Метрологическое обеспечение секвенирования молекул ДНК

С.С. Голубев[1)], С.А. Кононогов[1)], Ю.А. Кудеяров[1)], А.В. Марданов[2)], П.Ю. Николаева[1)], Н.В. Равин[2)], К.Г. Скрябин[2)]

1). ФГУП «Всероссийский научно-исследовательский институт метрологической службы»

2). Учреждение Российской академии наук Центр «Биоинженерия» РАН

В современной биологической науке и ее приложениях (геномика, генетическая инженерия, молекулярная диагностика и т.д.) часто возникает необходимость в прочтении генетической информации живых организмов. Материальным носителем генетической информации у всех живых организмов являются молекулы нуклеиновых кислот, в подавляющем большинстве случаев представляющие дезоксирибонуклеиновую кислоту – ДНК. Это длинная полимерная молекула, состоящая из последовательности пуриновых и пиримидиновых оснований. В молекуле ДНК имеется только четыре вида таких оснований – аденин, тимин, гуанин и цитозин, обозначаемых буквами А, Т, Г и Ц (A, T, G и C в английской транскрипции), т.е. генетическая информация любого организма на молекулярном уровне определяется порядком следования этих оснований в цепи ДНК. В реальных геномах длина таких полимерных цепочек насчитывает 10^6-10^{10} пар оснований.

Методы чтения нуклеотидной последовательности ДНК, т.е. определения порядка следования нуклеотидов в цепи ДНК, принято называть методами *секвенирования* (последовательного прочтения). Для этой цели применяются специальные приборы – *секвенаторы.* Почти все современные секвенаторы используют для определения нуклеотидной последовательности ДНК принцип «секвенирования путем синтеза», т.е. нахождения нуклеотидной последовательности исследуемой ДНК посредством определения порядка включения нуклеотидов во вновь синте-

зируемую по принципу *комплементарности* (дополнительности) нуклеотидную цепочку. Напомним, что под *комплементарностью* понимается одно из основных свойств цепочек ДНК (последовательностей «букв»), заключающееся в том, что в молекулах ДНК, как правило, являющихся двухцепочечными, каждому нуклеотиду прямой последовательности (A, T, G, C) соответствует комплементарный нуклеотид обратной последовательности по правилу (A↔T, G↔C). Эта комплементарная последовательность образует вторую (обратную) цепочку. Комплементарная цепочка записывается с противоположного конца из-за химического строения молекулы ДНК (например, к цепочке AACTGT комплементарная последовательность будет записана в виде ACAGTT).

Определение геномной последовательности (последовательности нуклеотидов в молекуле ДНК), является чрезвычайно актуальной задачей, над решением которой активно работают биохимические лаборатории всего мира. Практически ежемесячно появляется информация о частично или полностью расшифрованных геномах новых видов живых организмов. Это стало возможным благодаря прогрессу биологического приборостроения – автоматизации рутинных процедур, миниатюризации, объединению различных модулей в интегрированные многофункциональные системы и широкому использованию возможностей вычислительной техники, приведших к стремительному увеличению производительности отдельных биологических экспериментов и выведению исследований на новый качественный уровень.

В развитых странах активно финансируются масштабные проекты в области геномики. Одновременно с этим ведущие научные лаборатории и биотехнологические компании стремятся снизить стоимость расшифровки геномов. Если первый черновой вариант генома человека, законченный в 2001 году, стоил около трехсот миллионов долларов (вместе с технологиями, сделавшими его возможным, обошелся примерно в 3

миллиарда), то в 2004 г. Национальный институт здравоохранения США запустил программу исследований (с семидесятьюмиллионными грантами) по удешевлению стоимости расшифровки больших геномов до 100000 – 1000 долларов.

Одним из современных методов высокопроизводительного секвенирования ДНК является параллельное *пиросеквенирование* [1]. В работе [2] дано описание этого метода. Тем не менее, имея в виду его важность для понимания дальнейшего изложения, считаем необходимым более подробно рассмотреть принцип пиросеквенирования.

