УДК – 004.45

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ ПОШУКУ ТОЧОК ЗМІНИ ТРИПЛЕТНОЇ ПЕРІОДИЧНОСТІ

Кисляк С.В., ст. вик. каф. БМК
kisluak@ukr.net
Іллєнок М.П., студент
marisha.illenok@gmail.com
Факультет біомедичної інженерії
Національний технічний університет
«Київський політехнічний інститут імені Ігоря Сікорського»
м. Київ, Україна

Реферат *– з появою методів секвенування нового покоління спостерігається експоненційне збільшення молекулярно-біологічних даних. Банки даних біологічних послідовностей, такі як GenBank [1], UniProt [2], KEGG [3] та інші досягли великих розмірів. При цьому швидкість опису просеквенованих послідовностей значно відстає від швидкості їх накоплення. При такому збільшенні об'єму інформації, вирішити проблему «відставання» неможливо без застосування ефективних алгоритмів. Досить цікавим, складним та таким, що заслуговує уваги та детального дослідження є напрямок геноміки, що відповідає за пошук ефективних алгоритмів для вирішення задачі ідентифікації генів про – та еукаріот. Найбільш популярними є програмні продукти, що дозволяють знаходити гени відповідно до отриманого парного вирівнювання послідовностей методом динамічного програмування. Такий підхід не дозволяє дослідити всі ділянки нуклеотидної послідовності, оскільки більшість генів протягом еволюції змінювались за рахунок, наприклад, точкових мутацій або кодуюча ділянка гена була отримана за рахунок об`єднання декількох генів. В такій ситуації інформація про ділянки, що формують гібридні гени, буде відсутньою в базах даних. Заслуговує уваги нетривіальний алгоритм, що був описаний в роботах [4]. Пошук точок зміни триплетної періодичності для прямої та зворотно компліментарної послідовності дає можливість визначення координат, які можуть вказувати на місце об'єднання двох генів.*

Ключові слова – *нуклеотидна послідовність, триплетна періодичність,гібридні гени*

I.ВСТУП

Дослідження та опис просеквенованих послідовностей є однією з основних задач геноміки. Для цього використовують програмне забезпечення, що дозволяє знаходити кодуючи ділянки генів біологічних послідовностей. Базові алгоритми парного вирівнювання, що дозволяють вирішити задачу ефективного пошуку та ідентифікації генів засновані на методі динамічного програмування (наприклад, алгоритм Нидлмана-Вунша та Уотермана-Сміта). Такі методи, для деякого досліджуваного гена, знаходять найбільш схожу послідовність-гомолог (або сімейство послідовностей). Відповідно до їх подібності (парного вирівнювання) дослідники можуть зробити висновок про еволюційні та біологічні властивості досліджуваної біологічної послідовності. Алгорими парного вирівнювання вдосконалюються та модифікуються, їх точність та швидкість роботи досягли високих показників. Однак ці методи мають деякі суттєві недоліки. Основний недолік алгоритмів парного вирівнювання пов`язаний з тим, що, якщо для досліджуваної послідовності не вдається знайти подібну послідовність, то метод виявляється неефективним. Тому все частіше з'являються роботи, присвячені створенню альтернативних методів аналізу послідовностей, які не основані на вирівнювання (так звані, alignment-free методи) [5].

Властивість триплетної періодичності відомо вже досить довгий час [6-8]. Триплетна періодичність характеризуються нерівномірним розподілом символів в різних позиціях кодонів. Ця періодичність відсутня в некодуючих ділянках генома та деяких інтронів генів еукаріот. За час вивчення феномена триплетної періодичності були розроблені різні математичні методи для її дослідження, що базуються відповідно до декількох гіпотез, що пояснюють її існування [9]. Алгоритм пошуку триплетної періодичності знайшов своє застосування в комп'ютерних програмах, що можуть бути

використані для аналізу послідовностей ДНК. Триплетна періодичність також дозволяє розрізнити кодуючи ділянки геному від не кодуючих. Триплетна періодичність може використовуватися для вивчення однорідності біологічної послідовності.

Точки зміни триплетної періодичності відповідають позиціям зміни цієї властивості в послідовності можуть відображати еволюційну структуру даної послідовності. Цей факт дозволяє припустити, що, якщо деякий ген був сформований, наприклад, в результаті об'єднання послідовностей двох різних генів, триплетна періодичність яких значно відрізнялася, то на границі буде присутня точка зміни триплетної періодичності.

