Лаборатория математических методов и моделей
в биоинформатике
Институт проблем передачи информации
Российской академии наук

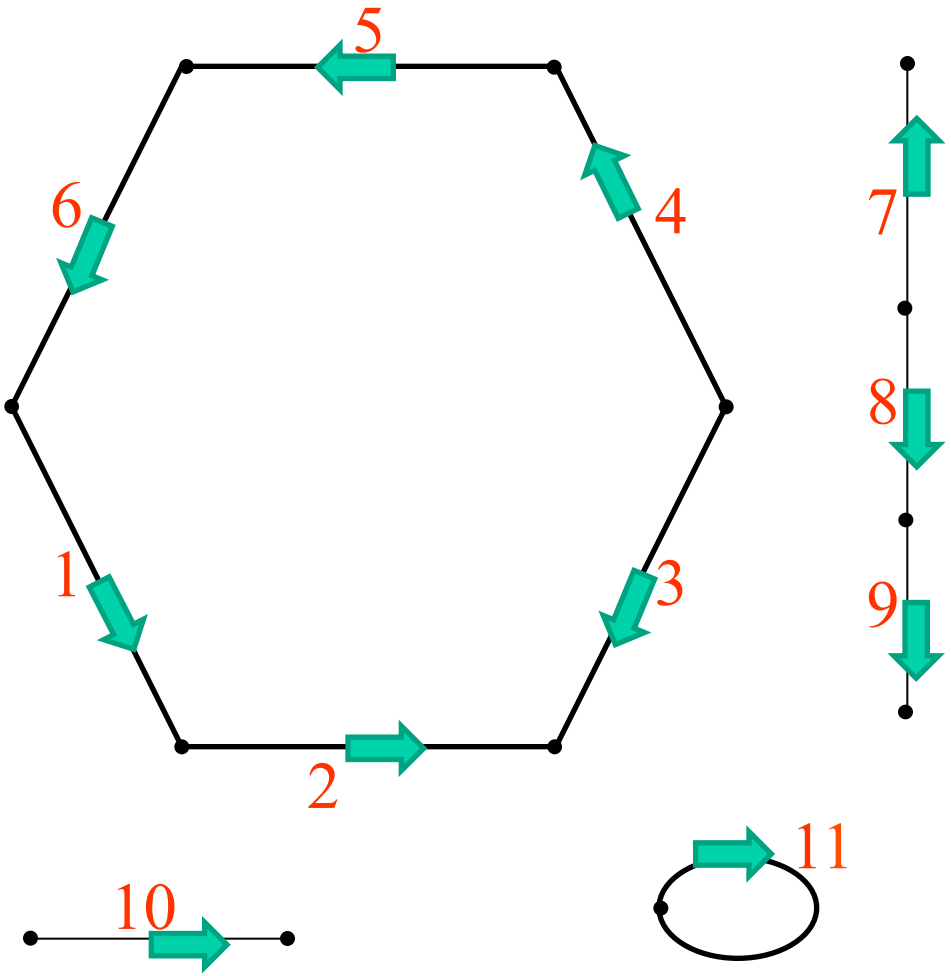
К.Ю. Горбунов

АЛГОРИТМ ПЕРЕСТРОЙКИ ХРОМОСОМНОЙ СТРУКТУРЫ

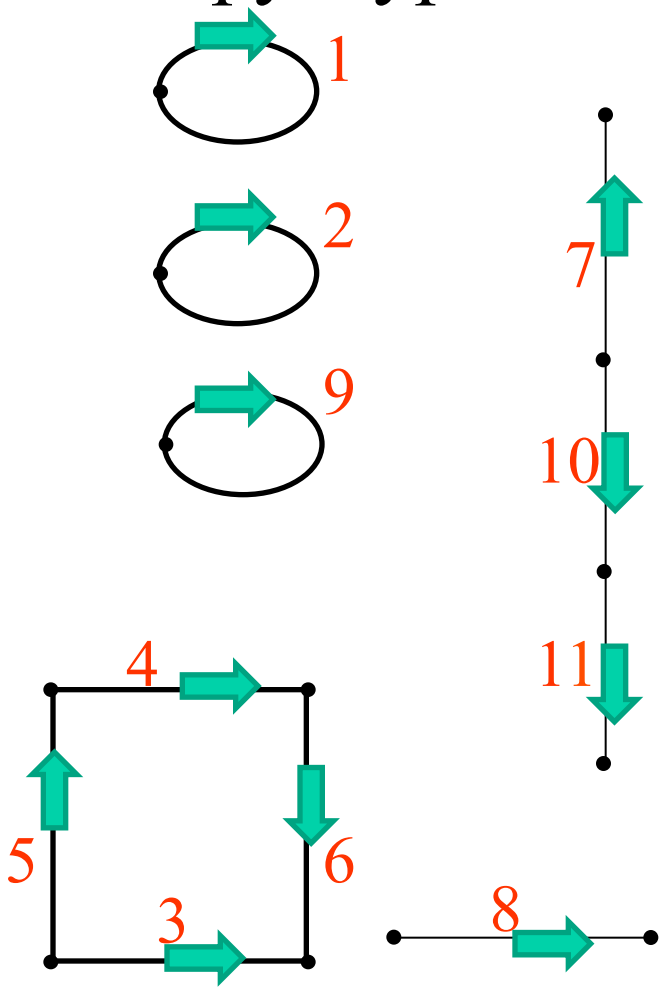
Накоплен значительный объем данных о хромосомных структурах геномов. Это позволило предложить следующую модель хромосомной структуры: пример двух хромосомных структур *a* и *b*.

Гены показаны направленными отрезками и занумерованы числами; структуры могут содержать **одинаковые или разные наборы генов**. Хромосомы линейные или циклические.

Структура *a*



Структура *b*



Задача: найти самую короткую (в варианте без цен) или самую дешевую (в варианте с ценами) последовательность операций («хромосомных перестроек»), которые переводят структуру *a* в структуру *b*.

Рассмотрим сначала случай **одинакового генного состава**.

Набор операций (Double Cut and Join) таков:

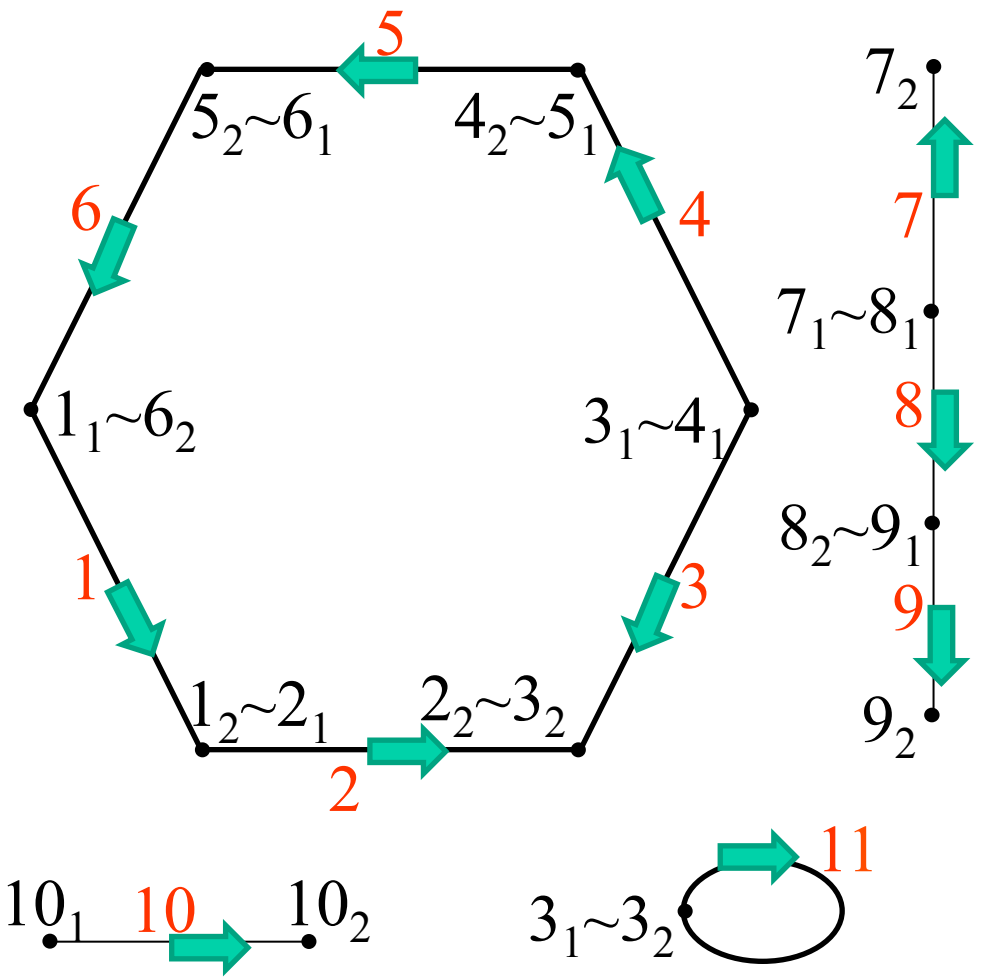
1) расклейка двух склеек структуры и склеивание четырёх освободившихся краёв генов по-другому;
2) расклейка одной склейки и склеивание одного из освободившихся краёв гена с каким-то свободным краем;
3) разрез одной склейки и обратная операция склейки двух свободных краёв.

- Эти операции позволяют выполнять все биологически содержательные перестройки хромосом: инверсия, транслокация, трансверсия, вырезание с зацикливанием, вставка цикла в цепь и другие.

