Выбор эталона для определения нуклеотидных последовательностей молекул ДНК

С. С. ГОЛУБЕВ*, С. А. КОНОНОГОВ*, Н. В. РАВИН, К. Г. СКРЯБИН****

**Всероссийский научно-исследовательский институт метрологической службы, Москва, Россия, e-mail: golubev@vniims.ru*

*** Центр «Биоинженерия» РАН, e-mail: office@biengi.ac.ru*

Предложено использовать в качестве эталонной последовательности для калибровки и метрологического обеспечения геномных анализаторов (секвенаторов) фрагмент последовательности ДНК плазмиды pUC18, состоящей из 271 пары оснований и получаемый по эталонной методике. Данная последовательность является оптимальной для определения метрологических характеристик геномных анализаторов из-за своей стабильности и наличия нескольких участков с повторяющимися нуклеотидами. Приведен пример вычисления характеристики геномного анализатора – вероятности ошибки при измерении геномной последовательности.

***Ключевые слова*:** ДНК, эталон, метрологическое обеспечение, генетическая последовательность, геномный анализатор.

Fragment of pUC18 plasmid is offered as reference sequence for calibration and metrological service of genome analyzers. This fragment consists of 271 nucleotides pairs and can be produced by reference method based on restriction of pUC19 DNA with RsaI enzyme and electrophoresis. This sequence is very useful for measuring of metrological characteristics of genome analyzers because of its stability and few places of occurrences of repeating nucleotides. It was shown how to calculate simple characteristics of genome analyzers – the probability of errors in measuring nucleotide sequence of DNA..

***Key words*:** DNA, etalon, reference sequence, metrological support, genetic code sequence, genome analyzer.

Носителем генетической информации у всех живых организмов являются молекулы нуклеиновых кислот, в подавляющем большинстве случаев дезоксирибонуклеиновой (ДНК). Она представляет собой длинную полимерную молекулу, состоящую из последовательности пуриновых и пиримидиновых оснований. При этом в ДНК живых организмов может быть только четыре вида таких оснований – аденин, тимин, гуанин и цитозин. Таким образом, генетическая информация организма на молекулярном уровне определяется последовательностью следования этих оснований в цепи ДНК. В реальных геномах количество оснований в этих полимерных цепочках насчитывает миллионы и миллиарды единиц.

В современной биологической науке принято записывать последовательность не в виде химической формулы ДНК, а в виде последовательности букв А, Т, Г и Ц (A, T, G и C в англоязычной записи), обозначающих соответствующие нуклеотидные основания. Нуклеотидная последовательность ДНК может иметь вид, например, AATGCA…

Необходимо пояснить сложившуюся терминологическую особенность. В классической химии принято давать названия соединений исходя из состава молекулы и ее структуры, изображаемой, как правило, структурной формулой. В данном случае под термином ДНК понимают множество структурно различных молекул. Каждый вид живого организма обладает своей уникальной последовательностью ДНК, которая и определяет его характеристики. Более того, ДНК даже у отдельных организмов одного вида (например, разных людей) могут отличаться. Однако в сложившейся терминологии любую последовательность оснований принято называть ДНК.

Для чтения нуклеотидной последовательности ДНК применяют геномные анализаторы (секвенаторы). Почти во всех современных приборах использован принцип секвенирования путем синтеза, т. е. нуклеотидная последовательность ДНК определяется последовательностью включения отдельных нуклеотидов во вновь синтезируемую по принципу комплементарности нуклеотидную цепь. Такой процесс аналогичен ферментативной репликации ДНК в клетке, осуществляемой ДНК-полимеразами. Здесь необходимо отметить, что речь идет не об образце, состоящем из одной цепочки, а о препарате, содержащем наработанные тысячи идентичных копий цепочки ДНК исследуемого организма. Современное состояние молекулярной биологии позволяет искусственно реализовывать процесс репликации – воспроизведения, или тиражирования молекул, свойственного всем живым организмам.

Один из методов высокопроизводительного секвенирования ДНК – параллельное пиросеквенирование [1]. Для его проведения микрочастицы, на которых иммобилизованы миллионы идентичных копий индивидуальных одноцепочечных фрагментов ДНК, помещают в микроячейки, расположенные на плоской пластине. Во время работы геномного анализатора растворы нуклеотидов A, C, G и T последовательно добавляются в проточную ячейку, содержащую пластину, и удаляются после реакции. Затем цикл внесения–удаления последовательно каждого из четырех нуклеотидов повторяется. Если на определенном цикле через микроячейку проходит нуклеотид, комплементарный матрице, иммобилизованной в этой микроячейке, то цепь удлиняется из-за встраивания этого нуклеотида специальным ферментом – полимеразой. Добавление нуклеотида приводит к высвобождению пирофосфата и далее к реакции, в результате которой генерируется световой сигнал, регистрируемый ПЗС-камерой прибора для каждой микроячейки. Интенсивность сигнала пропорциональна количеству нуклеотидов, встроенных в цепь ДНК. Таким

образом, последовательность растворов, которые дают хемилюминесцентный сигнал в конкретной микроячейке, позволяет определить нуклеотидную последовательность матрицы ДНК.