Для проведения пиросеквенирования микрочастицы, на которых иммобилизованы миллионы идентичных копий индивидуальных одноцепочечных фрагментов ДНК (это требует определенной технологии клонирования таких копий), помещают в микроячейки, расположенные на плоской пластине. Во время работы секвенатора растворы нуклеотидов A, C, G и T последовательно добавляются в проточную ячейку, содержащую пластину, и удаляются после реакции секвенирования. При прохождении через микроячейку нуклеотида, комплементарного матрице, иммобилизованной в этой микроячейке, происходит встраивание этого нуклеотида в цепочку нуклеотидов и ее удлинение. При этом высвобождается пирофосфат и активируется цепь ферментативных реакций, результатом которых является генерация светового сигнала, регистрируемого CCD-камерой прибора для каждой микроячейки, расположенной на пластине. Интенсивность светового сигнала пропорциональна количеству нуклеотидов, встроенных в цепь ДНК. Схема работы анализатора геномного изображена на рис.1.

Затем цикл внесения-удаления последовательно для каждого из четырех нуклеотидов повторяется. Таким образом, последовательность растворов, которые дают хемилюминесцентный сигнал в конкретной

микроячейке, позволяет определить нуклеотидную последовательность матрицы ДНК.

Из сказанного следует, что процесс секвенирования неизвестного генома состоит из следующих шагов:

1. Разбиение молекул ДНК, образующих геном и состоящих из миллионов пар нуклеотидов, на фрагменты по 300-500 нуклеотидов, пригодные для чтения. Связывание отдельных молекул ДНК с микрочастицами
2. Наработка большого количества идентичных копий исследуемой молекулы ДНК методами молекулярной биологии, аналогичными процессам, происходящим в живых клетках при их делении.
3. Подготовка их к процедуре чтения – помещение микрочастиц с ДНК в ячейки на специальной пластине секвенатора.
4. Определение тех ячеек (методом присоединения комплементарного основания и высвобождения пирофосфата), где первым идет один или несколько нуклеотидов A.
5. Повторение шагов 4-5 для трех остальных нуклеотидов (C, T, G).
6. Так как в результате выполнения шагов 4-5 цепочки в ячейках укоротились на один или более нуклеотидов, то возвращаются к шагу 4 для дальнейшего чтения цепочек.
7. Шаги 4-6 повторяют, пока цепочки в ячейках не будут полностью определены (прочтены), т.е. пока происходит присоединение нуклеотидов с высвобождением пирофосфата.

Реально это определение осуществляется с помощью программного обеспечения, когда компьютер «определяет» последовательность включенных в синтезируемую цепь нуклеотидов, связывая зарегистрированные от каждой микроячейки вспышки с типом нуклеотида, присутствующего в микроячейке на определенном цикле работы прибора, и фиксирует последовательность их появления, складывая «буквы» (нук-

леотиды) в «текст» (последовательность оснований в ДНК). При этом он делает это одновременно в миллионах находящихся на пластине микроячеек.

Как было уже отмечено в п. 1 описания процесса секвенирования, для расшифровки полной нуклеотидной последовательности генома он случайным образом разбивается на фрагменты, длина которых сопоставима с длиной чтения индивидуальной реакции пиросеквенирования (300-500 нуклеотидов). Для полного покрытия генома такой набор фрагментов при условии их случайного расположения в геноме должен быть избыточным, т.е. суммарная длина фрагментов должна в несколько раз превышать размер генома. При пиросеквенирвании нарабатывается большое количество коротких «чтений-текстов», которые поступают на вход специальной программы, запускаемой на компьютере. Программа находит места перекрывания «текстов» и, если такое перекрывание находится, то «тексты» объединяются в более длинный «контиг» (непрерывную последовательность, образуемую группой из нескольких перекрывающихся секвенированных участков ДНК). Упорядочивая перекрывающиеся «тексты», программа в идеальном случае выстраивает полную последовательность генома. Высокая достоверность расшифровки последовательности достигается посредством многократного прочтения одного и того же фрагмента. Отдельные прочтения одного и того же участка ДНК выравниваются относительно друг друга с учетом интенсивности световых сигналов, фиксируемых при прохождении через микроячейку того или иного нуклеотида. Основная проблема программного прочтения цепочки нуклеотидов сводится к тому, что отдельные фрагменты последовательности нуклеотидов могут повторяться в разных местах последовательности, что может приводить к неоднозначному и, следовательно, к неверному результату определения последовательности «контигов».

