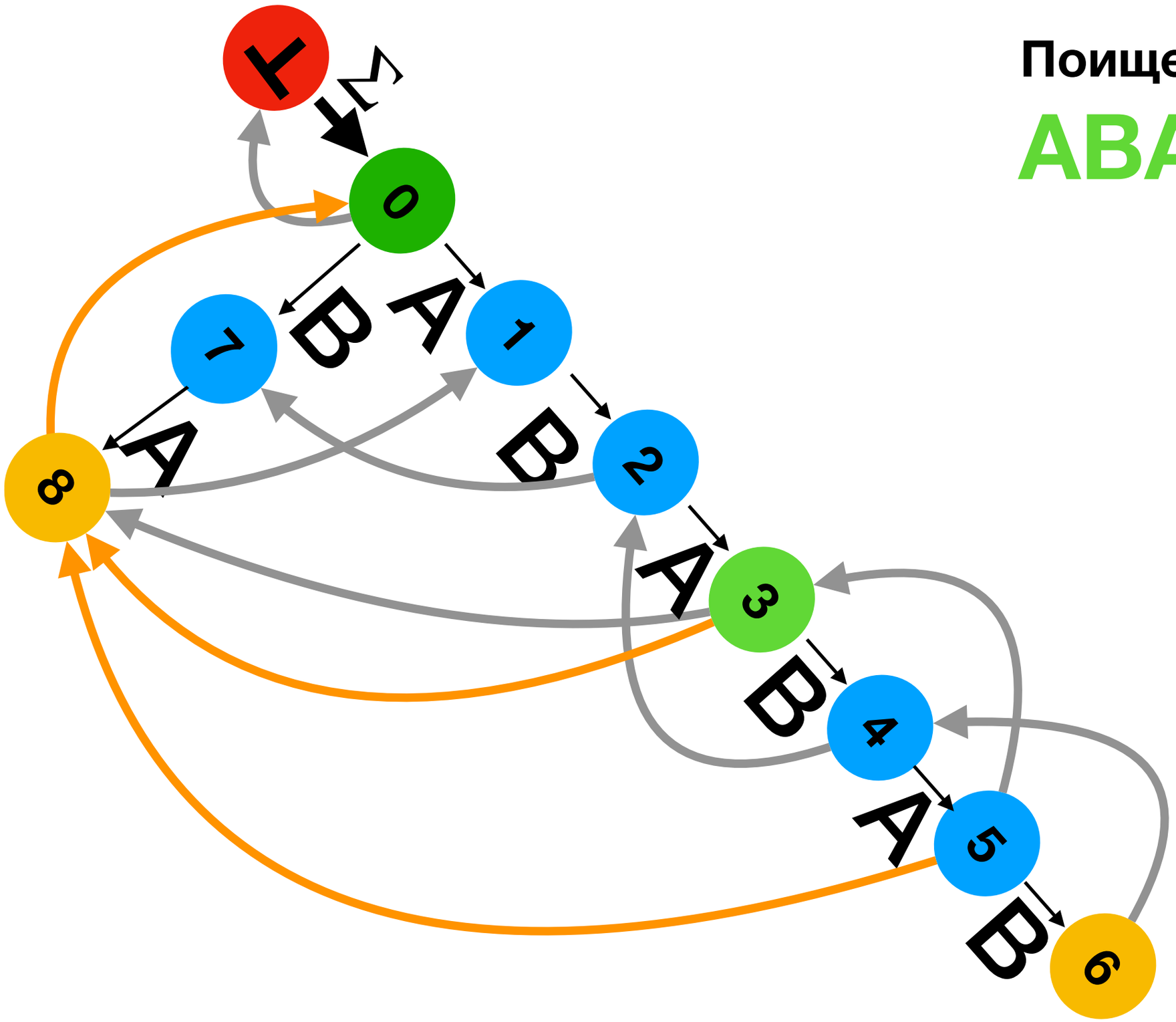
Поищем в строке

ABABABB

Нашли BA

Дошли до корня по хорошим суффиксным ссылка

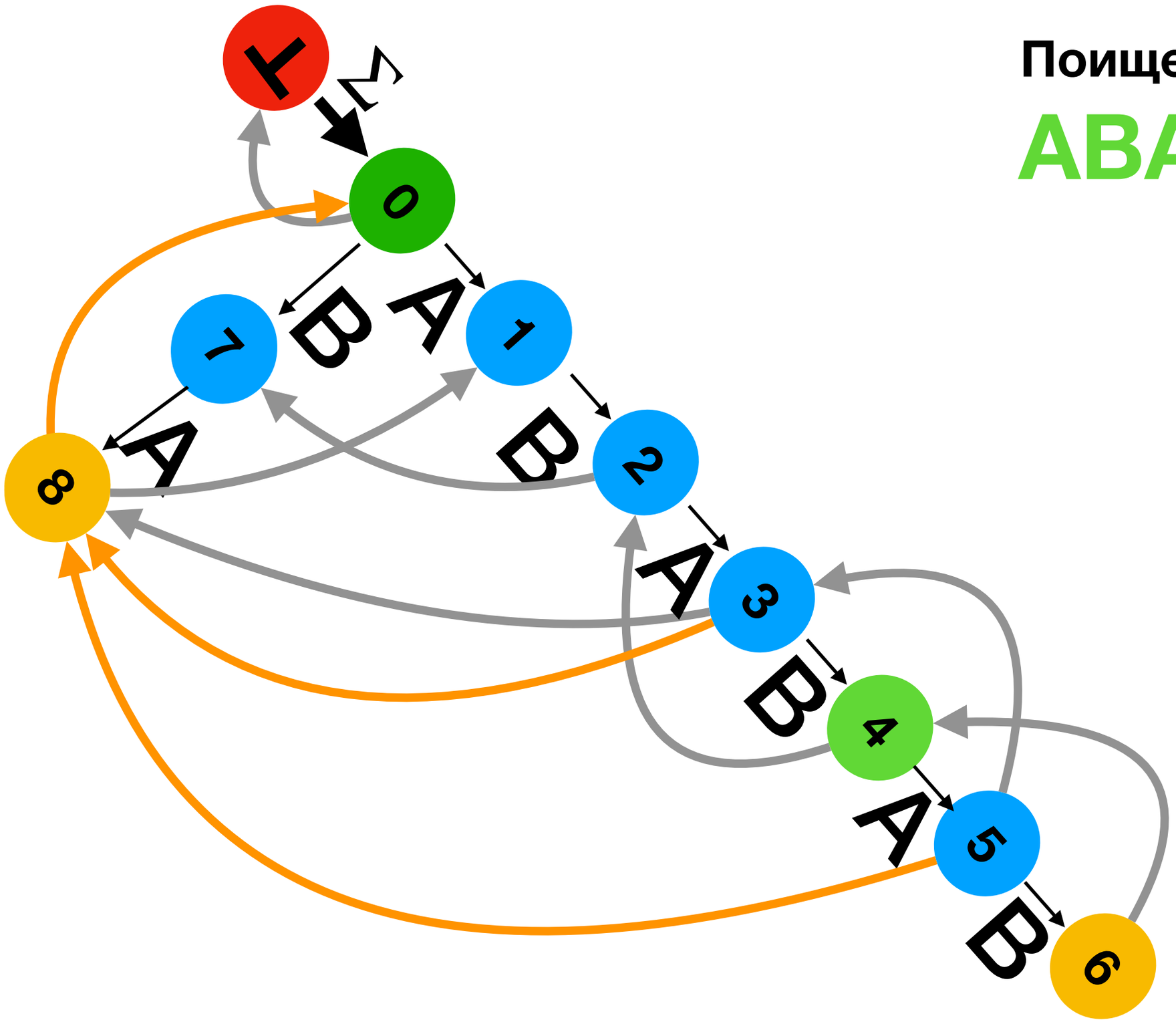
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

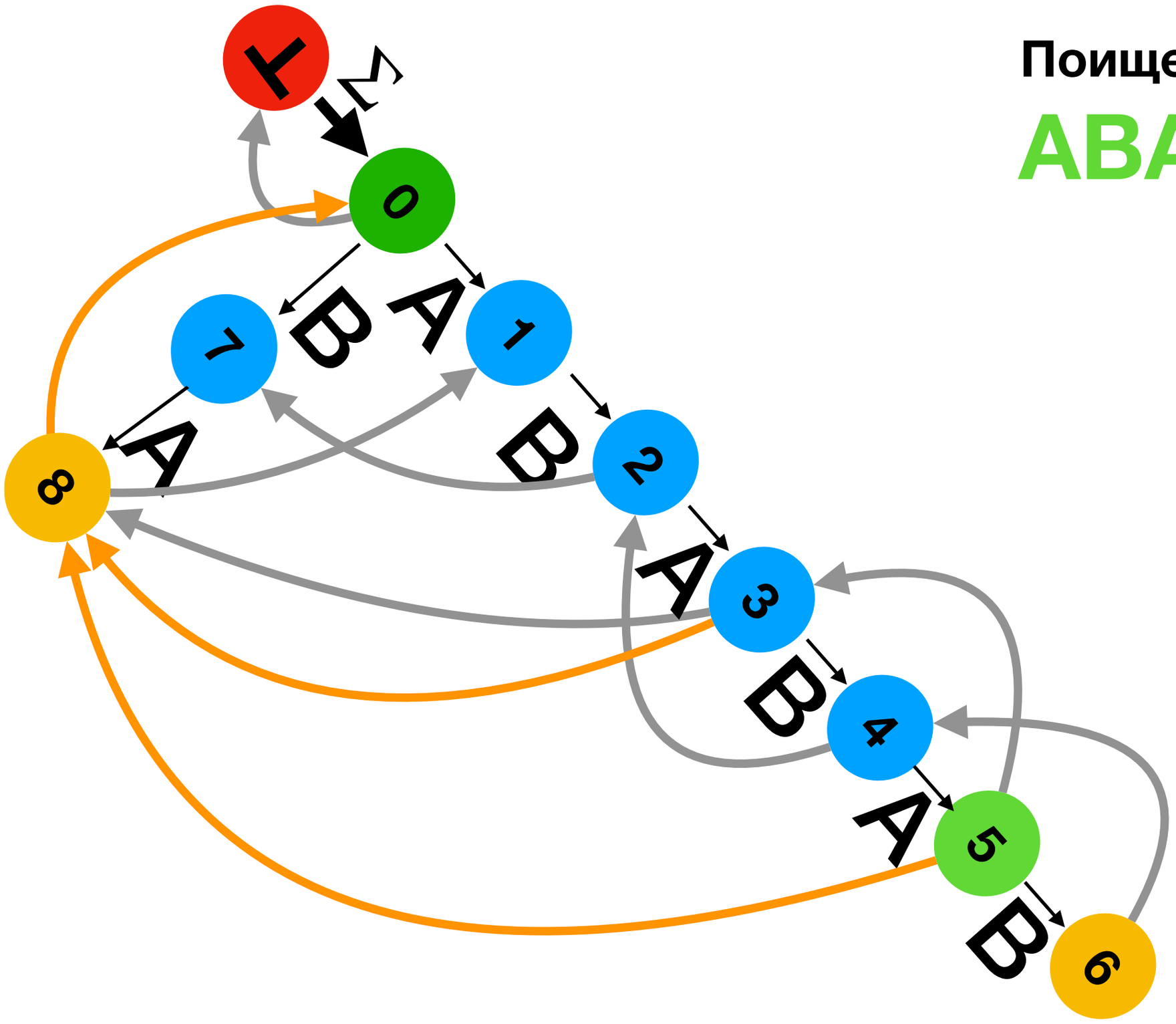
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

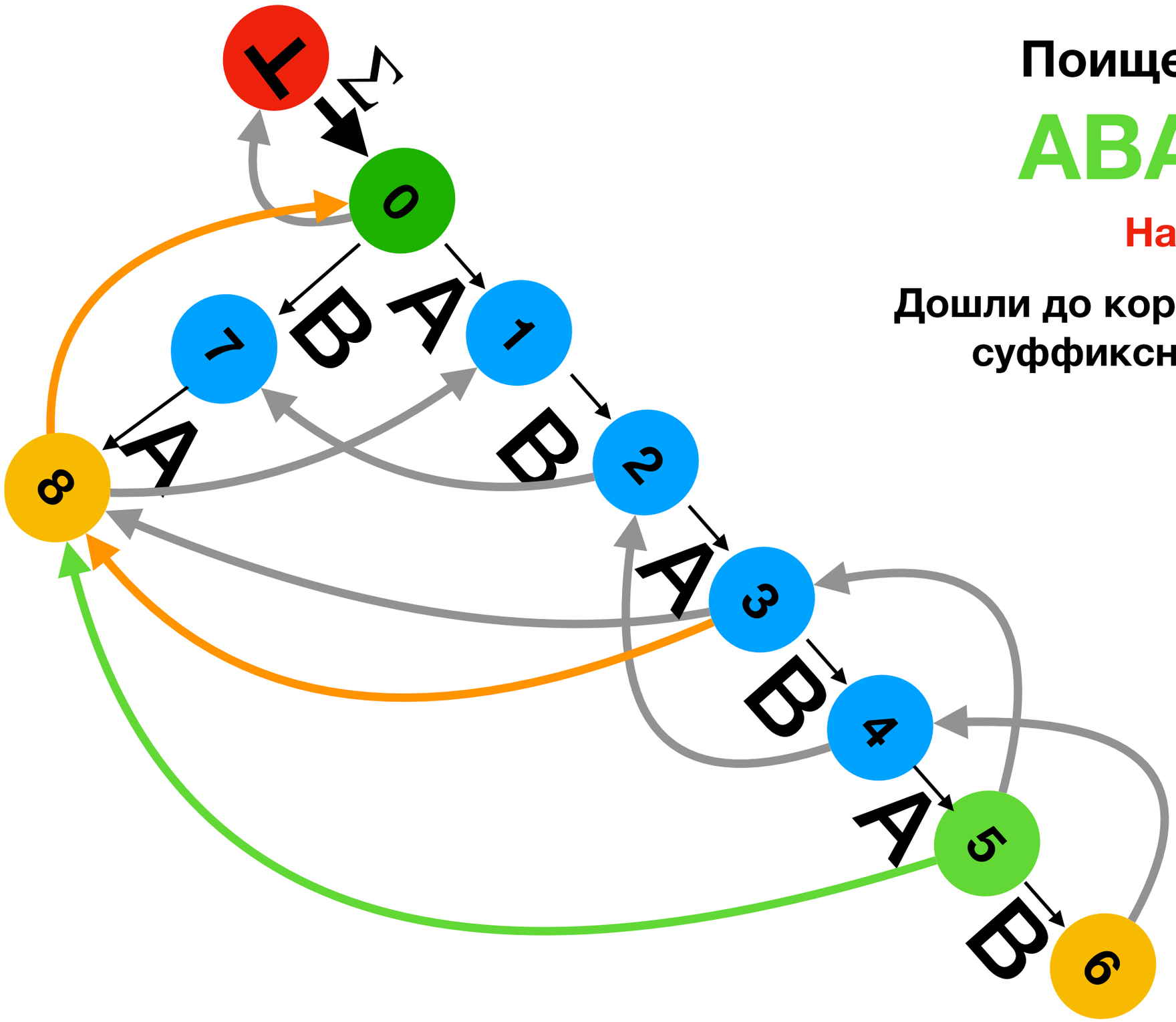
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

