МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. Ломоносова
ФАКУЛЬТЕТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И КИБЕРНЕТИКИ

ISSN 2411-1473

Современные информационные технологии и ИТ-образование

Научный журнал

Том 2 (№ 11)

Москва
2015

УДК [004:377/378](063)
ББК 74.5(0)я431+74.6(0)я431+32.81(0)я431
С 56

Современные информационные технологии и ИТ-образование. Т. 2 (№ 11), 2015. - 614 с. (ISSN 2411-1473)

В данном выпуске журнала представлены доклады X Юбилейной международной научно-практической конференции «Современные информационные технологии и ИТ-образование», прошедшей в Московском государственном университете имени М.В. Ломоносова 20-22 ноября 2015 года.

Журнал «Современные информационные технологии и ИТ-образование» включен в наукометрическую базу «Российский индекс научного цитирования» с размещением полнотекстовых версий в научной электронной библиотеке eLIBRARY.RU. URL: http://elibrary.ru/title_about.asp?id=52785

Издание осуществлено при финансовой поддержке Российского фонда фундаментальных исследований (Грант РФФИ № 15-07-20760_г)

Учредитель:
Фонд содействия развитию интернет-медиа, ИТ-образования, человеческого потенциала «Лига интернет-медиа»

Издатель:
Фонд содействия развитию интернет-медиа, ИТ-образования, человеческого потенциала «Лига интернет-медиа»

Адрес редакции:
119991, г. Москва, ГСП-1, Ленинские горы, д. 1, стр. 52, факультет ВМК МГУ имени М.В. Ломоносова, каб. 375. E-mail: sukhomlin@mail.ru, тел./факс (495) 939-46-26.

Журнал зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор).
Свидетельство о регистрации средства массовой информации ПИ № ФС77-61433 от 10 апреля 2015 г.
Издается с 2005 года. Выходит 1 раз в год.

Горбунов К.Ю.[1], Любецкий В.А.[2]

[1]ИППИ РАН, г. Москва, к.ф.-м.н., в.н.с., gorbunov@iitp.ru

[2]ИППИ РАН, г. Москва, д.ф.-м.н., зав.лаб., lyubetsk@iitp.ru

РЕКОНСТРУКЦИЯ ПРЕДКОВЫХ ХРОМОСОМНЫХ СТРУКТУР

КЛЮЧЕВЫЕ СЛОВА

Хромосомная структура, эволюционное дерево, предковый вид, реконструкция вдоль дерева, эффективный алгоритм, булево линейное программирование.

АННОТАЦИЯ

Рассматривается задача реконструкции хромосомных структур в нелистовых вершинах данного филогенетического дерева на основе заданных структур в листьях. Структура может иметь паралоги – гены с одинаковыми именами. Рассматриваемая задача сводится к задаче булева линейного программирования, решаемой эффективным алгоритмом.

Определения и постановка задачи. Будем рассматривать реконструкцию хромосомных структур на основе модели хромосомной структуры как произвольного набора цепей и циклов, состоящих из ориентированных рёбер: генов, которым приписаны имена – натуральные числа *i*. На рис.1 изображены две структуры: *a* и *b*.