Алгоритмы обработки генетической информации при секвенировании постоянно совершенствуются и модифицируются. Так, например, в работе [1] описан алгоритм, позволяющий вносить в процесс чтения последовательности нуклеотидов поправки, обусловленные необходимостью учета ряда физических процессов, приводящих к нежелательным сдвигам при считывании хемилюминесцентных сигналов из разных микроячеек. При этом, естественно, возникает вопрос о степени адекватности используемых алгоритмов обработки генетической информации решаемой задаче расшифровки генома, и хотя такая адекватность была неоднократно подтверждена независимой программной обработкой в разных лабораториях, тем не менее, такая задача продолжает оставаться актуальной с учетом многочисленных модификацией используемых программных продуктов.

В связи с широким применением секвенирования ДНК в научных исследованиях и в практических областях деятельности (биотехнология, медицина, сельское хозяйство и т.д.) остро встает вопрос о необходимости его метрологического обеспечения. Проблема становится особенно актуальной в связи с большим затратным характером процедуры секвенирования и высокими требованиями к достоверности результатов определения генома, обусловленными не в последнюю очередь возможными социальными последствиями такого определения (например, последовательности индивидуального генома человека могут указывать на предрасположенность к различным заболеваниям). Поскольку проблема метрологического обеспечения биологического приборостроения в нашей стране проявилась в самое последнее время, т.е. для отечественной метрологии является новой, то возникает вопрос о характере и месте этого бурно развивающегося вида деятельности в существующей системе обеспечения единства измерений.

В работе [2] предлагается рассматривать генетическую последовательность молекулы ДНК как новую величину для метрологии, и возникает задача нахождения для нее необходимой ниши. И если определение положения конкретного нуклеотида в последовательности в значительной степени носит качественный характер (есть – нет), то для всей последовательности нуклеотидов достоверность ее определения характеризуется, например, такой количественной характеристикой, как доля безошибочного определения (прочтения). С этой точки зрения, по аналогии с измерениями неархимедовых величин, *определение* последовательности нуклеотидов в молекуле ДНК можно рассматривать как вид *измерения*, целью которого является нахождение истинной последовательности нуклеотидов в молекуле ДНК и количественное определение правильности такого нахождения. Для этого используется специальный прибор – секвенатор, который с большой долей уверенности может быть отнесен к устройствам с измерительными функциями, на которые, в соответствии с Федеральным Законом РФ «Об обеспечении единства измерений» [3], распространяются процедуры метрологического обеспечения. Уместно также будет отметить, что последовательность нуклеотидов в ДНК, как объект метрологической деятельности, упоминается в таком хорошо известном международном документе по метрологии, как «Международный словарь по метрологии» [4] в разделе «Стандартные образцы», которые, как известно, играют роль «эталонов» для веществ (материалов).

В действительности биология давно уже оперирует такими последовательностями как новыми величинами, однако необходимое для этого метрологическое обеспечение – база для технологических процессов и измерений – в нашей стране на сегодняшний день полностью отсутствует. В этой ситуации было бы непростительно на основании формальных соображений игнорировать и оставлять без соответствующего метрологического обеспечения такую бурно развивающуюся сферу научно-

технической деятельности как молекулярная биология, в целом, и анализ последовательности нуклеотидов в геноме, в частности.

Когда говорят о системе метрологического обеспечения какого-то вида измерений, то, прежде всего, имеют в виду определение измеряемой величины и единицы ее измерения, разработку эталона этой величины и соответствующей нормативной базы (методики испытаний, калибровки и измерений). Все эти понятия в полной мере относятся к величинам, свойства которых описываются *шкалами отношений* [5], в том числе абсолютными шкалами.