В связи с широким использованием секвенирования ДНК как в научных исследованиях, так и в практических областях деятельности (биотехнологии, медицине, сельском хозяйстве и т. д.) возникает необходимость метрологического обеспечения данных процессов. Идет дискуссия о том, является ли процесс установления генетической последовательности измерением или же это определение качественного свойства и понятие измерения к нему не относится. С точки зрения авторов, генетическая последовательность – это новая величина, поэтому ее определение является измерением. Ее действительно нельзя выразить числом (конечно, буквы можно заменить цифрами, и последовательность AATGCT записать, например, как 114234, но к такой форме записи не будет применима привычная алгебра). Однако при условии сколь угодно длинной последовательности (а в природе реализуется именно эта ситуация, длины последовательностей кода хоть и конечны, но чрезвычайно велики) многообразие получающихся цепочек ДНК также сколь угодно велико. Более того, можно ввести понятие равенства как для всей цепочки в целом, так и для ее фрагментов. Последнее особенно важно с учетом механизмов синтеза органических молекул внутри живых организмов по имеющимся последовательностям ДНК.

Последовательность ДНК состоит из двух комплементарных цепочек. Комплементарность – одно из основных свойств цепочек ДНК (последовательности «букв»). Оно проявляется в том, что каждому нуклеотиду прямой последовательности A, T, G или C соответствует комплементарный нуклеотид обратной последовательности (по закону A↔T, G↔C), образующей вторую (обратную) цепочку. При этом комплементарная цепочка записывается с противоположного конца из-за химиче-

ского строения молекулы ДНК (например, комплементарная последовательность цепочки AACTGT будет иметь вид ACAGTT).

В действительности биологи давно уже оперируют такими последовательностями в качестве новой величины, однако необходимая для этого база для технологических процессов и измерений на сегодняшний день полностью отсутствует. Поэтому актуальной является разработка метрологического обеспечения процесса определения (измерения) цепочки последовательности генетического кода.

Важная задача современной метрологии – переопределение единиц физических величин через фундаментальные физические константы – численные параметры окружающего мира, значения которых определяются природой и (относительно) стабильны во времени [2]. В молекулярной биологии также существует возможность в качестве эталонной выбрать короткую последовательность, представляющую фрагмент генома хорошо изученного организма. Необходимость в такой эталонной последовательности обусловлена следующим.

Приборы для чтения нуклеотидных последовательностей – геномные анализаторы – при прочтении цепочек могут допускать ошибки, следовательно, найденные с их помощью последовательности могут отличаться от реального значения одним или несколькими нуклеотидами в определенных позициях. Соответственно, конечной задачей метрологического обеспечения этой области является установление вероятности ошибки при прочтении нуклеотидной последовательности. Таким образом, на выходе процесса секвенирования помимо цепочки AATGCT…, которую можно условно записать функцией $C(N)$ (в данном случае $C(1)$=A, $C(2)$=A, $C(3)$=T и т. д.), должна быть еще функция $P(N)$, значения которой представляют собой вероятность ошибки в N-м нуклеотиде.

Опыт работы и принцип действия пиросеквенатора показывают, что вероятность ошибок чтения повышается в тех местах, где подряд

расположено несколько одинаковых нуклеотидов. Поскольку интенсивность сигнала пропорциональна количеству нуклеотидов, встроенных в цепь ДНК на данном цикле секвенирования, с ростом числа идущих подряд одинаковых нуклеотидов относительная разница в интенсивности сигнала будет снижаться. Поэтому эталонная последовательность должна по возможности содержать участки повторов одинаковых нуклеотидов, что позволит с ее помощью экспериментально исследовать вероятности таких ошибок у существующих геномных анализаторов.

В качестве эталонной последовательности предложено выбрать один из фрагментов, на которые последовательность pUC18 [3] разделяется после обработки ферментом рестрикции RsaI (рис. 1). Конкретнее, речь идет о фрагменте, состоящем из 271 пары нуклеотидов, который представлен на рис. 2, где прямая цепочка отмечена серым фоном.

Кольцевая последовательность pUC18 имеет три участка (позиции №23, 294 и 2033, соответствующие стрелкам на рис. 1), где она расщепляется при обработке ферментом рестрикции RsaI на три линейных участка, содержащих 271, 676 и 1739 пар нуклеотидов. В качестве стандартного образца использован фрагмент плазмиды pUC18, состоящий из 271 пары нуклеотидов. Для его отделения от двух других образовавшихся линейных участков (676 и 1739 пар нуклеотидов) служит метод электрофореза в агарозном геле [4]. Из-за различной «длины» фрагментов скорость их движения в агарозном геле под действием электрического поля будет различна: «короткие» фрагменты двигаются быстрее (скорость движения обратно пропорциональная логарифму «длины» фрагмента). Это дает возможность разделить три фрагмента с отличающимся количеством пар оснований и выделить самый «короткий», состоящий из 271 пары нуклеотидов.