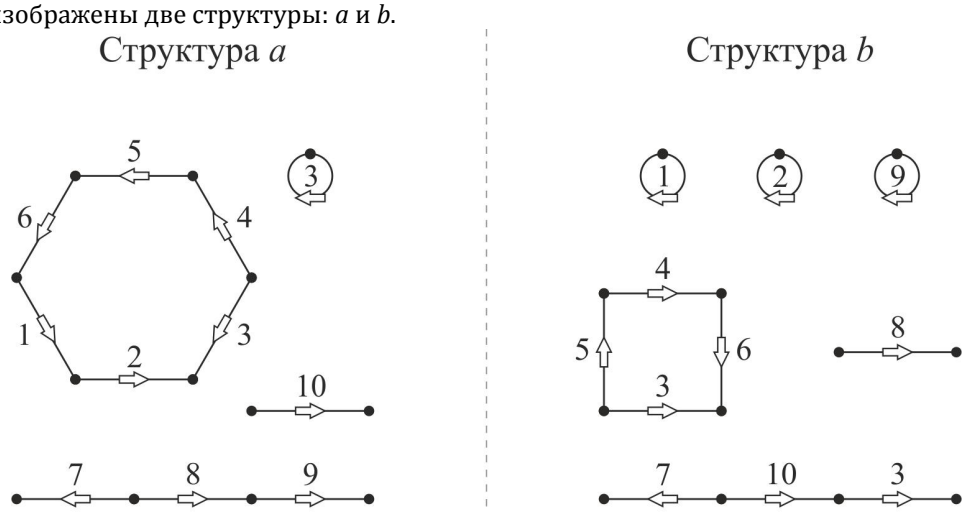

Рис.1. Две структуры a и b, каждая с двумя паралогами – генами, помеченными номером 3

Паралоги – гены из одной структуры, имеющие одинаковые номера *i*: для их идентификации применяется нумерация вида *i.j*, в которой число *i* назовём *первым номером* гена, а *j* – *вторым номером* (само выражение *i.j* будем называть *полным номером*; полные номера внутри одной хромосомной структуры не повторяются). В [1] мы рассматривали задачу реконструкции в предположении весьма существенного ограничения на модель – структуры должны иметь постоянный генный состав и не иметь паралогов. Здесь рассмотрим общий случай. Будем использовать понятие *брейкпоинтового расстояния* между двумя структурами. Для структур без паралогов (где есть лишь первые номера) оно определяется как число пар соответствующих краёв генов, соседних (или, как говорится в [1], *склеенных*) в одной структуре и не склеенных или отсутствующих в другой, сложенное с числом генов, присутствующих в одной структуре и отсутствующих в другой (под генами в разных структурах здесь понимаются гены с одинаковыми номерами).

Для структур с паралогами расстояние определяется как минимум расстояний между структурами, которые получаются установлением частичной биекции между паралогами каждого гена, представленного в обеих исходных структурах. Точнее, генов с первым номером *i* может быть разное число в структурах *a* и *b*, пусть это – соответственно множества $P_a(i)$ и $P_b(i)$; обозначим f_i

биекцию между частью множества $P_a(i)$ и частью множества $P_b(i)$; определяются *нумерации i.j* всех генов в $P_a(i)$ и $P_b(i)$ вторыми номерами, относительно которых f_i – тождественная функция; гены, не попавшие в области определения и значений f_i, имеют все разные вторые номера. Такая нумерация имеет естественную интерпретацию: если $y=f_i(x)$, то y «наследуется» от x; прочие гены из $P_a(i)$ и из $P_b(i)$ – самостоятельные, все различные между собой, поэтому они имеют различные значения j. Гены, не принадлежащие области определения или области значений f_i, можно считать *потерянными* и соответственно *возникшими* (на ребре, соединяющем a с b в филогенетическом дереве). Теперь *расстояние* между a и b определяется как минимум расстояний между a' и b', которые получаются из a и b при всевозможных нумерациях вторыми номерами всех генов в каждом множестве генов с одним и тем же первым номером (такую нумерацию будем называть добавлением вторых номеров). Например, для структур на рис.1 минимальное брейкпоинтовое расстояние достигается при нумерации в каждой структуре одного из паралогов гена 3 как 3.1 (неважно какого), а другого как 3.2. Оно равно 18.

