УДК 004.89

Мельников Б.Ф.[1] , Пивнева С.В.[2] , Трифонов М.А.[2]

[1] Центр информационных технологий и систем органов исполнительной власти, г. Москва, Россия
[2] Тольяттинский государственный университет, г. Тольятти, Россия

МУЛЬТИЭВРИСТИЧЕСКИЙ ПОДХОД К СРАВНЕНИЮ КАЧЕСТВА ОПРЕДЕЛЯЕМЫХ МЕТРИК НА МНОЖЕСТВЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК

Аннотация

В настоящей статье анализируется несколько метрик, определяющих различия в последовательностях ДНК разных видов. Рассматриваются несколько стандартных метрик, а также модификация оригинальной авторской метрики, предыдущие версии которой рассматривались в наших прежних публикациях. При определении качества метрик мы исходим из предположения о том, что для любых трёх далёких видов вычисляемые по этой метрике расстояния между ними должны образовывать треугольник, близкий к равнобедренному остроугольному. Мы считаем несколько вариантов отклонения треугольника от равнобедренного остроугольного, после чего считаем сумму таких отклонений для всех получающихся треугольников. На основании проведённых подсчётов делаем вывод о качестве первоначальных метрик. После этих вычислений применяем полученную методику к рассмотрению этих же метрик для близких видов (человекообразных обезьян и человека) - и на этих близких видах получаем немного иные результаты сравнительного анализа рассматриваемых метрик.

Ключевые слова

Мультиэвристический подход; метрика на множестве последовательностей ДНК; алгоритмы.

Melnikov B.F.[1], Pivneva S.V.[2], Trifonov M.A. [2]

[1] Center of Information Technologies and Systems for Executive Power Authorities, Moscow, Russia
[2] Togliatti State University, Togliatti, Russia

MULTIHEURISTIC APPROACH TO COMPARE THE QUALITY OF DEFINED METRICS ON THE SET OF DNA SEQUENCES

Abstract

In this article, we analyzed some several metrics that determine the differences in DNA sequences of different species. Several standard metrics are considered, as well as a modification of the original author's metric, the previous versions of which were considered in our previous publications. Determining the quality of some several metrics, we proceed from the assumption that for any three distant species, the distances between them computed from this metric should form a triangle close to an isosceles acute-angled triangle. We consider several variants of the deviation of a triangle from an isosceles acute-angled triangle, and then we consider the sum of such deviations for all the resulting triangles. Based on these calculations, we make a conclusion about the quality of the original metrics. After these calculations, we apply the obtained technique to the consideration of the same metrics for close species (anthropoid and human), and on these closely related species, we obtain slightly different results of a comparative analysis of the metrics under consideration.

Keywords

Multiheuristic approach, metrics on the set of DNA sequences, algorithms.

Введение

Задача определения схожести ДНК является частным случаем задачи неточного сопоставления последовательностей[1]. «Неточность» заключается в том, что при сравнении строк имеется возможность распознать схожие последовательности – даже несмотря на возможные ошибки и искажения в них, например, изменение, удаление или вставку нескольких символов. Количество таких искажений задаёт метрику на множестве строк, которая определяется по минимальному количеству

операций редактирования, позволяющих получить из одной строки другую. Эта задача встречается во многих областях. Например, сравнение генов, хромосом и белков является одной из важнейших задач и одним из основных инструментов молекулярной биологии и биоинформатики [1,2,3,4,5,6,7]. Точное сравнение цепочек нуклеотидов здесь неприемлемо из-за наличия ошибок в данных, а также из-за возможных мутаций. Неточное сопоставление осуществляется так же, как и при обработке обычного текста. Одна из метрик, получаемая при сравнении слов – расстояние Левенштейна – используется для исправления ошибок, для повышения качества распознавания отсканированных документов, для поиска в информационных системах и базах данных [1]. Для нахождения приближённого решения существуют различные алгоритмы в разных предметных областях, например, для поиска в базах данных генетической информации широко применяется алгоритм BLAST ([2] и др.), аппроксимирующий алгоритм Нидлмана-Вунша.

В разделе 1 данной статьи описывается применение к задаче определения схожести последовательностей ДНК т.н. мультиэвристического подхода [8,9], фактически являющегося развитием метода ветвей и границ. Отметим, что ранее, до наших работ, метод ветвей и границ к решению данной задачи, по-видимому, не применялся.