II. ОГЛЯД ЛІТЕРАТУРИ

Вирівнювання послідовностей - це процедура відображення символів послідовності, при яких досягається максимальний рівень подібністі (максимальна функція подібності). Ця процедура базується на методі динамічного програмування з використанням зважених матриць [10] та системи штрафів. Вагові матриці являють собою симетричну квадратну матрицю, клітини якої є вагами, що встановлюють рівень подібності між окремими символами алфавіту. При цьому дозволяється використовувати спеціальний «порожній» символ (gap або пропуск), що відповідає таким еволюційним подіям, при яких символ в першій послідовності був видалений або вставлений в іншу (другу) послідовність. При такому підході система оцінки (вагова матриця і система штрафів) відіграє велику роль, оскільки ця система повинна надавати перевагу біологічно правильному вирівнюванню. Існують методи глобального [11] (коли послідовності вирівнюються від початку до кінця) та локального [12] вирівнювання (що передбачає пошук найбільш подібних ділянок). Також для прискорення процесу пошуку гомологів у базах даних біологічних послідовностей (класичні методи динамічного програмування при вирівнюванні двох послідовностей довжиною m і n вимагають O (m×n) пам'яті та використовують таку ж кількість часу) використовують різні евристичні підходи.

Методи аналізу послідовностей, не основані на вирівнюванні. Програми вирівнювання не тільки історично є одні з перших програм аналізу біологічних послідовностей, але й лежать в основі багатьох інших алгоритмів та методів геноміки. Так, програма побудови вирівнювання заснована на евристичному алгоритмі, BLAST - найпопулярніший інструмент сучасної біоінформатики. Однак, незважаючи на все це, можливості вирівнювання обмежені. Так згідно з даними [14] програми анотації на основі аналізу послідовності можуть охоплювати до 70% білків представлені в банку даних амінокислот послідовностей UniProt. Це обмеження пояснюється тим, що в процесі еволюції послідовностей не збереглися предкові форми. Тому останнім часом все частіше намагаються створювати алгоритми аналізу послідовностей, що використовують так звані «alignment-free» методи [15-18]. Такі програми використовують для вирішення таких завдань тільки статичні властивості символьної послідовності. Такий підхід особливо часто використовується при пошуку регуляторних послідовностей. Існують такі програми і для філогенетичних досліджень [19], порівняння послідовностей [21,22]. Однак у порівнянні з методами, які використовують методи вирівнювання, їхня частка все ще мала.

Триплетна періодичність. З того часу як стали доступними для досліджень перші генетичні послідовності, вдалося виявити, що вони містять різні типи періодичності. Розмір періоду в яких, може бути дуже різним. Деякі біологічні послідовності мають зовсім короткий період - три, в кодуючих послідовностях [22]; періоди середньої довжини, наприклад, період 10-11 пар основ, пов'язані зі структурою молекули ДНК [23]; до дуже великих періодичних мотивів у геномах теплокровних хребетних, так звані ізохори [24]. Періодичність різної довжини і різного ступеня вираженості існують як на генному так і на протеомному рівнях, в кодуючих і некодуючих ділянках, представленні як явні, так і розмиті (приховані) повтори [25].

Періодичність можна розділити на дві категорії: тандемні повтори та періодичні

коливання [26]. Розпізнавання тандемних повторів в умовах «маскування» їх мутаціями типу вставок і делецій, особливо за умови, що сам патерн (або мотив) і навіть розмір періоду не відомий заздалегідь, виявляється досить складним завданням. Другий, більш складний тип періодичності. Для цього варіанту важко вибрати якийсь певний патерн, але існує статистично значуща кореляція між символами на певних позиціях послідовності. У цьому випадку, спільна ймовірність двох символів більше не дорівнює добутку їх ймовірностей. Властивість триплетної періодичності кодуючих послідовностей найчастіше можна віднести до другого типу.

Гіпотези, які описують властивості триплетної періодичності. За довгу історію вивчення феномена триплетної періодичності, було сформульовано декілька теорій, які пояснюють їх походження. Перше пояснення пов`язано з триплетною організацією генетичного коду. Відомо, що існують 64 кодони, три з них є стоп-кодонами, інші 61 кодують 20 амінокислот. Генетичний код є надлишковим (тобто кілька триплетів кодують одну й ту ж амінокислоту, такі кодони називаються синонімічними). Кількість триплетів відрізняється для різних амінокислот. Синонімічні кодони використовуються нерівно ймовірно [27,28]. Також відомо існування першочергових кодонів для кожної амінокислоти[29,30].