Итак, дано корневое (возможно, небинарное) дерево, и в каждом листе своя хромосомная структура. Требуется расставить хромосомные структуры по нелистовым вершинам, чтобы сумма по всем рёбрам расстояний между структурами на концах ребра была минимальна. Покажем что это эквивалентно более «простому» требованию: расставить хромосомные структуры с нумерацией генов полными номерами по нелистовым вершинам и добавить вторые номера генам в листьях так, чтобы сумма по всем рёбрам расстояний между структурами на концах ребра (уже фактически без паралогов) была минимальна. Пусть найдена расстановка, минимизирующая описанную функцию F при первом требовании. Опишем расстановку, дающую не большее её значение для второго требования. Рассмотрим биекции f_i между генами на концах прикорневых рёбер, соответствующие определению функции F. Добавим генам в корне дерева вторые номера произвольным образом. Добавим вторые номера в некорневом конце каждого прикорневого ребра в соответствии с f_i. Таким образом, идя от корня, мы добавим вторые номера генам во всех вершинах дерева, включая листья. Поскольку сами биекции f_i не меняются, получаем расстановку для второго требования с тем же значением F. Обратное соответствие (от второго требования к первому) получается «стиранием» вторых номеров у всех генов. Таким образом, далее рассматриваем задачу в последней формулировке.

Сведение задачи реконструкции к задаче булевого линейного программирования. Легко видеть, что в искомой расстановке имеются лишь гены, полные номера которых присутствуют хотя бы в одном листе. Кроме того, нам удобно считать, что каждая (тождественная) биекция f_i, определяемая искомой расстановкой, определена на объединении M всех присутствующих в листьях номеров $i.j$, а отсутствующие в структуре гены специальным образом помечены. Очевидно, это определение эквивалентно исходному. «Началом» ребра в дереве считается та вершина, которая ближе к корню; аналогично определяется «конец» ребра.

Добавим произвольным образом генам в данных листовых структурах вторые номера; полученную нумерацию обозначим как I. Для перехода к задаче линейного булевого программирования введём переменные, линейные ограничения и линейную функцию, которую нужно минимизировать.

Введём переменные, указывающие соответствие между нумерацией I и искомой нумерацией генов в листьях. Для каждого листа e дерева и каждой упорядоченной пары $k.i$, $k.j$ элементов из M введём переменную z_{kije} (первый индекс будем для краткости опускать); эта переменная равна 1, если ген $k.i$ в e имеет в искомой нумерации номер $k.j$, иначе $z_{kije}=0$. Соответствие между генами с первым номером k, определённое переменными z_{ije}, должно быть биективным, что выражается линейными равенствами: для фиксированного i и e сумма по j значений z_{ije} равна 1, и аналогично для фиксированных j и e.

Введём переменные, определяющие хромосомные структуры в вершинах.

Для каждой вершины v и каждой неупорядоченной пары k, l различных краёв генов введём переменную x_{klv}; эта переменная равна 1 если эти края склеены в вершине v, иначе она равна 0. Каждый край может быть склеен не более чем с одним краем, что выражается линейными неравенствами: при фиксированных k и v сумма по l значений x_{klv} не превышает 1, и аналогично для фиксированных l и v. В листьях значения переменных x_{kl} заданы.

Для каждой вершины v и каждого гена k введём переменную y_{kv}; эта переменная равна 1 если ген k отсутствует в v; иначе она равна 0. Края отсутствующих генов ни с чем не склеены, что выражается линейными неравенствами $x_{ijv} \le 1 - y_{kv}$, где i или j – край гена k. Для каждого гена k и

каждого его края i вместо этих неравенств можно использовать одно неравенство $\sum_{j \neq i} x_{ijv} \leq 1 - y_{kv}$, где j пробегает все края генов в v. В листьях значения переменных y_k заданы.

Введём переменные, вычисляющие брейкпоинтовое расстояние между структурами, приписанными началу и концу ребра дерева. Пусть края k и l генов лежат в конце ребра e, а края m, n – в его начале. Назовём четвёрку k, l, m, n краёв генов *брейкпойнтовой*, если пара возможных склеек – склейка краёв k и l и склейка краёв m и n, – учитывается при определении брейкпойнтового расстояния на нелистовом e или может учитываться на листовом e после какого-нибудь (возможно, отличного от I) добавления вторых номеров генам в листьях (эта четвёрка – упорядоченная пара двух неупорядоченных пар: (k,l) и (m,n)).