При определении места секвенирования молекул ДНК в системе обеспечения единства измерений следует учитывать то, что каждая молекула ДНК уникальна. Из приведенного выше материала следует, что свойства последовательности нуклеотидов в молекуле ДНК могут быть описаны *шкалой наименований* [5], т.е. такой шкалой, которая не имеет нуля и единицы измерений, в ней отсутствуют отношения сопоставления типа «больше – меньше», но, тем не менее, присутствуют отношения (оценки) эквивалентности. Известно, что при условии сколь угодно большой длины последовательности нуклеотидов, а в природе реализуется именно эта ситуация, когда длины последовательностей кода ДНК хоть и конечны, но чрезвычайно велики, многообразие получающихся цепочек нуклеотидов также сколь угодно велико. При этом возникает возможность введения понятия эквивалентности как для всей цепочки в целом, так и для ее фрагментов, характеризующих аминокислотную последовательность конкретного белка.

Под метрологическим обеспечением прочтения нуклеотидной последовательности ДНК, имеется в виду, прежде всего, технологическая и измерительная базы для проведения секвенирования.

В первую очередь необходима разработка стандартного образца «эталонной» последовательности нуклеотидов, позволяющего проводить

калибровку секвенаторов, т.е. должна быть достоверно известна последовательность нуклеотидов ДНК, чтение которой позволит проводить контроль правильности функционирования и метрологических характеристик секвенатора (анализатора геномного).

В работе [2] в качестве стандартного образца предлагается использовать «эталон» – один из фрагментов последовательности, обозначаемой в биологии как плазмида pUC18. Речь идет о фрагменте, состоящем из 271 пары нуклеотидов, который очень хорошо изучен и многократно прочитан в различных лабораториях мира. Отличительная его особенность заключается в свойстве природной «фундаментальности», поскольку, благодаря исключительно высокой точности копирования генетической информации в природе, последовательность нуклеотидов одинакова у всех экземпляров плазмиды pUC18. Кроме того, эта цепочка содержит участки из одинаковых нуклеотидов вплоть до пяти штук подряд, что позволяет исследовать секвенаторы на правильность прочтения таких «проблемных» участков. Дело в том, что при наличии в цепи нескольких одинаковых нуклеотидов подряд из-за нарушения линейной зависимости (пропорциональности) регистрируемой CCD-камерой интенсивности вспышек от числа включенных в растущую цепь нуклеотидов возможно неправильное прочтение секвенатором таких участков особенно тогда, когда друг за другом следуют три – четыре одинаковых нуклеотида и более.

В качестве метрологических характеристик секвенаторов предлагается рассматривать:

долю ошибочно прочтенных нуклеотидов при чтении нуклеотидной последовательности ДНК,

долю правильных прочтений всей последовательности стандартного образца, а также долю прочтений, содержащих различное число ошибок (1, 2 , 3 и более),

долю правильных прочтений участков стандартного образца с числом i одинаковых нуклеотидов подряд.

Например, доля A_3 правильного прочтения участка, состоящего из трех одинаковых нуклеотидов, определяется соотношением

$$A_3 = 1 - \frac{(K_{Z3} + N_{Z3})}{(X + Y) \cdot Z_3}, \qquad (1)$$

где

X – число прямых последовательностей нуклеотидов при чтении (имеется в виду число многократных прочтений прямой последовательности нуклеотидов в фрагменте плазмиды);

Y - число комплементарных последовательностей нуклеотидов при чтении (то же самое, но по отношению к комплементарной последовательности);

Z_3 – число участков в последовательности, состоящих из трех одинаковых нуклеотидов подряд;

K_{Z3} - общее число ошибочных прочтений по всем участкам, содержащим три одинаковых нуклеотида подряд при чтении прямых последовательностей;

N_{Z3} - общее число ошибочных прочтений по всем участкам, содержащим три одинаковых нуклеотида подряд при чтении комплементарных последовательностей.

Аналогично можно определить доли правильного прочтения секвенатором последовательности нуклеотидов в цепочке ДНК фрагмента плазмиды, состоящей из четырех и пяти одинаковых нуклеотидов. Правильное прочтение более длинных последовательностей нуклеотидов секвенатором становится проблематичным из-за, как уже отмечалось, нарушения линейной зависимости выходного сигнала от числа нуклеотидов.