Всі ці властивості та інші були враховані при поясненні триплетної періодичності в кодуючих послідовностях. Було показано, що штучні послідовності з нерівномірним використанням кодонів, з частотами близькими до природних, які зустрічаються в реальних геномах, мають виражені властивості триплетної періодичності [31]. В той час, як штучні послідовності, алгоритм побудови яких потребує рівномірного використання кодонів (в тому випадку і стоп-кодонів), такою періодичністю не володіють. Відповідно до проведеного моделювання, автори роботи [32], дійшли до висновку, що саме нерівномірне використання кодонів є причиною триплетної періодичності в екзонахкодуючих ділянок генів еукаріот.

В роботі [35] показано, що значення частот кодонів недостатньо для створення триплетної періодичності того рівня, який присутній в реальних геномах. В роботах [35,34] була показана кореляція між триплетною періодичністю та експресією відповідного гена. Нерівномірне використання амінокислот в послідовностях білків, також виявилось однією з причин появи триплетної періодичності [35].

III. МЕТА ДОСЛІДЖЕННЯ

Реалізувати алгоритм пошуку точок зміни триплетної періодичності для вивчаємого фрагменту ДНК.

IV. МАТЕРІАЛИ ДОСЛІДЖЕНЬ

Дослідження проводилося на попередньо підготованому файлі Fasta.file. Для тестування роботи програмного продукту було взято фрагмент нуклеотидної послідовності Escherichiacoli.

Fasta. Розширення fasta файлу використовується для збереження інформації про послідовності нуклеїнових кислот ДНК, РНК та амінокислотних залишків білків. У файлі, що має розширення fasta містяться заголовки, якімістять інформацію про послідовність. Зазвичай інформація записується в одному рядку, яка містить послідовність, символів та імен. Програмне забезпечення використовується для швидкого порівняння нуклеотидів та інших біологічних даних, і це може бути можливо тільки в тому випадку, коли файли знаходять yfasta форматі [36]

Для вирішення задачі пошуку точок зміни триплетної періодичності нуклеотидних послідовностей необхідно знайти на досліджуваній послідовності однорідні фрагменти таким чином, щоб всередині таких ділянок триплетна періодичність була схожа, а між ділянками спостерігалась її зміна (рисунок 1).

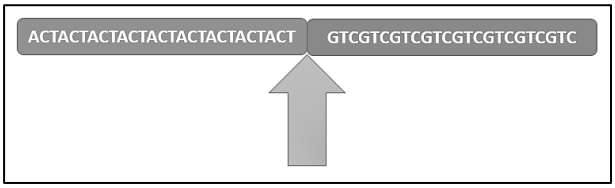

Рис. 1. Фрагмент нуклеотидної послідовності, що містить точку зміни триплетної періодичності

Матрицею триплетної періодичності - частотна матриця розміром 4x3: рядки такої матриці відповідають символам алфавіту, а стовпчики – трьом позиціям кодону. Елементи матриці відповідають числу нуклеотидів типу *i,* які знаходяться на позиціях *j* в досліджуваній послідовності. Матрицю триплетної періодичності, побудовану для ділянки послідовності від позиції 1 до позиції 2, буде позначатися, як $M(x_1, x_2)$.

Для того, щоб перевірити всі можливі варіанти триплетної періодичності ми використовуємо три рамки зчитування. Рамки зчитування визначаються шляхом зсуву фази послідовності на один або два символи. Тобто, відповідає циклічному зсуву матриці триплетної періодичності.

Одним із найважливіших параметрів в алгоритмах сегментації є вибір порогового значення. Він визначає наскільки мають відрізнятися дві ділянки, для того щоб їх можна було розділити. Очевидно, що вказавши занадто високе порогове значення, в результаті ми можемо не отримати точок зміни триплетної періодичності. І навпаки, при встановленні надто маленькому порогового значення призведе до появи точок зміни триплетної періодичності навіть на випадкових послідовностях.

