БІЯЛОГІЯ

УДК 595.752.2.51-76

ИЗМЕРЕНИЕ СЛОЖНОСТИ ФРАГМЕНТОВ ГЕНОМНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ ИНВАЗИВНОГО ВИДА ТЛИ *APHIS CRACCIVORA*

В.И. Чесалин, Н.В. Воронова-Барте, В.Л. Крук, А.В. Барышева, А.М. Шульгович, Р.С. Шулинский
Белорусский государственный университет

Актуальность исследования обусловлена тем, что современные методы аннотации опираются на уже известные данные и алгоритмы выравнивания, и это делает практически невозможным поиск и аннотацию принципиально новых генов без получения данных РНК-секвенирования.

Цель исследования – предложить новый способ измерения длин фрагментов геномной последовательнсти, учитывающий линейную сложность последовательности и уникальность ее строения.

***Материал и методы.** Программная реализация была осуществлена на языке программирования Python с использованием ряда ключевых технологий и библиотек этого языка.*

***Результаты и их обсуждение.** Компьютерный анализ строения геномной последовательности на примере Aphis craccivora показывает, что существует такое наименьшее и наибольшее значение длины, что для всех промежуточных расстояний между этими значениями экзонная и интронная части состоят из абсолютно различных фрагментов данной длины.*

***Заключение.** В результате компьютерных вычислений были найдены непересекающиеся множества большого количества фрагментов, содержащихся в экзонной и интронной частях геномной последовательности.*

Это может быть использовано в дальнейшем для обучения модели нейросети, задачей которой будет определение участков, характерных отдельно для экзонной и интронной областей. В данной работе предложен принципиально новый способ измерения длин фрагментов геномных последовательностей, которые в дальнейшем могут быть использованы для предсказания генов и их экзон-интронных моделей.

***Ключевые слова:** биоинформатика, аннотация, геномика насекомых, геномные последовательности.*

MEASURING THE COMPLEXITY OF FRAGMENTS OF THE GENOMIC SEQUENCE OF THE INVASIVE APHID *APHIS CRACCIVORA* SPECIES

V.I. Chesalin, N.V. Voronova-Bartet, V.L. Kruk, A.V. Barysheva, A.M. Shulgovich, R.S. Shulinsky
Belarusian State University

The relevance of the study is due to the fact that modern annotation methods rely on already known data and alignment algorithms, which makes it almost impossible to search for and annotate fundamentally new genes without obtaining RNA sequencing data.

The purpose of the study is to propose a new method for measuring the lengths of fragments of the genomic sequence, considering the linear complexity of the sequence and the uniqueness of its structure.

Material and methods. *The software implementation was carried out in the Python programming language, using a number of key technologies and libraries of this language.*

Findings and their discussion. *Computer analysis of the structure of the genomic sequence using the example of Aphis craccivora shows that there is such a smallest and largest length value that for all intermediate distances between these values, the exon and intron parts consist of absolutely different fragments of a given length.*

Conclusion. *As a result of computer calculations, disjoint sets of a large number of fragments contained in the exon and intron parts of the genomic sequence were found.*

This can be used in the future to train a neural network model, the task of which will be to identify such areas that are characteristic separately for the exon and intron regions. An absolutely new way of measuring lengths of genomic sequence fragments is proposed in the paper which can further be used to predict genes and their exon/intron models.

Key words: *bioinformatics, annotation, insect genomics, genomic sequences.*

Мониторинг животного мира представляет собой систему сбора, накопления и обработки большого количества биологической информации, в т.ч. данных полногеномного секвенирования. Для ее обработки необходимо наличие как вычислительных мощностей, так и актуального программного обеспечения. Современное ПО позволяет анализировать большой массив данных, однако множество разнообразных программ, которые могут требовать наличия стороннего ПО, значительно замедляет процесс получения и обработки данных.

На основе накопленных данных проводятся исследования, способствующие установлению хода эволюции геномов насекомых. Например, 76 сборок геномов членистоногих были использованы для улучшения понимания 500 миллионов лет эволюции путем характерных изменений в строении генов и белков во временном и филогенетическом контекстах. Получилось идентифицировать множество новых семейств генов, которые возникли на ранних стадиях эволюции членистоногих и во время диверсификации насекомых в современные отряды [1]. Аналогичным образом исследование 195 геномов насекомых выявило высокое разнообразие транспозонов у насекомых с различными уровнями консервативности в зависимости от филогенетического положения [2]. Появление сборок с использованием длинных прочтений, в частности, будет способствовать новому пониманию ранее трудных для характеристики аспектов генома (например, структуры часто повторяющихся областей).

