УДК 517.958:57 + 575.1/.2.08

ПРИМЕНЕНИЕ РАССТОЯНИЙ РЕДАКТИРОВАНИЯ ПРИ БИОИНФОРМАЦИОННОМ АНАЛИЗЕ ГЕНОМОВ ДЛЯ ЗАДАЧ ОЦЕНКИ СОСТОЯНИЯ РЕПРОДУКТИВНОЙ СИСТЕМЫ

Пономарева Н.С., Реброва Г.Н., Колина Е.А.

ФГАОУ ВО «Южный федеральный университет», Ростов-на-Дону, e-mail: nsponomareva@sfedu.ru

Настоящая статья посвящена особенностям применения расстояний редактирования при биоинформационном анализе генетических последовательностей для задач изучения и оценки состояния репродуктивной системы. В биоинформатике последовательность нуклеотидов ДНК представляется как последовательность символов, а генетические особенности строения допускают искажение этих последовательностей (вставки, удаления, замены и транспозиции символов) при поиске. Проведен сравнительный анализ использования расстояний редактирования Хэмминга, Левенштейна и метода n-грамм для генетических символьных последовательностей. Определены ограничения, особенности и критерии использования метрик. Разработан алгоритм выбора и применения расстояний редактирования для задач нечеткого поиска в генетических последовательностях в зависимости от возможных типов дефектов. При выборе метода учитывается длина анализируемых последовательностей, а перед анализом сходства предлагается нормирование по длине строки для сопоставления полученных результатов значений расстояний редактирования.

Ключевые слова: расстояния редактирования, биоинформационный анализ, генетические последовательности, персонифицированная медицина

APPLICATION EDIT DISTANCES BIOINFORMATIC GENOME ANALYSIS FOR ASSESSMENT OF REPRODUCTIVE SYSTEM

Ponomareva N.S., Rebrova G.N., Kolina E.A.

Southern Federal University, Rostov-on-Don, e-mail: nsponomareva@sfedu.ru

This article is devoted to the edit distance application features in bioinformatics analysis of genetic sequences for the reproductive system study and evaluation tasks. Nucleotides of DNA is represented as a sequence of characters, but the genetic characteristics of the structure allows the distortion of these sequences (insertion, deletion, substitution and transposition of characters) in bioinformatics search. A comparative analysis of genetic symbolic sequence use of Hamming and Levenstein distances editing and the n-grams method. Constraints, features and criteria for metrics use were determined. The selection and application distances for editing tasks fuzzy search in genetic sequence, depending on the possible types of defects algorithm were created. Method takes the length of the analyzed sequences into account. The string length of to match the normalization values of the edit distances results obtained for the proposed similarities analysis.

Keywords: edit distance, bioinformatic analysis, genetic sequence, personified medicine

Увеличение объема и повышение сложности обрабатываемых данных, структура которых представляет собой символьные последовательности требуют повышения эффективности их обработки и анализа. Задачи нечеткого поиска чаще всего возникают при коррекции ошибок, фильтрации нежелательных сообщений, обнаружении плагиата, поиске с учетом форм одного и того же слова и основаны на определении расстояния между строками. В настоящее время при сравнении геномных последовательностей активно используются алгоритмы поиска неструктурированных данных. Данные алгоритмы могут быть применены при анализе геномов для разрешении проблем репродуктивной системы и развитии персонифицированной медицины при сравнении нуклеотидных последовательностей, так как последовательность нуклеотидов ДНК может быть представлена как последовательность символов, содержащая 4 «буквы» – A, T, G, C, каждая из которых соответствует определенному нуклеотиду.

В настоящее время имеется огромное количество генетических данных, находящихся в открытом доступе на специализированных порталах. Математическая, аналитическая и программная обработка данных последовательностей имеет явное преимущество перед экспериментальными исследованиями в области трудоемкости, стоимости и времени и должна быть широко внедрена для изучения в различные сферы генетики и медицины, в том числе для задач оценки состояния репродуктивной системы.

Релевантность результатов нечеткого поиска зависит от учета искажений символов слов различных типов (вставки, удаления, замены и транспозиции символов), но генетическое строение молекулы ДНК допускает искажения. Ключевым элементом организации нечёткого поиска является выбор *меры сходства* слов или обратной функции – *функции расстояния* между словами,

часто называемой *метрикой*. Наибольшее распространение в случае нечеткого поиска подстроки в нуклеотидной строке получили *трансформационные метрики (принят термин расстояния редактирования – в* области текстового поиска) [2].

Наиболее известными расстояниями редактирования являются расстояния Хэмминга, Левенштейна, а также *n*-граммы [1, 3, 5].