Алгоритм пошуку точок зміни триплетної періодичності з використанням рамок зчитування був почергово використаний до кожної досліджуваної послідовності. Отримання потрібних ділянок можливе через зсув позиції x вздовж послідовності праворуч з кроком три.

Для кожної ділянки розраховувалася матриця триплетної періодичності, що далі порівнювалася з використанням міри відмінності з урахуванням можливого зсуву рамки зчитування.

На рис. 2 представлено блок-схему алгоритму пошуку точок зміни триплетної періодичності для однієї послідовності.

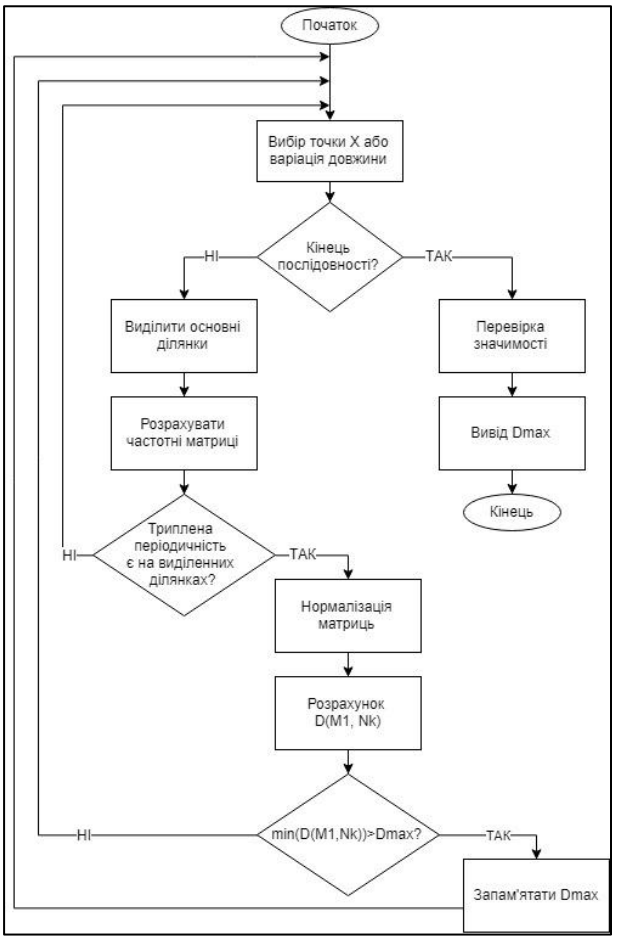

*Рис. 2.*Блок-схема алгоритму пошуку точок зміни триплетної періодичності для однієї послідовності

V. РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

Проаналізовано алгоритм пошуку точок зміни триплетної періодичності послідовностей. Алгоритм дозволяє ефективно знаходити гібридні гени, що були сформовані в процесі еволюції за рахунок об`єднання декількох генів. Алгоритм може працювати на невідомих послідовностях, інформація про які відсутня в базах даних, що є основною перевагою такого методу в порівнянні з класичними алгоритмами парного вирівнювання.

Реалізовано програмне забезпечення для пошуку точок зміни триплетної періодичності. Програмне забезпечення було розроблено за допомогою мови програмування Python версії 3 на базі програмного середовища «Jupyter Notebook» та за допомогою фреймворків реалізований інтерфейс користувача. На вхід програми подається текстовий файл у форматі Fasta, який містить в собі інформацію про досліджувану послідовність. Початкове вікно користувача показано на рисунку 3.

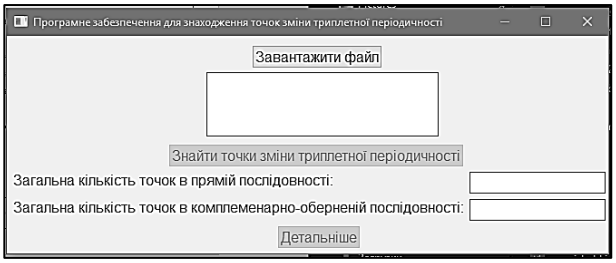

Рис. 3. Початковий екран користувача.

Спочатку вікно має лише одну активну кнопку – «Завантажити файл». Натиснувши на неї користувач переходить до вибору досліджуваного файлу. Слід зазначити, що обрати користувач може лише текстові файли з розширенням fna та fasta (рис. 4).