Для каждого ребра e и каждой брейкпойнтовой четвёрки k, l, m, n краёв введём переменную s_{klmne}; эта переменная равна 1, если эти края принадлежат соответствующим генам (напомним: на листовом e соответствие определяется переменными типа z) и склеены в начале ребра и не склеены (или хотя бы один ген отсутствует) в конце ребра или наоборот. Это условие для s_{klmne} выражается линейными неравенствами; приведём их ниже для случая, когда e листовое, все четыре края – начала генов (или все – концы) и все четыре гена имеют одинаковые первые номера (другие случаи аналогичны): если e=(u,v), гены i_1 и j_1 имеют, соответственно, края k и m, а гены i_2 и j_2 – края l и n, то

$s_{klmne} \geq x_{klv} - x_{mnu} - (1 - z_{i1j1e}) - (1 - z_{i2j2e})$, $s_{klmne} \geq x_{mnu} - x_{klv} - (1 - z_{i1j1e}) - (1 - z_{i2j2e})$,

$s_{klmne} \geq x_{klv} - x_{mnu} - (1 - z_{i1j2e}) - (1 - z_{i2j1e})$, $s_{klmne} \geq x_{mnu} - x_{klv} - (1 - z_{i1j2e}) - (1 - z_{i2j1e})$.

Если правая часть неравенства равна 0, переменная s_{klmne} принимает значение 0 за счёт того, что она входит в качестве слагаемого в минимизируемую функцию.

Для каждого ребра e=(u,v) и каждых двух генов i и j, лежащих, соответственно, в конце и в начале ребра e, введём переменную s_{ije}; эта переменная равна 1, если эти гены соответствуют друг другу и один из них присутствует в структуре, а другой отсутствует. Это условие выражается (для листового e) линейными неравенствами $s_{ije} \geq y_{iv} - y_{ju} - (1 - z_{ije})$ и $s_{ije} \geq y_{ju} - y_{iv} - (1 - z_{ije})$.

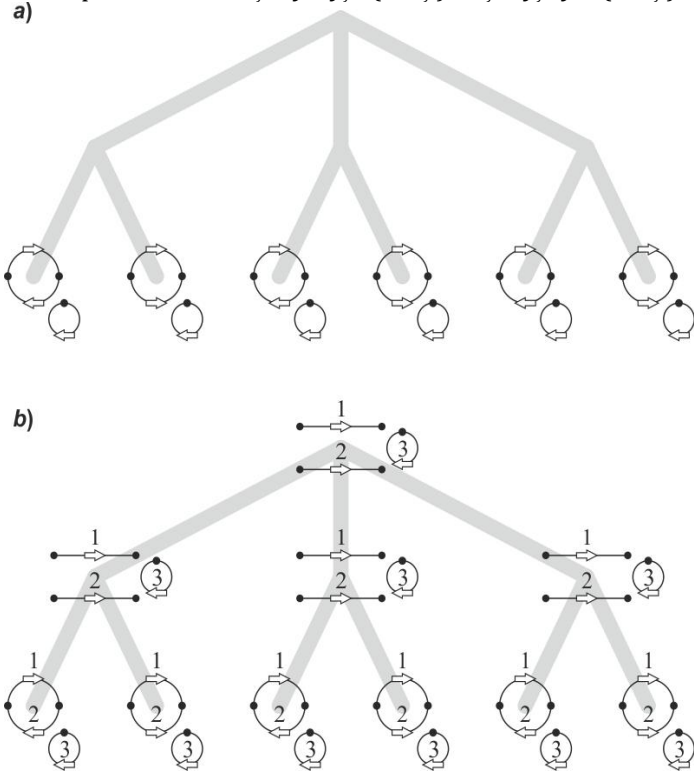

Рис.2. Пример реконструкции для искусственных данных. a) В каждом листе дерева дана хромосомная структура с тремя паралогами одного и того же гена. b) Реконструированные структуры, построенные алгоритмом в нелистовых вершинах, и добавленные вторые номера генов в листьях

Минимизируемая функция – сумма всех переменных s_{klmne} и s_{ije}, линейные ограничения описаны выше. Решение описанной задачи линейного булева программирования даёт искомую расстановку хромосомных структур по нелистовым вершинам дерева и вторые номера генам в

листьях.