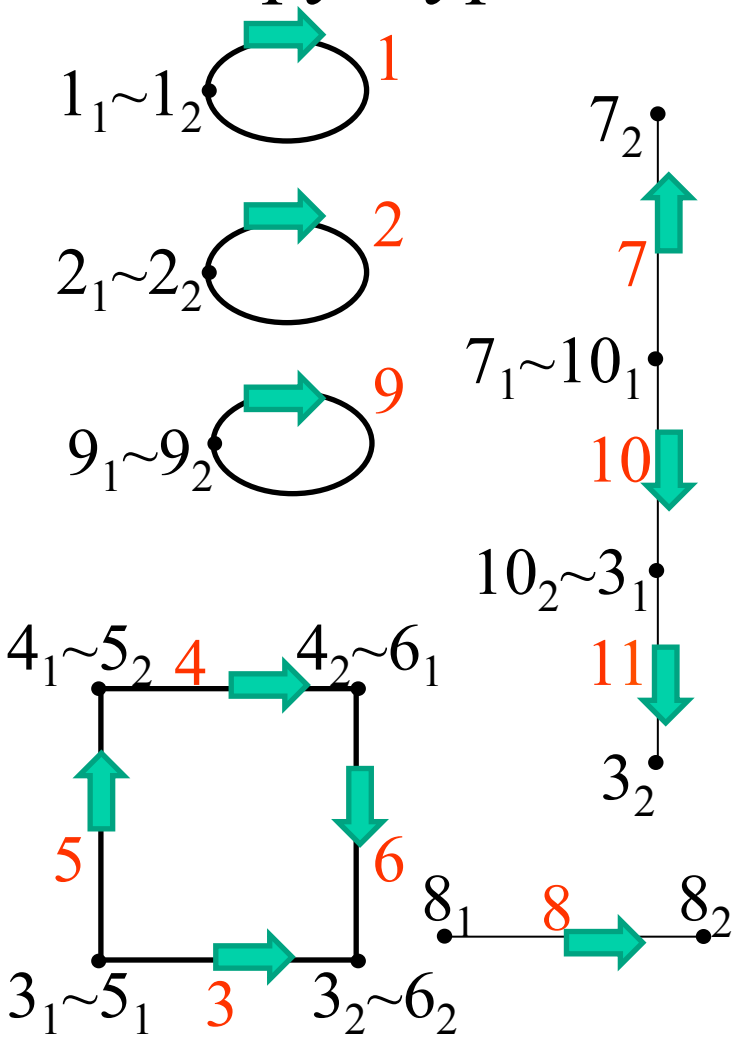
Понятие общего графа

Показаны те же две структуры, но теперь края генов склеены.
Для гена *i* его начало обозначено i_1 , его конец обозначен i_2.

Структура *a*

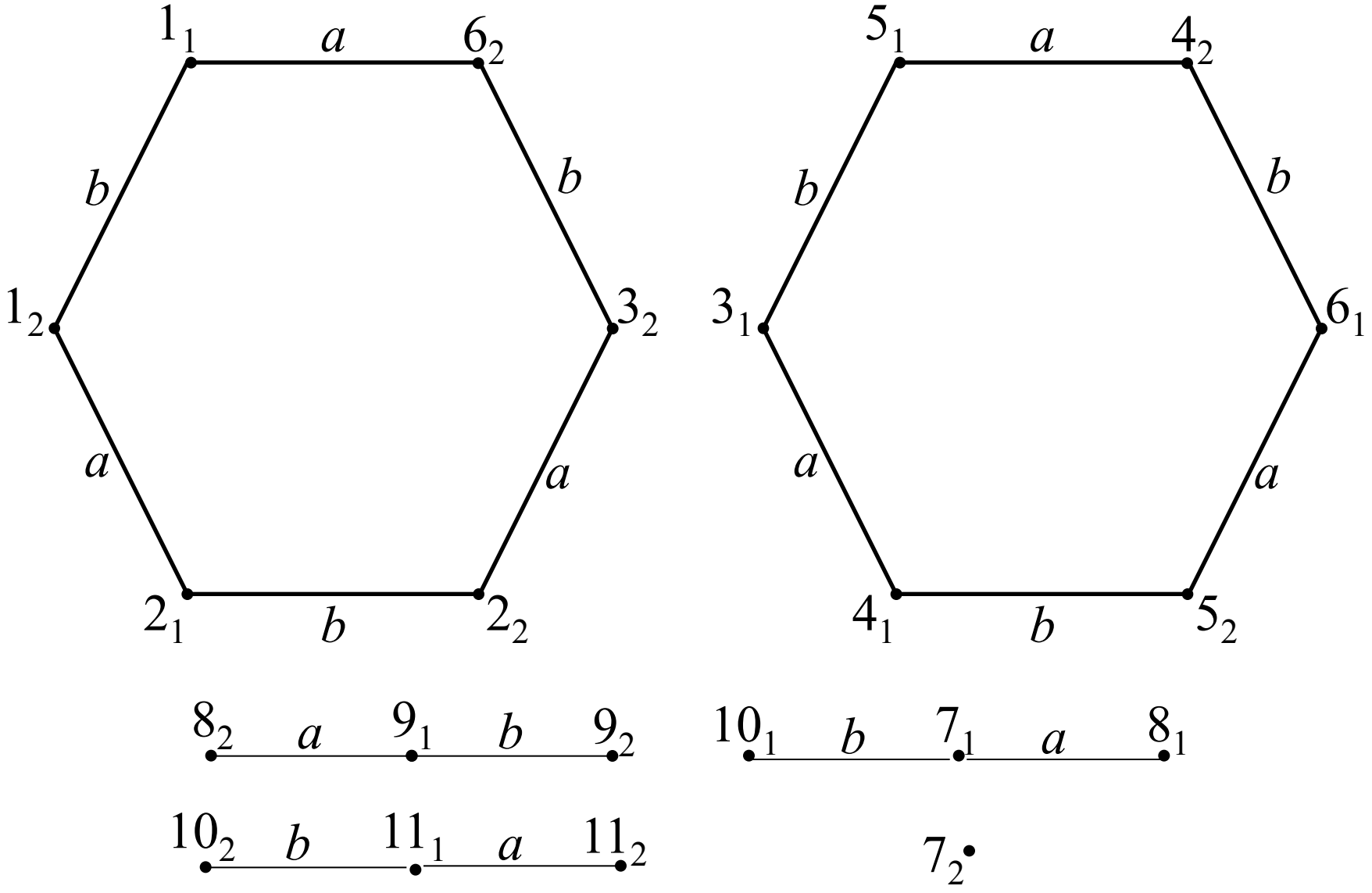


Структура *b*



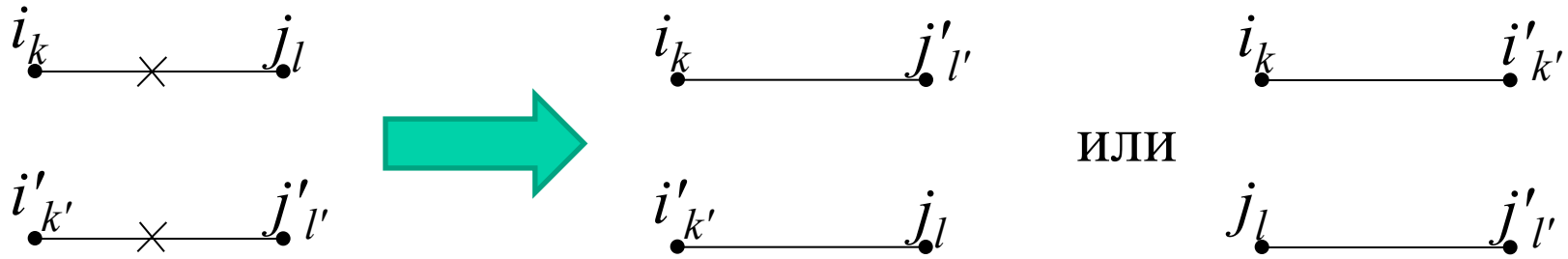
Мы предложили способ представить данную пару структур a и b в виде одного графа $a+b$: в нем вершины – все края всех генов, ребра – склейки краёв, которые заданы в двух данных структурах a и b, ребра помечаются именем структуры, из которой взяты края. По $a+b$ легко восстановить исходные структуры a и b. Граф вида $a+b$ (для любых a и b) назовём общим.

Общий граф $a+b$ структур a и b

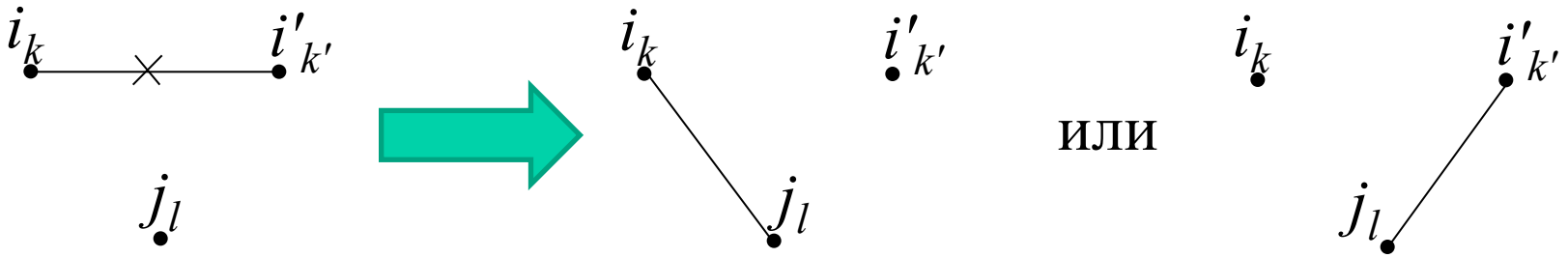


Наши операции выглядят на общем графе так:

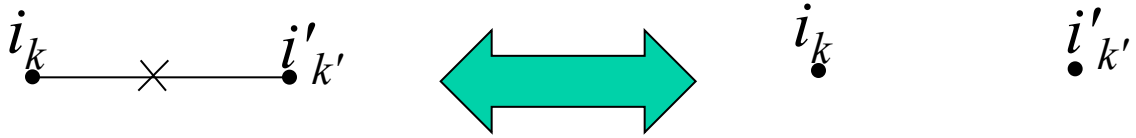
1)



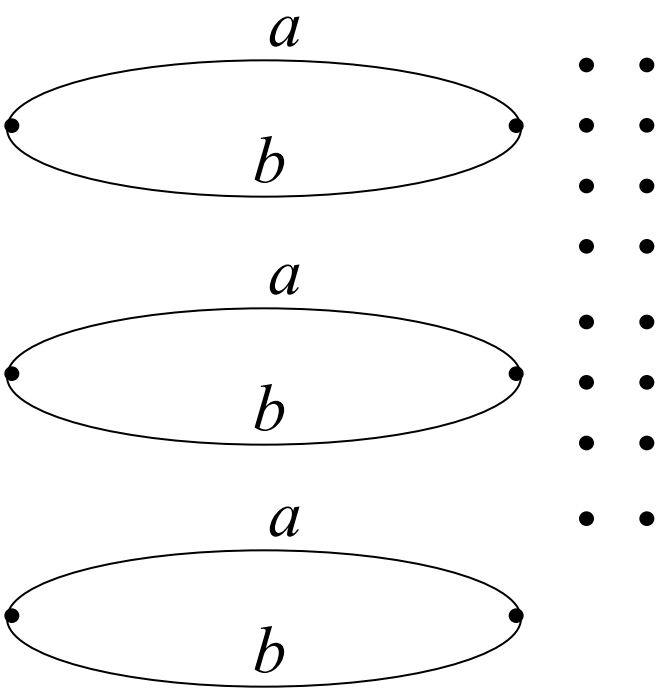
2)



3)



Исходная задача про a и b эквивалентна задаче приведения структур a и b к общей структуре c, что эквивалентно приведению графа $a+b$ к «финальному» виду $c+c$, состоящему из циклов длины 2 и изолированных вершин:

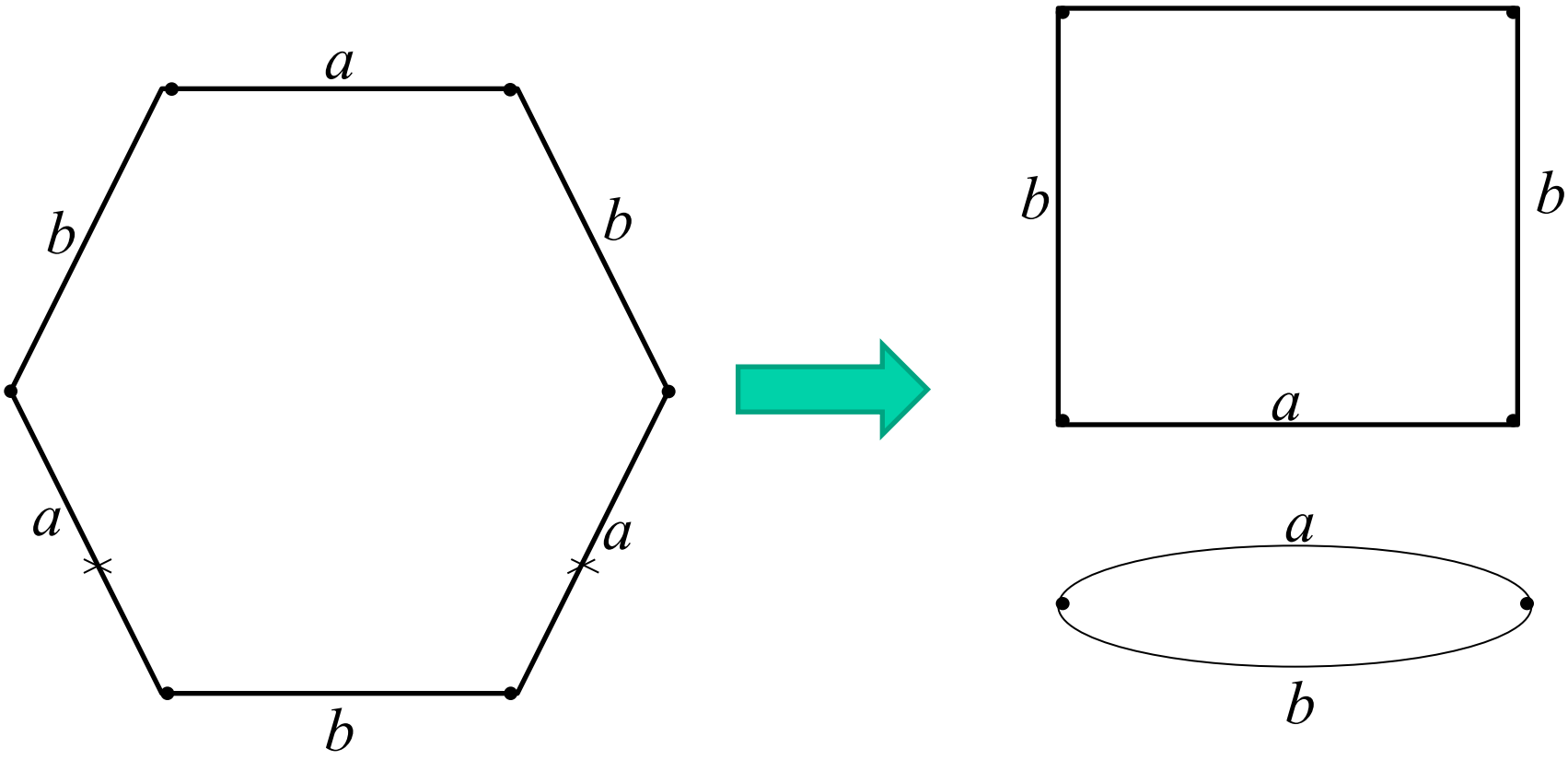


Обозначим цены: разреза c_1, склейки c_1', полуторной переклейки $c_{1.5}$, двойной переклейки c_2. Рассмотрим два варианта соотношения цен: $c_2 \le c_1 \le c_1' \le c_{1.5}$ («циклический») и $c_1 \le c_1' \le c_{1.5} \le c_2$ («линейный»).

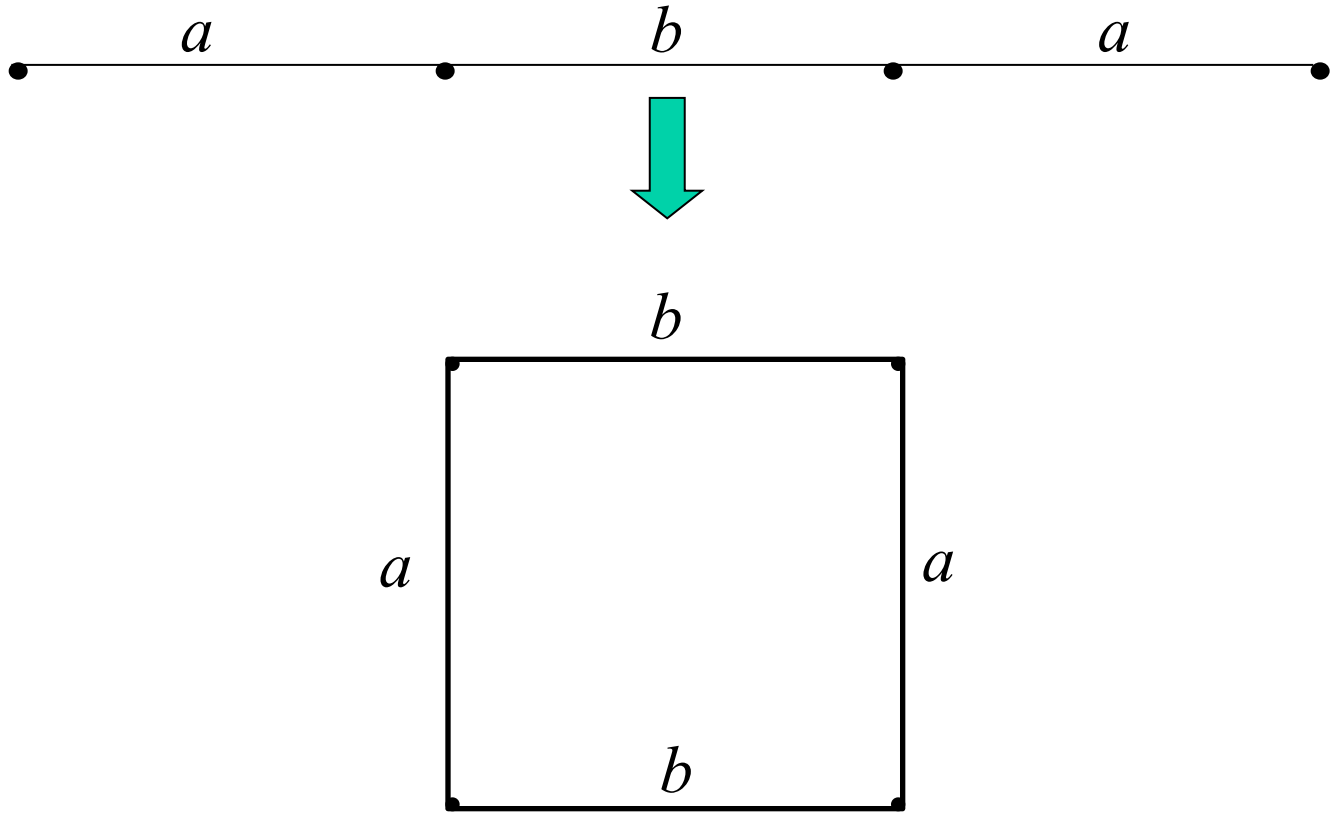
Решена задача условной оптимизации: при каждом из двух соотношениях цен и дополнительном условии: наиболее дешевая последовательность ищется среди всех самых коротких последовательностей.