Подсчитаем суффиксные ссылки



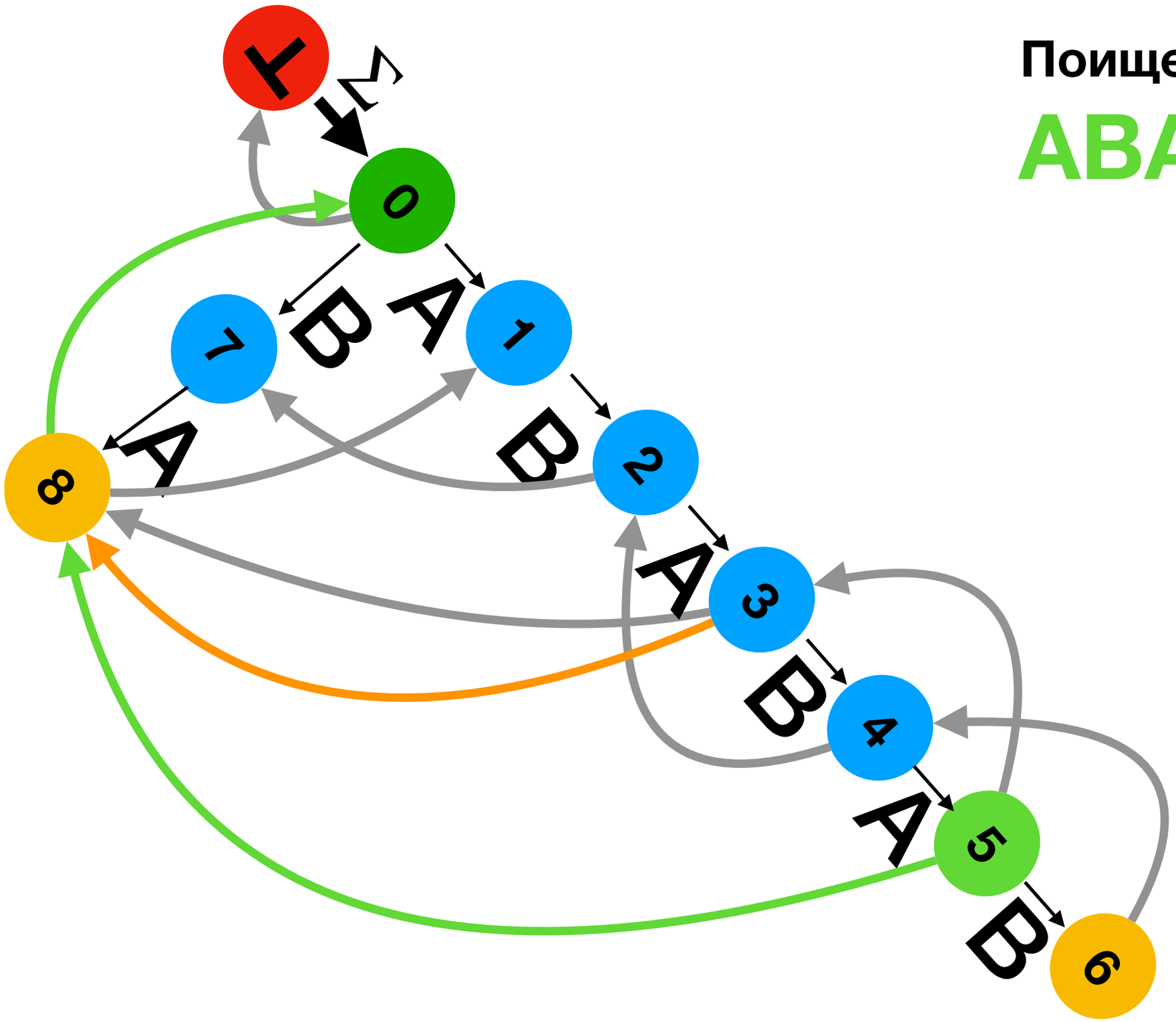
Поищем в строке

ABABABB

Нашли BA

Дошли до корня по хорошим суффиксным ссылка

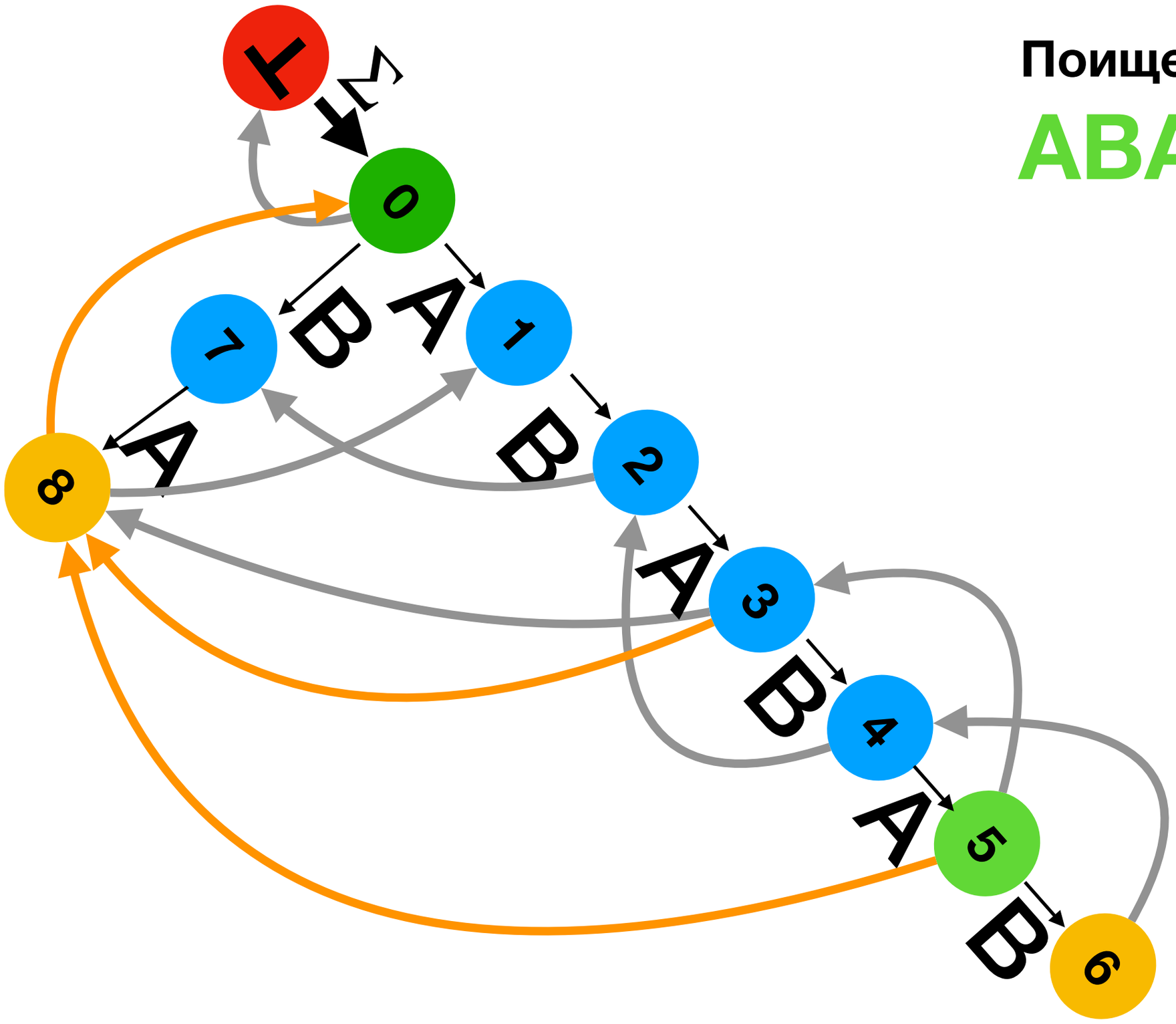
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

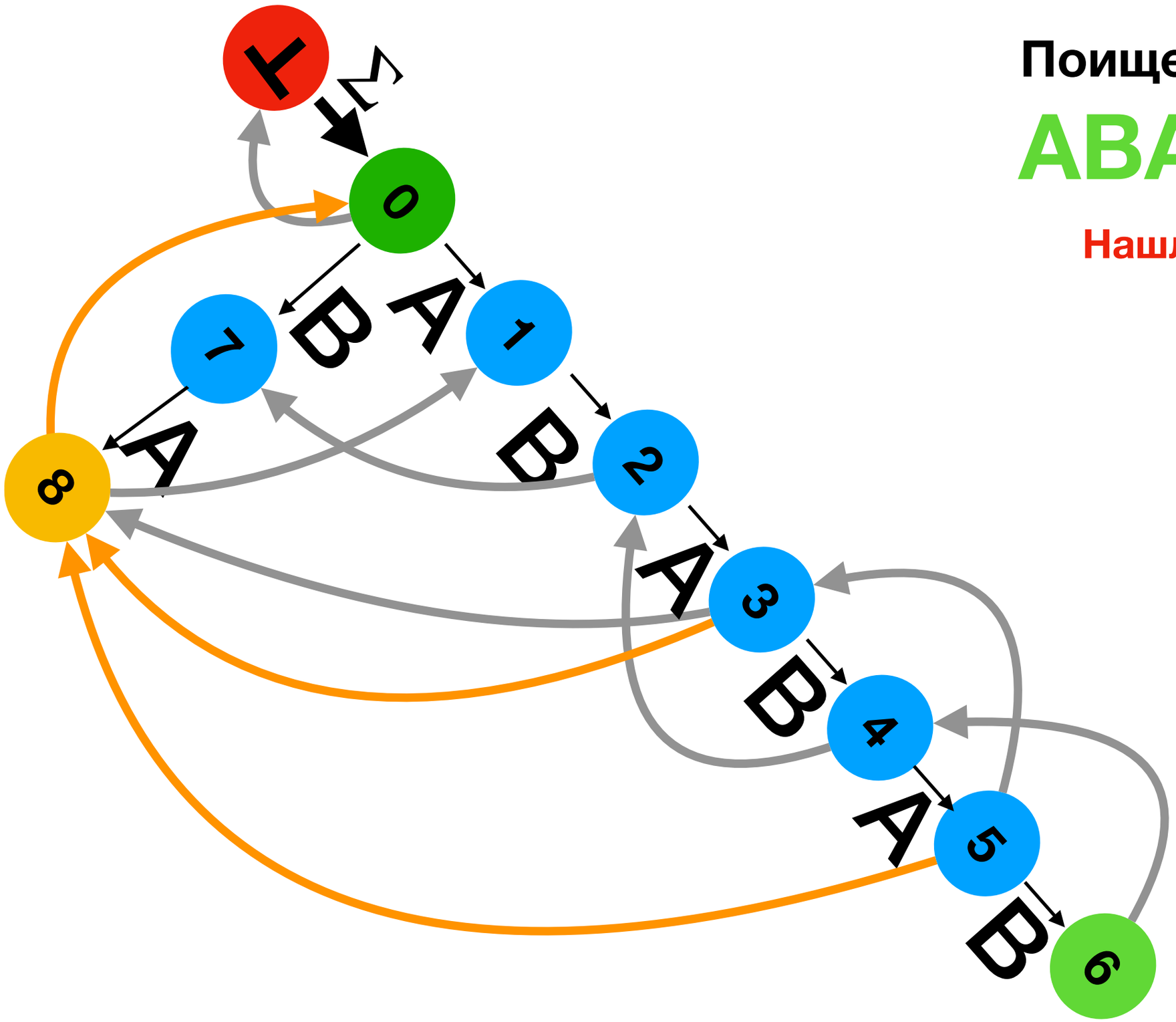
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

Подсчитаем суффиксные ссылки

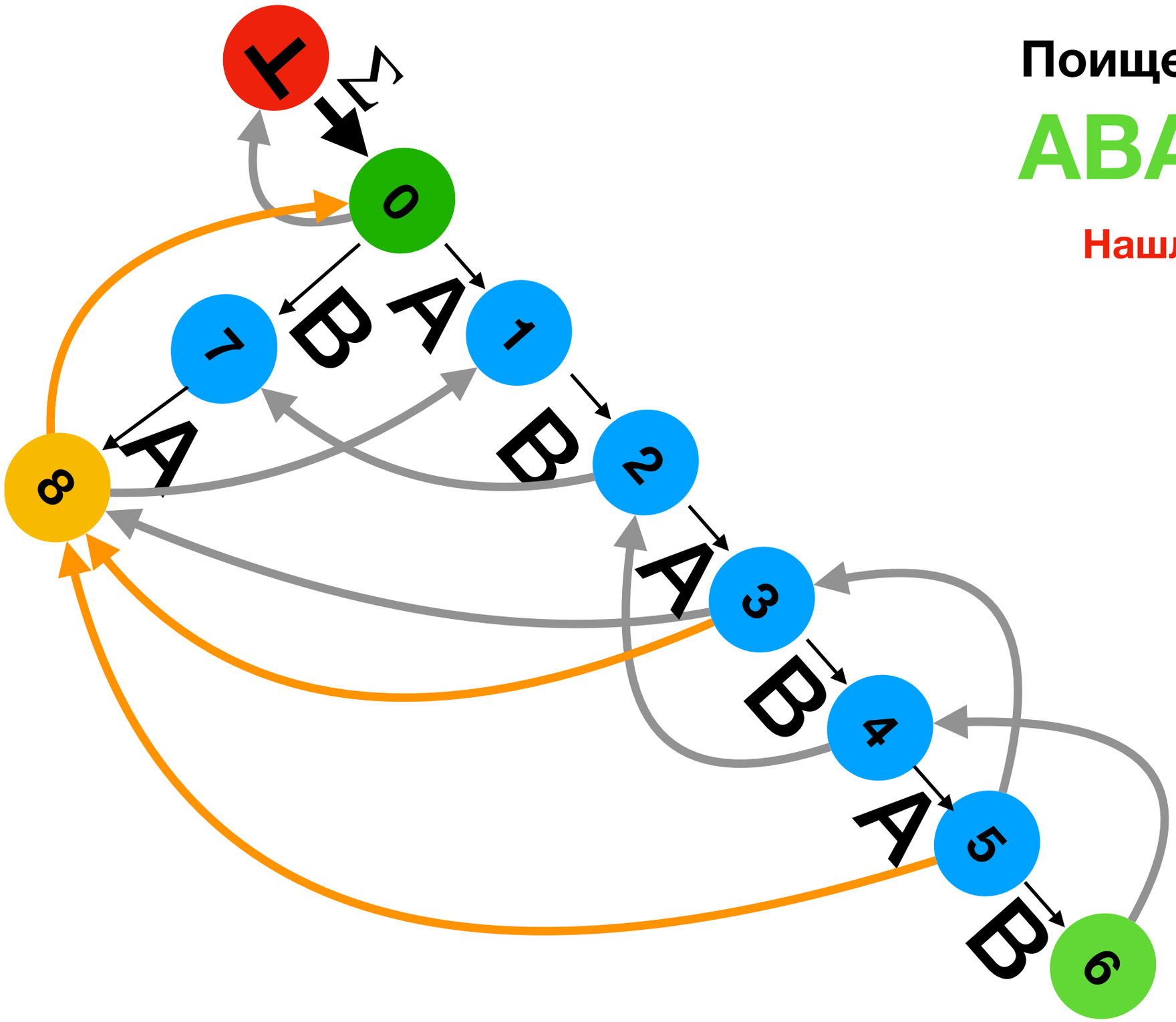


Поищем в строке

ABABABB

Нашли ABABAB

Подсчитаем суффиксные ссылки

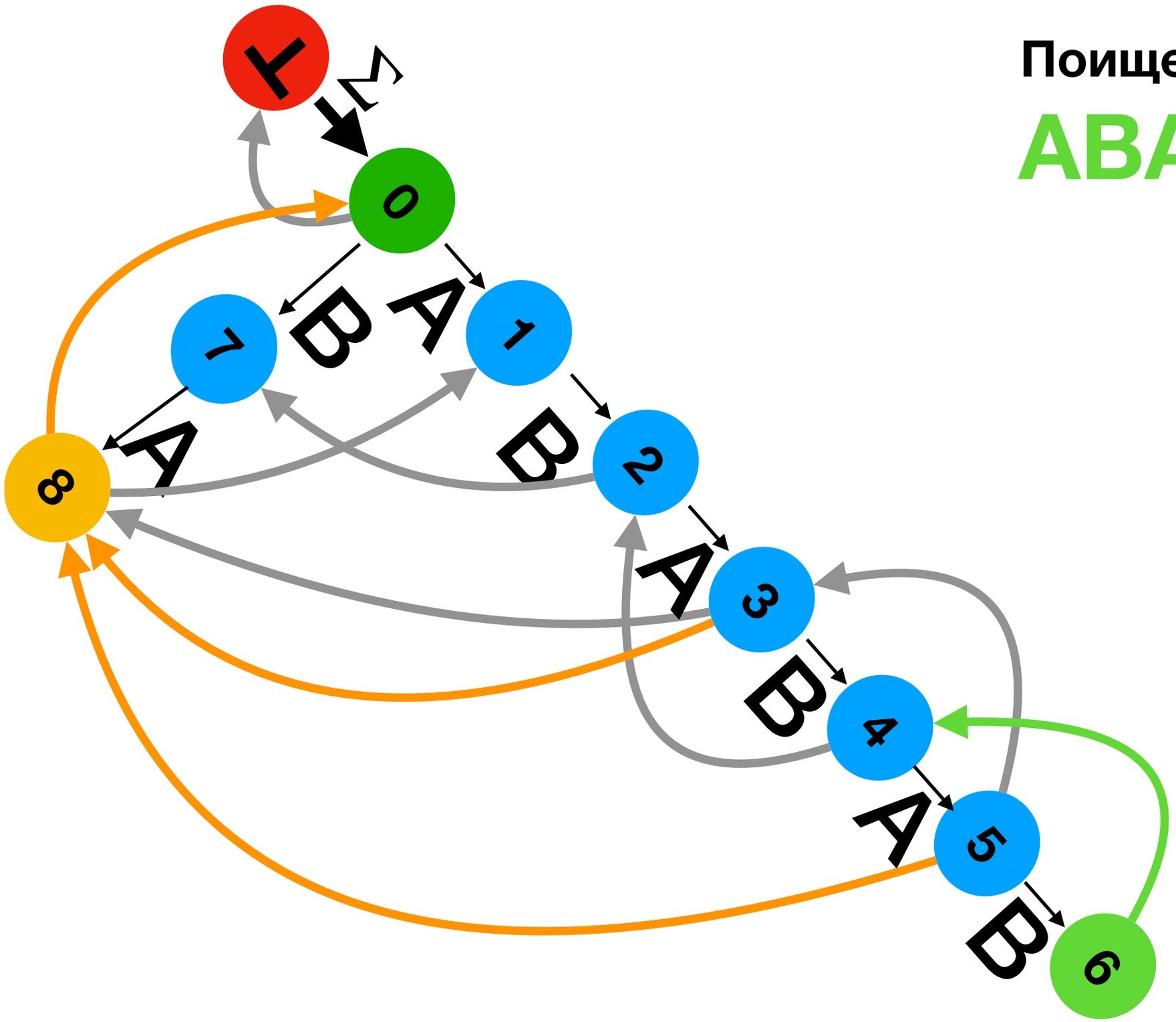


Поищем в строке

ABABABB

Нашли ABABAB

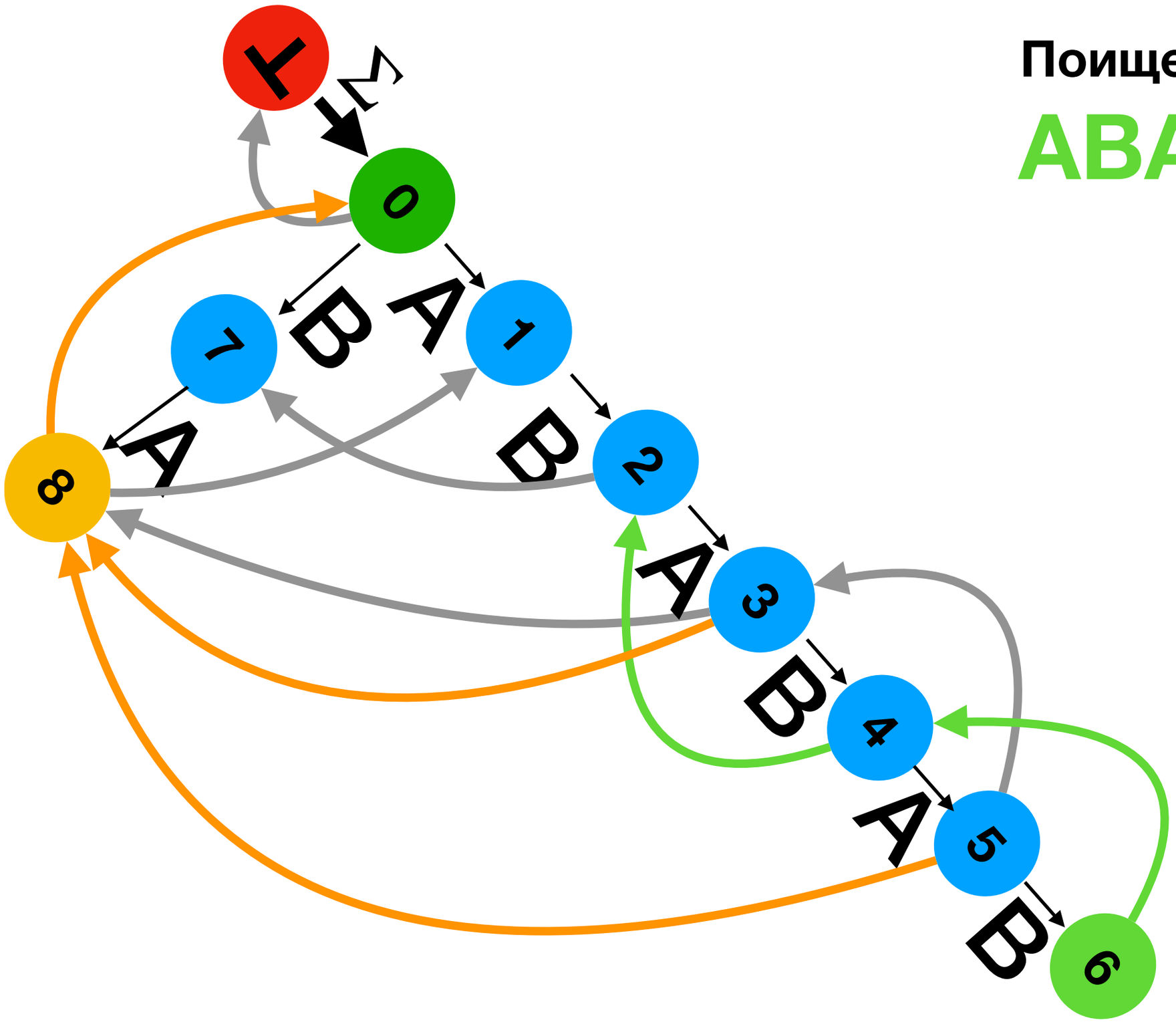
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