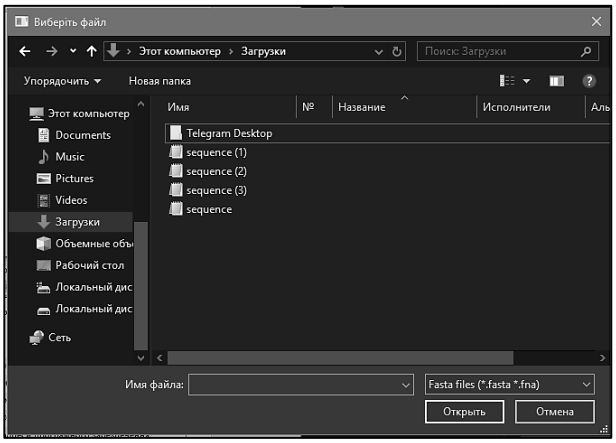

Рис. 4. Екран завантаження текстового файлу, який містить інформацію про послідовність, з розширенням .fasta або .fna.

З завантаженого файлу програма зчитує дані про досліджувану послідовність та виводить їх на початкове вікно користувача. Інформація про назву досліджуваної послідовності записується у активну область (рис. 5).

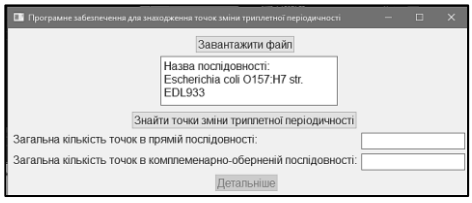

Рис. 5. Вигляд початкового екрану після завантаження файлу.

Одразу після завантаження файлу стає активною кнопка – «Знайти точки зміни триплетної періодичності». Натиснувши на неї, програма запускає алгоритм, по якому здійснюється пошук всіх можливих точок зміни триплетної періодичності.

Результат роботи програми можна отримати в загальному вигляді (виведення кількості знайдених точок зміни триплетної періодичності в прямій послідовності та комплементарно-оберненій) та детальний (виведення інформації про обране порогове значення, координати всіх знайдених точок зміни триплетної періодичності та значення функції відмінності (D)).Загальні результати виводяться на початковий екран користувача (рис. 5).

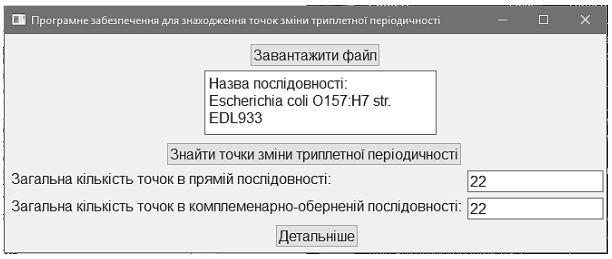

Рис. 5. Вигляд початкового екрану після завершення пошуку точок зміни триплетної періодичності.

Для отримання детального виводу результатів потрібно натиснути кнопку «Детальніше», яка стає активною одразу після виведення загального результату. Детальні результати виводяться окремо для прямої послідовності (рис. 6) та комплементарно-оберненої послідовності (рис. 7).

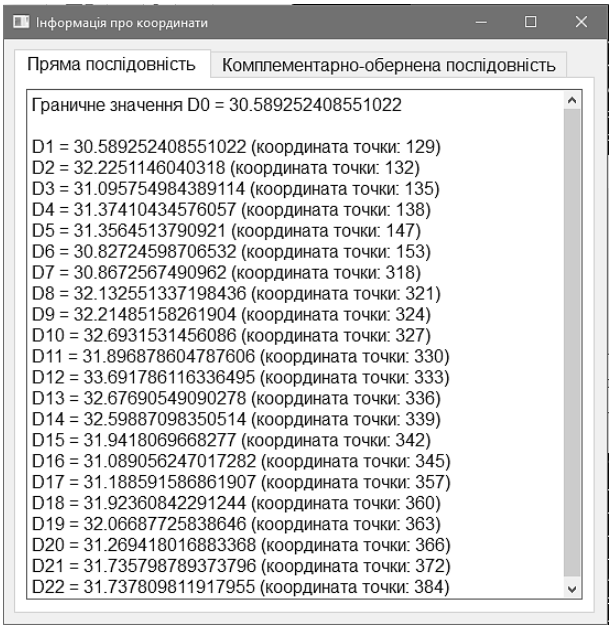

Рис. 6. Детальний вивід результатів для прямої послідовності.