Аннотация же генома необходима для характеристики функциональных элементов в геноме. Этот процесс можно разделить на два этапа: структурную аннотацию и функциональную. Структурная аннотация устанавливает, какие области сборки соответствуют определенным элементам, таким как гены (включая границы интрон–экзон) и транспозоны. Функциональная аннотация направлена на выявление функции и идентификацию генов и других элементов на основе сходства последовательностей, обычно с применением программного обеспечения blast.

1. Методы идентификации повторяющихся последовательностей можно разделить на две категории: поиск гомологии и предсказание ab initio. Поиск гомологии идентифицирует гомологичные повторяющиеся последовательности посредством сходства последовательностей. Метод прогнозирования ab initio использует структурные особенности повторяющейся последовательности для идентификации новых повторяющихся последовательностей. Этот метод имеет большие преимущества в прогнозировании повторяющейся последовательности с отчетливыми структурными особенностями,

такими как миниатюрные транспозоны с длинными инвертированными концевыми повторами. Для большинства геномов насекомых применяются как поиск гомологии, так и метод ab initio, позволяющий получить полный набор данных о повторяющихся последовательностях.

2. Идентификация некодирующей РНК. Некодирующая РНК – это класс генов РНК, которые не продуцируют белковые продукты, такие как транспортная РНК (тРНК), рибосомная РНК (рРНК), микроРНК (миРНК), малая ядрышковая РНК (мяРНК). Некодирующие РНК играют важную регуляторную роль в различных биологических процессах [3]. Соответственно, идентификация некодирующей РНК представляет важную задачу при аннотации генома.

3. Предсказание генов, кодирующих белок. Идентификация генов, кодирующих белок, является ключевой частью структурной аннотации. Существует три подхода к прогнозированию генов, кодирующих белок, из генома: идентификация гомологов известных генов, кодирующих белок, посредством сходства последовательностей; *de novo* прогнозирование генов, кодирующих белок, с помощью программного обеспечения, разработанного благодаря машинному обучению структур генов, кодирующих белок; определение экзонных областей путем прямого секвенирования транскриптома или метки экспрессируемой последовательности и выравнивание по собранным скэффолдам.

Однако вышеперечисленные методы опираются на уже известные данные и алгоритмы выравнивания, что делает практически невозможным поиск и аннотацию принципиально новых генов без получения данных РНК-секвенирования, или позволяет с низким предсказательным эффектом, при помощи скрытых марковских моделей. В данной работе мы предлагаем принципиально новый способ измерения длин фрагментов геномных последовательностей, которые в дальнейшем могут быть использованы для предсказания генов и их экзон-интронных моделей.

Материал и методы. Для измерений сложности фрагментов геномной последовательности был выбран геном инвазивного вида тли *A. craccivora.* Данная сборка характеризуется следующими параметрами: параметр N50 – 47498 п.н., сумма длин всех полученных контигов составила 336,445 Mb, что соответствует среднему размеру генома тли. Процессинг генома по бенчмаркам CEGMA и Busco показывает результаты в 93,47% и 98,31% соответственно. Итоговая аннотация представляет собой консенсус между аннотациями, проведенными по гомологии и ab initio, что позволяет предсказать и более 20000 генов со средней длиной 5430 п.н.

Программная реализация была осуществлена на языке программирования Python, с использованием ряда ключевых технологий и библиотек этого языка. Ниже представлено описание некоторых применяемых технологий.

NumPy: в процессе вычислений использовалась библиотека NumPy для работы с многомерными массивами и выполнения математических операций над ними. NumPy обеспечивает высокую производительность при выполнении операций над массивами данных.

Numba: для оптимизации времени выполнения критических участков кода на Python использовалась библиотека Numba. Декоратор @jit (Just-In-Time компиляция) был применен к некоторым функциям, что способствует ускорению их выполнения за счет компиляции в машинный код во время исполнения. В частности, Dc вычислялась с использованием данной библиотеки.

Matplotlib: для визуализации результатов анализа и вычислений применялась библиотека Matplotlib. Она предоставляет широкий спектр инструментов для создания графиков и диаграмм, что позволяет наглядно представить полученные результаты.

Результаты и их обсуждение. Для более глубокого понимания законов и принципов строения фрагментов геномной последовательности в данной статье предложен новый способ измерения длин таких фрагментов. В отличие от описания длины фрагмента как количества символов, участвующих в его записи, введенная метрика учитывает линейную сложность геномной последовательности и уникальность ее строения.