«Интегральным» показателем точности результатов прочтения секвенатором произвольной последовательности нуклеотидов можно считать долю ошибочно прочтенных нуклеотидов, определяемую формулой

$$S = \frac{S_{fX} + S_{fY}}{S_X + S_Y}, \qquad (2)$$

где S_X (S_Y) - общее число нуклеотидов в прочтенных прямых (обратных) последовательностях, S_{fX} (S_{fY}) - число неправильно прочтенных нуклеотидов в прямых (обратных) последовательностях.

На основании изложенной выше методологии нами была разработана методика испытаний с целью утверждения типа анализатора геномного GS FLX, изготовленного фирмой Roche Diagnostics GmbH (Германия), и проведены необходимые испытания. В результате испытаний были, в частности, установлены следующие характеристики секвенатора:
доля прочтений стандартного образца, не содержащих ошибок: >75%,
доля прочтений стандартного образца, содержащих 1 ошибку: <15%,
доля прочтений стандартного образца, содержащих 2 ошибки: <10%,
доля прочтений стандартного образца, содержащих 3 и более ошибок: <5%,
доля ошибочно прочтенных нуклеотидов при чтении стандартного образца: не более 2%.

Было также установлено, что
средняя доля прочтений участка стандартного образца из трех одинаковых нуклеотидов подряд, в которых он прочтен без ошибок: > 93 %,
средняя доля прочтений участка из четырех одинаковых нуклеотидов подряд, в которых он прочтен без ошибок: > 85%,
средняя доля прочтений участка из пяти одинаковых нуклеотидов подряд, в которых он прочтен без ошибок: > 75%.

Для нахождения длины прочтения индивидуальной реакции секвенатором использовался препарат геномной ДНК бактерии Escherichia

coli. При этом более 50% чтений индивидуальных реакций имело длину более 350 нуклеотидов, а общая длина прочитанных нуклеотидных последовательностей за один рабочий цикл составила 350-500 млн. нуклеотидов.

По материалам испытаний анализатор геномный GS FLX был занесен в Государственный реестр средств измерений (№ 47873-11).

Доля ошибочно прочитанных нуклеотидов при чтении индивидуальной реакции секвенирования стандартного образца в значении не более 2% на практике обеспечивает высокую достоверность прочтений геномных последовательностей. Дело в том, что, как уже отмечалось выше, для прочтения геномной последовательности ее разбивают на перекрывающиеся короткие последовательности, набор которых «избыточен», т.е. каждая точка генома читается в 10-40 независимых реакциях, по которым строится «консенсусная» последовательность. Такое последовательное прочтение, как уже отмечалось, приводит к существенному уменьшению вероятности ошибочного определения последовательности. На практике распределение ошибок является неравномерным (например, их частота выше в районах, содержащих гомополимерные участки), однако, многократная избыточность все же позволяет читать геномные последовательности с высокой степенью достоверности.

Разработка системы метрологического обеспечения секвенирования молекул ДНК в нашей стране находится в начальной стадии развития. Изложенный материал описывает только первые шаги, предпринятые в этом направлении. Необходимо дальнейшее развитие метрологического обеспечения в этой новой и бурно развивающейся научно-технической деятельности, попадающей в сферу государственного регулирования в области обеспечения единства измерений.

В первую очередь требуется разработка и внедрение в измерительную практику новых стандартных образцов, содержащих пять и более

повторяющихся нуклеотидов в последовательности для повышения надежности результатов поверки (калибровки) и расширения функциональных возможностей секвенаторов.

Необходимо также разработка и обоснование принципиальных вопросов метрологии этих приборов и измерений, начиная с уточнения понятия погрешности (неопределенности) измерений, определения их функций распределения, процессов обработки результатов измерений и согласования всех этих понятий и процедур с понятиями и процедурами, принятыми в обычной метрологической практике. С учетом трудностей и неоднозначности процедуры прочтения нуклеотидных последовательностей в молекуле ДНК необходима разработка методов оценки используемых в секвенаторах программных продуктов. Наконец, требуется разработка необходимого для метрологического обеспечения секвенирования набора нормативных документов в виде методик испытаний и поверки (калибровки), методик измерений и контроля метрологических характеристик секвенаторов.