Итак, расчёт расстояния (метрики) между строками ДНК различных видов организмов является одной из важнейших современных задач биоинформатики. Как уже было отмечено, на сегодняшний день существует множество алгоритмов, позволяющих произвести приближенный расчёт за полиномиальное время ([4,5,6,7,10] и мн. др.). Очевидным недостатком при расчёте расстояния между одной и той же парой строк ДНК является получение различных результатов при использовании различных алгоритмов для расчёта метрик. Однако авторам неизвестны работы, в которых сравнивались бы различные алгоритмы решения данной задачи. В связи с этим одна из задач, рассматриваемых в настоящей статье, заключалась в разработке метода сравнительной оценки таких алгоритмов. Более того, эта задача, по-видимому, является наиболее важной из рассматриваемых. В результате авторами был предложен метод оценки с использованием свойств равнобедренного треугольника в метрическом пространстве (раздел 2, т. н. «треугольная норма», по результатам которой считается т. н. badness, относящаяся к некоторой метрике для нескольких рассматриваемых видов).

В статье также рассматриваются варианты улучшения уже имеющихся метрик. При этом ни для одного из рассматриваемых методов построения расстояний между цепочками ДНК не является недостатком использование его для оценки расстояний в двух разных ситуациях: как между ближними видами (пары «человек – шимпанзе» и «человек – бонобо» и т.п.), так и между более дальними (пары «человек – крокодил» и «шимпанзе – крокодил» и т.п.). Это объясняется тем, что мы рассматриваем, в первую очередь, углы треугольников в евклидовом пространстве. Однако некоторые вычислительные эксперименты, связанные с применением для преобразования метрик непрерывных монотонных функций, нами проделаны были (раздел 3).

Краткие результаты вычислительных экспериментов над геномами 100 видов приведены в разделе 4. Среди этих результатов стоит отметить следующее. Во-первых, для «далёких» видов badness очень мала – что свидетельствует об удачном выборе наших подходов и соответствующих конкретных алгоритмов; при этом данный факт верен для нескольких разных норм. Во-вторых, (также для «далёких» видов) предложенный нами подход к определению метрики даёт наилучшие результаты (по всем рассматриваемым «треугольным» нормам) – среди 5 рассматривавшихся метрик [4,5,6,7,10]. Для «близких» видов (человек и человекообразные обезьяны) результаты несколько хуже (увеличивается значение badness, и, кроме того, наш вариант метрики даёт 2-й по качеству результат). В-третьих, вряд ли какая-нибудь из этих метрик приемлема для определения расстояния между подвидами: так, при применении данных алгоритмов к человеческим расам иногда даже возникает нарушение неравенства треугольника. Точные объяснения последних фактов, по-видимому, должны приводить специалисты-биологи, однако мы ниже также пытаемся объяснить их, с нашей точки зрения.

Возможные направления дальнейшей работы, уже ведущиеся нашей группой в настоящее время, кратко описаны в заключении (раздел 5).

1. Алгоритм определения расстояний между последовательностями нуклеотидов на основе мультиэвристического подхода

Как уже было сказано ранее, мультиэвристический подход к задачам дискретной оптимизации был нами рассмотрен в [8,9] и многих других последующих публикациях. В этом разделе описын вариант его применения к задаче определения схожести последовательностей ДНК. Для решения данной задачи он был применён следующим образом[1].

1 Здесь мы немного изменили описание алгоритма, приведённое в [10]. Авторы готовы выслать исходный код программы при запросе по электронной почте.

Пусть x, y– исходные строки, i, j – индексы символов строк x и y соответственно, r – значение метрики, которое требуется найти. Под сдвигом строки будем понимать увеличение на единицу соответствующего индекса. Общую схему алгоритма можно описать следующим образом.

Вход: Строки x и y.
Шаг 1: i := 0, j := 0, r := 0;
Шаг 2: if *x[i]* = *y[j]* then begin
сдвигаем обе строки;
r :+ стоимость совпадения символов *x[i]* и *y[j]*;
end
elsebegin
применяем эвристики для генерации возможных «траекторий» сдвига в позиции *i'* и *j'*, таких что *x[i']* = *y[j']*;
оцениваем их с помощью других эвристик;
усредняем полученные оценки, используя функцию риска;
осуществляем сдвиг (при этом может измениться значение *r*);
end;
Шаг 3: повторяем второй шаг до тех пор, пока не достигнут конец одной из строк.

Мы считаем «стоимость» совпадения двух символов, которую в простейшем случае полагаем равной 1; для ДНК можно определять её с помощью матрицы весов аминокислотных замен BLOSUM [1,2,11] или какой-либо иной подобной матрицы.

В процессе разработки и реализации алгоритма нами были использованы следующие вспомогательные эвристики.

Выбираем траектории, для которых выражение (i' – i) + (j' – j) принимает минимальное либо близкое к минимальному значение. Например, сначала мы рассматриваем все траектории со сдвигом только одной из строк на один символ; затем – со сдвигом одной из строк на два символа или обеих на один символ; и т.д.