Замечание 1. Описанное сведение задачи реконструкции хромосомных структур к задаче булевого линейного программирования легко обобщается на случай, когда стоимости брейкпоинтовых операций (переход от двух склеенных краёв к расклеенным или наоборот, и переход от присутствия гена к отсутствию или наоборот) различны. Для этого каждую переменную s_{klmne} следует заменить на две переменных $s1_{klmne}$ и $s2_{klmne}$. Для первой из них действуют линейные ограничения вида $s1_{klmne} \geq x_{klv} - x_{mnu} - \ldots$; для второй – ограничения вида $s2_{klmne} \geq x_{mnu} - x_{klv} - \ldots$. Аналогично, переменную s_{ije} следует заменить на $s1_{ije}$ и $s2_{ije}$. Теперь можно в минимизируемой функции умножать новые переменные на коэффициенты, отражающие веса событий.

Замечание 2. Задачу вычисления брейкпойнтового расстояния между двумя данными хромосомными структурами с паралогами можно рассматривать как частный случай задачи реконструкции. Здесь "дерево" состоит из одного ребра с данными структурами на концах. Очевидно, остаются лишь переменные типа z и s.

На рис.2*b* приведено решение задачи реконструкции хромосомных структур для искусственных данных, показанных на рис.2*a*. Все гены в данных листовых структурах имеют один и тот же первый номер 1 (не показан), в реконструированных структурах показаны лишь вторые номера.