Пусть $G_n = \alpha_1\alpha_2 \dots \alpha_n$ – некоторая геномная последовательность, содержащая в записи n символов, где $\alpha_i \in \{A, T, C, G\}$.

На множестве G_n зададим метрику следующим образом:

$$d_C(\alpha_i, \alpha_j) = \begin{cases} 0, i = j, \\ d, i \neq j, \end{cases}$$

где d – наименьшее натуральное число, такое, что фрагменты геномной последовательности, состоящие из d символов, встречаются только один раз во фрагменте геномной последовательности $\alpha_i\alpha_{i+1}\dots\alpha_{j-1}\alpha_j$.

Заметим, что фрагмент $\alpha_i\alpha_{i+1}\dots\alpha_{j-1}\alpha_j$ содержит $j-i+1$ символов и в нем имеется ровно $j-i-d+2$ различных фрагментов, содержащих d символов. При увеличении количества символов во фрагменте $\alpha_i\alpha_{i+1}\dots\alpha_{j-1}\alpha_j$ на единицу длина $d_C(\alpha_i,\alpha_{j+1})$ может только увеличиться, но не более чем на 1.

Введенное понятие длины обобщает понятие линейной сложности кодовой последовательности, которое играет важную роль в практических задачах генерации последовательностей с помощью регистров сдвига с обратной связью, в криптографии и многих других задачах [4; 5]. Всего существует 64 последовательности длины $d_c=1$ и самая длинная из них состоит из 4 символов, например $ATCG$. Самая длинная последовательность длины $d_c=2$ содержит 17 символов, например $AATTACCAGGTCGCTGA$. В общем случае, если обозначить через g_l фрагмент геномной последовательности из l символов, то имеют место следующие оценки:

$$k \le l \le k-1+4^k,\ d_c(g_l)=k.$$

Компьютерная программа вычисления длины $d_c(G_n)$ геномной последовательности G_n основывается на адаптированном методе половинного деления. Для каждой последовательности G_n, содержащей n символов, определяем функцию $F_G(l)=1$, $1\le l\le n$, если все ее фрагменты g_l, состоящие из l символов, различны и $F_G(l)=0$ в противном случае. Для удобства полагаем $F_G(0)=0$. Очевидно, что $F_G(n)=1$. Введенная функция $F_G(l)$ является неубывающей и существует единственное число d, такое, что: $0<d\le n$, $F_G(d-1)=0$ и $F_G(d)=1$. Для нахождения числа d можно использовать следующий алгоритм:

1. Задаем такие числа d_{left} и d_{right}, что $F_G(d_{left})=0$, а $F_G(d_{right})=1$. Тогда, очевидно, что $d_{left}<d\le d_{right}$.

2. Вычисляется среднее $d_{mean}=\left[\frac{d_{left}+d_{right}+1}{2}\right]$, где [ ] – целая часть числа.

3. Если $F_G(d_{mean})=0$, то переменной d_{left} присваиваем значение d_{mean}. В противном случае переменной d_{right} присваиваем значение d_{mean}.

4. Если $d_{right}-d_{left}=1$, значит, $d=d_{right}$, останавливаем алгоритм. В противном случае переходим к шагу 1.

Приведем пример вычисления числа d. Пусть последовательность $G_n=AAATTT$, для нее $d_{left}=0, d_{right}=6$.

1. $d_{mean}=\left[\frac{6+0+1}{2}\right]=3$.

2. Среди фрагментов AAA, AAT, ATT, TTT нет повторений, следовательно, $F_G(3)=1$. Присваиваем d_{right} значение 3.

3. $d_{right}-d_{left}=3-0=3$.

4. $d_{mean}=\left[\frac{3+0+1}{2}\right]=2$.

5. Среди фрагментов AA, AA, AT, TT, TT есть повторения, следовательно, $F_G(2)=0$. Присваиваем d_{left} значение 2.

6. $d_{right}-d_{left}$=3−2=1, $d=d_{right}=3$.

В рассмотренном примере длина последовательности $AAATTT$ в новой метрике равна 3.

Далее исследуются множества всех фрагментов, содержащихся в экзонной и интронной частях геномной последовательности G_n, и устанавливается зависимость количества таких общих фрагментов от расстояния d. Компьютерный анализ строения геномной последовательности на примере Aphis craccivora показывает, что существует такое наименьшее d_{min}(экзон, интрон) и наибольшее d_{max}(экзон, интрон) значение длины, что для всех промежуточных расстояний d между этими значениями экзонная и интронная части состоят из абсолютно различных фрагментов данной длины.