Сдвигаем ту строку, текущий символ которой реже встречается в другой строке. Для этой эвристики желательно знать вероятности появления символов в каждой из строк. Если они заранее неизвестны, мы считаем их равными. В процессе работы мы динамически корректируем вероятности или (в простой версии алгоритма) используем т.н. алгоритм старения [12], чтобы вероятность символа определялась по некоторому фрагменту перед текущим символом, а не по целой строке. Если вероятности для текущих символов окажутся равными, сдвигается строка, в которой осталось больше символов.

Комбинация двух предыдущих эвристик: результирующая оценка складывается как её оценки первой и второй эвристиками. При этом для определения оценки второй эвристики суммируются вероятности появления в другой строке для всех символов, которые придётся пропустить при сдвиге.

Используем алгоритм для поиска наибольшей общей подпоследовательности строк x[i..i+k] и y[j..j+k], где эмпирически выбрано значение k ~ 15. Для сдвига выбираем такие индексы i', j', в которых заканчивается наибольшая общая подпоследовательность. Если не будет найдено ни одной пары одинаковых символов, область поиска увеличивается. Отметим, что при использовании этой эвристики результат будет близок к значению наибольшей общей подпоследовательности.

Комбинация третьей и четвёртой эвристик: оценка ситуации складывается из её оценок обеими эвристиками. Оценка ситуации (i', j') четвёртой эвристикой является отношением длины наибольшей общей подпоследовательности строк x[i..i'] и y[j..j'] к средней длине сдвига строк из позиции (i, j) в позицию (i', j').

Используем алгоритм Нидлмана-Вунша [1] для подстрок x[i..i+k] и y[j..j+k], где k ~ 15. Сдвигаем строки в позицию (i', j'), для которой соответствующее значение в таблице алгоритма Нидлмана-Вунша является наибольшим.

Комбинация третьей и шестой эвристик: оценка ситуации складывается из её оценок этими обеими эвристиками. Оценка ситуации (i', j') шестой эвристикой является отношением значения в таблице алгоритма Нидлмана-Вунша, соответствующего этой позиции, к средней длине сдвига строк из позиции (i, j) в позицию (i', j').

При этом мы практически не отмечали специально применение функций риска (см. [13], а варианты их применения конкретно в задачах дискретной оптимизации – в [14]). Они применяются как вспомогательные эвристики практически в каждом из упомянутых здесь пунктов, несколько более подробно см. ниже.

Реализация алгоритма дала приемлемые результаты – однако, как уже было отмечено во введении, подробное обсуждение результатов вычислительных экспериментов приведено далее.

2. Варианты «треугольной» нормы определения качества метрики расстояния

Итак, существуют различные алгоритмы определения расстояний между геномами – их можно назвать алгоритмами определения метрики на множестве геномов. Однако при этом возникают не только обычные вопросы об адекватности соответствующих математических моделей (которые, с точки зрения авторов, в данной предметной области обычно решаются специалистами-биологами, [15] и др.), но и о сравнительной оценке этих моделей. Самый важный вопрос при этом, по-видимому, такой: можно ли говорить об эффективности таких алгоритмов и об адекватности этих моделей на

основе только лишь одного анализа матриц близости (расстояния) между геномами без привлечения специалистов-биологов? Авторы данной статьи считают, что на этот вопрос следует ответить положительно.

Для нескольких разных алгоритмов [4,5,6,7,10] рассматриваем матрицы расстояний между соответствующими геномами; в наших вычислительных экспериментах мы применяли 5 различных алгоритмов[2] и строили соответствующие матрицы расстояний, в которых количество геномов достигало 100.

При этом использовали следующее естественное соображение (аналога которому в литературе мы не нашли). Приведём его на примере человека (Ч), шимпанзе (Ш) и бонобо (Б). Согласно данным биологов, Ш и Б разошлись (имели общего предка), по разным оценкам, около 2–2.5 млн. лет назад(недаром альтернативное название Б –«карликовый Ш», [16]), а Ч с ними обоими – 5.5–7 млн. лет назад[3]. В связи с этим возникает вопрос: почему Ч должен быть ближе к Б чем к Ш? Или наоборот – почему он должен быть ближе к Ш чем к Б? Очевидно, что ответ на оба этих вопроса отрицательный, т.е., иными словами, объяснения большей близости существовать не может. Поэтому в матрице расстояний между геномами все получаемые треугольники в идеале должны быть остроугольными равнобедренными.