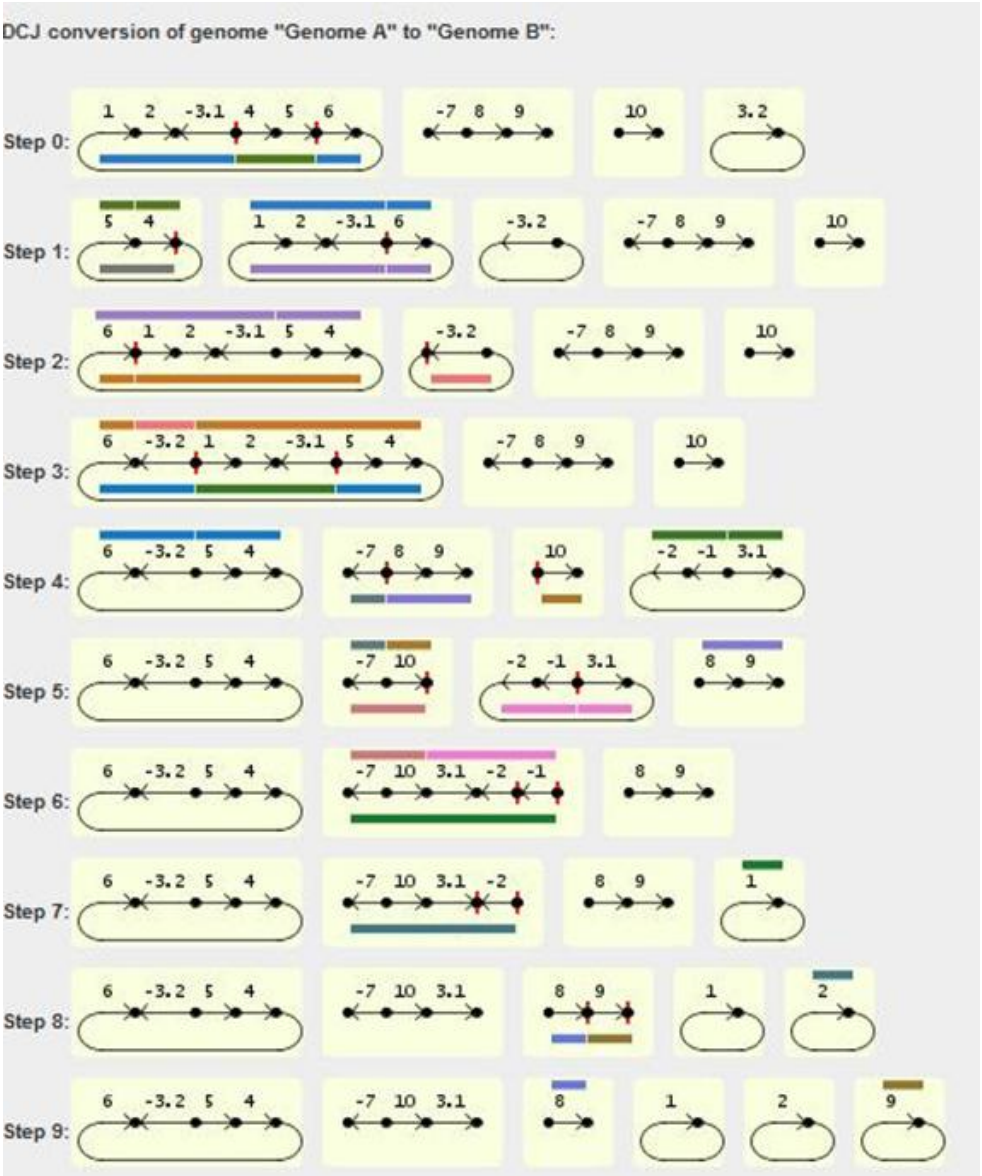

Рис.3. Кратчайшая последовательность операций, переводящая структуру a в структуру b (рис.1) для случая, когда верхний ген 3 в структуре a соответствует нижнему гену 3 в структуре b.

Обобщение понятия расстояния. Пусть, для простоты, даны две структуры с одинаковым набором генов. Брейкпоинтовое расстояние между ними можно рассматривать как минимальное число операций расклейки двух склеенных краёв генов или склейки двух свободных краёв, необходимых для перевода одной структуры в другую. Добавим к этим операциям следующие *операции* над хромосомной структурой: *двойная переклейка* – расклейка двух склеек краёв генов и переклейка четырёх краёв, приводящая к новой структуре и *полуторная переклейка* – расклейка двух склеенных краёв и склеивание одного края с каким-то несклеенным краем, второй край остаётся свободным. Получим обобщение брейкпойнтового расстояния, называемое *биологическим расстоянием*. Для случая двух хромосомных структур с одинаковым набором генов (без паралогов) реализация алгоритма вычисления биологического расстояния описана в [3]. Пример с двумя структурами на рис.1 показывает, что биологическое расстояние "чувствительнее" брейкпоинтового: теперь два возможных соответствия между паралогами гена 3 дают различные значения расстояния. Если верхний ген в структуре *a* соответствует нижнему гену в структуре *b*, перевести *a* в *b* можно самое меньшее за девять операций – рис.3; при альтернативном соответствии достаточно семи операций – рис.4 (рисунки получены программой, описанной в [3]).