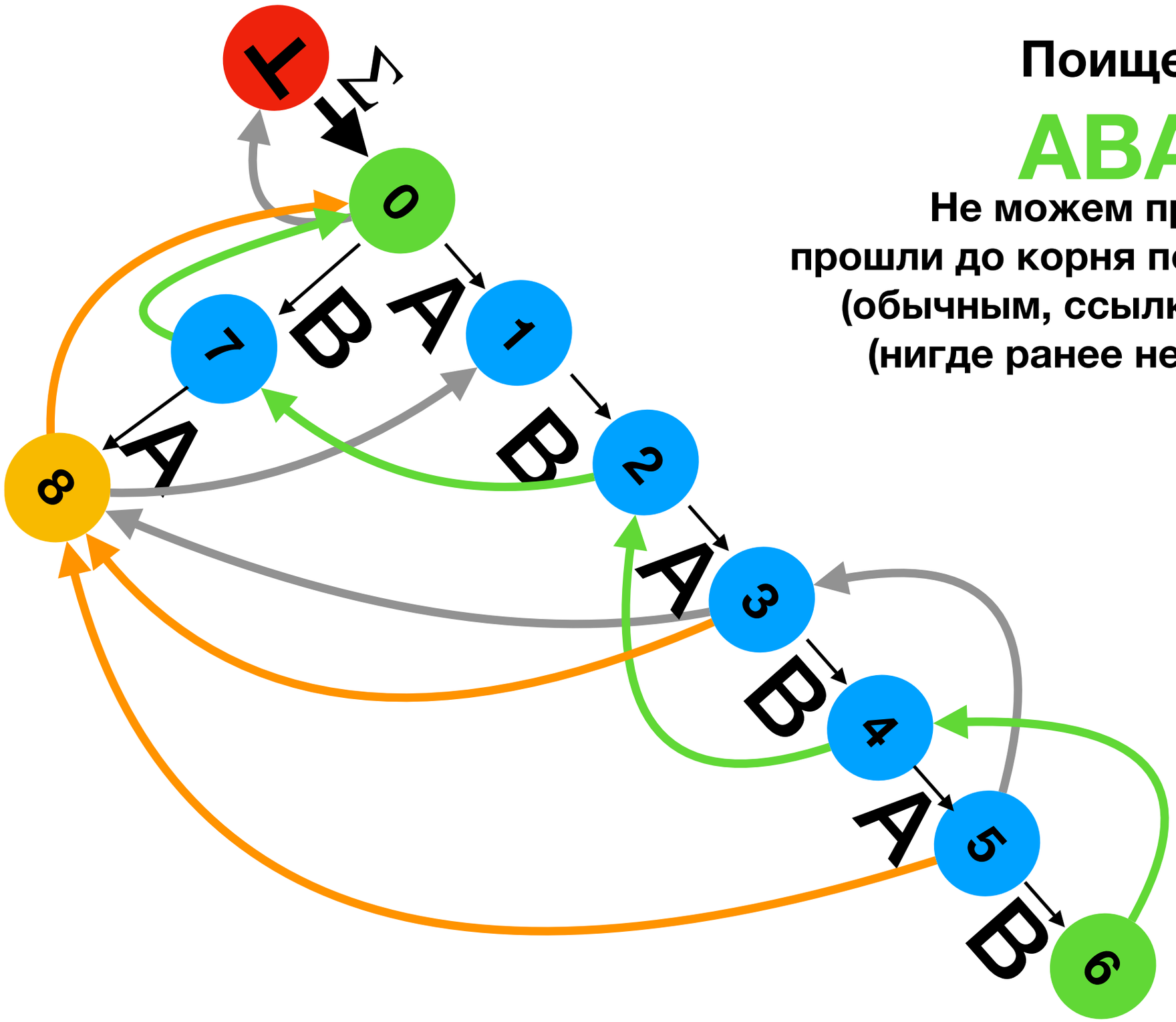
Подсчитаем суффиксные ссылки



Поищем в строке

ABABABB

Подсчитаем суффиксные ссылки

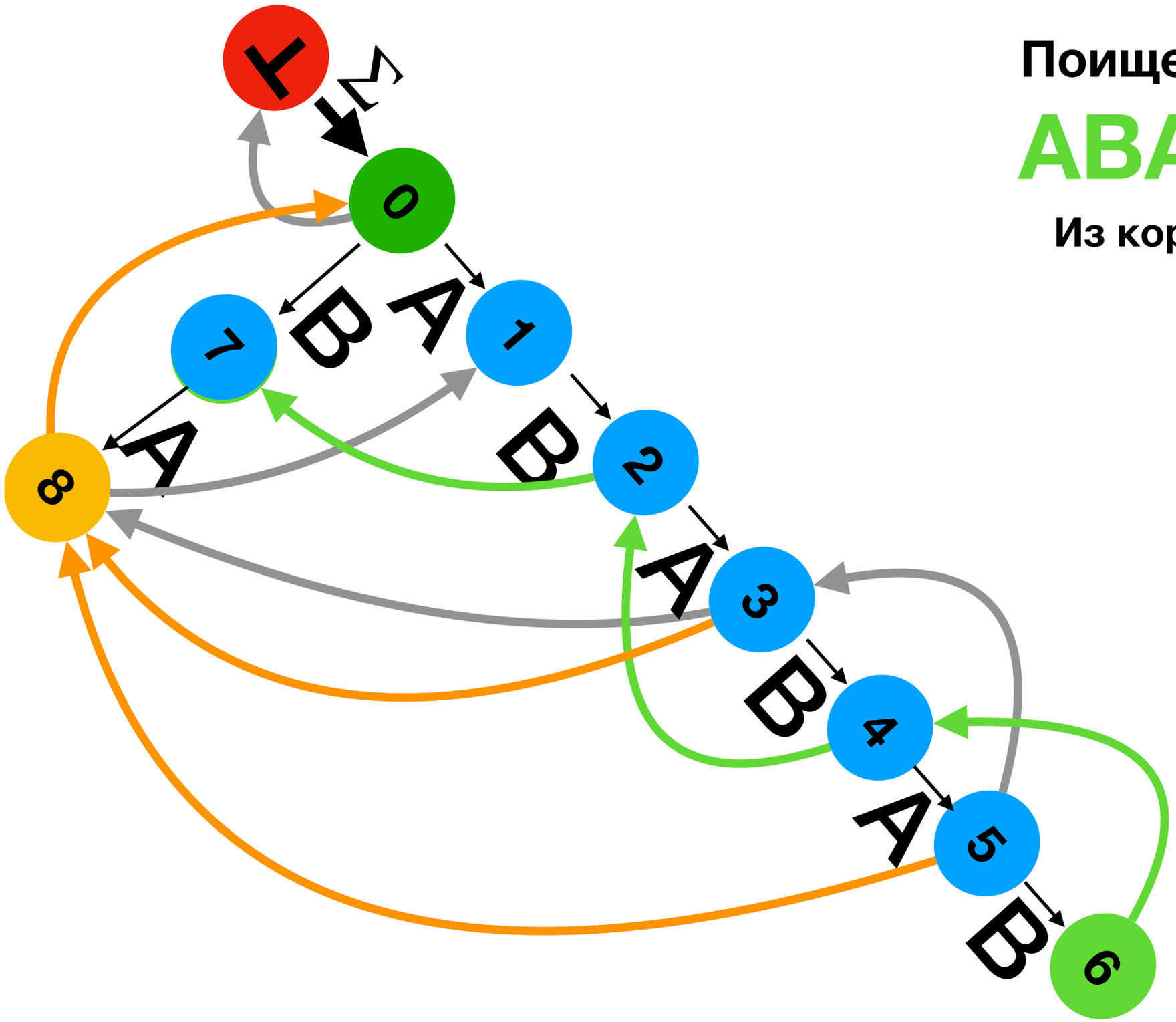


Поищем в строке

ABABABB

Не можем прочитать символ,
прошли до корня по суффиксным ссылкам
(обычным, ссылку до корня показали)
(нигде ранее не смогли прочесть B)

Подсчитаем суффиксные ссылки



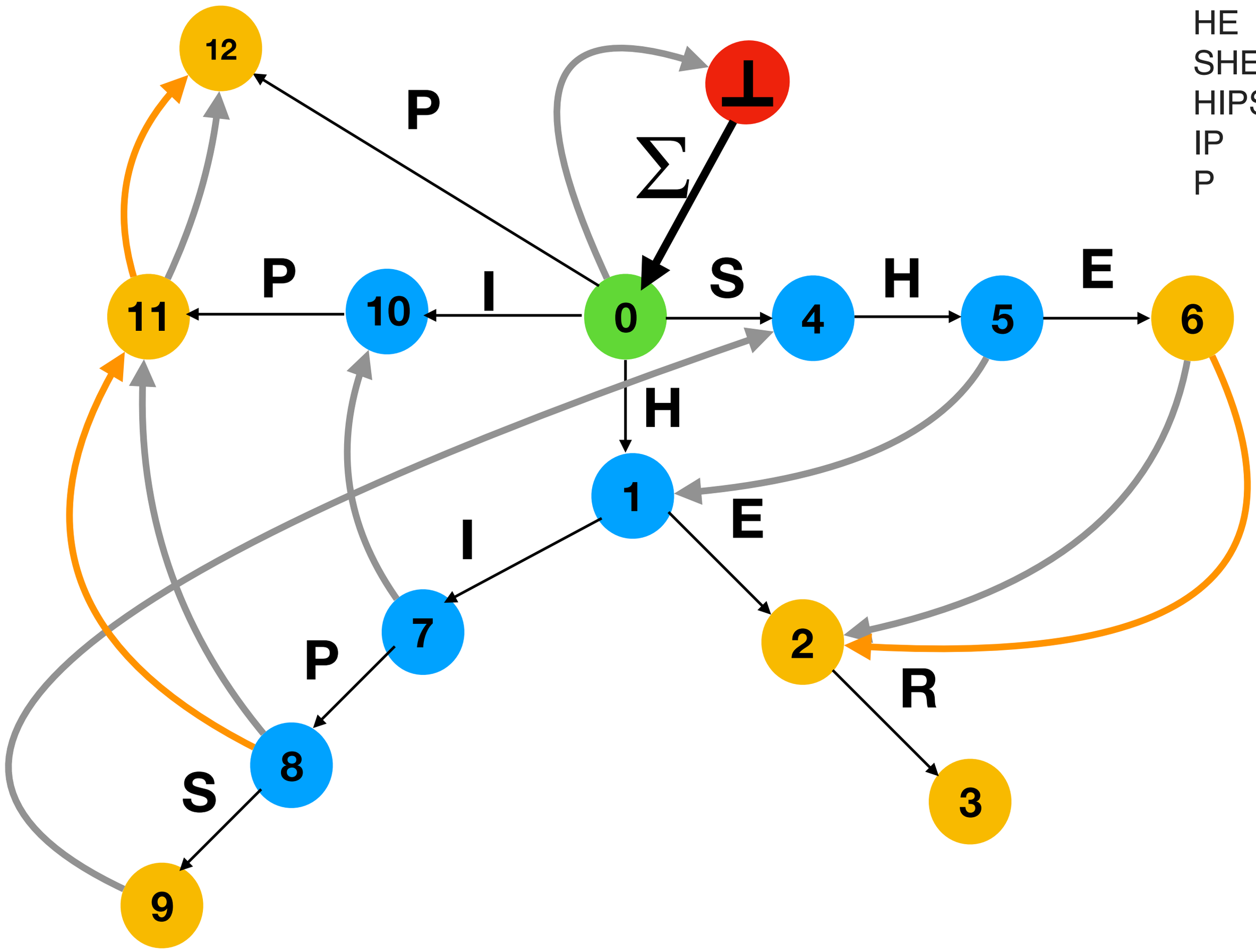
Поищем в строке

ABABABB

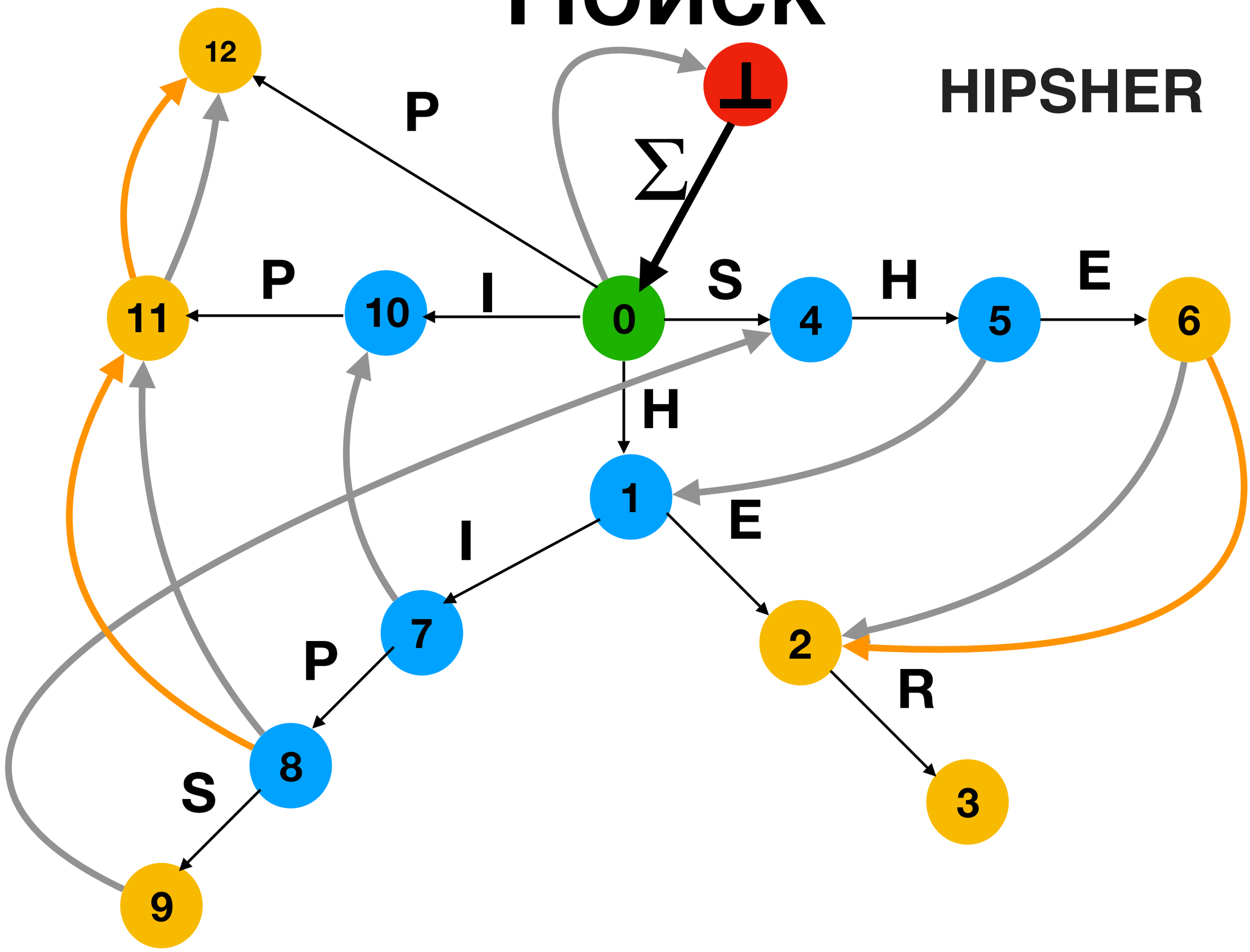
Из корня прочли

Строим для паттернов:

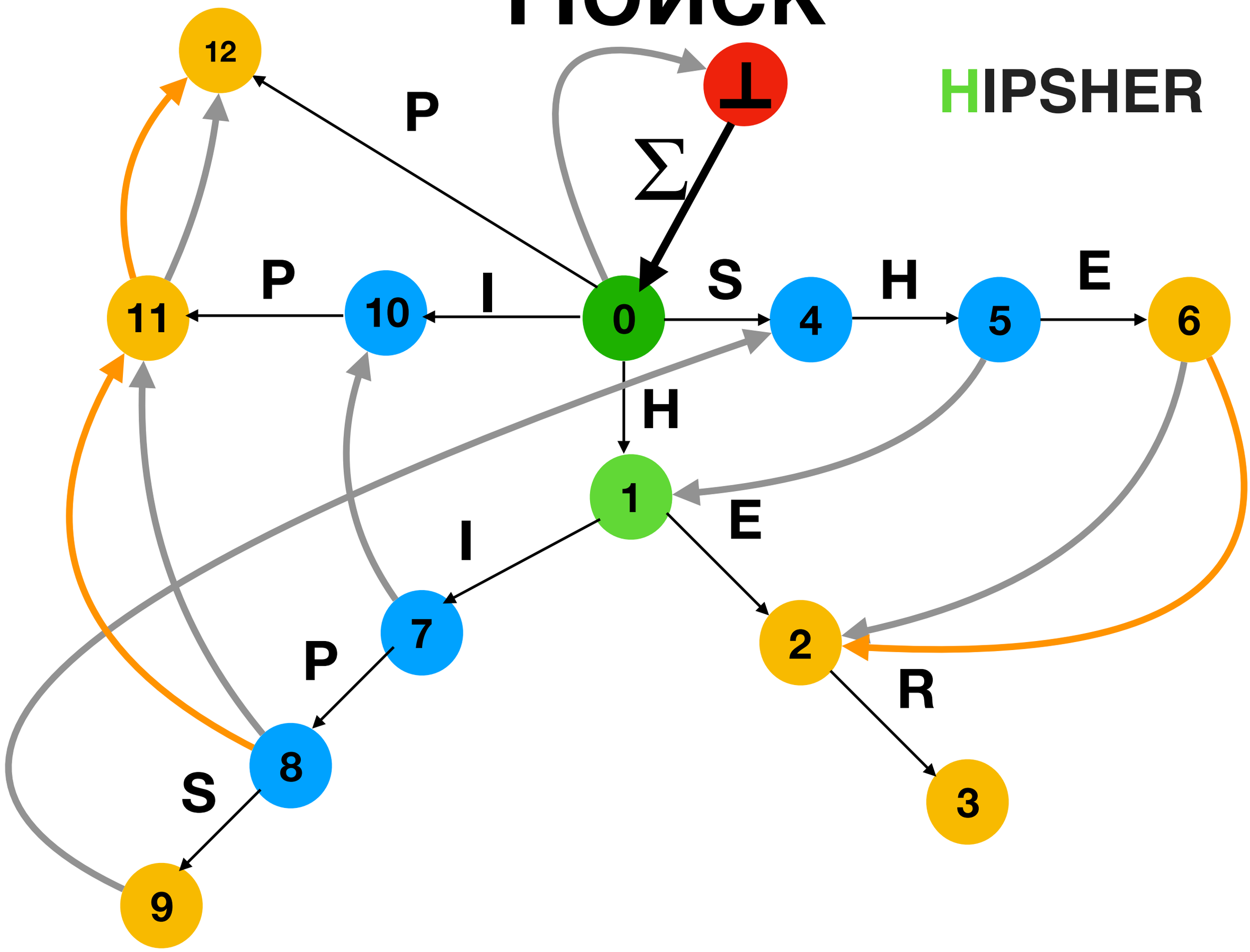
HER
HE
SHE
HIPS
IP
P



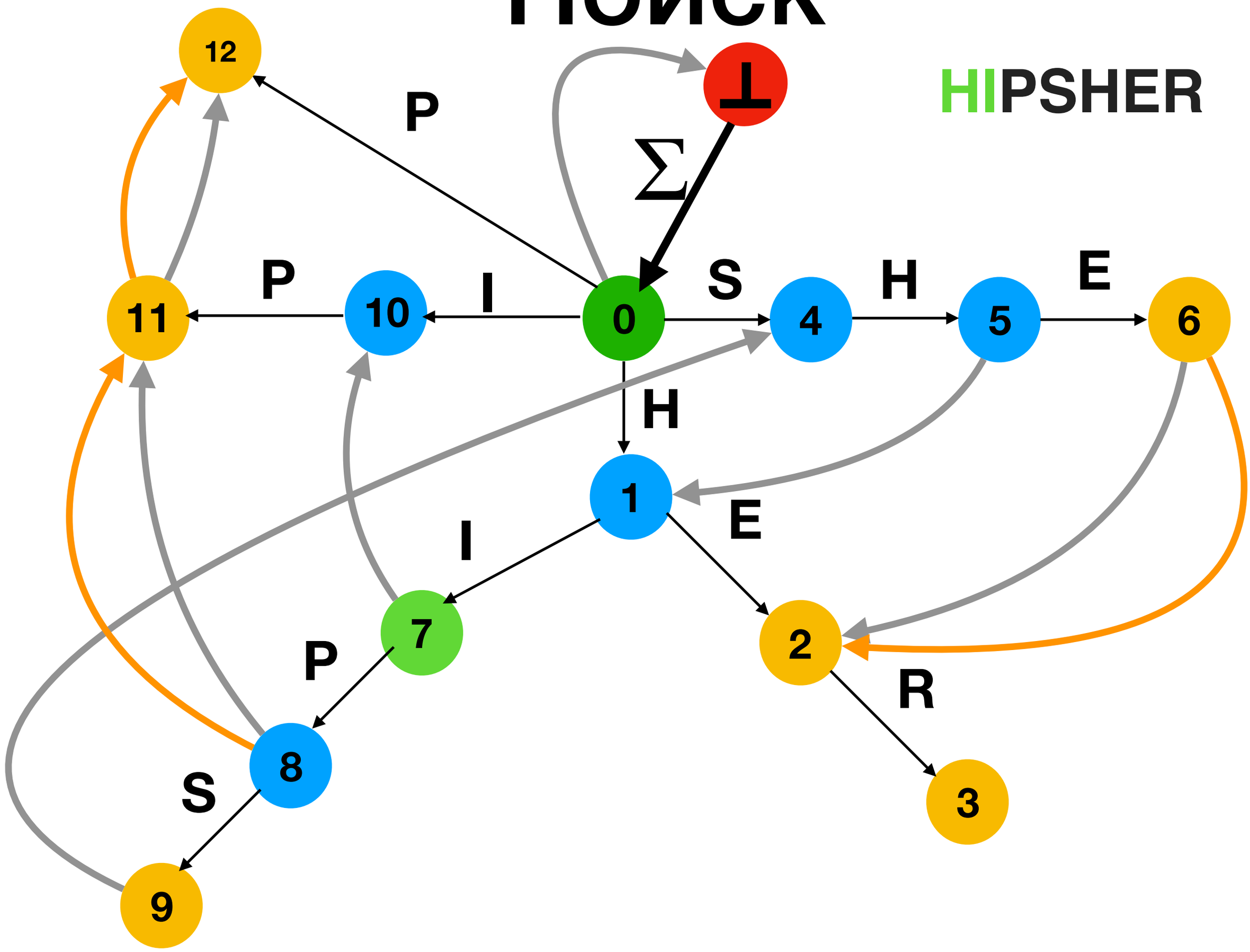
Поиск



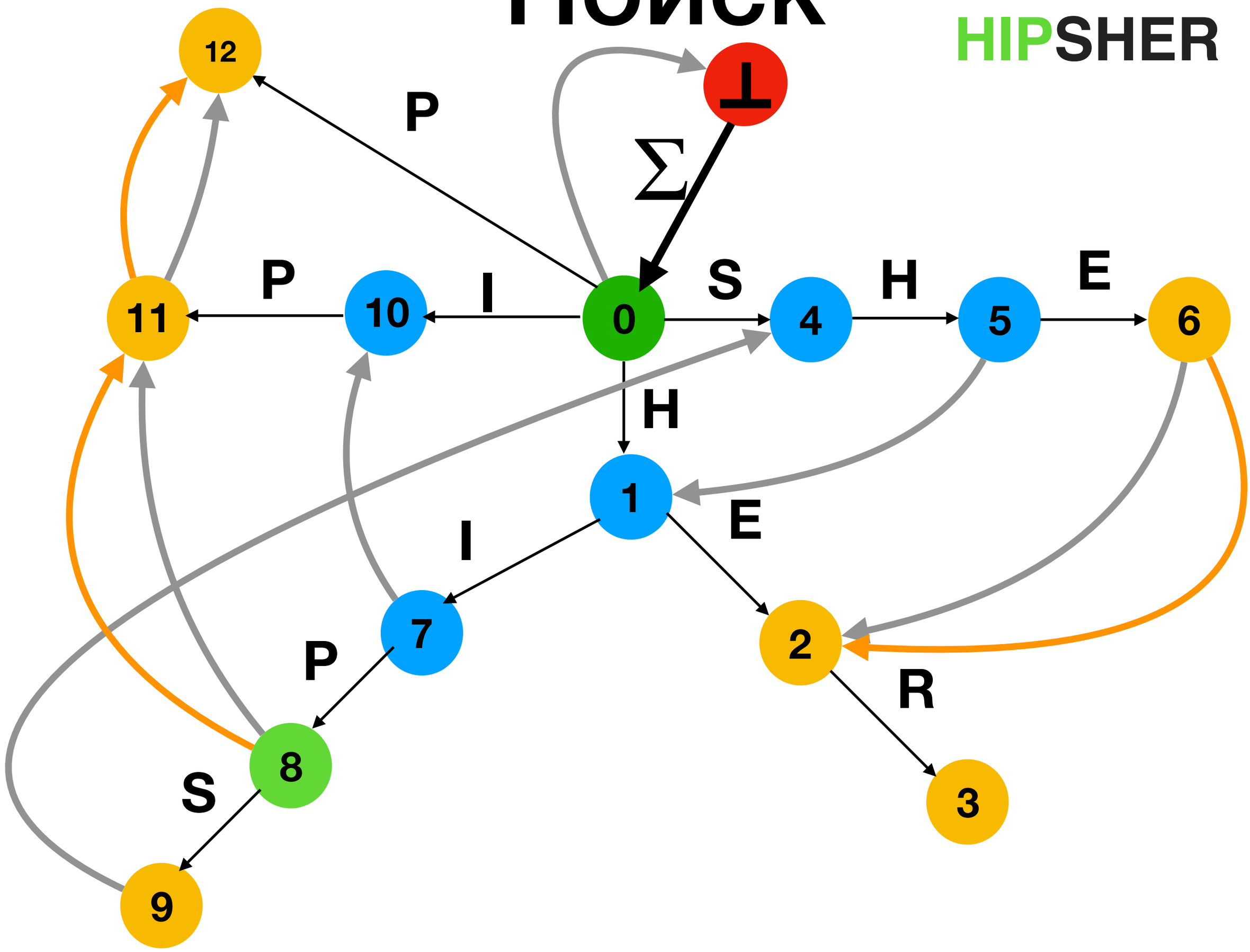
Поиск



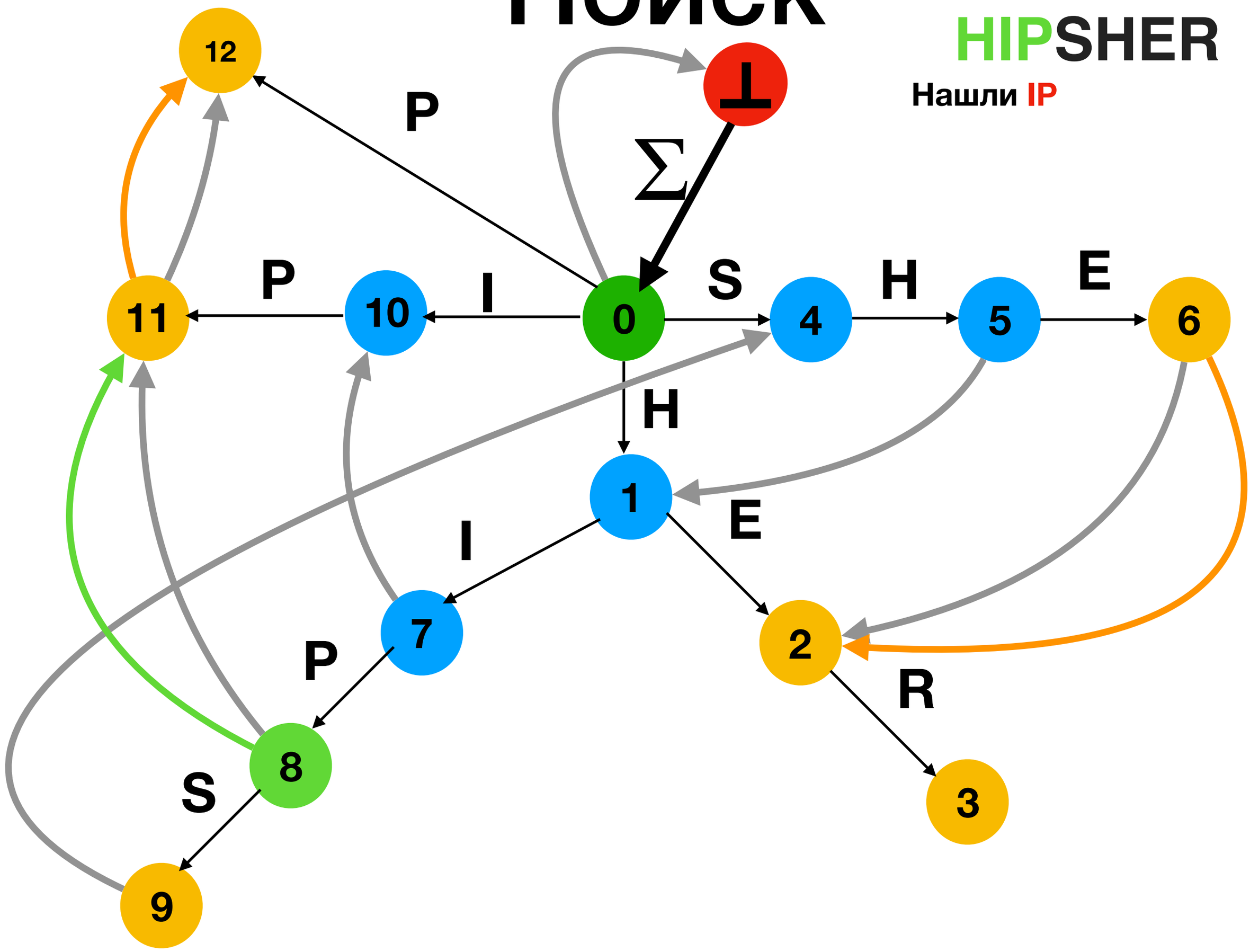
Поиск



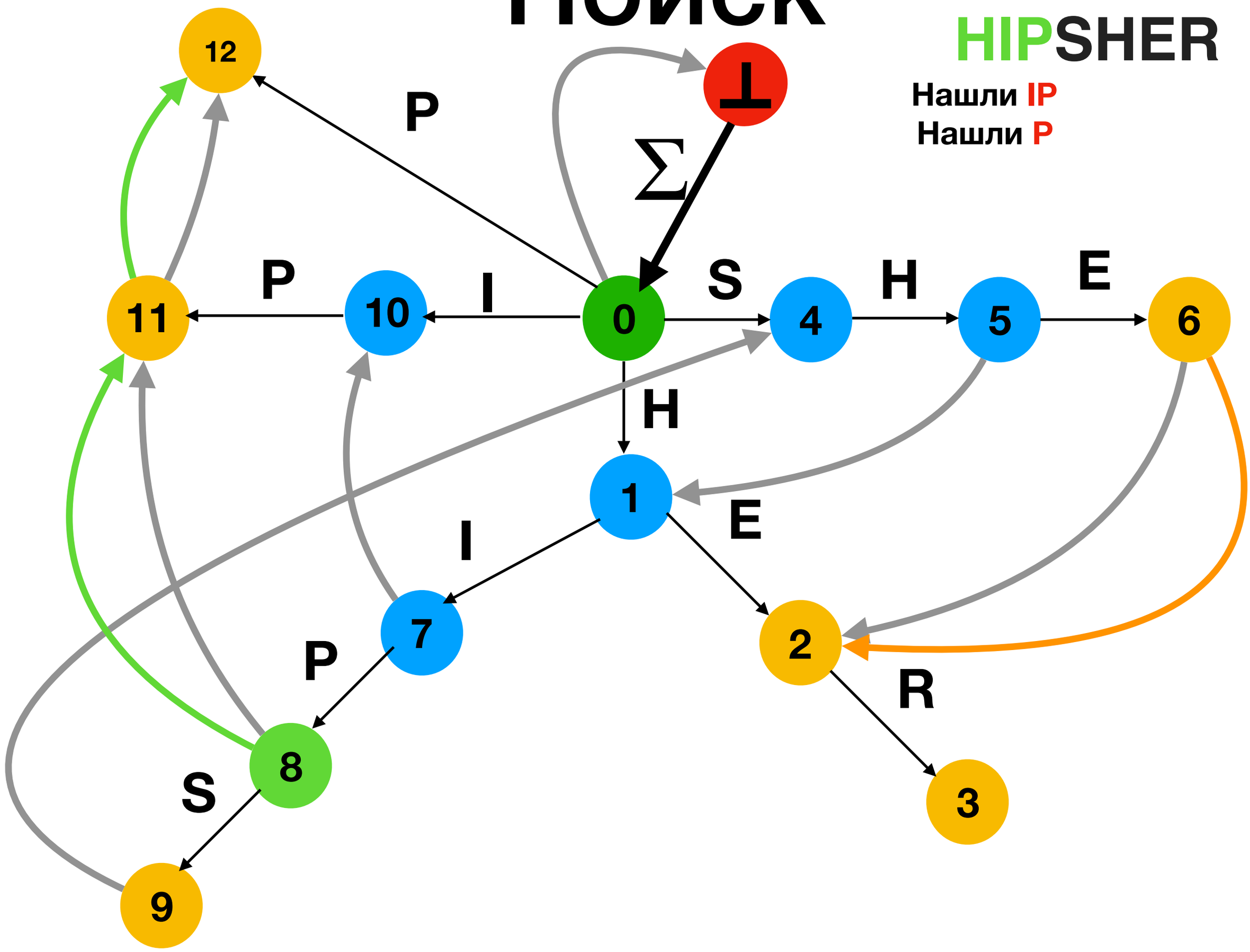
Поиск



Поиск



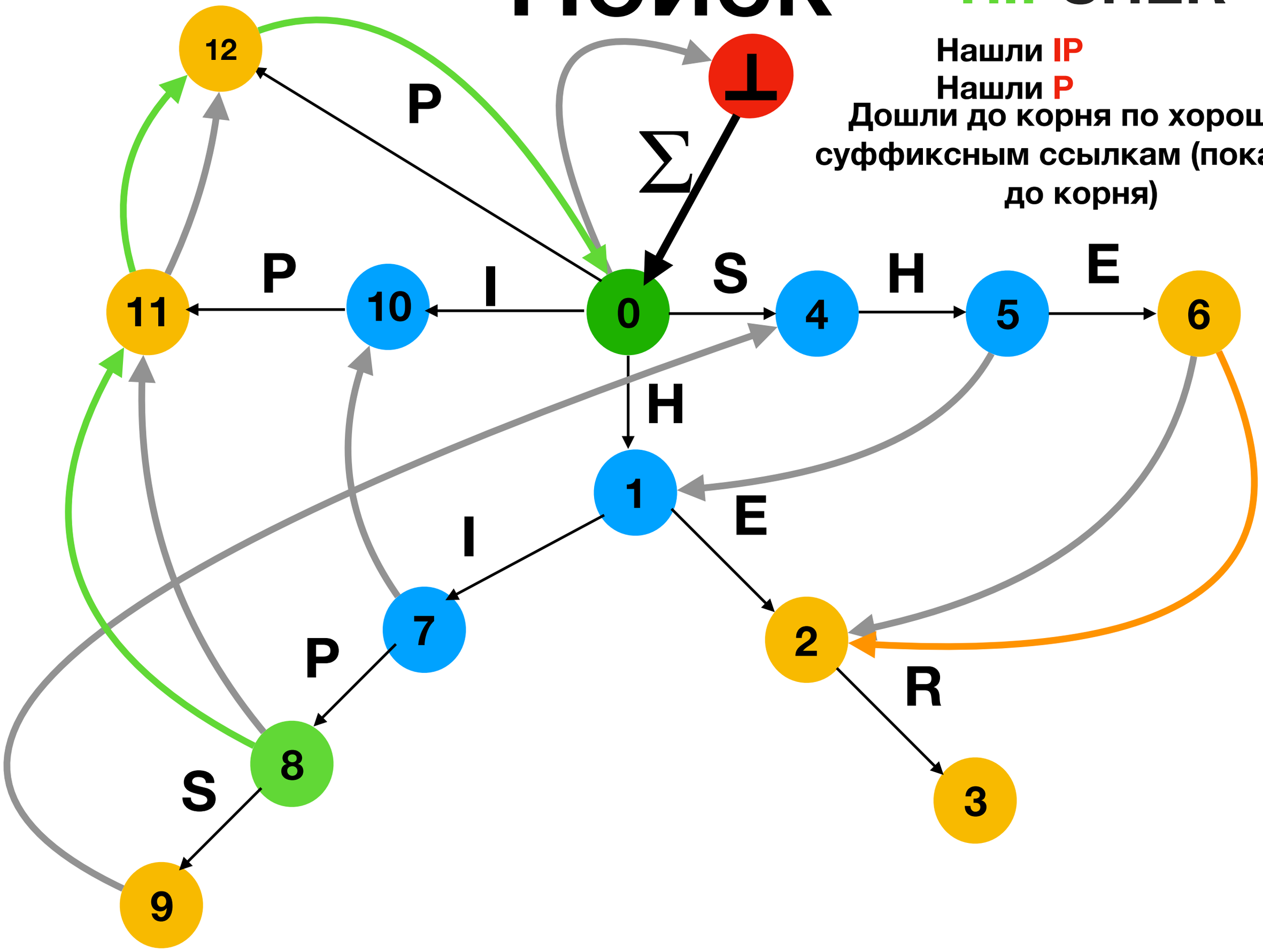
Поиск



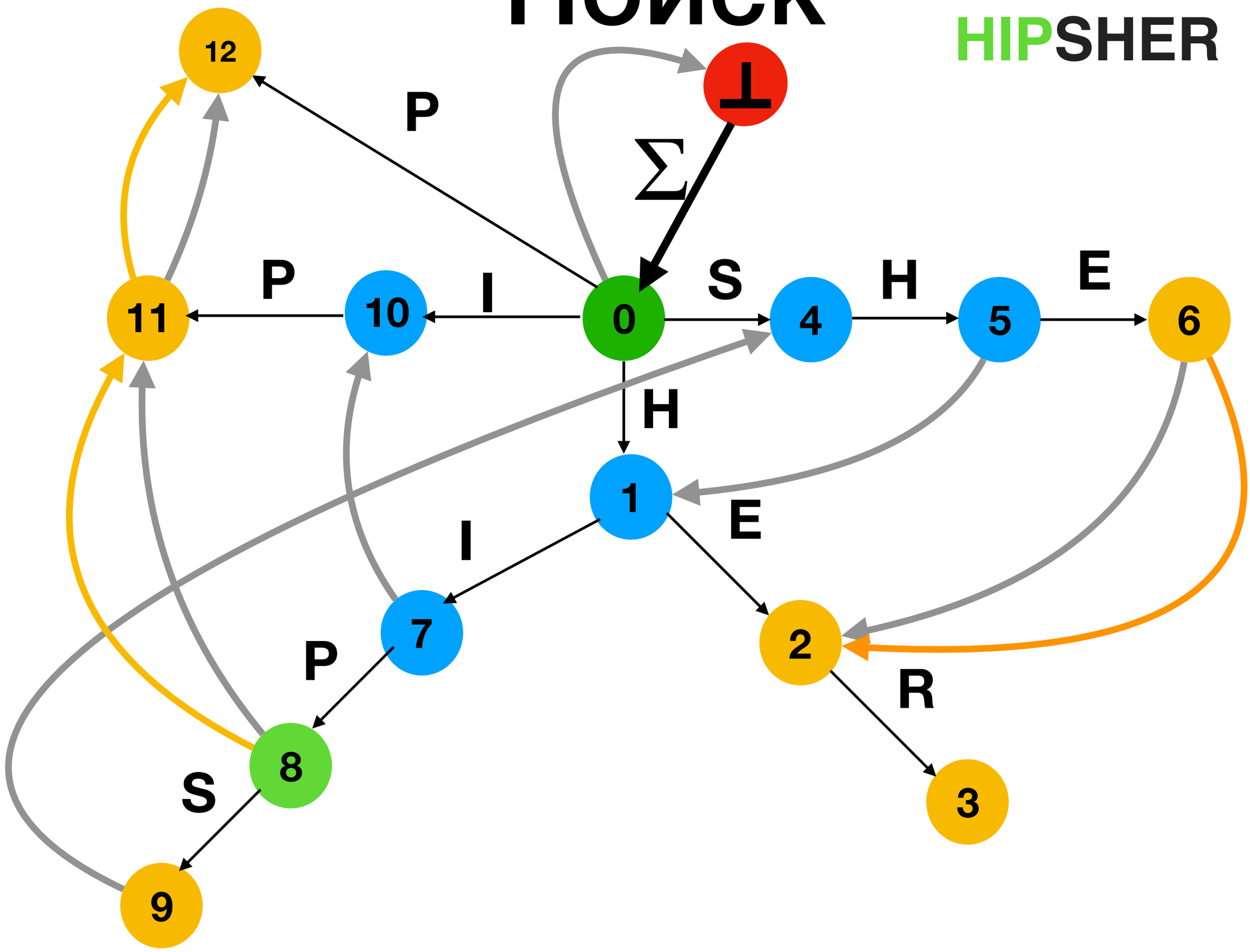
Поиск

HIPSHER

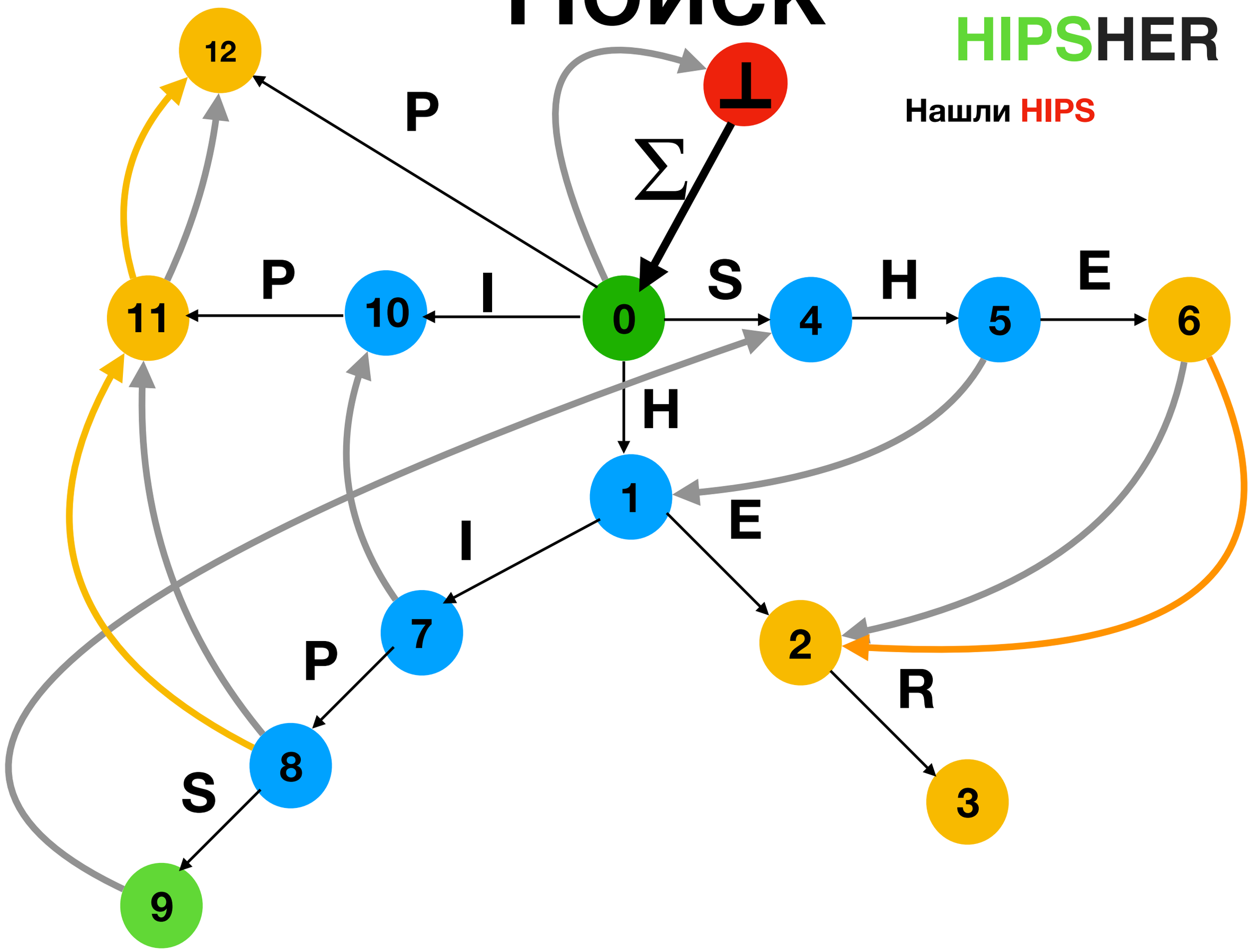
Нашли IP