Для сравнения качества алгоритмов построения расстояний было предложено несколько вариантов «отхода» (т.н. badness) от этих «вытянутых равнобедренных» треугольников. По-видимому, при расчёте badness всей матрицы для каждого варианта всегда нужно суммировать соответствующие badness по всем возможным треугольникам рассматриваемых матриц – что мы и делаем в данной работе[4].

Итак, мы в простых случаях[5] будем считать badness (норму) всей матрицы расстояний суммой, а для badness каждого треугольника будем применять один из следующих 4 вариантов. (Всюду считаем, что в рассматриваемом треугольнике стороны – a, b и c, причёма ≥ b ≥ c; углы – α, βи γ, причёмα ≥ β ≥ γ.)

(α–β) / π.

(α–β) / α.

(a–b) / a.

В последней норме «нарушение равнобедренности» и «нарушение остроугольности» рассмотрим отдельно:

(A) 1 – min (b/a, c/b) ;

(B) max (a–π/3, 0) / (2π/3) ;

общий ответ –(A+B) / 2 .

При этом максимальное значения badness (в каждом из этих 4 случаев) для некоторого треугольника может быть равно 1. В самом же плохом случае работы алгоритмов построения метрики – т.е. при возникающем нарушении неравенства треугольника – мы полагаем это значение равным от 1 до 2 (также в зависимости от количественных характеристик этого нарушения).

3. Специальные версии нормализации («препроцессорные» вычисления)

В этом разделе рассмотрим ещё одну эвристику, которую можно считать дополнительной для всех эвристик «нарушения остроугольной равнобедренности». Для неё рассмотрим некоторую функцию вида f(x)=xα, где значение α (обычно 0<α<1) выбирается своим для каждой рассматриваемой матрицы расстояний между геномами. При этом каждый из элементов x матрицы расстояний заменяется на f(x).

Для выбора конкретных значений α, улучшающих, с нашей точки зрения, качество выбора метрик, применялось следующее соображение. Ниже, при описании результатов вычислительных экспериментов, будет показано, что различные эвристики выбора метрик находятся в относительно разном приоритете для геномов «далёких» и «близких» видов; причём стоит отметить, что такой приоритет мало изменяется при его исследовании по различным нормам, описанным выше. Попытки улучшить значения этих норм (badness) с помощью применения какой-либо указанной выше функции вида f(x)=xα оказываются неудачными: решения соответствующих минимизационных задач дают либо максимальное, либо минимальное значение α (среди допускаемых нами). Несложно понять, что именно в этом случае получаемые в матрице расстояний между геномами треугольники «наиболее близки к остроугольным равнобедренным». Поэтому если действительно пытаться улучшить качество метрик, то надо воспользоваться принципиально иной эвристикой. Для этого мы пытались найти такую функцию указанного выше вида, при которой набор значений расстояний матрицы, рассматриваемый как распределение некоторой случайной величины, получается как можно более близким к равномерному распределению[6]. Заранее отметим,

2 Специально отметим ещё раз, что среди этих алгоритмов имеется один наш, оригинальный.

3 Очень важно отметить, что *точные* значения времени в такой модели не являются важными!

4 Интересно отметить следующее обстоятельство. *Специально, искусственно* построить таблицу расстояний размерности NxN, в которой все образуемые треугольники были бы остроугольными равнобедренными, вряд ли возможно без применения моделей, связанных с (N–1)-мерным пространством. (См. на http://habrahabr.ru/post/238107/ решение задачи, в которой применяются подобные алгоритмы.)

5 Отметим заранее, что мы иногда рассматриваем и несколько более сложные варианты, которые, однако, в настоящей статье не описаны.

6 Что неформально можно объяснить, например, следующим образом. Мы *уже* знаем, что в нашей модели геномы человека

что для разных задач (т.е. для разных конкретных матриц расстояний) значения величины α, получаемых псевдооптимальными алгоритмами реального времени (реализуемыми согласно описанному нами в [8,9,14] и др.) получаются разными.

Целевую функцию при этом выбирали на основе метода максимизации энтропии ([17] и мн. др.). Конкретные результаты, связанные с применением данной эвристики, приведены далее.

4. Некоторые результаты вычислений

Далее мы будем называть:

наш оригинальный алгоритм построения метрики между геномами – первым (ниже – №1, см. [10]);

вторым – один из алгоритмов M. van der Loo и др. (ниже – № 2, см. [5], используемая функция – jarowinkler());

третьим – другой алгоритм M. van der Loo и др. (ниже – № 3, также см. [5], используемая функция – stringdist());

четвёртым – один из алгоритмов H. Pages и др. (ниже – № 4, см. [6], используемая функция – stringDist());

пятым – другой алгоритм H. Pages и др. (ниже – № 5, также см. [6], используемая функция – pairwiseAlignment()).