Расстояние Хэмминга определяется как число позиций, в которых для двух слов одинаковой длины соответствующие символы длины [8]. В [6] приводится альтернативное определение: если две строки *Ai* и *Aj* имеют одинаковую длину *n*, расстояние Хемминга *dH* (*Ai*, *Aj*) определяется как минимальное количество подстановок (замен), необходимых для преобразования строки *Ai* в строку *Aj*.

Удовлетворяя следующим условиям, расстояние Хэмминга обладает свойствами метрики:

$$d(x, y) \geq 0;$$

$$d(x, y) = 0 \Leftrightarrow x = y;$$

$$d(x, y) = d(y, x);$$

$$d(x, z) \leq d(x, y) + d(y, z).$$

Расстояние Левенштейна позволяет сравнивать строки различной длины с учетом таких искажений, как замены, вставки, и удаления. Расстояние Левенштейна равно минимальному числу элементарных операций редактирования, необходимых для преобразования одной строки в другую.

В [4] расстояние Левенштейна $d_L(Ai, Aj)$ определяется как минимальное количество операций вставки, удаления либо замены одного символа на другой, необходимых для превращения одной строки в другую, а [6] определяет расстояние Левенштейна учитывающим только операции удаления и вставки, а расстояние, учитывающее еще и замену (подстановку), называется расстоянием преобразования $d_E(Ai, Aj)$.

Пусть S_1 и S_2 – две строки (длиной M и N соответственно) над некоторым алфавитом, тогда расстояние $d_L(S_1, S_2)$ можно посчитать по следующей рекуррентной формуле:

$$d_L(S_1, S_2) = D(M, N),$$

где

$$D(i,j)=\begin{cases} 0; & i=0,\ j=0; \\ i; & j=0,\ i>0; \\ j; & i=0,\ j>0; \\ \min\begin{pmatrix} D(i,j-1)+1, \\ D(i-1,j)+1, \\ D(i-1,j-1)+m(s_1\|i|, s_2\|j|) \end{pmatrix}; & j>0,\ i>0, \end{cases}$$

здесь шаг по *i* символизирует удаление (*D*) из первой строки, по *j* – вставку (*I*) в первую строку, а шаг по обоим индексам символизирует замену символа (*R*) или отсутствие изменений (*M*).

Очевидно, справедливы следующие утверждения:

$$d(S_1, S_2) = \|S_1| - |S_2\|;$$

$$d(S_1, S_2) \leq \max(|S_1|, |S_2|);$$

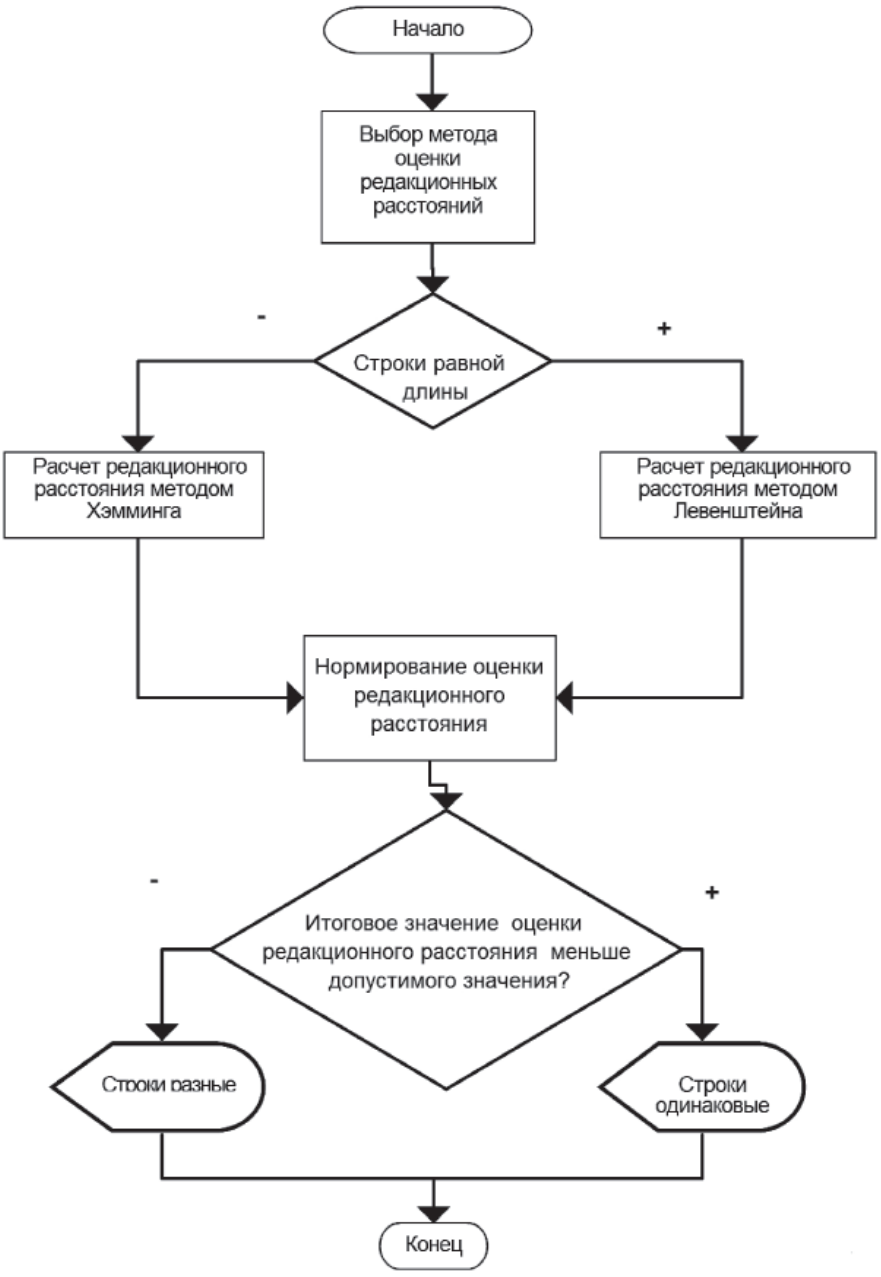
$$d(S_1, S_2) = 0 \Leftrightarrow S_1 = S_2.$$