Для кожної досліджуваної послідовності обраховується порогове значення (D_0), яке є

індивідуальним. Обраховане порогове значення, яке використовувалося і за допомогою якого обиралися ймовірні точки зміни триплетної періодичності, вказане у детальному виводі результатів. Саме від цього значення залежить точність дослідження, а саме – кількість точок зміни триплетної періодичності які були знайдені та збережені.

Значення функції відмінності матриць триплетної періодичності (D) обраховуються для кожної рамки зчитування окремо та порівнюється з отриманим раніше пороговим значенням (D_0). Отриманні значення D записані у детальному описі результатів для кожної знайденої точки зміни триплетної періодичності. Координати точок зміни триплетної періодичності зберігають та виводяться для кожної точки. Координати точки визначає позицію символу в нуклеотидній послідовності послідовності.

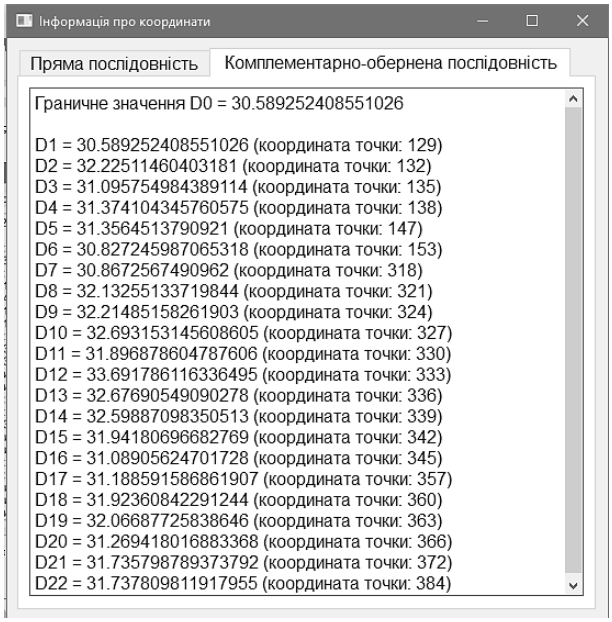

Рис.7. Детальний вивід результатів для обернено-комплементарної послідовності.

VI. ВИСНОВКИ

Реалізовано алгоритм для пошуку точок зміни триплетної періодичності за допомогою мови програмування Python версії 3 на базі програмного середовища «JupyterNotebook» За допомогою фреймворків реалізований користувацький інтерфейс.

Алгоритм пошуку точок зміни триплетної періодичності реалізований як для прямої так і для обернено-комплементарної послідовностей.

Алгоритм може працювати на невідомих послідовностях, інформація про які відсутня в базах даних, що є основною перевагою такого методу в порівнянні з класичними алгоритмами парного вирівнювання.

ПЕРЕЛІК ПОСИЛАНЬ

[1] Benson D.A. и др. GenBank: update. // Nucleic Acids Res. 2004. Т. 32. № Databaseissue. С. D23–D26.

[2] Consortium T.U. The Universal Protein Resource (UniProt) 2009 // Nucleic Acids Res. 2009. Т. 37. № Data baseissue. С. 169–174.

[3] Ogata H. и др. KEGG: Kyoto Encyclopedia of Genesand Genomes // Nucleic Acids Res. 1999. Т. 27. № 1. С. 29–34.

[4] Коротков Е.В., Суворова Ю.М. // Изучение одиночных и парных точек разладки в кодирующих последовательностях ДНК // V съезд биофизиков Росии, Нижний Новгород. 2012. Том 1.

[5] Vinga S., Almeida J. Alignment-free sequence comparison-a review. // Bioinformatics. 2003. Т. 19. № 4. С. 513–523.

[6] Konopka A.K. и др. Distance analysis helps to establish characteristic motifsinintron sequences. // GeneAnal. Tech. 1987. Т. 4. № 4. С. 63–74.

[7] Shepherd J.C. Periodic correlationsin DNA sequence sandevidence suggestingtheire volutionary originin a comma-less genetic code. // J. Mol. Evol. 1981a. Т. 17. № 2. С. 94–102.

[8] Tsonis A.A., Elsner J.B., Tsonis P.A. Periodicityin DNA codings equences: implications ingenee volution. // J. Theor. Biol. 1991. Т. 151. № 3. С. 323–31.