Отметим, что алгоритмы № 4 и № 5 являются «несимметричными» алгоритмами, и при заполнении матрицы расстояний мы использовали полусуммы двух получаемых значений. Также отметим, что случаи нарушения неравенства треугольника были зафиксированы только в результате работы алгоритмов № 4 и № 5 – однако в случае «далёких» видов подобных результатов было совсем немного: примерно 1 случай на 2000 рассматриваемых потенциальных треугольников.

Для дальнейших подсчётов, во-первых, более-менее случайно выбрали геномы 100 представителей видов, приведённые в [18] (случай рассмотрения «далёких» видов)[7]. Некоторые результаты вычислительных экспериментов (таблица 100x100, всего 100·99/2=4950 значений, образующих (100·99·98) / (2·3) = 161700 треугольников) приведены далее в табл. 1, в которой:

по строкам – номера алгоритмов (как указано выше);

по столбцам: примерное время работы создания матрицы (для получения всех 4950 значений, тактовая частота процессора ~2 ГГц); число нарушений неравенства треугольника (в промилле, т.е. в среднем на 1000 запусков); средняя badness – вычисленная для каждого из алгоритмов 1–4 расчёта badness каждого треугольника.

Все значения badness приводим с точностью до 3 десятичных знаков (время работы алгоритмов построения матриц фиксировалось менее точно). Во всех таблицах мы отмечали наилучшую метрику по рассматриваемой норме (она выделена дважды) и вторую по значению (выделена жирным шрифтом).

Как мы видим, алгоритм, реализованный нашей группой, по большинству норм оказывается оптимальным. При этом очень важно отметить, что эвристики, применявшиеся для создания этого алгоритма, не имели совершенно никакой связи с эвристиками, применявшимися для описания норм.

Во-вторых (случай рассмотрения «близких» видов), мы также случайно выбрали геномы человека и человекообразных обезьян (бонобо, шимпанзе, горилла, орангутанг, гиббон), также приведённые в [18]. При этом у каждого вида мы брали по 4-5 представителей (всего выбрали 28 геномов), причём у человека взяли геномы представителей различных рас. Некоторые результаты вычислительных экспериментов приведены далее в табл. 2, в которой, в отличие от табл. 1, не приведено время построения. Кроме того, в связи с небольшим общим числом треугольников (менее 5000) привели число нарушений неравенства треугольника (а не относительные значения этой величины).

Как видим, существенно возрастает относительное число нарушений неравенства треугольника. Кроме того, наша оригинальная метрика расстояний между геномами теперь не является оптимальной.

В-третьих, мы применили «препроцессорные» алгоритмы, согласно описанному ранее методу. Стоит отметить, что при применении этих вспомогательных алгоритмов практически во всех клетках уменьшилось значение badness – что, однако, не было целью применения этого алгоритма. Кроме того, «немного поменялись лидеры», т.е. наш алгоритм построения метрики (строка № 1) показывает более хорошие результаты (чем без применения вспомогательных алгоритмов). Впрочем, последний факт как раз и можно объяснить «настройкой» алгоритма № 1 на его применения для большего диапазона значений. Результаты вычислительных экспериментов приведены в табл. 3.

И, в-четвёртых, мы применили те же самые алгоритмы к геномам представителей

(Ч), шимпанзе (Ш) и крокодила (К) образуют «вытянутый» остроугольный треугольник, близкий к равнобедренному. При этом *точные* значения длин Ч-К и Ш-К вряд ли представляют интерес – важно то, что они примерно равны. Также вряд ли представляет интерес отношение длины Ч-Ш к длине Ч-К.

7 Все конкретные списки видов, соответствующие им геномы, взятые, в первую очередь, с сайта [18], получаемые значения матриц расстояний, а также исходные коды программ авторы готовы выслать (при вашем запросе) по электронной почте. Мы также готовы выслать подробные результаты вычисления badness, включающие не только усреднённые, но и все получаемые в процессе работы значения.

человеческих рас (белый, жёлтый, негр, бушмен, австралиец). При этом у каждой расы брали по 3-4 представителя (всего 18 геномов). Некоторые результаты вычислительных экспериментов приведены далее в табл. 4, в которой смысл столбцов аналогичен столбцам табл. 2. К значениям, приведённым в табл. 4, можно было бы сделать очень много комментариев; приведём только самый важный из них. Относительно большое число нарушений неравенство треугольника (и, что связано с последним, существенно бóльшие значения badness, при её подсчёте по любой из норм), по-видимому, связано с большим числом скрещиванием конкретных людей уже после разделения человечества на расы. Т.е., по-видимому, к подвидам данные алгоритмы (без дальнейшей их модификации) применять не стоит.