Данная цепочка, во-первых, очень хорошо изучена, последовательность pUC18 прочтена многократно в различных лабораториях во всем

мире и обладает свойством природной «фундаментальности» – она одинакова у всех экземпляров этой плазмиды благодаря исключительно высокой точности копирования генетической информации в природе [5]. Во-вторых, она содержит участки, в которых одинаковые нуклеотиды повторяются до 5 раз подряд, , что позволяет исследовать геномные анализаторы на прочтение таких участков. В-третьих, ее достаточно просто получить например, по следующей методике. Вначале из бактерий, содержащих плазмиду, выделяют препарат ДНК pUC18. Затем приготавливают эталонную последовательность (стандартный образец [6]). Эта процедура состоит из двух этапов: обработки препарата ДНК pUC18 ферментом рестрикции (рестриктазой) RsaI и его очистки – получения фрагмента, содержащего 271 пару нуклеотидов. На первом этапе к образцу pUC18 добавляют рестриктазу RsaI и выполняют инкубирование смеси при температуре 37 °С на протяжении нескольких часов. Второй этап осуществляется в электрофорезной камере за счет разделения движущихся в агарозном геле под действием электрического поля фрагментов ДНК. «Короткие» фрагменты в таких условиях движутся быстрее «длинных» и могут быть от них отделены. После процедуры очистки от геля образец эталонной последовательности готов для дальнейшей работы.

Из экспериментальных исследований геномного анализатора GS FLX (Roche) с помощью данной цепочки уже можно сделать некоторые выводы о его метрологических характеристиках. Во-первых, вычислена средняя вероятность ошибки. Пусть цепочка, состоящая из 271 пары нуклеотидов, прочтена 1000 раз. Тогда всего прочтено 271·1000=271 000 нуклеотидов. Поскольку известна истинная последовательность данных цепочек (она соответствует приведенной на рис. 2), можно определить количество E ошибочно прочтенных нуклеотидов. Средняя вероятность ошибочного прочтения нуклеотида в данном случае P = E / 271 000. Во-вторых, установлено количество цепочек (из 1000 прочтений), не содер-

жащих ни одной ошибки E_0; содержащих только одну; только одну, две; только две, три, только три; а также четыре и более ошибок: E_1, E_2, E_3, E_{4+}. И, наконец, в-третьих, найдено, с какой вероятностью геномный анализатор ошибается при чтении стоящих подряд одинаковых двух, трех, четырех и пяти нуклеотидов. Для этого выбирают место цепочки с двумя подряд расположенными нуклеотидами (например, см. рис. 2, 13 позиция). Затем определяют долю ошибочных прочтений данного места из 1000 прочтений и рассчитывают вероятность такой ошибки $P_2 = E_{c2}/1000$. Аналогично находят вероятности P_3, P_4, P_5. Вероятности ошибок для более длинных гомологичных участков с помощью данного стандартного фрагмента определить невозможно, поскольку таких участков он не содержит.

Заключение. В качестве эталонной последовательности для метрологического обеспечения геномных анализаторов предложено выбрать фрагмент вектора pUC18, состоящий из 271 пары нуклеотидов, получаемый после обработки ДНК pUC18 ферментом рестрикции RsaI. Представлен способ определения метрологических характеристик геномных анализаторов (достоверности чтения генетических последовательностей) по рассмотренной эталонной последовательности.

Решение данной задачи представляет значительный интерес для биоинженерии и биотехнологии и должно лечь в основу метрологического обеспечения процессов измерения генетической последовательности.

Работа выполнена в рамках Федеральной целевой программы «Развитие инфраструктуры наноиндустрии в Российской Федерации на 2008-2011 годы» (государственный контракт 16.648.11.3005).

Литература

1. Margulies M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.J., Chen Z., Dewell S.B., Du L., Fierro J.M., Gomes X.V., Godwin B.C., He W., Helgesen S., Ho C.H., IrzykG.P., JandoS.C., Alenquer M.L.I., Jarvie T.P., Jirage K.B., Kim J.B., J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 2005. 437: 376-380

2. **Кононогов С. А.** Метрология и фундаментальные физические константы. М.: Стандартинформ, 2008.

3. **Yanisch-Perron C. e. a.** Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors // Gene. 1985. N 33. P. 103–119.

4. **Рыбчин В. Н.** Основы генетической инженерии. Учебник для вузов. СПб.: Изд-во СПбГТУ. 2002. С. 463.

5. **Kunkel T. A.** DNA replication fidelity. J Biol Chem. 2004. 279(17):16895-8

6. **Международный** словарь по метрологии. Основные и общие понятия и соответствующие термины. С.-Пб.: НПО «Профессионал», 2010.

Дата принятия 21.09.2011 г.

Подрисуночные подписи

Рис. 1. Рестрикционная карта плазмиды pUC18

Рис. 2. Фрагмент плазмиды pUC18, состоящий из 271 пары нуклеотидов; темным фоном отмечена прямая цепочка