[9] Herzel H. и др. Interpreting correlations inbio sequences // Phys. A Stat. Mech. itsAppl. 1998. Т. 249. № 1-4. С. 449–459.

[10] Durbin R. и др. Biological Sequence Analysis: Probabilistic Modelsof Proteinsand Nucleic Acids.: Cambridge University Press, 1998. 356 с.

[11] Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the aminoacid sequence of two proteins. // J. Mol. Biol. 1970. Т. 48. № 3. С. 443–453.

[12] Smith T.F., Waterman M.S. Comparison of bios equences // Adv. Appl. Math. 1981b. Т. 2. № 1 981. С. 482–489.

[13] Altschul S.F. и др. Basiclocalal ignment search tool. // J. Mol. Biol. 1990. Т. 215. № 3. С. 403–410.

[14] Loewenstein Y. и др. Protein functionan notation by homology-based inference // GenomeBiol. 2009. Т. 10. № 2. С. 207.

[15] Bonham-Carter O., Steele J., Bastola D. Alignment-free genetic sequence comparisons: a review of recent approach esby word analysis. // Brief. Bioinform. 2013. Т. 15. № 6. С. 890–905.

[16] Kantorovitz M.R., Robinson G.E., Sinha S. A statistical method for alignment-free comparison of regulatory sequences. // Bioinformatics. 2007. Т. 23. № 13. С. 249–255.

[17] Vinga S. Editorial: Alignment-free methods in computational biology. // Brief. Bioinform. 2014. Т. 15. № 3. С. 341–2.

[18] Vinga S., Almeida J. Alignment-freesequencecomparison-a review. // Bioinformatics. 2003. Т. 19. № 4. С. 513–523.

[19] Stuart G.W., Moffett K., Baker S. Integrated gene and species phylogenies from unaligned whole genomeprotein sequences. // Bioinformatics. 2002. Т. 18. № 1. С. 100–108.

[20] Borozan I., Watt S., Ferretti V. Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. // Bioinformatics. 2015. С. btv006.

[21] Yin C., Chen Y., Yau S.S.-T. A measureof DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. // J. Theor. Biol. 2014. Т. 359С. С. 18–28.

[22] Korotkov E. V, Korotkova M.A., Kudryashov N.A. Informationd ecomposition of symbolic sequences // Phys. Lett. A. 2003. Т. 312. № 3. С. 198–210.

[23] Fickett J.W., Tung C.S. Assessment of protein coding measures. // Nucleic Acids Res. 1992. T. 20. № 24. C. 6441–6450.
[24] Herzel H., Weiss O., Trifonov E.N. 10-11 bp period icities incompletegeno mesreflectprote in structure and DNA folding. // Bioinformatics. 1999. T. 15. № 3. C. 187–193.
[25] Bernardi G. и др. Themosaicgenome of warm-blooded vertebrates. // Science. 1985. T. 228. № 4702. C. 953–958.
[26] Trifonov E.N., Sussman J.L. The pitch of chromatin DNA isreflected inits nucleotide sequence. // Proc. Natl. Acad. Sci. 1980. T. 77. № 7. C. 3816– 3820.
[27] Zhang M. и др. Mining periodic patterns with gapr equirement from sequences // ACM Trans. Knowl. Discov. Data. 2007. T. 1. № 2. C. 7–es.
[28] Plotkin J.B., Kudla G. Synonymous but not the same: thecauses and consequences of codonbias. // Nat. Rev. Genet. 2011. T. 12. № 1. C. 32–42.
[29] Iriarte A. и др. General trend sinselective lydrivencodonusagebiases in the domain archaea. // J. Mol. Evol. 2014. T. 79. № 3-4. C. 105–10.
[30] Sharp P.M. и др. Codonusage patterns in Escherichiacoli, Bacillussubtilis, Saccharomy cescerevisiae, Schizosaccharomy cespombe, Drosophilamel anogaster and Homosapiens; a review of the considerable with in-species diversity. // Nucleic Acids Res. 1988. T. 16. № 17. C. 8207–8211.
[31] Suzuki H. и др. Comparison of Correspondence Analysis Methods for Synonymous Codon Usagein Bacteria // DNA Res. 2008. T. 15. № 6. C. 357–365.
[32] Eskesen S.T. и др. Periodicity of DNA inexons // BMC Mol. Biol. 2004. T. 5. C. 12.
[33] Sánchez J., López-Villaseñor I. A simple model to explain three-base 33 periodicity incoding DNA. // FEBS Lett. 2006. T. 580. № 27. C. 6413–6422.
[34] López-Villaseñor I., José M. V, Sánchez J. Three-base periodicity pattern sand self-similarity in wholebacterial chromosomes. // Biochem. Biophys. Res. Commun. 2004. T. 325. № 2. C. 467–478.
[35] Trotta E. The 3-Base Periodicity and Codon Usage of Coding Sequences Are Correlated with Gene Expressionat the Level of Transcription Elongation // PLoSOne. 2011. T. 6. № 6. C. 11.
[36] Електронний ресурс – режим доступу – URL: https://www.openthefile.net/ru/extension/fasta