Таблица 1. «Далёкие» виды

№	время (ч)	нарушения	**badness-1, (α–β) / π**	**badness-2, (α–β) / α**	**badness-3, (a–b) / a**	**badness-4, (A+B) / 2**
1	27	0	0,0372	0,0822	0,0416	0,196
2	2.1	0	0,0954	0,197	0,0926	0,252
3	2.3	0	0,345	0,476	0,163	0,468
4	28	0.37	0,0416	0,0907	0,0469	0,176
5	28	0.38	0,0549	0,116	0,0556	0,214

Таблица 2. «Близкие» виды

№	нарушения	**badness-1, (α–β) / π**	**badness-2, (α–β) / α**	**badness-3, (a–b) / a**	**badness-4, (A+B) / 2**
1	0	0,0757	0,152	0,0645	0,364
2	1	0,0333	0,0687	0,0302	0,215
3	1	0,514	0,622	0,170	0,582
4	32	0,0595	0,122	0,0496	0,341
5	39	0,0741	0,151	0,0615	0,350

Таблица 3. «Близкие» виды (после предварительного применения «препроцессорного» алгоритма)

№	нарушения	**badness-1, (α–β) / π**	**badness-2, (α–β) / α**	**badness-3, (a–b) / a**	**badness-4, (A+B) / 2**
1	0	0,0522	0,121	0,0527	0,351
2	0	0,0314	0,0692	0,0290	0,205
3	0	0,501	0,600	0,154	0,580
4	12	0,0527	0,122	0,0482	0,323
5	14	0,0732	0,150	0,0608	0,320

Таблица 4. Расы человека

№	нарушения	**badness-1, (α–β) / π**	**badness-2, (α–β) / α**	**badness-3, (a–b) / a**	**badness-4, (A+B) / 2**
1	17	0,140	0,243	0,0924	0,325
2	29	0,119	0,173	0,0359	0,342
3	30	0,420	0,527	0,187	0,493
4	30	0,119	0,218	0,0880	0,313
5	26	0,129	0,229	0,0881	0,306

Однако, несмотря на последний факт, в целом представляется перспективным как дальнейшее улучшение нашего оригинального алгоритма построения метрики, так и предлагаемого нами подхода к описанию норм. Кроме того, в разных ситуациях могут оказаться относительно более адекватными разные алгоритмы построения метрик.

5. Заключение. Направления дальнейшей работы

В данном разделе рассмотрим некоторые алгоритмы, которые собираемся опубликовать в последующих статьях.

Во-первых, авторская метрика на ДНК-последовательностях «в соревновании с другими метриками» показала весьма приемлемые результаты. По-видимому, это является достаточным основанием по продолжению работ, связанных с её улучшением.

В качестве возможной связи между нашими подходами к решению задач биокибернетики и задачи коммивояжёра (прежде всего – т. н. её псевдогеометрической версии, [8,19] и др.) можно назвать не только упомянутый выше мультиэвристический подход к задачам дискретной оптимизации, но и т.н. алгоритмы псевдоразмещения точек в k-мерном евклидовом пространстве[19]. Эти вспомогательные

алгоритмы улучшают работу других рассматриваемых нами алгоритмов. Подобными вспомогательными алгоритмами, также применяющимися для улучшения описанных в данной статье основных алгоритмов, можно считать алгоритмы, аналогичные применявшимися нами в [13,14]. К этому направлению можно отнести описанное в этих статьях применение функций риска, а также специальные варианты применения широко известного «правила 3 сигм».

Кроме того, одной из часто рассматриваемых в биокибернетике задач является задача восстановления матрицы расстояний, когда известно только часть заполненных элементов [11,20]. Нами – с помощью той же самой «треугольной нормы» – предлагается оригинальный алгоритм подобного восстановления.

Информация о финансовой поддержке

Работа выполнена при поддержке Российского фонда фундаментальных исследований, соглашение по проекту №16-47-630829.