В модификации расстояния редактирования, предложенной Дамерау [7], в множество элементарных операций включены операции перестановки (транспозиции) двух соседних символов, но при этом требуется, чтобы к транспонированным символам не применялись другие операции редактирования.

Основным недостатком расстояния редактирования Хэмминга является требование одинаковой длины строк, таким образом, расстояние Хэмминга подходит для расчета расстояния редактирования с учетом таких искажений, как замена и транспозиция, но не подходит при вставках и удалениях. Учитывать все названные искажения позволяют расстояния Левенштейна и Дамерау – Левенштейна, но при перестановке местами слов или частей слов получаются сравнительно большие расстояния [5]. Ввиду того что описанные расстояния редактирования не нормированы по длине строк, строки с одинаковым количеством допустимых искажений символов, но различной длины будут иметь одинаковое значение d_H расстояния Хэмминга, а значения расстояний Левенштейна $d_L(A_i, A_j)$ и Дамерау – Левенштейна $d_E(A_i, A_j)$ между совершенно разными короткими словами оказываются меньшими, чем расстояния между очень похожими длинными словами. При необходимости метрику можно нормированием по длине строки: d_H/n. Известные меры сходства Джаро [9] и Джаро – Уинклера [10] для сравнения коротких строк – коэффициенты, включающие нормирования.

Существуют различные модификации *n*-граммных расстояний, основанные на подсчете числа общих подстрок равной длины (*n-грамм*), *n*-граммами (*q*-граммами) называют множество подстрок длины *n* исходной строки. Оценка расстояния производится на основе подсчета количества различающихся *n*-грамм данного множества. Обобщением расстояния Левенштейна $dw(A_i, A_j)$ является использование матрицы весовых коэффициентов для замены символа *i* символом *j*. Это расстояние будет являться метрикой, только если матрица весовых коэффициентов симметрична [6]. Частным случаем будет вариант, учитывающий вес для каждой из операций, вне зависимости от заменяемого символа.

Задача поиска структурного или функционального элемента (мотива) в последовательности межгенного пространства, сводится к поиску подстроки в строке [1], и, несмотря на то, что размерность алфавита небольшая, равна 4, задача осложняется возможностью вставок, делеций и замен в конкретной ключевой позиции.



Алгоритм применения расстояний редактирования для задач нечеткого поиска в генетических последовательностях

Например, в последовательности

ctctagtggtcagtgttagcactgcatccagctgactcaggtgggc

необходимо определить наличие следующего мотива:

accactgcactccagcctgag.

Результаты поиска:

ctctagtggtcagtgttagcactgca.tccag.
ctgactcaggtgggc

accactgcactccagcctgag.

Сравним найденное и искомое, большими буквами обозначены замены, а точками делеции:

aGcactgca.tccag.ctgaC

accactgcactccagcctgag.

Разработан алгоритм применения расстояний редактирования для задач нечеткого поиска в генетических последовательностях (рисунок). На данный момент для оценки редакционных расстояний может быть использовано 2 метода – Хэмминга и Левенштейна. Для каждого из них есть ряд случаев, когда их рекомендуется применять, например метод Хэмминга не применим к последовательностям разной длины.