Работа выполнена при финансовой поддержке Министерства образования и науки РФ (государственный контракт 16.648.11.3005 в рамках Федеральной целевой программы «Развитие инфраструктуры наноиндустрии в Российской Федерации на 2008-2011 годы).

Авторы.

Голубев Сергей Сергеевич – начальник лаборатории ФГУП «ВНИИМС», к.т.н.

119361, Москва, Озерная, 46

Кононогов Сергей Алексеевич – директор ФГУП «ВНИИМС», д.т.н.

119361, Москва, Озерная, 46

Кудеяров Юрий Алексеевич – главный научный сотрудник ФГУП «ВНИИМС», д.ф.-м.н., профессор

119361, Москва, Озерная, 46

Марданов Андрей Владимирович – старший научный сотрудник Учреждения Российской академии наук Центр «Биоинженерия» РАН, к.б.н.

117312, Москва, пр-т 60-летия Октября д.7, корп.1

Николаева Полина Юрьевна – инженер ФГУП «ВНИИМС»

119361, Москва, Озерная, 46

Равин Николай Викторович – заместитель директора Учреждения Российской академии наук Центр «Биоинженерия» РАН, д.б.н.

117312, Москва, пр-т 60-летия Октября д.7, корп.1

Скрябин Константин Георгиевич - директор Учреждения Российской академии наук Центр «Биоинженерия» РАН, д.б.н., профессор, академик РАН и РАСХН

117312, Москва, пр-т 60-летия Октября д.7, корп.1

Список литературы

1. M. Margulies, M. Edholm, W.E. Altman *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 2005, 376 – 380.
2. С.С. Голубев, С.А. Кононогов, Н.В. Равин, К.Г. Скрябин. Генетическая последовательность как природный эталон новой биологической величины. Измерительная техника, № 12, 2011.
3. Федеральный Закон РФ № 102-ФЗ «Об обеспечении единства измерений», глава 1, ст. 2.
4. Международный словарь по метрологии. Основные и общие понятия и соответствующие термины – С.- Петербург, НПО «Профессионал», 2010.
5. В.А. Кузнецов, Л.К. Исаев, И.А. Шайко. Метрология. – М.: Стандартинформ, 2005 год. 298 стр.

АННОТАЦИЯ

к статье С.С. Голубева, С.А. Кононогова, Ю.А. Кудеярова, А.В. Марданов, П.Ю. Николаевой, Н.В. Равина, К.Г. Скрябина «Метрологическое обеспечение секвенирования молекул ДНК».

Изложен принцип действия секвенатора – прибора для последовательного прочтения нуклеотидной последовательности молекул ДНК при пирофосфатном секвенировании. Сформулированы основы метрологического обеспечения секвенирования молекул ДНК и приведены результаты исследований метрологических характеристик анализатора геномного GS FLX (секвенатора). Перечислены основные задачи создания системы метрологического обеспечения секвенирования молекул ДНК: разработка и внедрение в измерительную практику новых стандартных образцов, дальнейшая разработка и обоснование принципиальных вопросов метрологии секвенирования, разработка необходимого для метрологического обеспечения секвенирования набора нормативных документов и т.п.

Ключевые слова:

Молекула ДНК, последовательность нуклеотидов, комплементарность, пирофосфатное секвенирование, метрологическое обеспечение, стандартный образец, плазмида pUC18.

Annotation

of the article of S.S. Golubev, C.A. Kononogov, Yu.A. Kudeyarov, A.V. Mardanov, P.Yu. Nikolaeva, N.V. Ravin, K.G. Skryabin «The metrological support of DNA sequencing»

The principle of DNA pyrophosphate sequencing is set out. The basics of metrological support of sequencing DNA are formulated and the results of investigations of the metrological characteristics of the analyzer genomic GS FLX (sequencer) are presented. The main tasks of creating a system of metrological support of DNA sequencing are recounted: development and implementation in the measurements of new standard samples, development and justification of fundamental metrology issues of sequencing, development a set of regulations, etc.

Key words:

Molecule of DNA, nucleotide sequence, complementarity, pyrophosphate sequencing, metrological support, standard sample, plasmid pUC18