Литература

1. Гасфилд, Д. (2003), Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология, СПб: Невский Диалект, БХВ-Петербург, 654 с.
2. Toppi, J., De VicoFallani, F.,Petti, M., Vecchiato, G., Maglione, A. G., Cincotti, F., Salinari, S., Mattia, D., Babiloni, F., Astolfi, L. (2013), "A new statistical approach for the extraction of adjacency matrix from effective connectivity networks", "IEEE Engineering in Medicine and Biology Society (EMBC)", No 3-7, pp. 2932-2935.
3. Torshin, I. Yu. (2006), "Bioinformatics in the Post-Genomic Era: The Role of Biophysics", Nova Biomedical Books, NY,ISBN 1-60021-048-1.
4. Winkler, W. E. (1990), String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, Proceedings of the Survey Research Methods Sections, American Statistical Association, pp. 354-359.
5. Van der Loo, M. P. J. (2014), "The Stringdist Package for Approximate String Matching",The R Journal, vol. 6, pp. 111-122.
6. Pages, H., Aboyoun, P., Gentleman, R., DebRaoy, S. (2009), "Biostrings: String Objects Representing Biological Sequences and Matching Algorithms", R package version 2.10.1.
7. Morgan, M., Lawrence, M. (2009), "ShortRead: Base classes and methods for high-throughput short-read sequencing data", R package version 1.0.6.
8. Melnikov, B. F., (2006) "Multiheuristic approach to discrete optimization problems", Cybernetics and Systems Analysis, No. 3, pp. 335-341.
9. Melnikov, B.F. (2005), "Discrete optimization problems some new heuristic approaches", Proceedings – Eighth International Conference on High-Performance Computing in Asia-Pacific Region, HPC Asia 2005 8th International Conference on High-Performance Computing in Asia-Pacific Region, China Computer Federation, Beijing, pp. 73-80.
10. Makarkin, S., Melnikov, B., PaninA. (2013), "On the metaheuristics approach to the problem of genetic sequence comparison and its parallel implementation", Applied Mathematics (Scientific Research Publishing), Vol. 04, No. 10, pp. 35-39.
11. Eckes, B.,Nischt, R.,Krieg, T. (2010), "Cell-matrix interactions in dermal repair and scarring", Fibrogenesis Tissue Repair., No. 3:4, doi:10.1186/1755-1536-3-4.
12. Carr R.W., HennessyJ. L., (1981),"WSCLOCK – a simple and effective algorithm for virtual memory management", SOSP '81 Proceedings of the eighth ACM symposium on Operating systems principles, pp. 87-95.
13. Melnikov, B.F. (2001), "Heuristics in programming of nondeterministic games", Programming and Computer Software., No. 5, pp. 277-288.
14. Melnikov, B., Radionov, A.,Moseev, A.,Melnikova, E., (2006),"Some specific heuristics for situation clustering problems", ICSOFT, Technologies, Proceedings 1st International Conference on Software and Data Technologies, pp. 272-279.
15. Foley, J. (2011),"Fossil Hominids: mitochondrial DNA", available at: http://www.talkorigins.org/faqs/homs/mtDNA.html
16. Frans, B. M., (1997), "Bonobo: The Forgotten Ape", University of California Press, ISBN 0-520-20535-9; trade paperback, October, 1998, pp. 224.
17. Popkov , Y. S. (1995), "Substantiation of the entropy maximization method for problems of image restoration from projections", Automation and Remote Control, 56:1, pp. 77-82.
18. (2014) NCBI:nucleotidedatabase,availableat:http://www.ncbi.nlm.nih.gov/nuccore.
19. Макаркин, С. Б., Мельников, Б. Ф., (2013), "Геометрические методы решения псевдогеометрической версии задачи коммивояжера", Стохастическая оптимизация в информатике, Т. 9., № 2., сс. 54-72.
20. Midwood, K. S., Williams, L. V., Schwarzbauer, J. E. (2004), "Tissue repair and the dynamics of the extracellular matrix", The International Journal of Biochemistry & Cell Biology, Vol. 36, Issue 6, pp. 1031–1037
21. Shao, M., Lin, Y., Moret, B. , (2014), "An Exact Algorithm to Compute the DCJ Distance for Genomes with Duplicate Genes", Research in Computational Molecular Biology, Lecture Notes in Computer Science Volume 8394, pp. 280-292.

References

1. Gaefild, D. (2003), Stroki, derevia i posledovatelnosti v algoritmah. Informatika I vichislitelnaia biologia, SPb: Nevski Dialekt, BHV-Peterburg, 654 p.
2. Toppi, J., De VicoFallani, F.,Petti, M., Vecchiato, G., Maglione, A. G., Cincotti, F., Salinari, S., Mattia, D., Babiloni, F., Astolfi, L. (2013), "A new statistical approach for the extraction of adjacency matrix from effective connectivity networks", "IEEE Engineering in Medicine and Biology Society (EMBC)", No 3-7, pp. 2932-2935.
3. Torshin, I. Yu. (2006), "Bioinformatics in the Post-Genomic Era: The Role of Biophysics", Nova Biomedical Books, NY,ISBN 1-60021-048-1.
4. Winkler, W. E. (1990), String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, Proceedings of the Survey Research Methods Sections, American Statistical Association, pp. 354-359.
5. Van der Loo, M. P. J. (2014), "The Stringdist Package for Approximate String Matching",The R Journal, vol. 6, pp. 111-122.
6. Pages, H., Aboyoun, P., Gentleman, R., DebRaoy, S. (2009), "Biostrings: String Objects Representing Biological Sequences and Matching Algorithms", R package version 2.10.1.
7. Morgan, M., Lawrence, M. (2009), "ShortRead: Base classes and methods for high-throughput short-read sequencing data", R