УДК – 004.45

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ПОИСКА ТОЧЕК ИЗМЕНЕНИЯ ТРИПЛЕТНОЙ ПЕРИОДИЧНОСТИ

Кисляк С.В., ст. преп.
kisluak@ukr.net
Ильенок М.П., студент
marisha.illenok@gmail.com
Факультет биомедичной инженерии
Национальный технический университет «Киевский политехнический інститут шимени Игоря Сикорского»
г. Киев, Украина

Реферат- *с появлением методов секвенирования нового поколения наблюдается экспоненциальное увеличение молекулярно-биологических данных. Банки данных биологических последовательностей, такие как GenBank [1], UniProt [2], KEGG [3] и другие достигли больших размеров. При этом скорость описания просеквенованих последовательностей значительно отстает от скорости их накопления. При таком увеличении объема информации, решить проблему «отставания» невозможно без применения эффективных алгоритмов. Достаточно интересным, сложным и заслуживающим внимания и тщательного исследования является направление геномики, отвечающий за поиск эффективных алгоритмов для решения задачи идентификации генов про- и эукариот. Наиболее популярными являются программные продукты, позволяющие находить гены в соответствии к полученному парному выравниванию последовательностей методом динамического программирования. Такой подход не позволяет исследовать все участки нуклеотидной последовательности, поскольку большинство генов на протяжении эволюции изменялись за счет, например, точечных мутаций или кодирующая участок гена была получена за счет объединения нескольких генов. В такой ситуации информация об участках, которые формируют гибридные гены, будет отсутствовать в базах данных. Заслуживает внимания нетривиальный алгоритм, который был описан в работе [4]. Поиск точек изменения триплетной периодичности для прямой и обратно комплементарной последовательности дает возможность определения координаты, которые могут указывать на место объединение двух генов.*

Ключевые слова *– поиск точек изменения триплетной периодичности, нуклеотидна последовательность, методы выравнивания.*

UDC – 004.45

SOFTWARE FOR FINDING POINTS OF CHANGE IN TRIPLET PERIODICITY

Kisluak S., senior lecturer
kisluak@ukr.net
Illienok M.
Marisha.illenok@gmail.com
Faculty of Biomedical Engineering
National Technica University
"Igor Sikorsky Kyiv Polytechnic Institute"
Kyiv, Ukraine

Abstract

With the advent of new generation sequencing methods, an exponential increase in molecular-biological data is observed. Data banks of biological sequences, such as GenBank [1], UniProt [2], KEGG [3] and others have reached large sizes. At the same time, the velocity of the description of the sequenced sequences is considerably lagging behind the rate of their accumulation. With such an increase in the volume of information, solving the problem of "lagging" is impossible without the use of effective algorithms. A rather interesting, complex and deserving attention and detailed study is the direction of genomics responsible for finding efficient algorithms for solving the problem of pro and eukaryote gene identification. The most popular are software products that allow finding genes in accordance with the received pairwise alignment of the sequences using the dynamic programming method. This approach does not allow to explore all areas of the nucleotide sequence, since most of the genes have evolved during evolution, for example, by point mutations or the encoding region of the gene was obtained due to the incorporation of several genes. In such a situation, information about the sites forming hybrid genes will not be available in the databases. Noteworthy is the non-trivial algorithm that was described in [4]. Finding the points of triplet periodicity change for a direct and backward complementary sequence enables the definition of coordinates that may indicate the location of the combination of the two genes.

Key words – *search for points of change of triplet periodicity, nucleotide sequence, alignment methods.*