Рассмотрим результаты вычислений на примере гена *Dual specificity tyrosine-phosphorylation-regulated kinase 2* класса *Cell growth and death* суперкласса *Cellular Processes.* Ген содержит 4609 букв 4-буквенного алфавита, область интрона – 2854, область экзона состоит из двух фрагментов,

первый фрагмент экзон[1] – 1065 букв и второй фрагмент экзон[2] – 690 букв. В этом примере d_{min}(экзон[1], интрон) = 8, d_{min}(экзон[2], интрон) = 6 и d_{max}(экзон, интрон) = 10. Полученные результаты отражены на рис. 1–2.

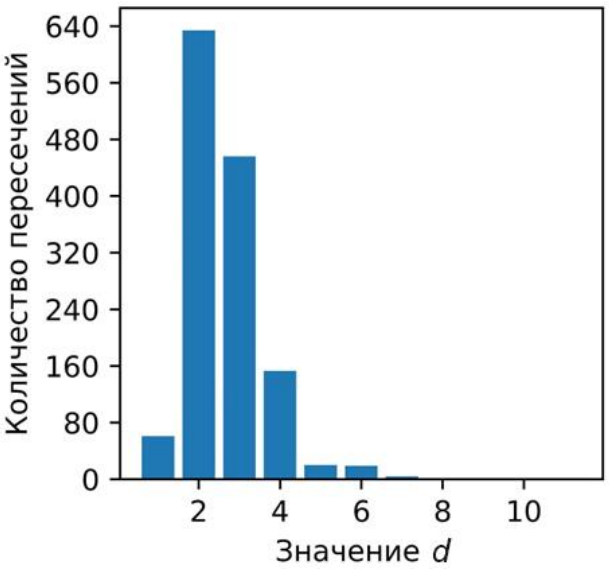

Рис. 1. **Зависимость количества пересечений от значения d для пары экзон[1] – интрон**

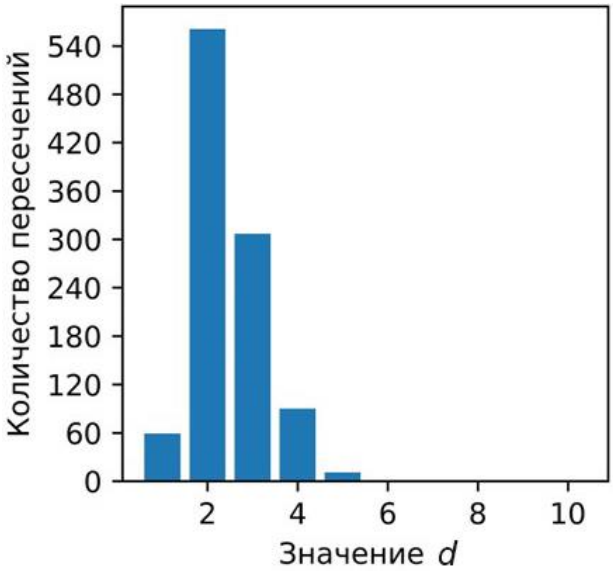

Рис. 2. **Зависимость количества пересечений от значения d для пары экзон[2] – интрон**

Следует отметить, что с увеличением значения d количество пересечений уменьшается.

Приведем пример фрагментов области экзон[1] со значением $d = 8$:

TCAGCCACCTCTTGTCTTCAGCCAAT, GCTGCTGCTG, TAACGCTGCTGCTG.

Количество таких фрагментов – 186159.

Для области экзон[2] со значением $d = 8$:

CTGAATCAAATCATCATCAAACATTGCCC ATCATCATCA, TCAAATCATCATCA.

Количество таких фрагментов – 52350.

Для области интрон со значением $d = 8$:

GATTTAAATTTAAAAAT, GTAGAAAAAAAATCTAATG, TAATAATATGATTTAAATTTAAA.

Количество таких фрагментов – 128236.

Для эффективного анализа больших объемов генетической информации требуется использование эффективных методов предварительной обработки. Например, рассмотрим применение алгоритма BPE (Byte Pair Encoding, Парного Кодирования Байтов).

Этот алгоритм, изначально разработанный для сжатия текстовых данных, успешно используется в области обработки геномных последовательностей. Принцип BPE заключается в последовательном объединении наиболее часто встречающихся пар символов, что позволяет создавать новый алфавит, в котором последовательность становится короче.