package version 1.0.6.
8. Melnikov, B. F., (2006) “Multiheuristic approach to discrete optimization problems”, Cybernetics and Systems Analysis, No. 3, pp. 335-341.
9. Melnikov, B.F. (2005), “Discrete optimization problems some new heuristic approaches”, Proceedings – Eighth International Conference on High-Performance Computing in Asia-Pacific Region, HPC Asia 2005 8th International Conference on High-Performance Computing in Asia-Pacific Region, China Computer Federation, Beijing, pp. 73-80.
10. Makarkin, S., Melnikov, B., PaninA. (2013), “On the metaheuristics approach to the problem of genetic sequence comparison and its parallel implementation”, Applied Mathematics (Scientific Research Publishing), Vol. 04, No. 10, pp. 35-39.
11. Eckes, B.,Nischt, R.,Krieg, T. (2010), “Cell-matrix interactions in dermal repair and scarring”, Fibrogenesis Tissue Repair., No. 3:4, doi:10.1186/1755-1536-3-4.
12. Carr R.W., HennessyJ. L., (1981),“WSCLOCK – a simple and effective algorithm for virtual memory management”, SOSP '81 Proceedings of the eighth ACM symposium on Operating systems principles, pp. 87-95.
13. Melnikov, B.F. (2001), “Heuristics in programming of nondeterministic games”, Programming and Computer Software., No. 5, pp. 277-288.
14. Melnikov, B., Radionov, A.,Moseev, A.,Melnikova, E., (2006),“Some specific heuristics for situation clustering problems”, ICSOFT, Technologies, Proceedings 1st International Conference on Software and Data Technologies, pp. 272-279.
15. Foley, J. (2011),”Fossil Hominids: mitochondrial DNA”, available at: http://www.talkorigins.org/faqs/homs/mtDNA.html
16. Frans, B. M., (1997), “Bonobo: The Forgotten Ape”, University of California Press, ISBN 0-520-20535-9; trade paperback, October, 1998, pp. 224.
17. Popkov , Y. S. (1995), “Substantiation of the entropy maximization method for problems of image restoration from projections”, Automation and Remote Control, 56:1, pp. 77-82.
18. (2014) NCBI:nucleotidedatabase,availableat:http://www.ncbi.nlm.nih.gov/nuccore.
19. Makarrin, S., Melnikov, B., (2013), “Geometricheskie metodi reshenia psevdogeometricheskoi versii zadachi kommiviashera”, Stohasticheskaia optimizacia v informatike, T. 9., № 2., pp. 54-72.
20. Midwood, K. S., Williams, L. V., Schwarzbauer, J. E. (2004), “Tissue repair and the dynamics of the extracellular matrix”, The International Journal of Biochemistry & Cell Biology, Vol. 36, Issue 6, pp. 1031–1037
21. Shao, M., Lin, Y., Moret, B. , (2014), “An Exact Algorithm to Compute the DCJ Distance for Genomes with Duplicate Genes”, Research in Computational Molecular Biology, Lecture Notes in Computer Science Volume 8394, pp. 280-292.

Поступила: 25.06.2017

Об авторах:

Мельников Борис Феликсович, доктор физико-математических наук, главный научный сотрудник, Центр информационных технологий и систем органов исполнительной власти, bf-melnikov@yandex.ru;

Пивнева Светлана Валентиновна, кандидат педагогических наук, доцент кафедры высшей математики и математического образования, Тольяттинский государственный университет, tlt.swetlana@rambler.ru;

Трифонов Максим Андреевич, аспирант, Тольяттинский государственный университет, trifonov_max@mail.ru.

Note on the authors:

Melnikov Boris F., Doctor of Physical and Mathematical Sciences, Chief Researcher, Center of Information Technologies and Systems for Executive Power Authorities, bf-melnikov@yandex.ru;

Pivneva Svetlana V., Candidate of Pedagogical Sciences, Associate Professor of the Department of Higher Mathematics and Mathematical Education, Togliatti State University, tlt.swetlana@rambler.ru;

Trifonov Maksim A., Postgraduate, Togliatti State University, trifonov_max@mail.ru.