Решение безусловной задачи может не быть самой короткой последовательностью, но если разница в ценах мала, то будет. Не известно, существует ли полиномиальный по времени алгоритм решения безусловной задачи для какого-нибудь варианта.

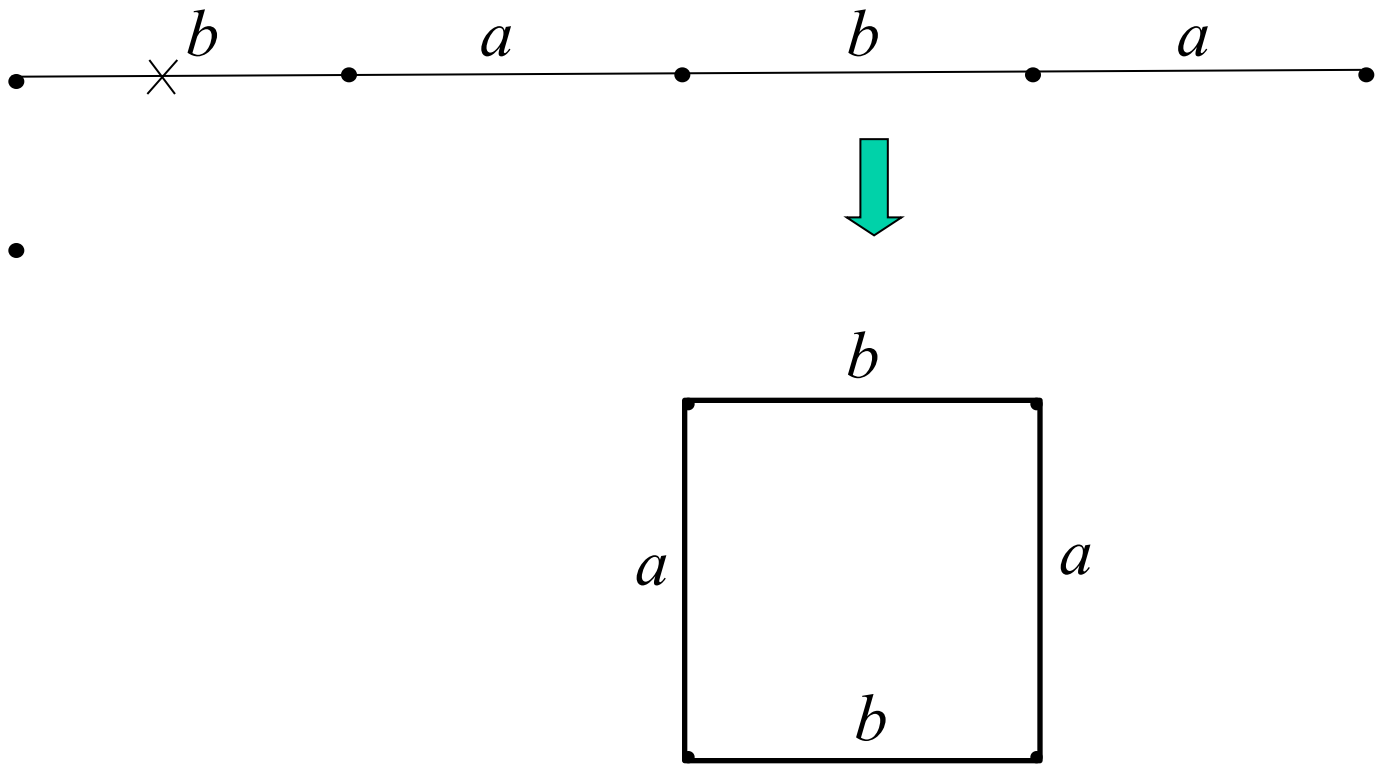
Укорачивание циклов одно и то же в обоих вариантах



В циклическом варианте все цепи замыкаются в цикл. Нечетная – склейкой концов:

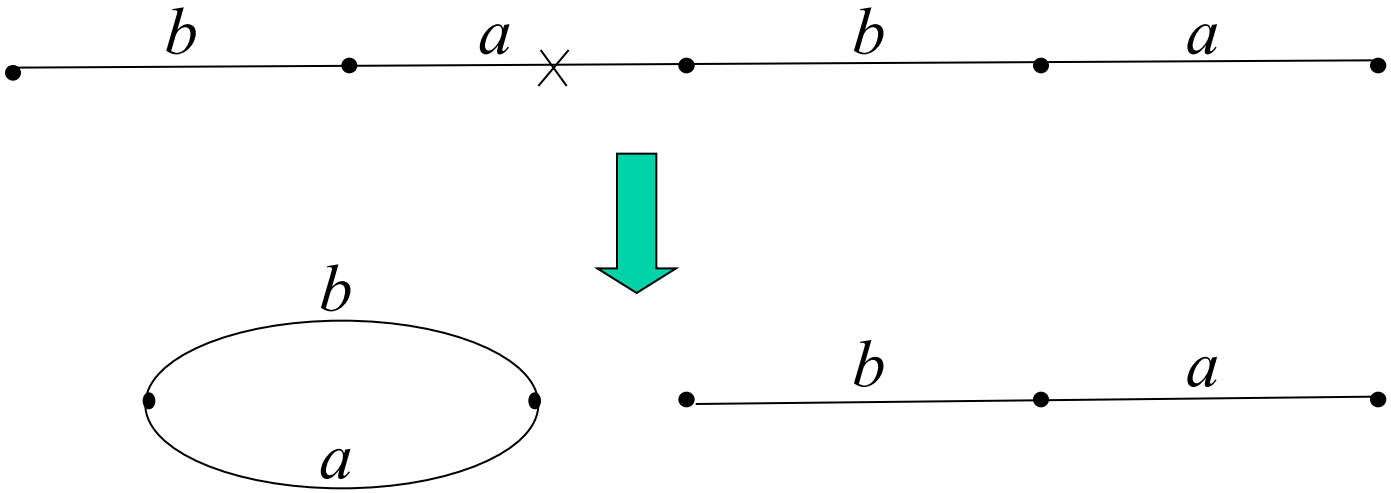


Четная – полуторной переклейкой:



В линейном варианте каждая нечетная цепь разрезается на две четные, одна из которых нулевая:

Затем каждая четная цепь последовательно укорачивается на 2 полуторной переклейкой:



Случай различного генного состава структур *a* и *b*

Добавляются две операции.

Операция удаления – удаление связного участка генов, принадлежащих структуре *a,* но не *b*. Такой участок может быть удалён, если

а) он находился строго внутри линейной или циклической хромосомы, два конца внешних генов склеиваются;

б) он находился с края линейной хромосомы, конец внешнего гена становится не склеенным;

в) он являлся отдельной хромосомой.

Операция вставки – обратная к операции удаления, но вставлять разрешается только связные участки генов, принадлежащих структуре *b,* но не *a*.

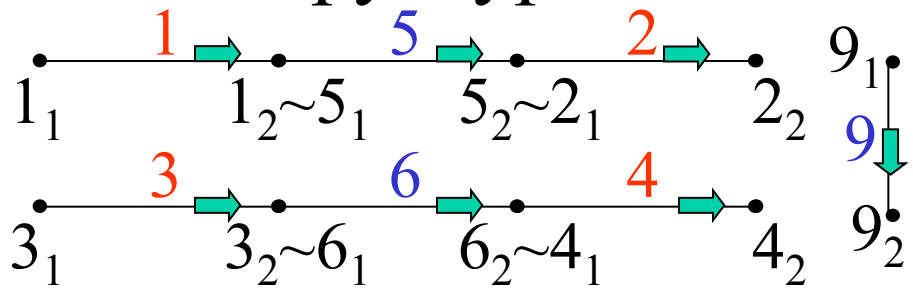
Сформулируем задачу перехода от a к b в случае **различного генного состава** структур a и b в терминах общего графа. Доказано, что формулировки в терминах хромосомных структур и общего графа эквивалентны по крайней мере для случая, когда все цены равны, т.е. когда ищется самая короткая последовательность операций.

Понятие общего графа нуждается в некоторой модификации. Вершинами общего графа $a+b$ разумно объявить лишь края генов, общих для структур a и b. Что же касается необщих генов, то нас интересует лишь их расположение относительно общих генов.

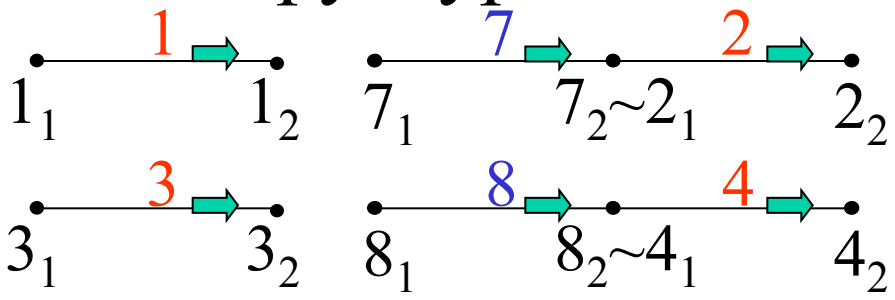
Поэтому, ребро между двумя вершинами проводится не только в случае непосредственной склейки соответствующих краев, но и в случае, когда эти два края склеены с краями одного и того же отрезка необщих генов, называемого **блоком**. Это ребро помечается синим прямоугольником, говорящим о наличии блока и называется **блоковым ребром.** Если блок расположен с краю цепи, то из склеенного с ним конца общего гена проводится висячее (т.е. инцидентное лишь одной вершине) ребро. Наконец, если блок составляет целую цепь или цикл, он показывается плавающим ребром или петлей, не инцидентным никаким вершинам.