В контексте нуклеотидных последовательностей символы могут представлять отдельные нуклеотиды. Применение BPE позволяет выделить значимые элементы последовательностей и снизить длину последовательности, сохраняя при этом информацию о ее структуре.

Для составления словаря алгоритма BPE было использовано множество из 13847184 кодирующих частей геномных последовательностей насекомых из открытой базы данных RefSeq, содержавших 242 известных консервативных домена. Большинство геномов принадлежали насекомым отряда Чешуекрылые (1968 образцов), Двукрылые (949 образцов) и Перепончатокрылые (497 геномов). Полный состав выборки приведен в табл.

Таблица

Состав генеральной совокупности последовательностей геномов насекомых

Таксон	Число геномов	Таксон	Число геномов
Thysanoptera	5	*Coleoptera*	267
Phasmatodea	18	*Psocoptera*	3
Zygentoma	1	*Mantodea*	1
Hemiptera	150	*Lepidoptera*	1968
Megaloptera	5	*Archaeognatha*	1
Odonata	14	*Siphonaptera*	1
Strepsiptera	1	*Diptera*	949
Orthoptera	21	*Phthiraptera*	9
Neuroptera	3	*Hymenoptera*	497
Blattodea	11	*Trichoptera*	49
Dermaptera	4	*Plecoptera*	1
Ephemeroptera	9		

Проведен ряд численных экспериментов по выявлению зависимости коэффициента сжатия от размера словаря. В результате численного анализа для последующей обработки выбрана длина словаря 4096 как оптимальная для имеющегося аппаратного обеспечения, поскольку BPE с такой длиной словаря позволяет снижать длину последовательности примерно в 4 раза, пусть и за счет существенного увеличения размера словаря.

Заключение. Следует особо отметить, что в результате компьютерных вычислений были найдены непересекающиеся множества большого количества фрагментов, содержащихся в экзонной и интронной частях геномной последовательности. Это может быть использовано в дальнейшем для обучения модели нейросети, задачей которой будет определение таких участков, характерных отдельно для экзонной и интронной областей. Подобная модель сможет упростить последующую сборку геномной последовательности из множества ридов, полученных путем его секвенирования.

ЛИТЕРАТУРА

1. Gene content evolution in the arthropods / G.W.C. Thomas [et al.] // Genome Biol. – 2020. – Vol. 21, № 1. – P. 15.
2. Gilbert, C. Transposable Elements and the Evolution of Insects / C. Gilbert, J. Peccoud, R. Cordaux // Annu. Rev. Entomol. – 2021. – Vol. 66, № 1. – P. 355–372.
3. Huntzinger, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay / E. Huntzinger, E. Izaurralde // Nat. Rev. Genet. – 2011. – Vol. 12, № 2. – P. 99–110.
4. Massey, J.L. Linear Complexity of Periodic Sequences / J.L. Massey, S. Serconek // Lecture Notes in Computer Science. – 1996. – Vol. 1109. – P. 358–371.
5. Логачёв, О.А. Булевы функции в теории кодирования и криптологии / О.А. Логачёв, А.А. Сальников, В.В. Ященко. – М.: МЦНМО, 2004. – 470 с.

REFERENCES

1. Gene content evolution in the arthropods / G.W.C. Thomas [et al.] // Genome Biol. – 2020. – Vol. 21, № 1. – P. 15.
2. Gilbert, C. Transposable Elements and the Evolution of Insects / C. Gilbert, J. Peccoud, R. Cordaux // Annu. Rev. Entomol. – 2021. – Vol. 66, № 1. – P. 355–372.
3. Huntzinger, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay / E. Huntzinger, E. Izaurralde // Nat. Rev. Genet. – 2011. – Vol. 12, № 2. – P. 99–110.
4. Massey, J.L. Linear Complexity of Periodic Sequences / J.L. Massey, S. Serconek // Lecture Notes in Computer Science. – 1996. – Vol. 1109. – P. 358–371.
5. Logachev O.A., Salnikov A.A., Yashchenko V.V. *Bulevy funktsii v teorii kodirovaniya i kriptologii* [Bull Functions in the Theory of Coding and Cryptology], M.: MTsNMO, 2004, 470 p.

Поступила в редакцию 27.12.2022

Адрес для корреспонденции: e-mail: Vladimir.chesalin@gmail.com – Чесалин В.И.