Нашли P
Дошли до корня по хорошим суффиксным ссылкам (показали до корня)



Поиск

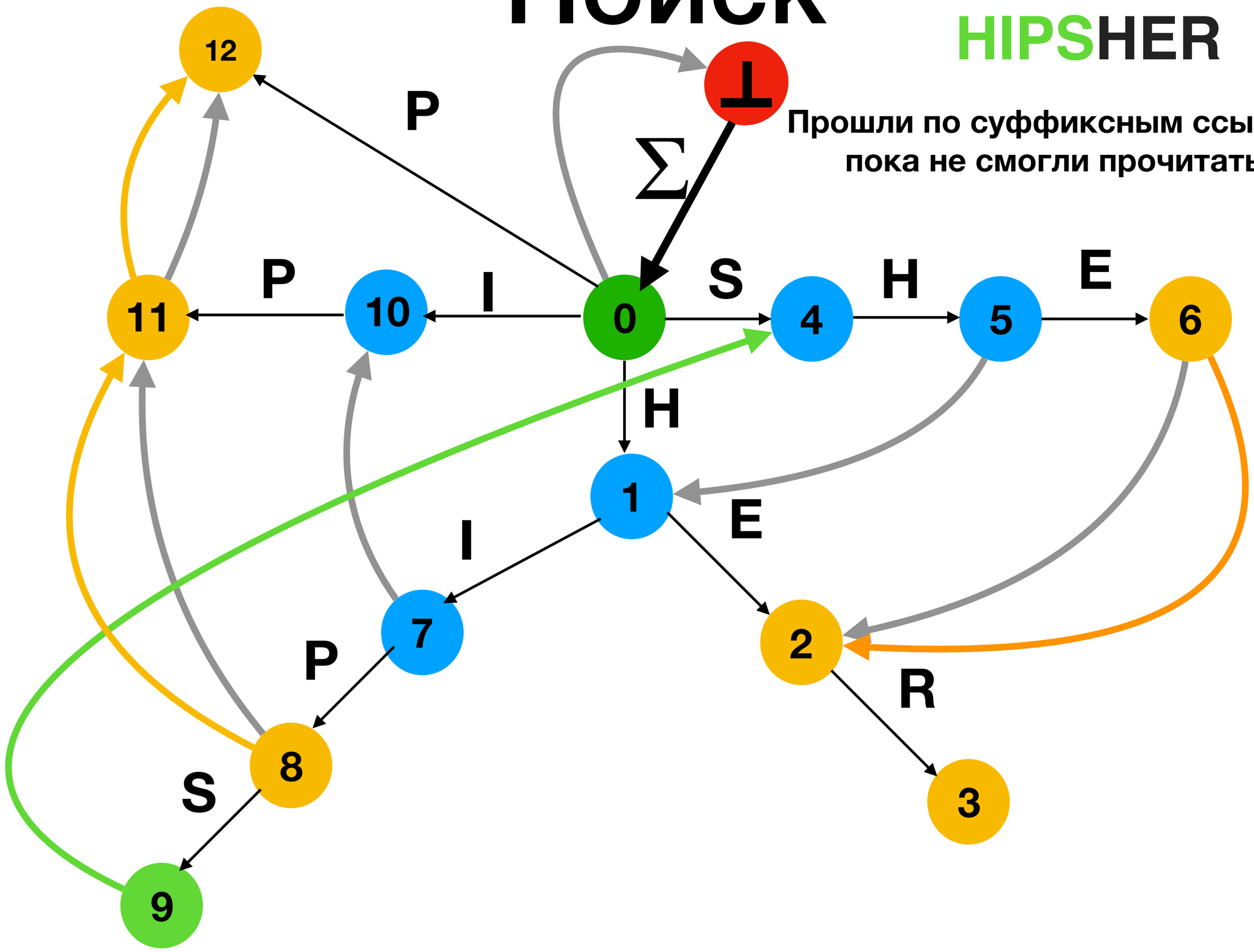


Поиск



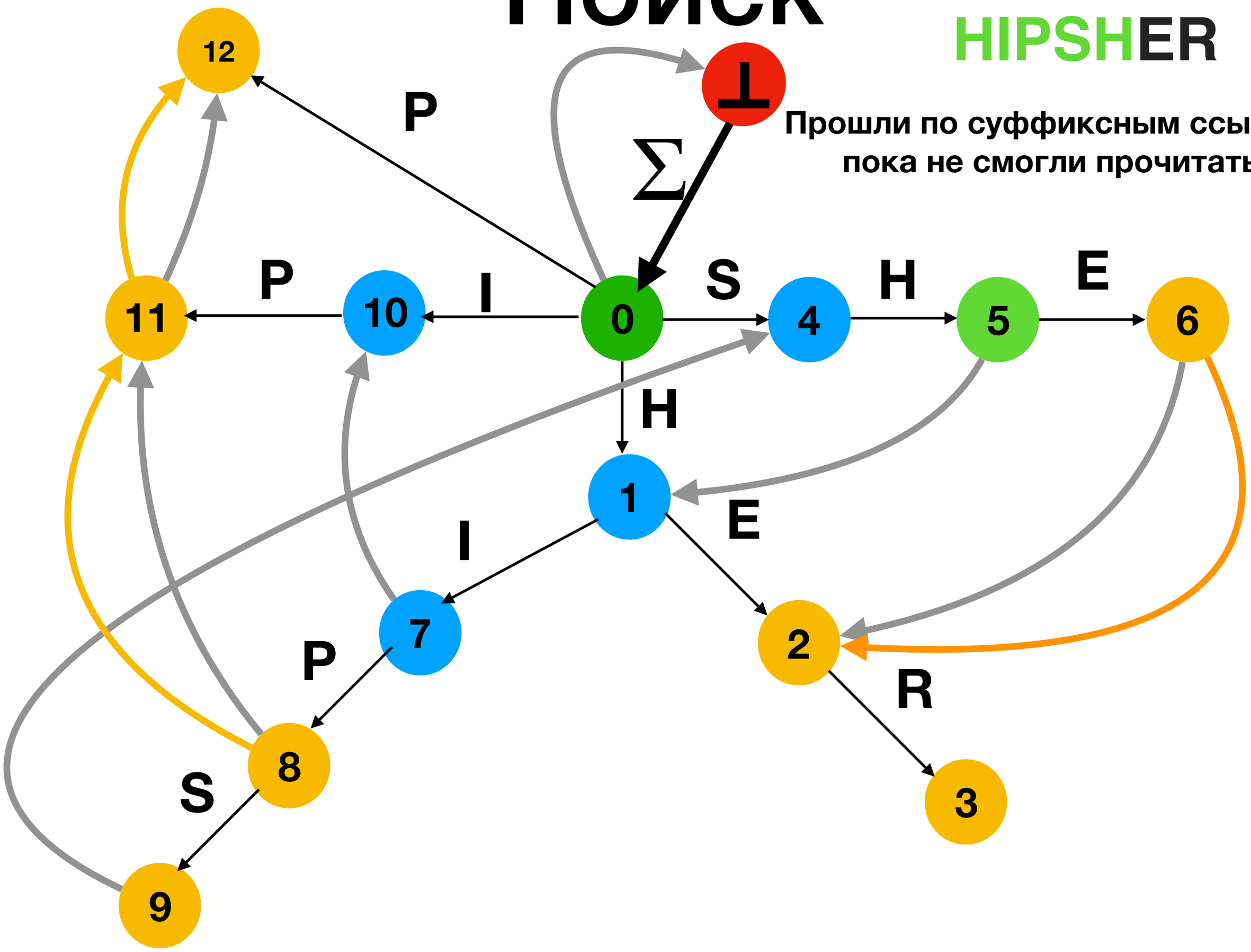
Поиск

HIPSHER

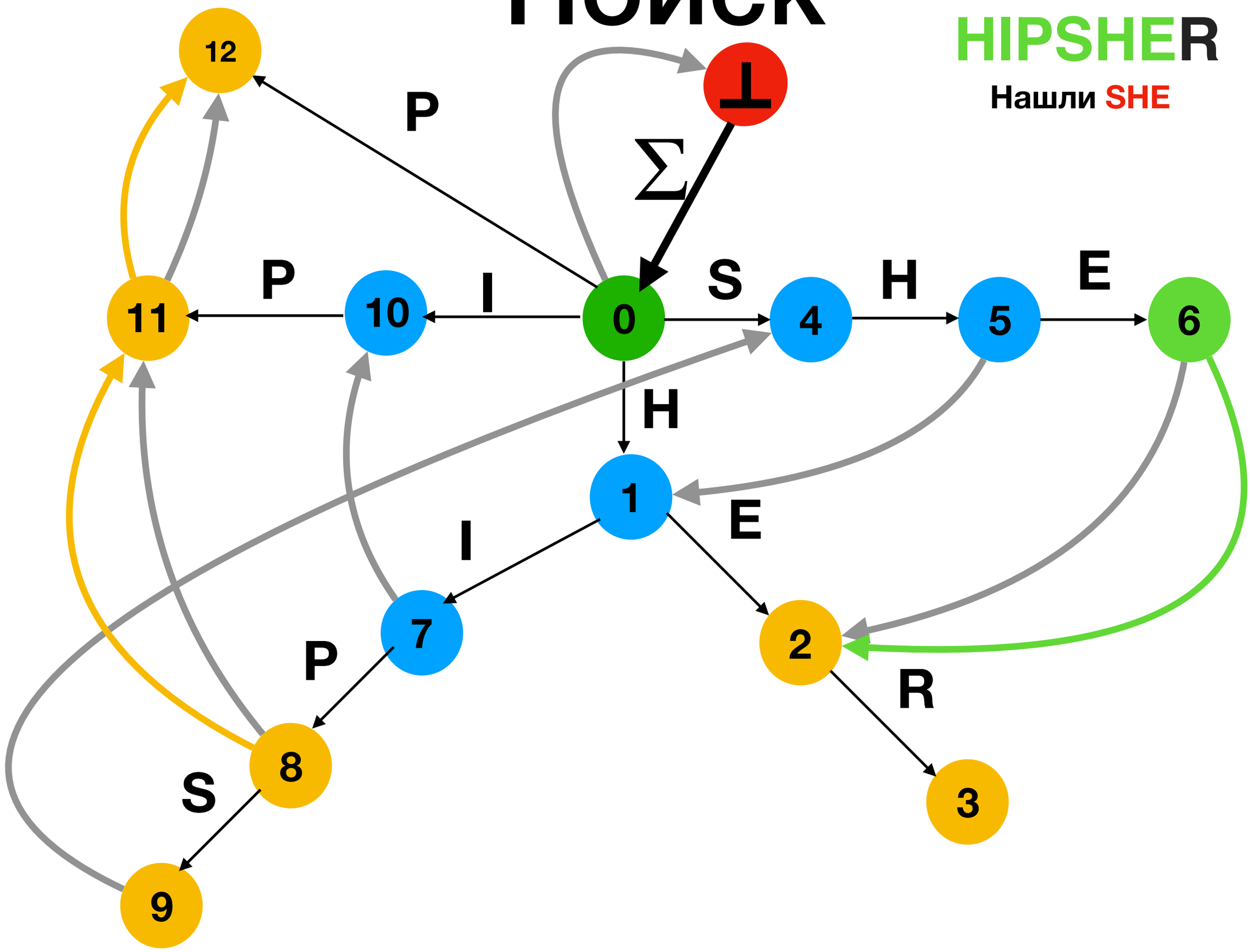


Прошли по суффиксным ссылкам, пока не смогли прочитать H

Поиск


HIPSHER
Прошли по суффиксным ссылкам,
пока не смогли прочитать H
⊥
Σ
P
P
I
S
H
E
H
E
I
P
R
S
0
1
2
3
4
5
6
7
8
9
10
11
12

