ОТЗЫВ

члена диссертационного совета на диссертацию Дворкиной Татьяны Евгеньевны на тему «Разработка алгоритмов для анализа графов геномной сборки и геномных сборок» на соискание ученой степени кандидата физико-математических наук по специальности 1.5.8 – Математическая биология, биоинформатика.

Диссертация Т.Е.Дворкиной посвящена разработке новых алгоритмов для работы с графами сборки, а также проблемам автоматического поиска повторов высокого порядка (HOR, high order repeats) в области активного центромера. Сборки центромерных районов, обогащенных тандемными повторами (ТП) районов, начали появляться, но пока только для человека.

В первой главе описаны основные принципы работы нового алгоритма, которые закончились созданием программы SPAligner, предназначенной для извлечения потенциальных последовательностей кодирующих белки генов из графов метагеномной сборки. Сравнение общепринятых программ и SPAligner показало, что последняя успешно борется с фрагментацией генов, попавших на разные ветви графа (рис. 1.3). Во 2й главе подробно описана успешная работа конвейера ORFograph для поиска генов-кандидатов для инсектицидных белков. В этом случае проблема состояла в локализации генов на плазмидах. Таким образом, результаты работы новых программ не вызывают сомнений в их эффективности. Качество проделанной работы отражено в результатах, которые демонстрируют существенные преимущества разработанных алгоритмов и программ на их основе. О важности и новизне диссертации можно судить по информации, представленной в введении и заключении. Приведены примеры использования созданных программ в реальных проектах. Результаты работы представлены в пяти серьезных статьях, опубликованных в журналах, индексируемых в базах данных Web of Science Core Collection и Scopus. В четырех статьях-Т.Е.Дворкина – первый автор.

Третья часть посвящена анализу недавно сделанной сборки центромерного (ЦЕН) района хромосом человека. Видимо, эта задача появилась в связи с недавно осознанной научным и медицинским сообществом важности вариабельности ЦЕН для человеческой популяции. Создание автоматической программы для анализа ЦЕН – нетривиальная задача. Отдельные мономеры в составе поля ТП могут различаться по нуклеотидной последовательности заменами, делециями или инсерциями в несколько нуклеотидов. Различные варианты повторов высшего порядка (HOR, high order repeat) характерны для отдельных хромосом и могут формировать длинные блоки в составе одной хромосомы. Т.Е.Дворкина, как автор программы HORmon, ввела строгие ограничения – изучен только «живой» ЦЕН, т.е. область кинетохора без учета пери-ЦЕН областей, HORы в ЦЕН и вставки не ТП. Даже для такой строго ограниченной задачи, чтобы сделать ее решение автоматическим, понадобились изощренные методы информатики. Теперь, с созданием программы HORmon можно будет автоматически обрабатывать вновь прочитанные сиквенсы ЦЕН людей.

Взаимодействие информатиков и «мокрых» биологов нуждается в общем языке. Надо отметить некоторые недостатки диссертации в этом смысле

1. В первых 2х частях постоянно происходит путаница между понятиями «ген, кодирующий белковый продукт» и «аминокислотная последовательность». Например, стр.6 – «…предложенных методов к задаче извлечения белков непосредственно из графов сборки» или «выравнивание последовательности нуклеотидов или аминокислот на граф сборки». Как можно выровнять аминокислоты на нуклеотидный граф?
2. В разделе 3.1.2. речь идет об эволюции ЦЕН, дано определение. Однако, для «мокрого» биолога эволюционные отношения – это отношения между видами. В работе же речь идет лишь о вариабельность HORов в разных хромосомах внутри одной сборки. Отлично, что этот давно известный факт приобрел конкретное наполнение, однако никакого отношения к эволюции видов это не имеет.

3. Стр 70 «консенсусный мономер для всех последовательностей альфа-сателлитов в геноме человека»- у альфа сателлита 2 консеснсуса αI и αII. Коль скоро Вы занимаетесь «живым» ЦЕН, то в нем αI. Тогда не надо говорить «всех»

В порядке дискуссии хотелось бы задать автору несколько вопросов и высказать некоторые пожелания.

1. стр. 41 «Мы попытались использовать plasmidSPAdes для извлечения полных плазмидных последовательностей, и, как следствие, полных последовательностей генов. Было обнаружено, что у этого метода есть ограничения, которые не позволяют нам использовать его в нашем тестировании.» В чем именно состоят ограничения plasmidSPAdes для Вашей конкретной задачи?
2. Стр.46. Сначала нашли 419 потенциальных кандидатов в гены. Потом ослабили критерии отбора, а стало меньше - 232 кандидата. Почему так получилось?
3. Стр 64 – «36 подряд идущих символов пробелов, что указывает на присутствие в регионе LINE-элемента» – означает ли это, что LINE тоже тандемно организованы? Почему подряд? Или из рис.3.5 – после каждых 12 альфа, организованных в HOR, идет один LINE?

Следует отметить, что возникшие вопросы ни в коей мере не умаляют высокого качества диссертации.

Одна из самых больших проблем сборок геномов состоит в том, что прочитаны и собраны только эухроматические части генома, тогда как гетерохроматиновые (ГХ) области остаются несобранными. Прогресс в секвенировании не помогает решить задачу, проблема кроется в собственно составе ГХ и отсутствии методов сборки для районов с высоко повторяющимися последовательностями. Основным компонентом ГХ являются тандемные повторы (ТП) (сателлитная ДНК). ТП представляют собой класс ДНК, который появляется у эукариот, но отсутствует у прокариот. Становится ясно, что глобальная организация ядра зависит именно от ассоциации ТП. Очень хочется надеяться, что замечательные метода анализа ТП и их полей, которые разрабатывают в группе Павла Певзнера и, в частности, настоящая работа, помогут в дальнейшем собрать пери-ЦЕН поля αII, а может быть даже и поля больших сателлитов (HS1-4) и выяснить особенности их организации.

Диссертация Дворкиной Татьяны Евгеньевны на тему «Разработка алгоритмов для анализа графов геномной сборки и геномных сборок**»** соответствует основным требованиям, установленным Приказом от 01.09.2016 № 6821/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель Дворкина Татьяна Евгеньевна заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 1.5.8 – Математическая биология, биоинформатика. Пункты 9 и 11 указанного Порядка диссертантом не нарушен.

Член диссертационного совета
доктор биологических наук, проф., внс Института Цитологии РАН

13 июля 2023 года _____ Подгорная О.И.