Пример двух структур с различным множеством генов:

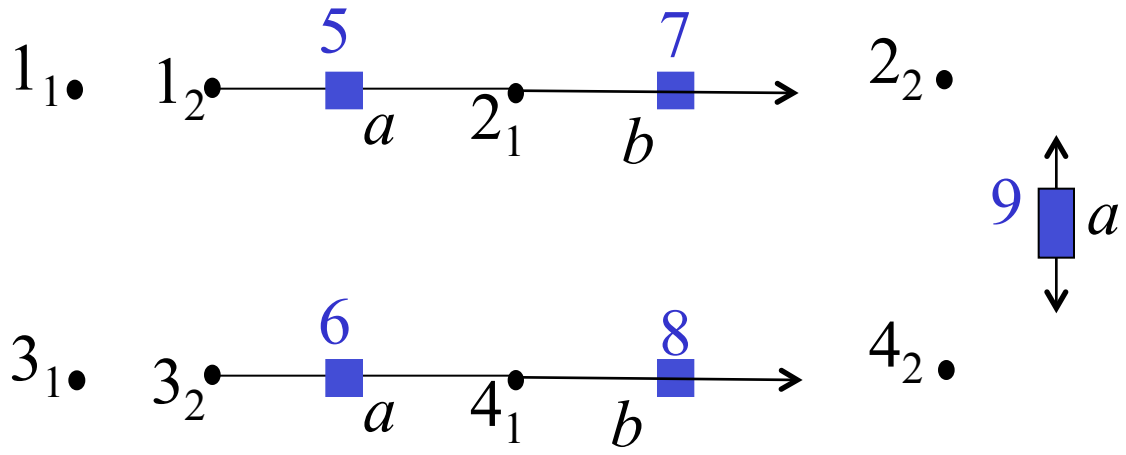
Структура *a*



Структура *b*



Общий граф *a*+*b* этих двух структур:

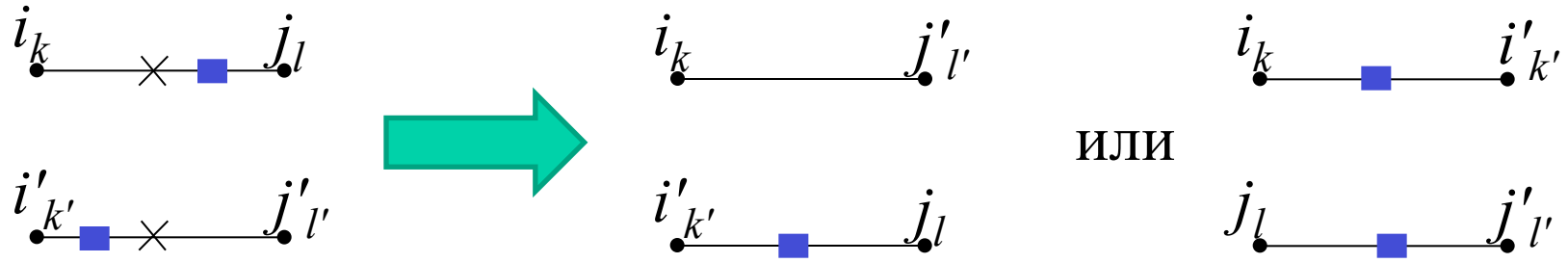


Здесь все ребра блоковые.

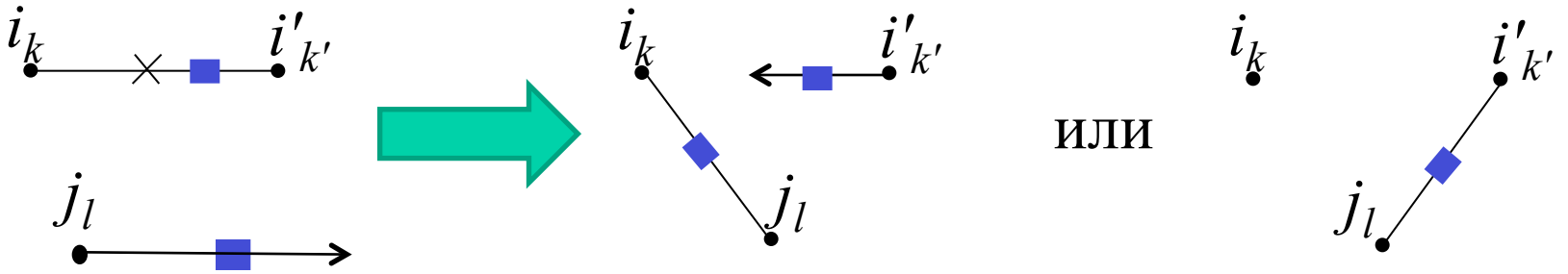
Задача: привести общий граф $a+b$ к финальному виду.

Набор операций расширен операцией удаления блока:

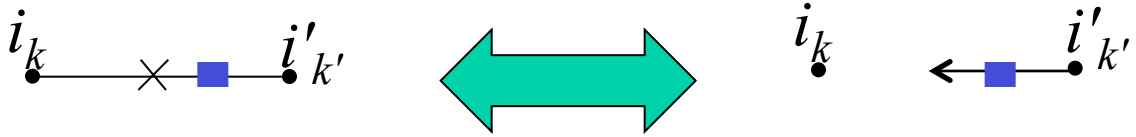
1)



2)



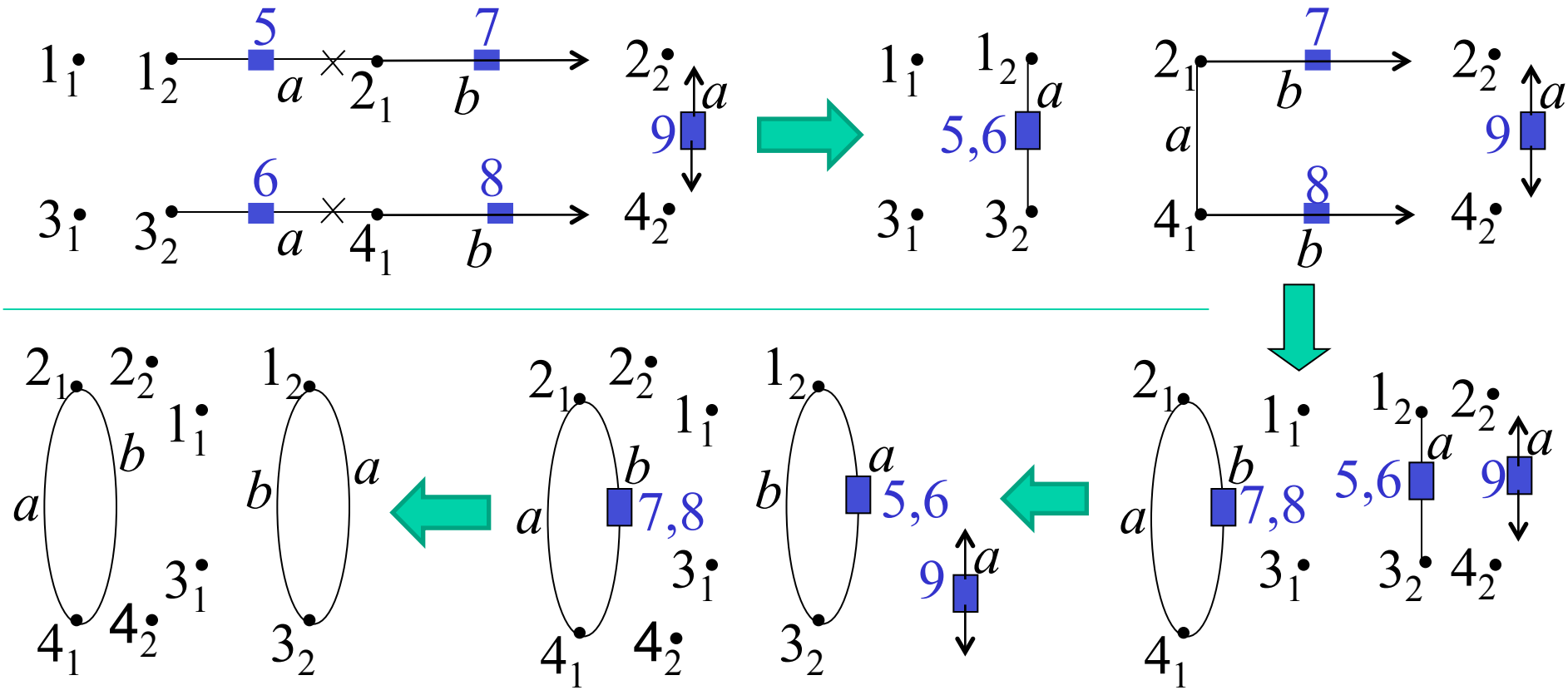
3)



4)

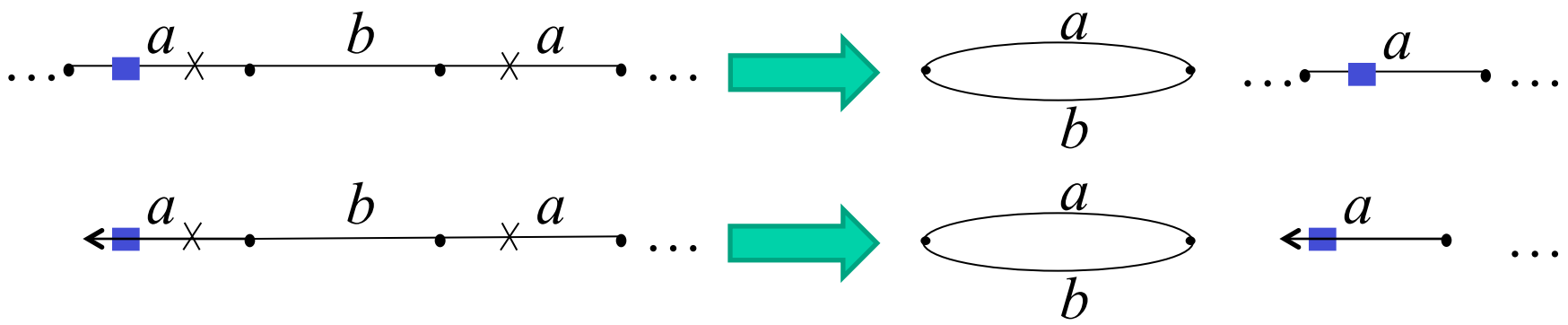
Блоковая операция – операция, которая уменьшает число блоков в структуре.