Поиск

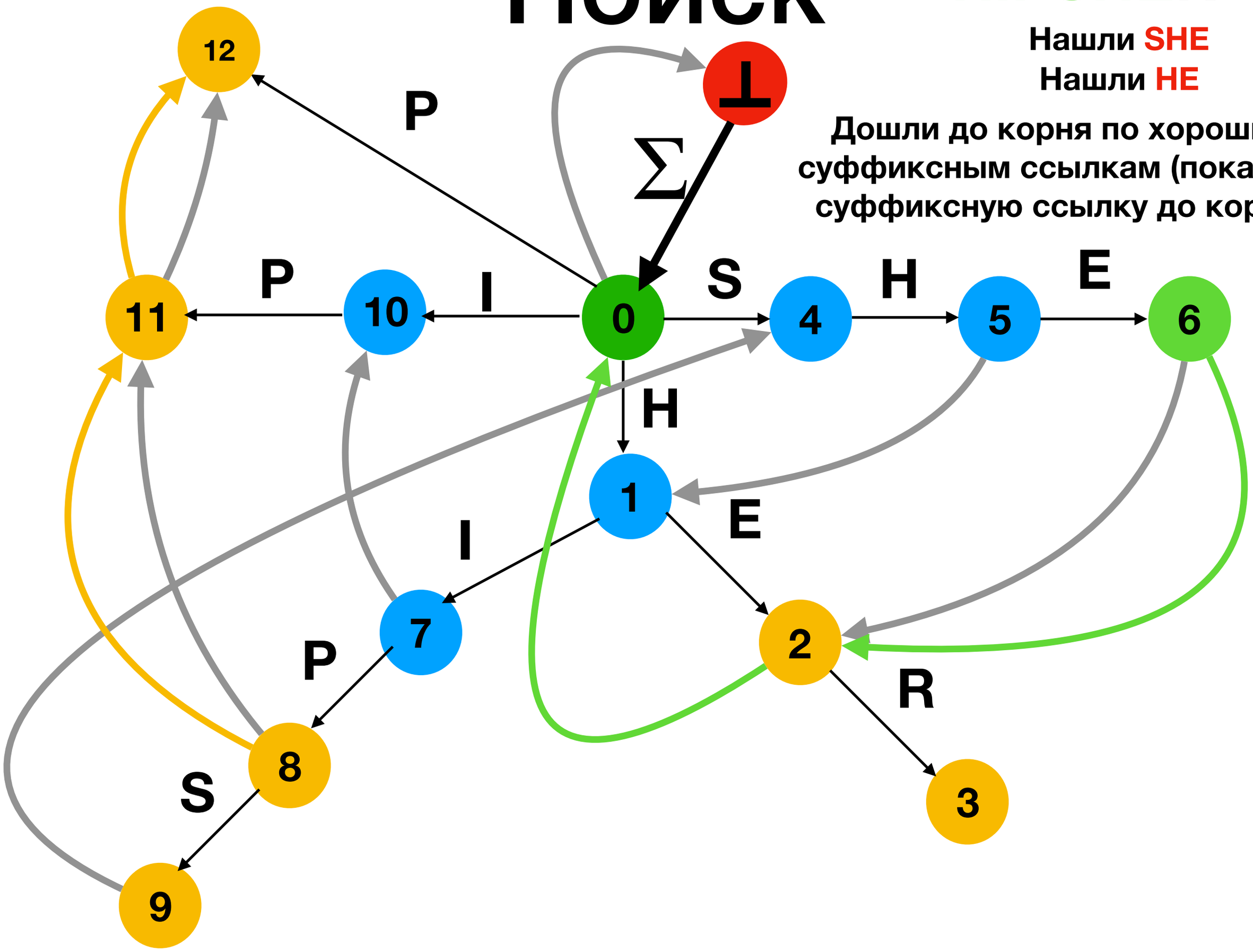


Поиск

HIPSHER

Нашли SHE
Нашли HE

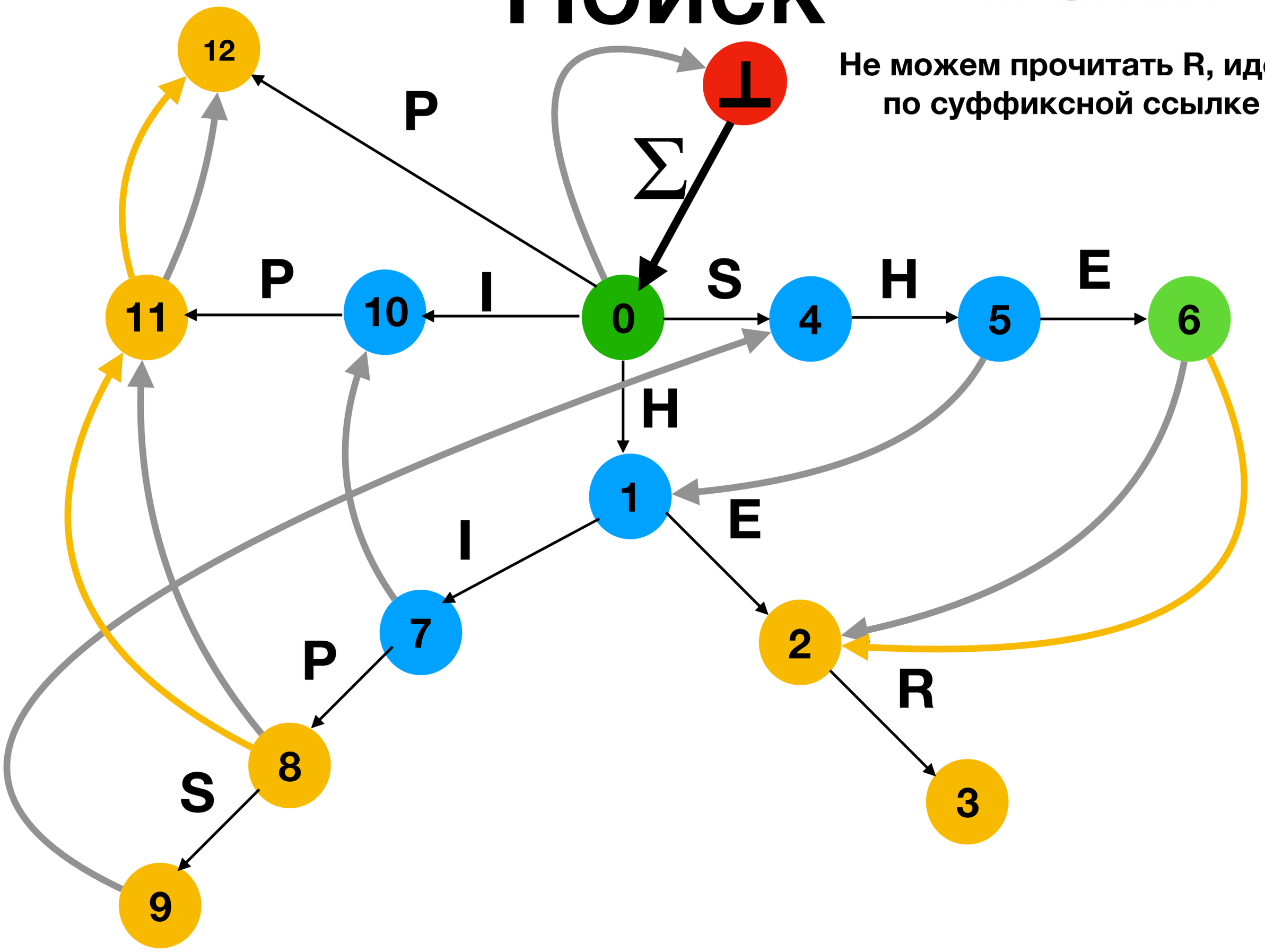
Дошли до корня по хорошим суффиксным ссылкам (показали суффиксную ссылку до корня)



Поиск

HIPSHER

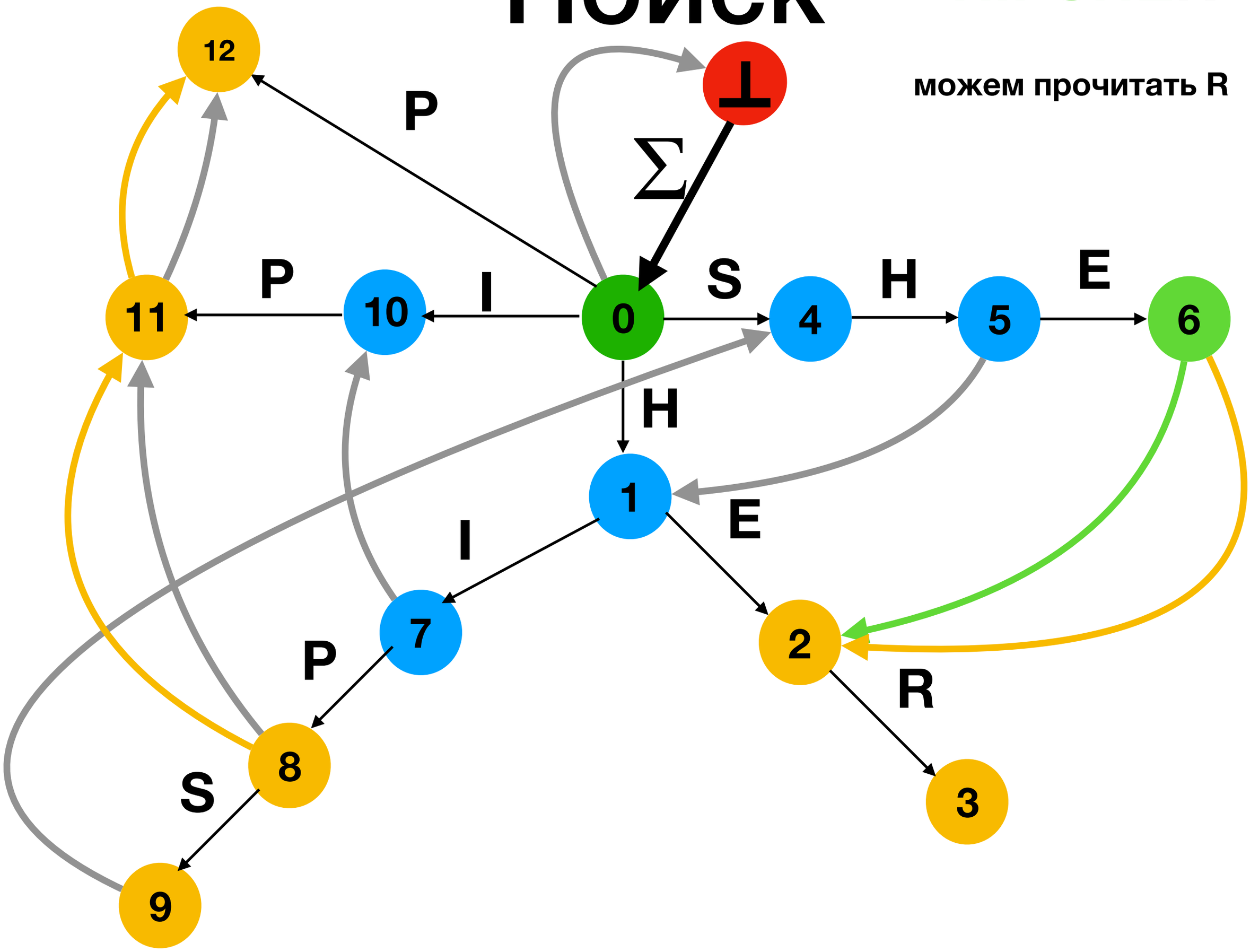
Не можем прочитать R, идем по суффиксной ссылке



Поиск

HIPSHER

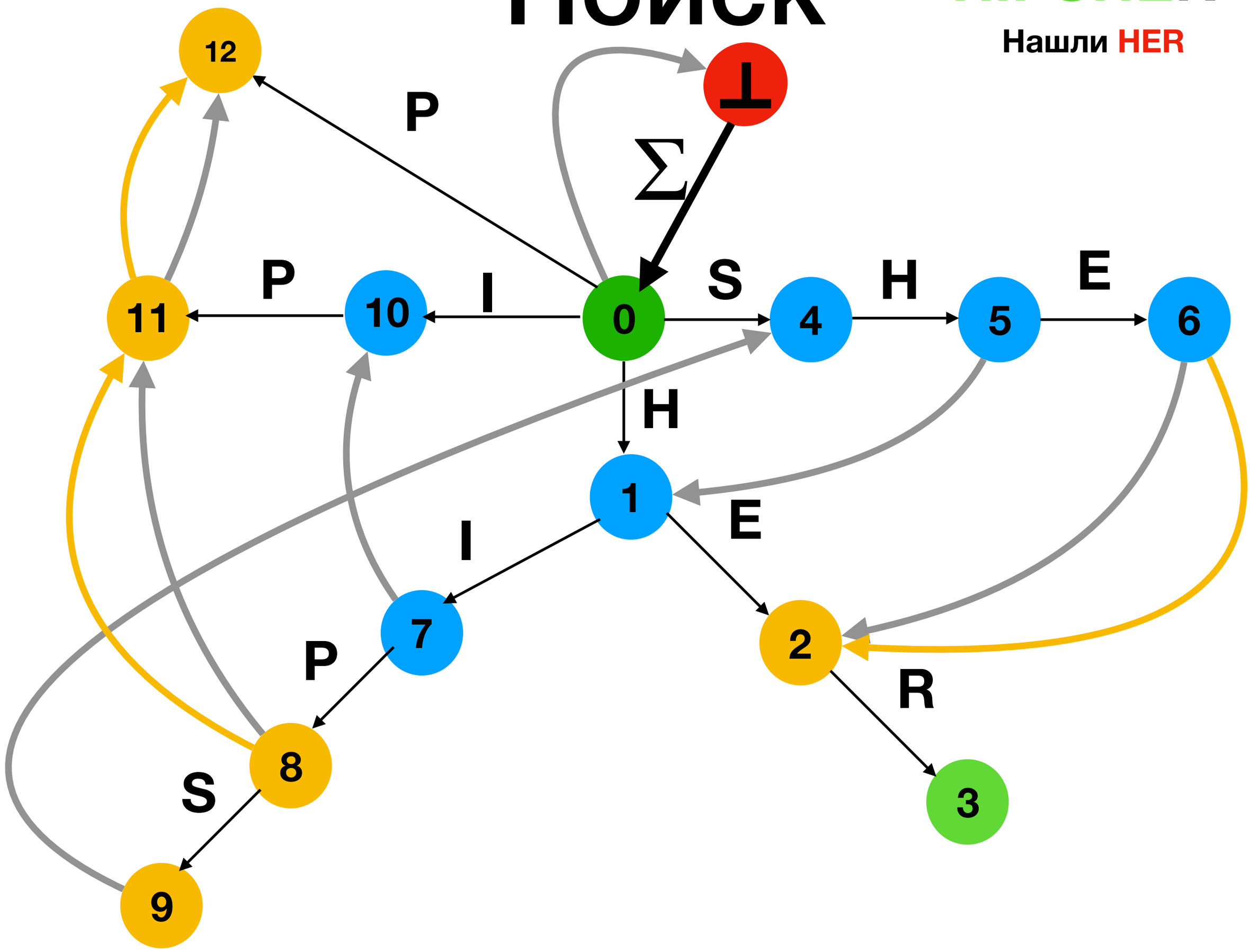
можем прочитать R



Поиск

HIPSHER

Нашли HER



Ахо-Корасик

Сколько времени занимает?

Время работы на тексте длины N состоит из времени прохождения по основным переходам, времени прохождения суффиксы ссылок и времени хождения по хорошим суффиксным ссылкам

Ахо-Корасик

Сколько времени занимает проход по тексту?

Время работы на тексте длины N состоит из времени прохождения по основным переходам, времени прохождения суффиксы ссылок и времени хождения по хорошим суффиксным ссылкам

На каждый символ выполняется один проход по основному переходу - значит, будет затрачено O(N) времени. Мы предполагаем, что куда идти по данному символу мы узнаем за O(1) (переходы хранятся в хэш-таблице)

Сколько тратим на проходы по суффиксным ссылкам?

Ахо-Корасик

Сколько времени занимает проход по тексту?

Время работы на тексте длины N состоит из времени прохождения по основным переходам, времени прохождения суффиксы ссылок и времени хождения по хорошим суффиксным ссылкам

На каждый символ выполняется один проход по основному переходу - значит, будет затрачено O(N) времени. Мы предполагаем, что куда идти по данному символу мы узнаем за O(1) (переходы хранятся в хэш-таблице)

Сколько тратим на проходы по суффиксным ссылкам?

Воспользуемся методом потенциалов - потенциалом будет высота вершины, в которой мы находимся. Высота $\perp$ - 0, высота корня - 1

Ахо-Корасик

Сколько времени занимает проход по тексту?