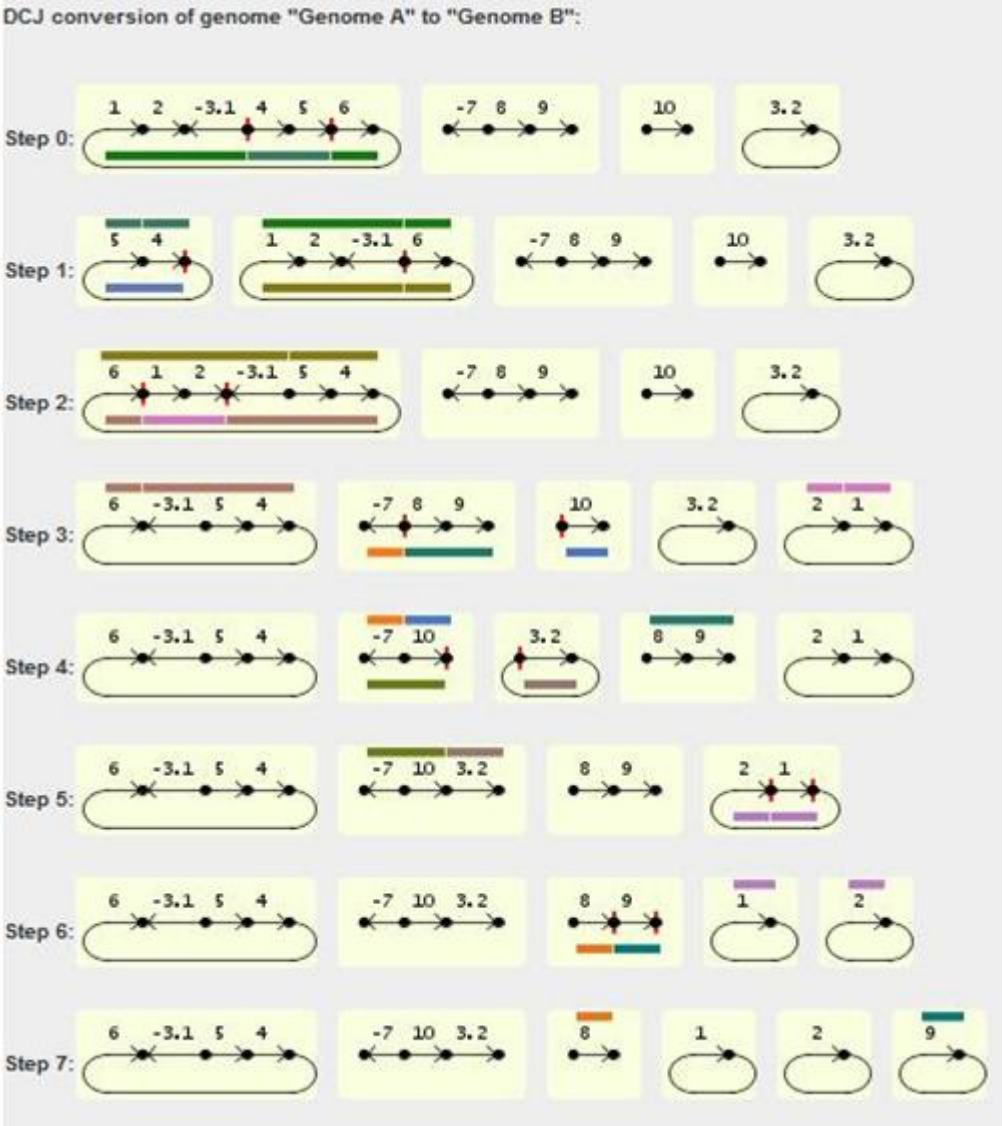

Рис.4. Кратчайшая последовательность операций, переводящая структуру a в структуру b (рис.1) для случая, когда верхний ген 3 в структуре a соответствует верхнему гену 3 в структуре b

Для биологического расстояния задачи его вычисления (в том числе для структур с различным набором генов и с паралогами) и соответствущей реконструкции хромосомных структур рассматриваются в [1, 2], а также в нескольких готовящихся к публикации работах

авторов. Реализации соответствующих алгоритмов доступны на сайте http://lab6.iitp.ru/ru/chromo/.

Работа выполнена за счёт гранта Российского научного фонда (проект № 14–50–00150).

Литература

1. Горбунов К.Ю., Гершгорин Р.А., Любецкий В.А. Перестройка и реконструкция хромосомных структур // *Молекулярная биология*, 2015, том 49, № 3, стр. 372–383.
2. Gershgorin R.A., Gorbunov K.Yu., Seliverstov A.V., Lyubetsky V.A. Evolution of Chromosome Structures // "Information Technology and Systems 2015" An IITP RAS Interdisciplinary Conference & School (ITaS'15), Sochi, Russia, Sep 7–11 2015, pp. 105–120.
3. Hilker R., Sickinger C., Pedersen C., and Stoye J. UniMoG – a unifying framework for genomic distance calculation and sorting based on DCJ // *Bioinformatics*, 2012, vol. 28, pp. 2509–2511.