Например, возможное приведение графа $a+b$ к финальному виду для последнего примера выглядит так:

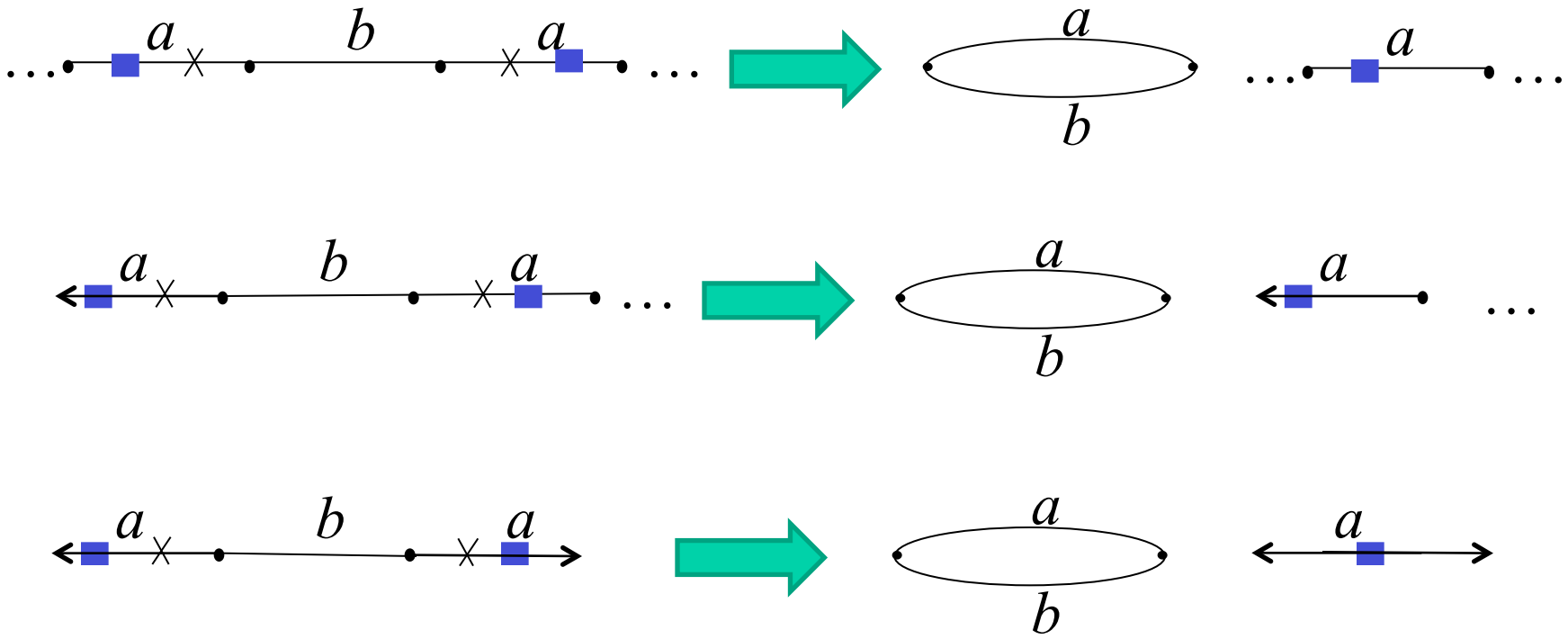


Но вначале следует удалить все неблоковые рёбра из нефинальных компонент, т.е. привести граф к **блоковому виду**.

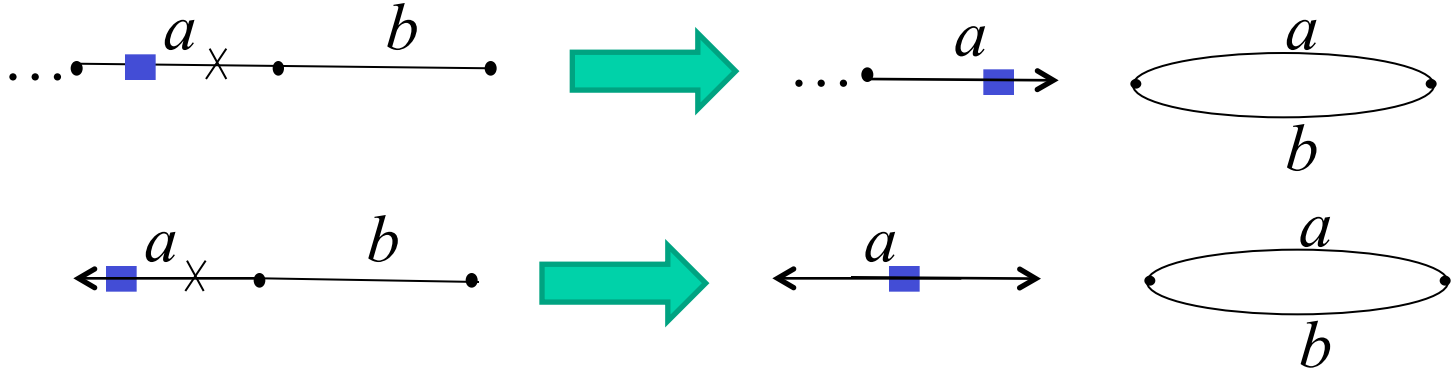
Удаление пары соседних неблоковых ребер:



Удаление одинарных некрайних неблоковых ребер блоковой операцией:

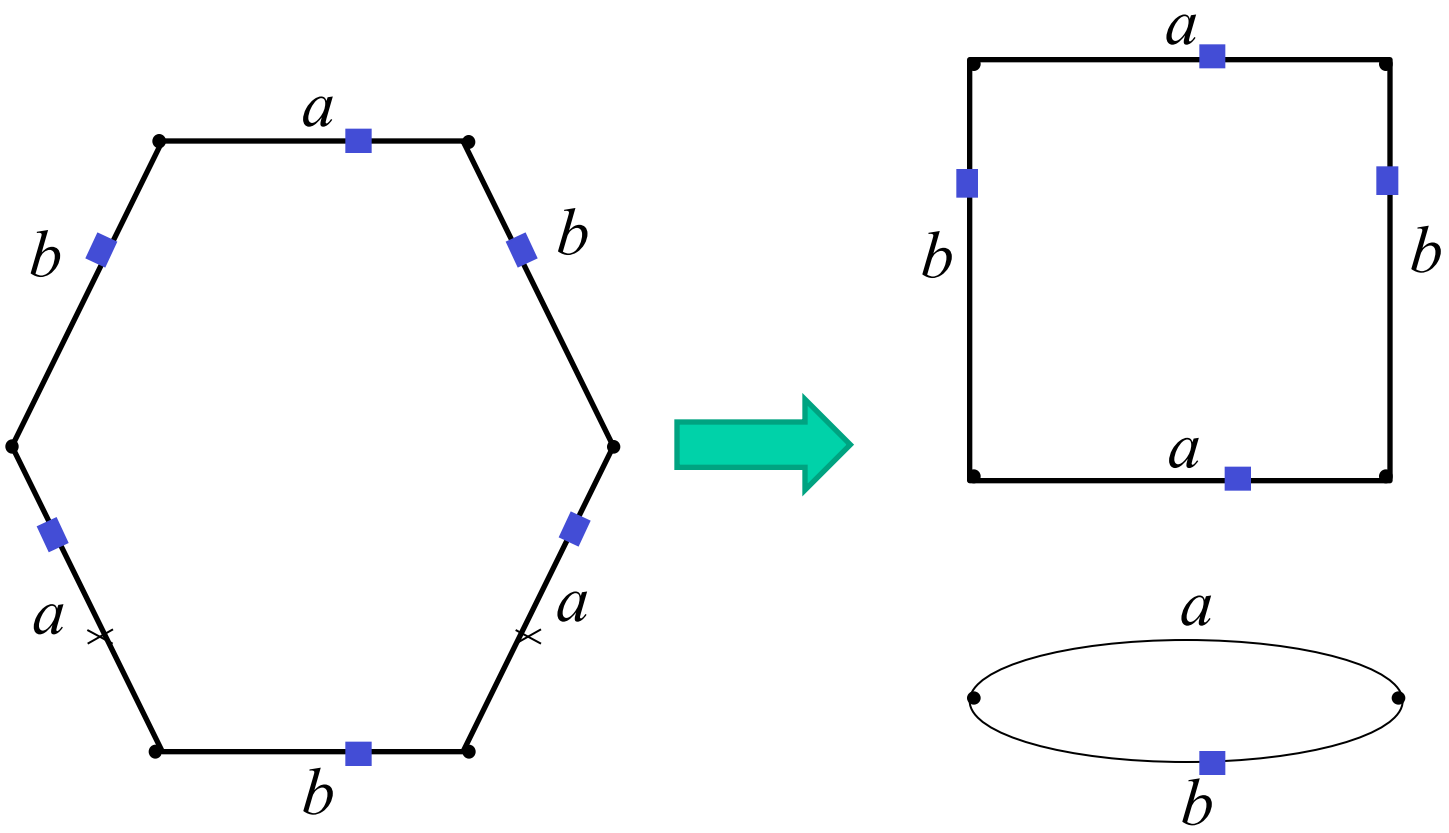


Удаление крайних одинарных неблоковых ребер неблоковой операцией:

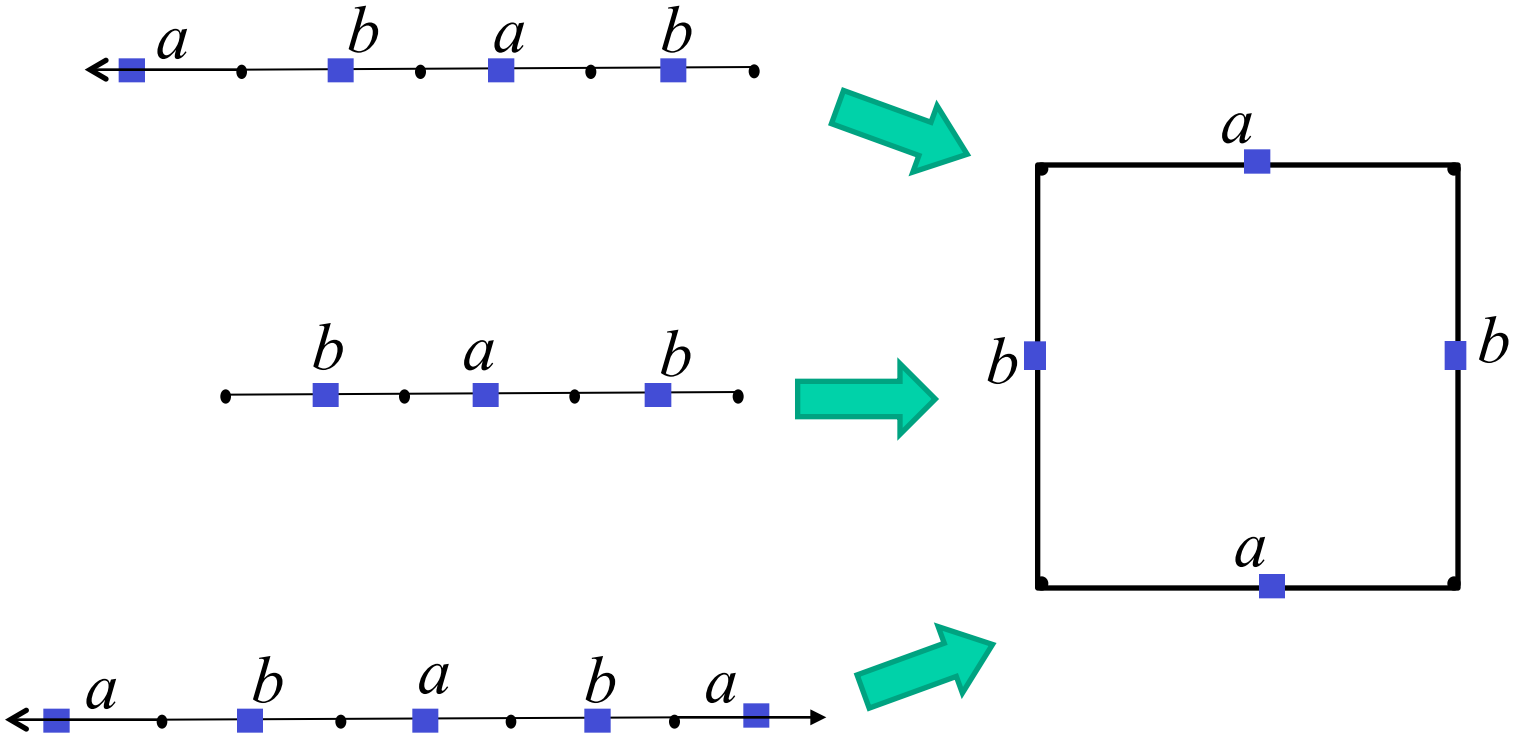


Рассмотрим вариант покомпонентного приведения к финальному виду. Отметим: количество блоковых операций при приведении графа к финальному виду всегда равно числу блоков в нём, так что мы стремимся минимизировать число неблоковых операций. Напомним: все компоненты приведены к блоковому виду.

Цикл: все операции блоковые:

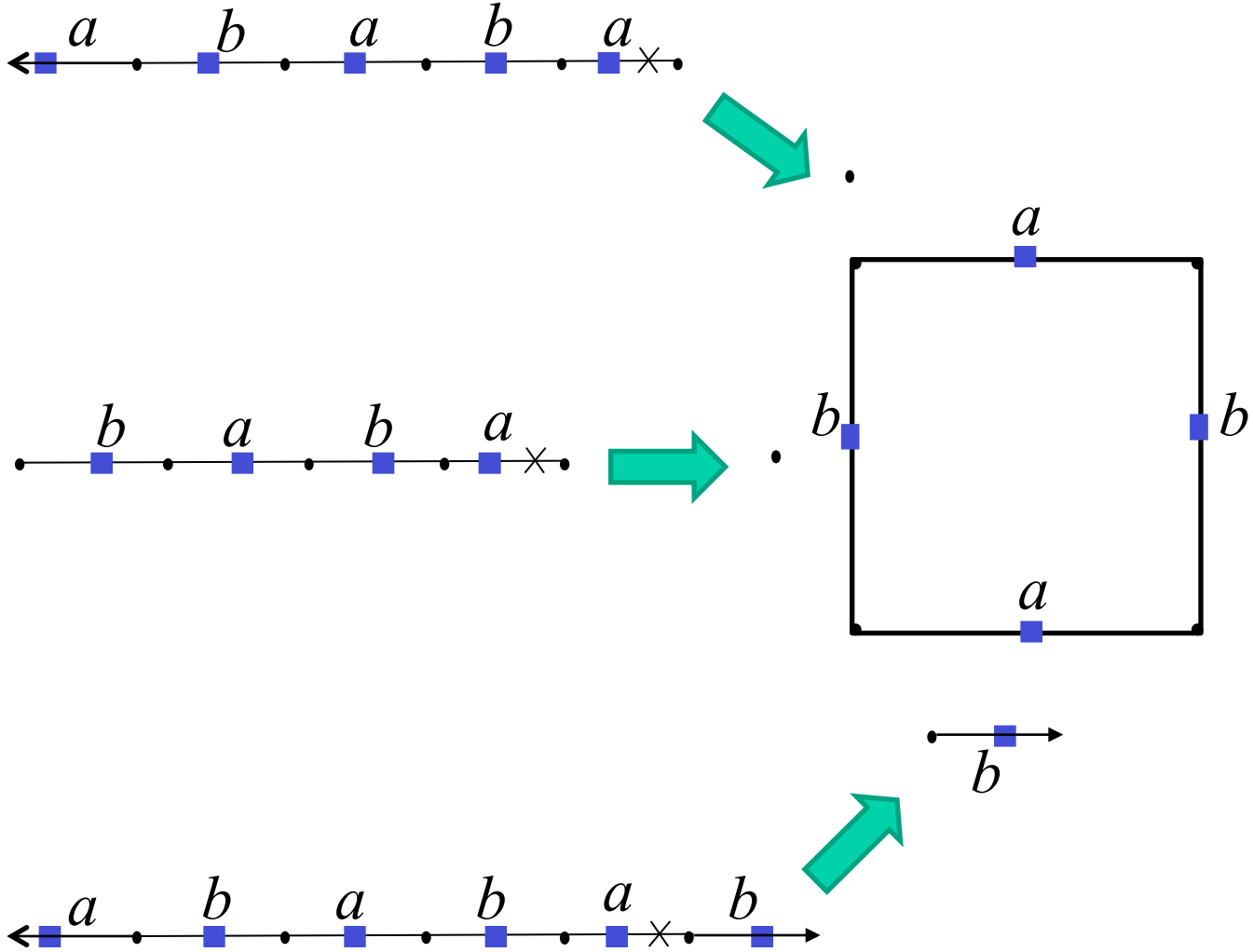


Нечетные цепи: если висячих ребра два, то все операции блоковые, иначе – все, кроме одной:



Для среднего случая верхний блок цикла отсутствует, но наличие в цикле одного неблокового ребра не мешает привести его к финальному виду блоковыми операциями.