Время работы на тексте длины N состоит из времени прохождения по основным переходам, времени прохождения суффиксы ссылок и времени хождения по хорошим суффиксным ссылкам

На каждый символ выполняется один проход по основному переходу - значит, будет затрачено O(N) времени. Мы предполагаем, что куда идти по данному символу мы узнаем за O(1) (переходы хранятся в хэш-таблице)

Сколько тратим на проходы по суффиксным ссылкам?

Воспользуемся методом потенциалов - потенциалом будет высота вершины, в которой мы находимся. Высота $\perp$ - 0, высота корня - 1

Высота вершины, где мы находимся, увеличивается только при переходе по основному переходу. Значит, она не может увеличиться больше чем на N

Ахо-Корасик

Сколько времени занимает проход по тексту?

Время работы на тексте длины N состоит из времени прохождения по основным переходам, времени прохождения суффиксы ссылок и времени хождения по хорошим суффиксным ссылкам

На каждый символ выполняется один проход по основному переходу - значит, будет затрачено O(N) времени.

Сколько тратим на проходы по суффиксным ссылкам?

Воспользуемся методом потенциалов - потенциалом будет высота вершины, в которой мы находимся. Высота $\perp$ - 0, высота корня - 1. Очевидно, высота всегда больше или равна 0

Высота вершины, где мы находимся, увеличивается только при переходе по основному переходу. Значит, она не может увеличиться больше чем на N

Высота вершины уменьшается каждый раз, когда мы проходим по суффиксной ссылке, минимум на 1. Значит, переходов по суффиксной ссылке не больше N + 1 - O(N)

Ахо-Корасик

Сколько времени занимает проход по тексту?

Время работы на тексте длины N состоит из времени прохождения по основным переходам, времени прохождения суффиксы ссылок и времени хождения по хорошим суффиксным ссылкам

На каждый символ выполняется один проход по основному переходу - значит, будет затрачено O(N) времени.

Переходов по суффиксной ссылке не больше N + 1 - O(N)

Сколько тратим на проходы по хорошим суффиксным ссылкам?

Каждый раз, заходя в вершину по основному переходу, мы проверяем ее хорошую ссылку. Это N раз. Кроме того, иногда эта вершина не ведет в корень, тогда мы сообщаем о вхождении одного или нескольких паттернов. Такое случается не более z раз - число встреч наших паттернов в тексте. То есть суммарно O(N + z)

Ахо-Корасик

Сколько времени занимает проход по тексту?

Время работы на тексте длины N состоит из времени прохождения по основным переходам, времени прохождения суффиксы ссылок и времени хождения по хорошим суффиксным ссылкам

На каждый символ выполняется один проход по основному переходу - значит, будет затрачено O(N) времени.

Переходов по суффиксной ссылке не больше N + 1 — O(N)

на проходы по хорошим суффиксным ссылкам — O(N + z)

Суммарно проход по тексту занимает

$$O(N) + O(N) + O(N + z) = O(N + z)$$

Ахо-Корасик

Сколько времени занимает всего?

Суммарно построение автомата занимает не более

$$O(k \cdot M)$$

k - число паттернов, M - размер наибольшего

Суммарно проход по тексту занимает

$$O(N) + O(N) + O(N + z) = O(N + z)$$

Суммарно

$$O(k \cdot M) + O(N + z) = O(k \cdot M + N + z)$$

kM можно заменить на сумму длин паттернов (кому как нравится)

Ахо-Корасик

$$O(k \cdot M + N + z)$$

Можно ли быстрее?

Асимптотически нет, но можно считать ссылки только тогда, когда они понадобятся. Тогда константа будет лучше (редко считаем ссылки для редких паттернов).

https://habr.com/ru/post/198682/

Ахо-Корасик

$$O(k \cdot M + N + z)$$

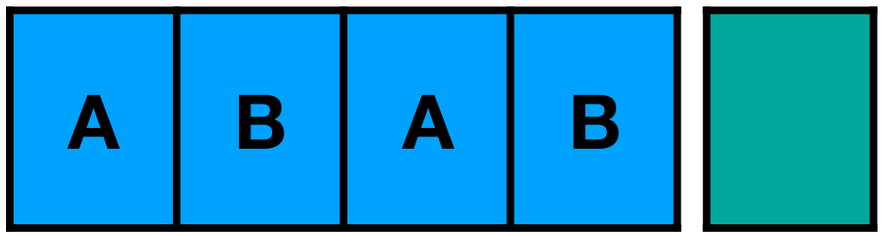
Можно ли быстрее?

Асимптотически нет, но можно считать ссылки только тогда, когда они понадобятся. Тогда константа будет лучше (редко считаем ссылки для редких паттернов).

https://habr.com/ru/post/198682/

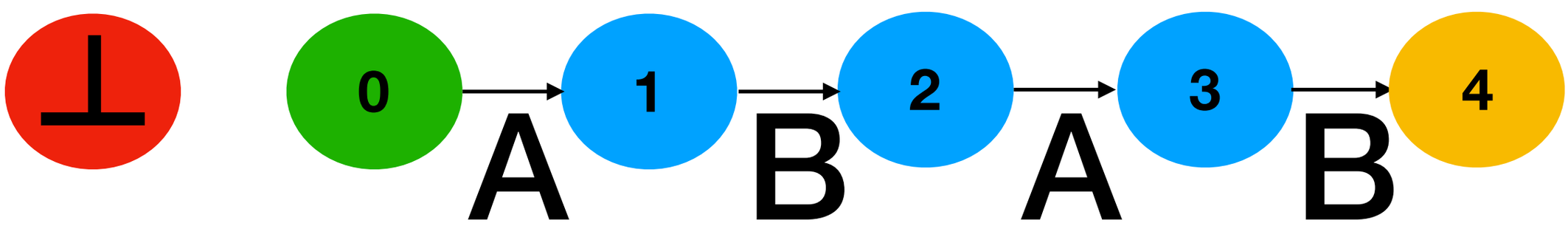
Так же можно преобразовать его в конечный автомат по аналогии с тем, как мы делали с КМП на прошлом семинаре (построить таблицу и посчитать все переходы при помощи суффиксных ссылок).

КМП и недетерминированный автомат (дополнительно)

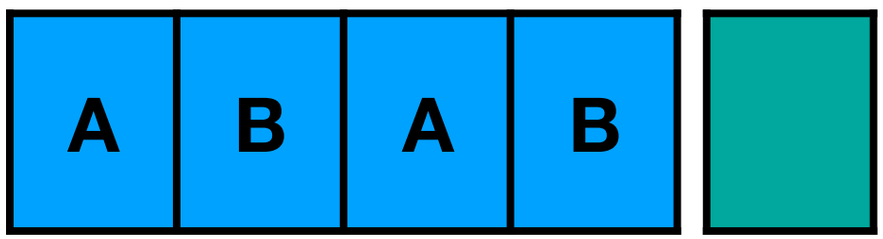


Добавим еще одно состояние - $\perp$
Проведем эпсилон-переход из каждого состояния state,
начиная с 1, в p[state - 1].

ind	0	1	2	3
π	0	0	1	2

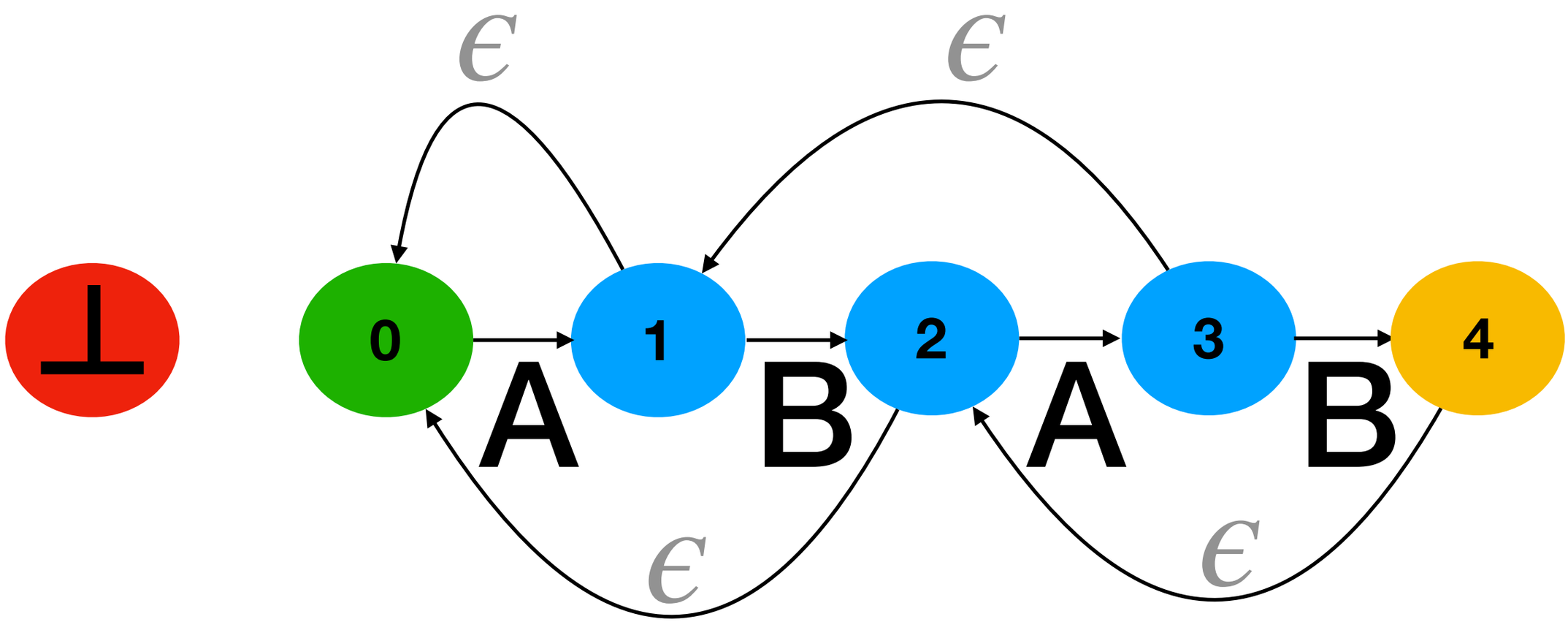


КМП и недетерминированный автомат

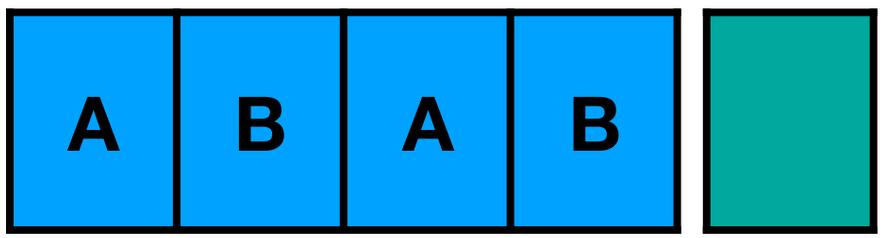


ind	0	1	2	3
π	0	0	1	2

Добавим еще одно состояние -
Проведем эпсилон-переход из каждого состояния state,
начиная с 1, в p[state - 1].