Данный алгоритм реализован и модифицирует программу для ЭВМ «Анализ нуклеотидных последовательностей ДНК с помощью точечной матрицы гомологии» при расчете сходства. Программа позволяет выявлять значимые последовательности, а также анализировать последовательности ДНК на предмет наличия прямых и инвертированных повторов заданной степени сходства при проведении визуального анализа генома и целенаправленной идентификации данных секвенирования ДНК с помощью точечной матрицы гомологий для решения задач анализа и оценки состояния репродуктивной системы.

По мере изучения строения ДНК и понимания механизмов и генетического значения вставок, удалений, замены и транспозиции нуклеотидов будут корректироваться алгоритм применения и расчет расстояний редактирования при биоинформационном анализе геномов. В то же время биоинформационные исследования генетических данных уже сейчас позволяют найти особенности и нацелить лабораторные исследования на изучение определенных участков, тем самым сократить время и стоимость исследований в целом.

Работа выполнена в рамках базовой части внутреннего гранта ЮФУ в 2015 году по проекту 213.01–2015/003ВГ «Изучение ДНК-элементов некодирующих белок в структуре различных геномов».

Список литературы

1. Бойцов Л.М.Классификация и экспериментальное исследование современных алгоритмов нечеткого словарного поиска // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 6ой Всероссийской научной конференции. – RCDL2004. – Пущино, Россия, 2004.

2. Бондаренко А.В. Формальный метод нечеткого поиска персональной информации // Препринты ИПМ им. М.В. Келдыша. – 2009. – 45 с.

3. Гасфилд Д. Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология: пер. с англ. И.В. Романовского. – СПб.: Невский Диалект; БХВ-Петербург, 2003. – 654 с: ил.

4. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. АН СССР. – 1963. – № 4. – С. 845–848, 1965.

5. Прытков В.А. Функция расстояния между строками на основе кусочно-постоянной модели // Доклады БГУИР. – 2013. – URL: http://plib.unibel.by/index.php/dbsuir/article/view/1028/1017.

6. Смит У. Методы и алгоритмы вычислений на строках. – М., 2006.

7. Damerau Fred J. A technique for computer detection and correction of spelling errors // In Communications of ACM. – 1964. – Vol. 7(3). – P 171–176.

8. Hamming R.W. Error detecting and error correcting codes // The Bell System Technical Journal. – 1950. – Vol. XXIX, № 2. – P. 147–160

9. Jaro M.A. Advances in record linking methodology // Journal of the American Statistical Society. – 1989. – Vol. 84, № 406.

10. Winkler W.E. Overview of Record Linkage and Current Research Directions // Research Report Series. – 2006.

References

1. Bojcov L.M.Klassifikacija i jeksperimentalnoe issledovanie sovremennyh algoritmov nechetkogo slovarnogo poiska // Trudy 6oj Vserossijskoj nauchnoj konferencii «Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii» RCDL2004, Pushhino, Rossija, 2004.

2. Bondarenko A.V. Formalnyj metod nechetkogo poiska personalnoj informacii // Preprinty IPM im. M.V. Keldysha, 2009. 45 p.

3. Gasfild D. Stroki, derevja i posledovatelnosti v algoritmah. Informatika i vychislitelnaja biologija Per. s angl. I.V. Romanovskogo // SPb.: Nevskij Dialekt; BHV-Peterburg, 2003. 654 p: il.

4. Levenshtejn V.I. Dvoichnye kody s ispravleniem vypadenij, vstavok i zameshhenij simvolov // Dokl. AN SSSR, 1963, 4, pp. 845–848, 1965.

5. Prytkov V.A. Funkcija rasstojanija mezhdu strokami na osnove kusochno-postojannoj modeli // Doklady BGUIR, 2013. URL: http://plib.unibel.by/index.php/dbsuir/article/view/1028/1017.

6. Smit U. Metody i algoritmy vychislenij na strokah. M., 2006.

7. Damerau Fred J. A technique for computer detection and correction of spelling errors // In Communications of ACM, Vol. 7(3), pp. 171–176, 1964.

8. Hamming R.W. Error detecting and error correcting codes // The Bell System Technical Journal. 1950. Vol. XXIX, no. 2. pp. 147–160

9. Jaro M.A. Advances in record linking methodology // Journal of the American Statistical Society, vol. 84, no. 406, 1989.

10. Winkler W.E. Overview of Record Linkage and Current Research Directions // Research Report Series, 2006.

Рецензенты:

Соловьев А.Н., д.ф.-м.н., профессор, заведующий кафедрой «Теоретическая и прикладная механика», ФГБОУ ВПО «Донской государственный технический университет», г. Ростов-на-Дону;

Елсуков В.С., д.т.н., профессор кафедры «Автоматика и телемеханика», ФГБОУ ВПО «Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова», г. Новочеркасск.