Четные цепи: если висячие ребра есть, то все операции блоковые, иначе – все, кроме одной:



Есть разные типы межкомпонентных взаимодействий, позволяющих сэкономить неблоковые операции. Есть множество базовых типов взаимодействий. В каждом из них участвует от 2 до 4 компонент, которые совместными усилиями экономят от 1 до 3 неблоковых операций. Это множество обладает двумя важными свойствами:

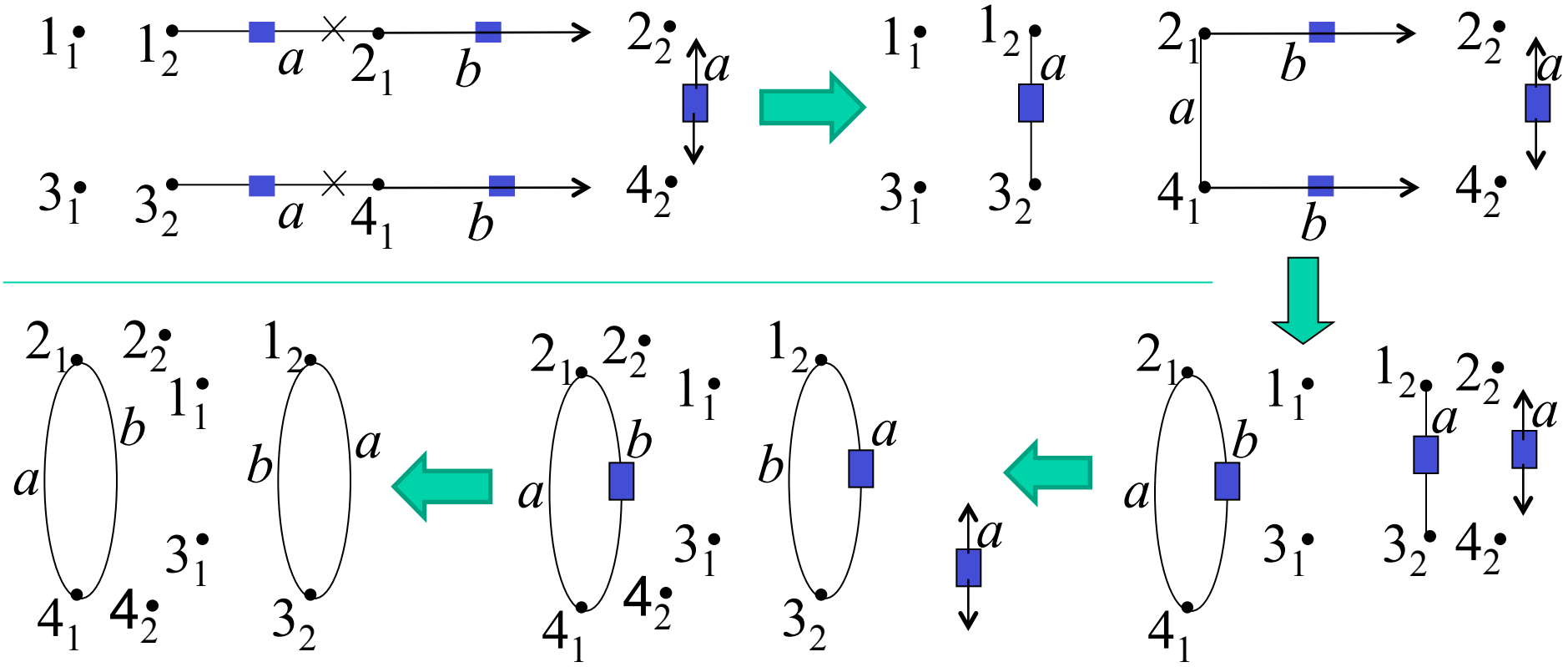
1) Выполнение базовых взаимодействий в определенном порядке позволяет достичь максимально возможной экономии.

2) При этом результаты каждого взаимодействия не используются в последующих взаимодействиях.

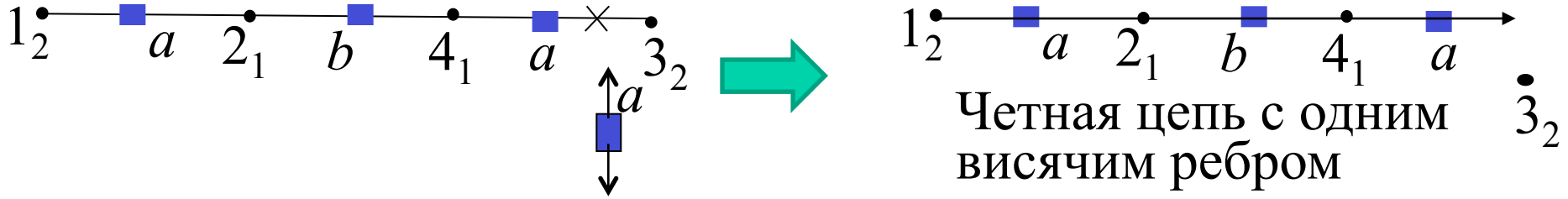
Экономят одну операцию следующие совокупности:
{1*a*,1*a*}; {1*b*,1*b*}; {1*a*,2*b*}; {1*b*,2*a*}; {1*a*,3*b*}; {1*b*,3*a*)}; {1*a*, 2}; {1*b*,2}; {1*a*,3}; {1*b*,3}; {2*a*,3*b*}; {2*b*,3*a*}; {2,3}; {2*a*,2*b*,3}; {2*a*,3,3};{2*b*,3,3}; {3*a*,3*b*,2}; {3*a*,2,2}; {3*b*,2,2}.
Экономят две операции следующие совокупности: {1*a*, 1*b*};
{1*a*,1*a*,2*b*}; {1*b*,1*b*,2*a*}; {1*a*,1*a*,3*b*}; {1*b*,1*b*,3*a*}; {1*a*,2*b*,3}; {1*b*,2*a*,3}; {1*a*,3*b*,2}; {1*b*,3*a*,2};
{2*a*,2*b*,3,3}; {3*a*,3*b*,2,2}.
Экономят три операции следующие совокупности: {1*a*,1*a*, 2*b*,3*b*}; {1*b*,1*b*,2*a*,3*a*}.

Обозначения:
1*a*,1*b,*2*a,*2*b*,3*a,*3*b* – нечетные цепи с одним висячим ребром (1*a*,1*b*), с двумя (2*a,*2*b*) и без таковых (3*a,*3*b*). Буквы *a* или *b* – пометка крайних невисячих ребер.
1,2,3 - четные цепи с одним висячим ребром (1), с двумя (2) и без таковых (3).
Плавающее ребро с пометкой *a* имеет тип 2*b*, с пометкой *b* – тип 2*a*.

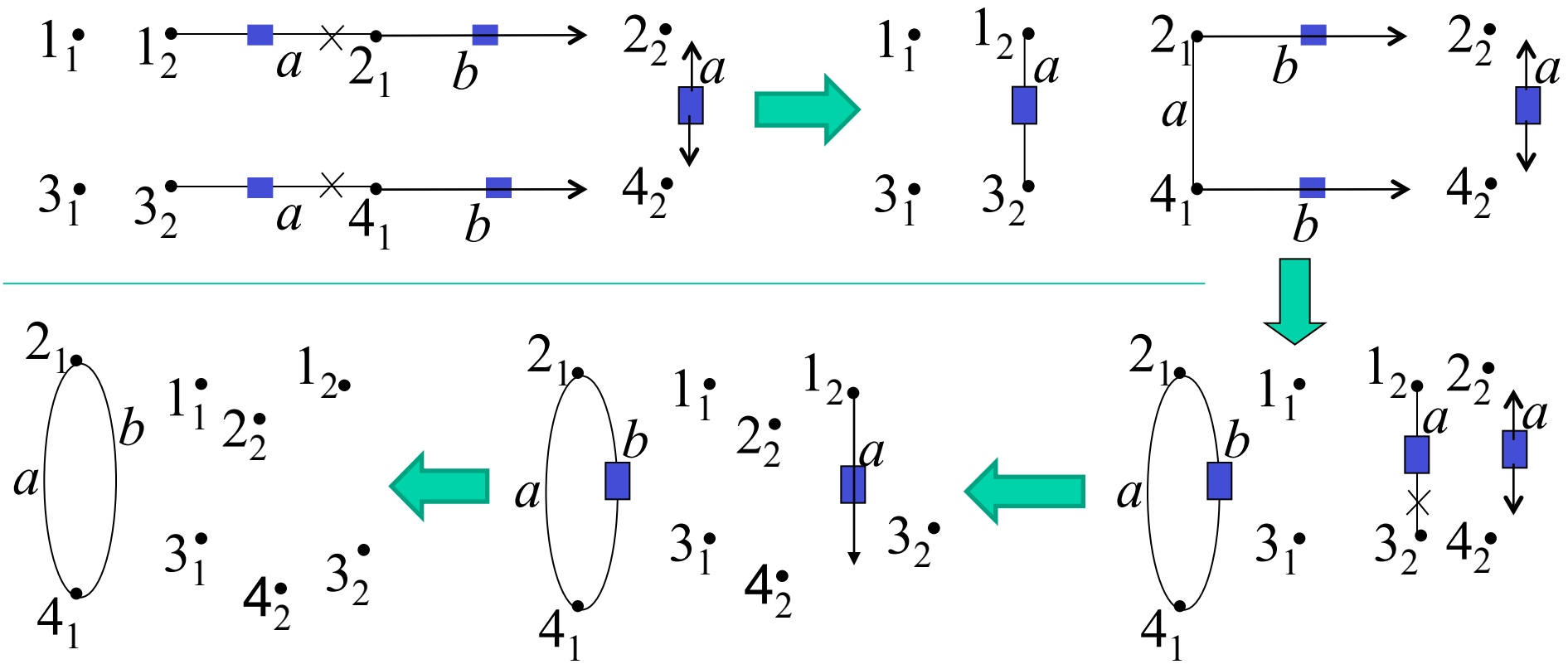
Напомним, приведение графа $a+b$ к финальному виду для последнего примера выглядит так:



Типы компонент: 1a, 1a, 2b. Сначала взаимодействуют 1a и 1a. Висячие ребра склеиваются блоковой склейкой, образуя нечетную цепь. Затем блоковая полуторная переклейка с присоединением 2b:



Таким образом, **самое короткое** приведение графа $a+b$ к финальному виду для последнего примера выглядит так:



Спасибо за внимание!