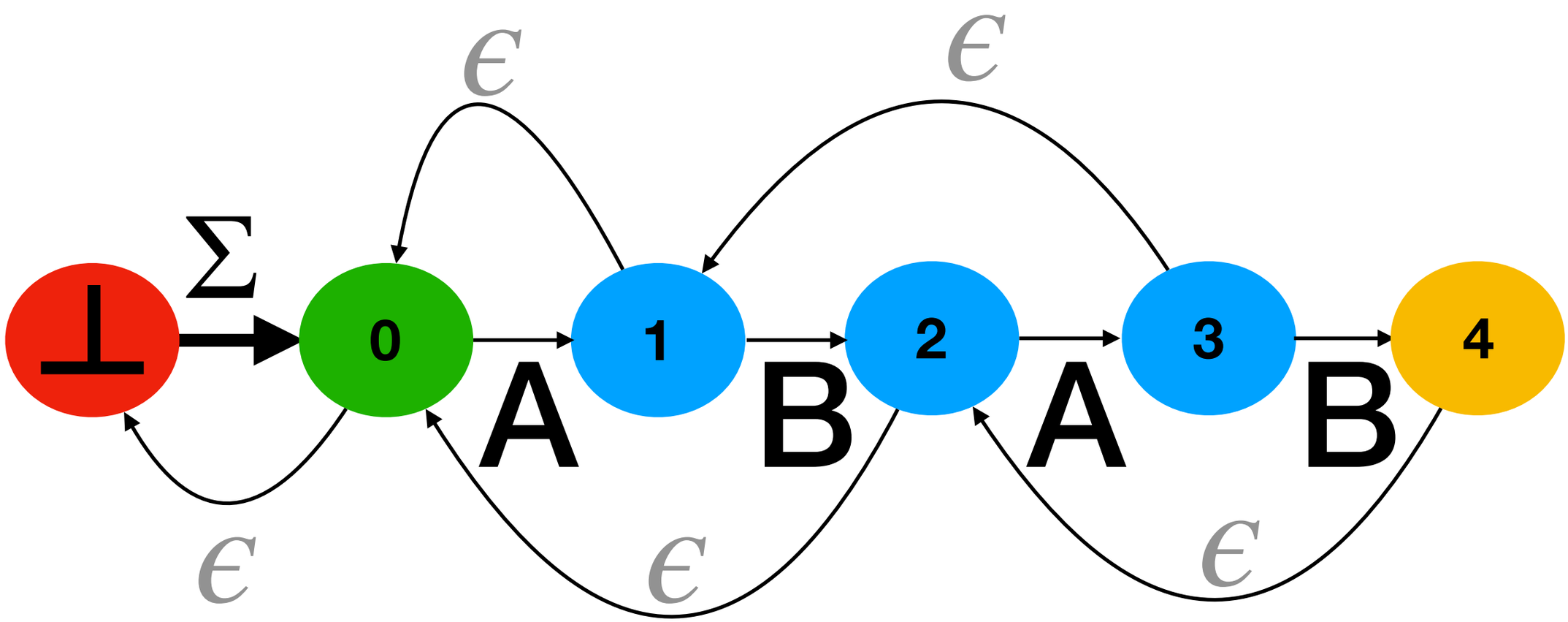
КМП и недетерминированный автомат



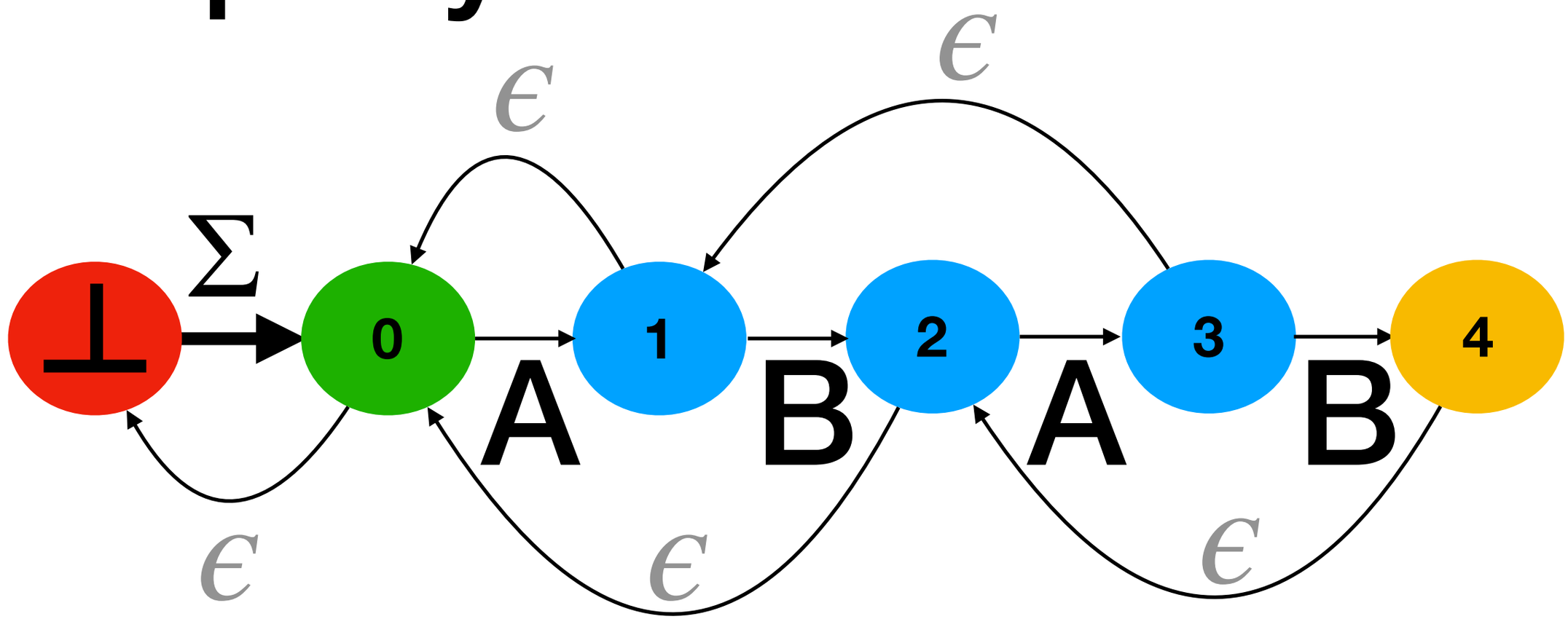
ind	0	1	2	3
π	0	0	1	2

Добавим из $\perp$ переход по-любому символу из алфавита Σ в состояние 0

И добавим эпсилон переход из состояния 0 в $\perp$

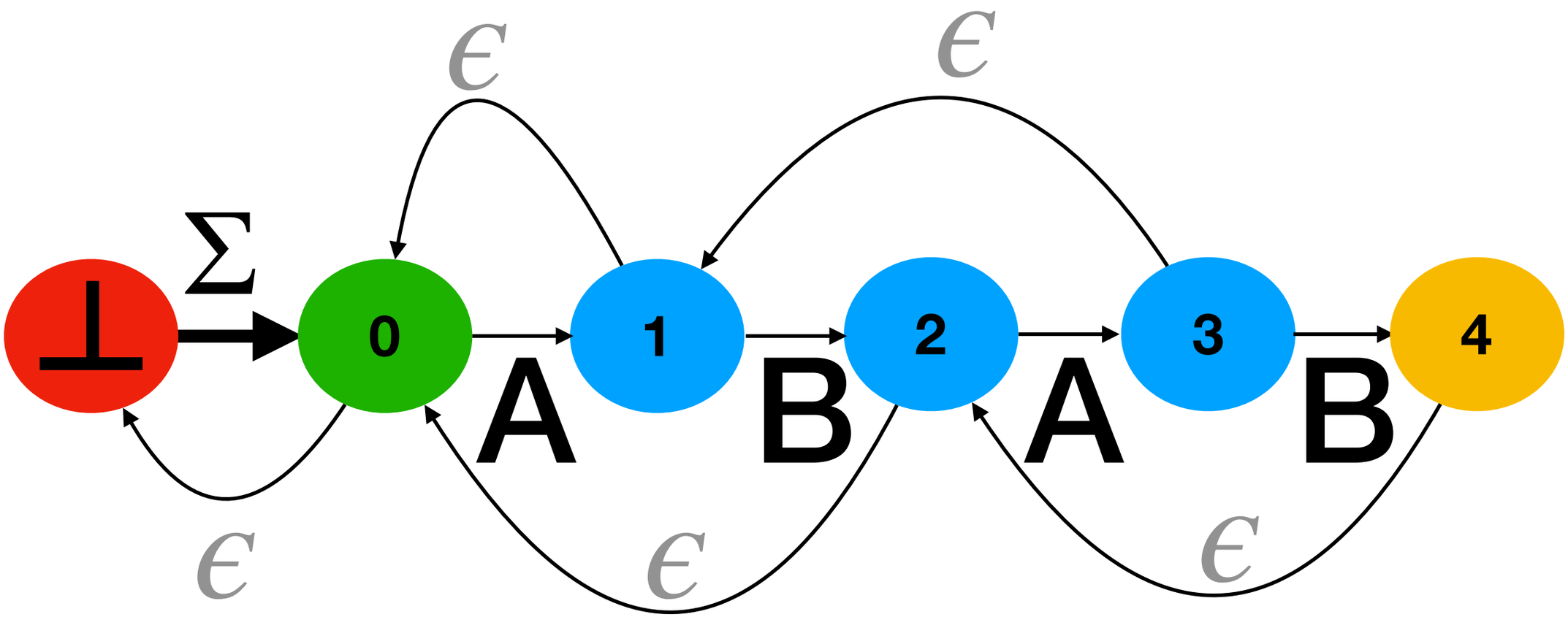


Попробуем поискать им

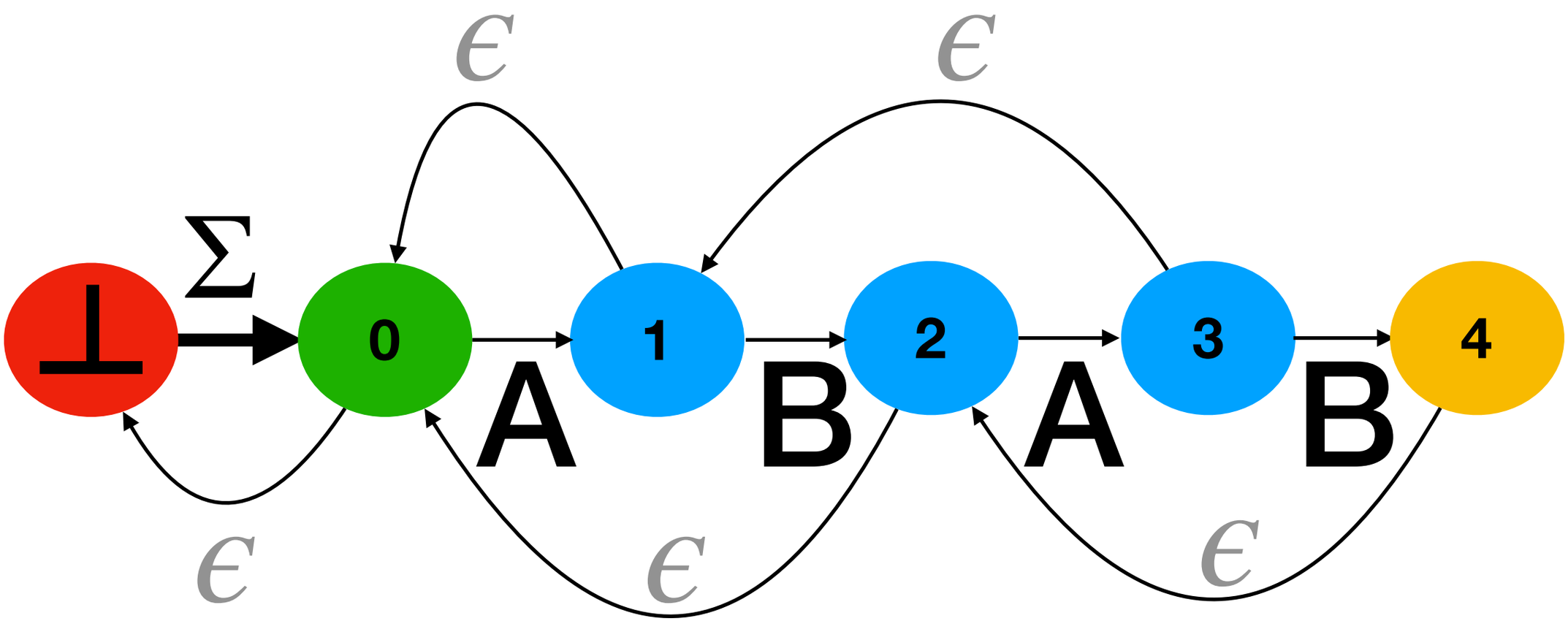


Символ текста	ϵ	A	ϵ	B	ϵ	A	ϵ	B	ϵ	A	ϵ	B
Начинаем в 0	0 ⊥	0 1	0 1 ⊥	2 0	0 2 ⊥	1 3 0	0 1 3 ⊥	2 4 0	0 2 ⊥	1 3 0	0 1 3 ⊥	4 2 0

За сколько найдет все вхождения?

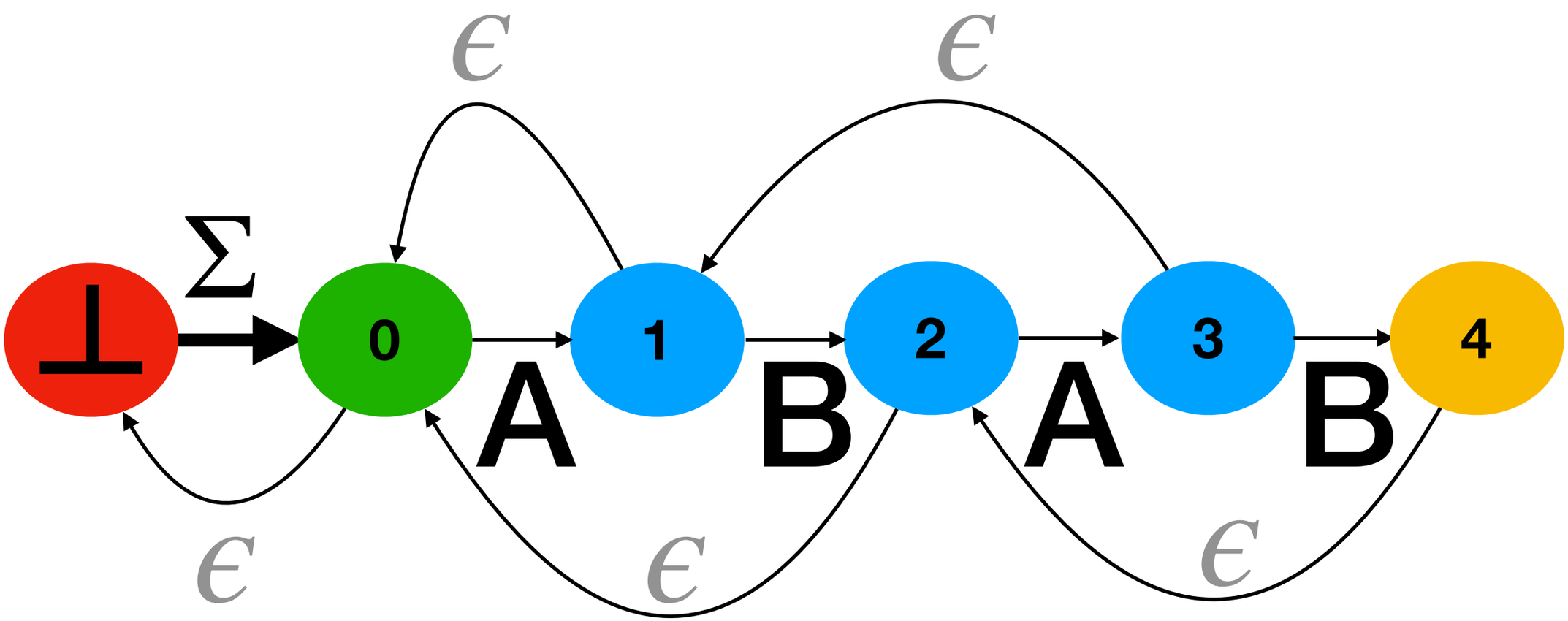


За сколько найдет все вхождения?



Недетерминированный автомат - за O(MN)

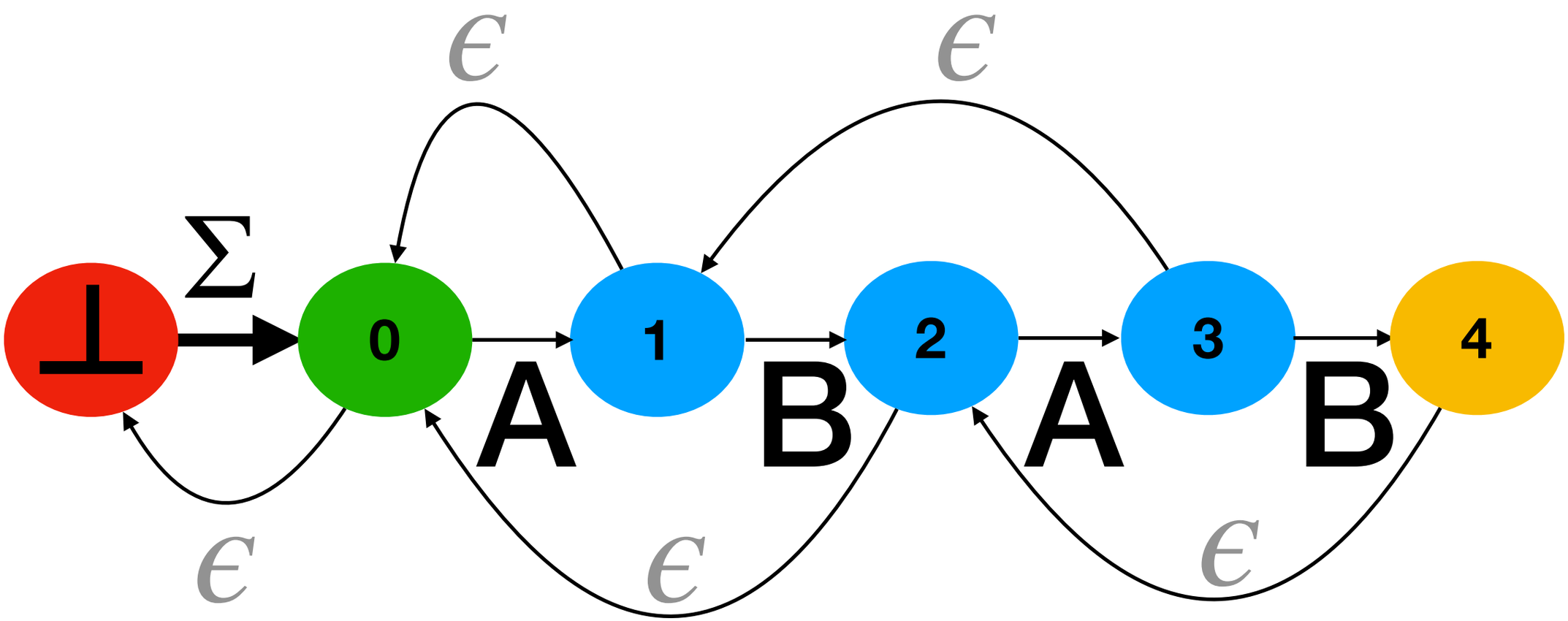
За сколько найдет все вхождения?



Недетерминированный автомат - за O(MN)

А если будем переходить по эпсилон-переходам только когда читаем “не тот” символ (либо дошли до конца)?

КМП

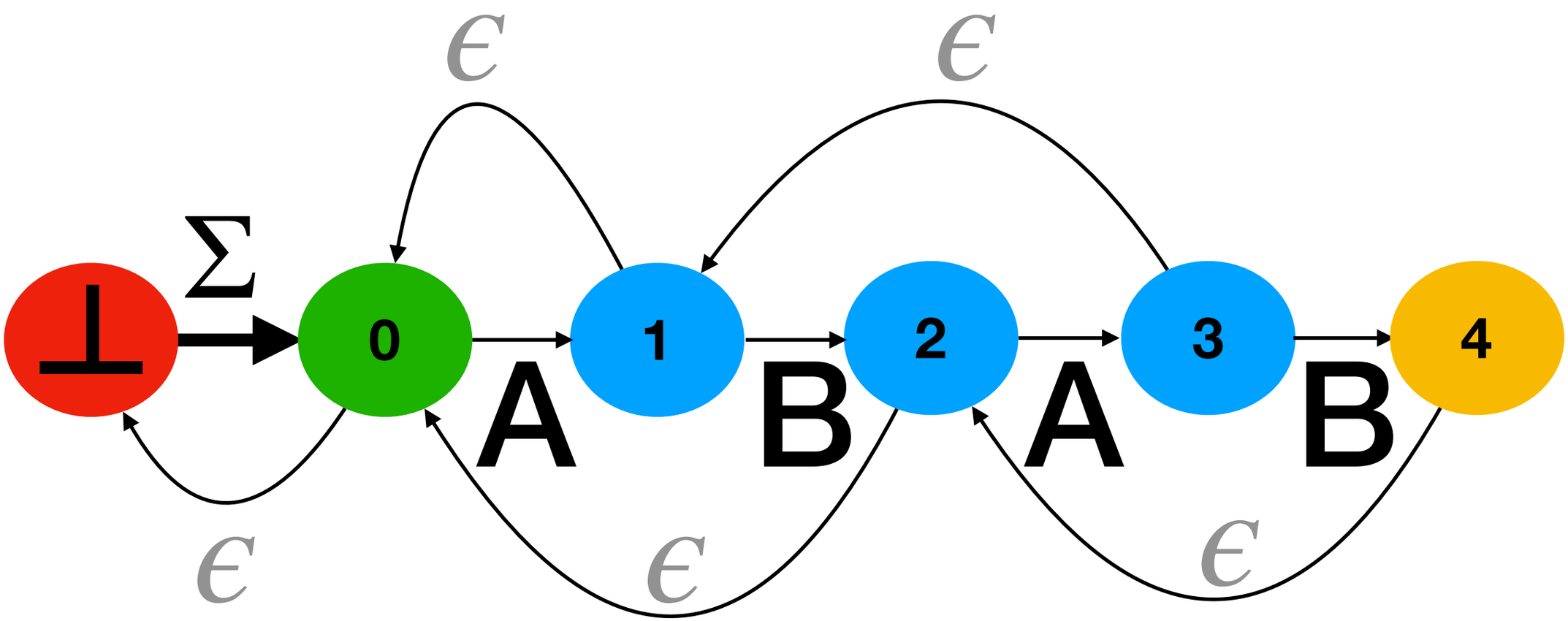


Недетерминированный автомат - за O(MN)

А если будем переходить по эпсилон-переходам только когда читаем “не тот” символ (либо дошли до конца)?

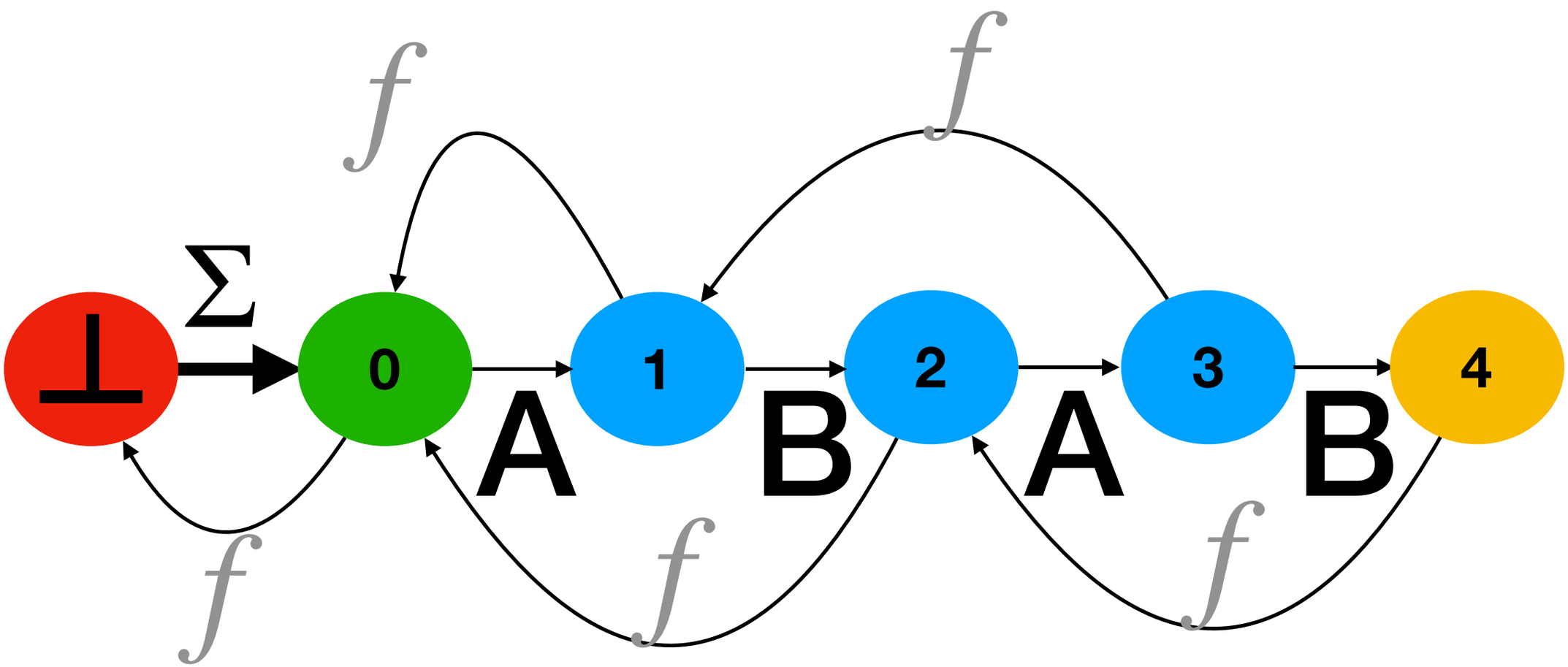
Это будет в точности Кнут-Моррис-Пратт - за O(M + N)

За сколько найдет все вхождения?



А на что это еще похоже?

Ахо-Корасик



А на что это еще похоже?

Это автомат Ахо-Корасик для одного паттерна