Тимкин П.Д., мнс
Пензин А.А., мнс

Всероссийский научно-исследовательский институт сои, Россия, Благовещенск
paa@vniisoi.ru

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ОБРАТНОЙ ТРАНСЛЯЦИИ ДЛЯ БЕЛКОВОЙ ИНЖЕНЕРИИ *IN SILICO*

Реферат. Белковая инженерия – раздел биотехнологии, который занимается разработкой полезных или ценных белков. Технологии, предоставляемые данной отраслью, с каждым годом, все активнее внедряется в работу селекционеров и генетиков. Лабораториям биотехнологического профиля, которые занимаются выделением генов, а также их редактированием, с целью выяснения новых признаков либо улучшением уже изученных, технологии белковой инженерии позволят сократить процесс поиска, ввиду того что уже известному белку задаются необходимые свойства. То есть подход заключается не во внесение изначальных модификаций в геном, а изменению в первую очередь белка, после чего, синтезу новой последовательности нуклеотидов, которая в дальнейшем будет встроена в геном взамен старой. Сложность такого подхода заключается в избыточности генетического кода. Свойство избыточности приводит к тому, что от трех до четырех разных кодонов, могут кодировать одну и ту же аминокислоту. Это свойство дает возможность при получении генетических повреждений и изменений нуклеотидов в геноме, сохранять исходный смысл триплета, то есть при изменении нуклеотидов, не изменяется аминокислота. В случае, когда мы будем пытаться закодировать полипептидную последовательность в нуклеиновую кислоту, возникнет выбор между разными триплетами. В

биоинформатике такой процесс называется обратной трансляцией и для него написаны специальные алгоритмы для перевода протеина обратно в ДНК. Эти методы работают по принципу подбора наиболее вероятного нуклеотида, наиболее часто встречающегося на той или иной позиции. Исходя из этих данных, каждому нуклеотиду в кодоне присваивается свой номер и вероятность наличия каждого нуклеотида. В данной работе описывается опыт применения подобного алгоритма в случае перевода полипептида WIN с внесенным однонуклеотидным полиморфизмом в 86 положении. Измененный ген *WIN*, полученный в результате обратной трансляции, был сравнен с референсным геном из базы данных национального центра биотехнологической информации(NCBI). В результате была получена нуклеотидная последовательность, предсказанная с точностью около 75% по идентичности нуклеотидного состава. Так же, несмотря на отличия при обратной трансляции, генетический смысл каждого триплета в новой последовательности нуклеотидов с точностью до 100% был сохранен. При сохранности смысла новой цепи, можно сделать вывод об успешном проведении конвертации аминокислотной последовательности белка в ДНК. Это послужит основой для синтеза измененной нуклеотидной последовательности для внедрения в геном.

Ключевые слова: WIN, *Glycine max,* обратная трансляция, биоинформатика, белковая инженерия, биотехнология, in silico

Abstract. Protein engineering is a branch of biotechnology that develops useful or valuable proteins. The technologies provided by this industry are being increasingly introduced into the work of breeders and geneticists every year. Biotechnological laboratories that are engaged in the isolation of genes, as well as their editing, in order to find out new features or improve those already studied, protein engineering technologies will reduce the search process, due to the fact that the necessary properties are set for an already known protein. That

is, the approach is not to make initial modifications to the genome, but to change, first of all, the protein, after which, the synthesis of a new sequence of nucleotides, which will later be embedded in the genome instead of the old one. The complexity of this approach lies in the redundancy of the genetic code. The redundancy property leads to the fact that from three to four different codons can encode the same amino acid. This property makes it possible, when receiving genetic damage and changes in nucleotides in the genome, to preserve the original meaning of the triplet, that is, when changing nucleotides, the amino acid does not change. In the case when we try to encode a polypeptide sequence into a nucleic acid, there will be a choice between different triplets. In bioinformatics, this process is called reverse translation and special algorithms have been written for it to translate the protein back into DNA. These methods work on the principle of selecting the most likely nucleotide most commonly found in a particular position. Based on these data, each nucleotide in the codon is assigned its own number and the probability of the presence of each nucleotide. This paper describes the experience of using such an algorithm in the case of translation of a WIN polypeptide with an introduced single-nucleotide polymorphism in position 86. The modified WIN gene obtained as a result of reverse translation was compared with a reference gene from the database of the National Center for Biotechnology Information (NCBI). As a result, a nucleotide sequence was obtained, predicted with an accuracy of about 75% by the identity of the nucleotide composition. Also, despite the differences in reverse translation, the genetic meaning of each triplet in a new sequence of nucleotides with an accuracy of 100% was preserved. With the preservation of the meaning of the new chain, it can be concluded that the conversion of the amino acid sequence of the protein into DNA has been successfully carried out. This will serve as the basis for the synthesis of an altered nucleotide sequence for insertion into the genome.

Keywords: *WIN, Glycine max,* reverse translation, bioinformatics, protein engineering, biotechnology, *in silico*

Введение

Обратная трансляция - это процесс декодирования последовательности аминокислот в соответствующие кодоны. Все системы синтетического генного дизайна включают модуль обратной трансляции. Избыточность генетического кода делает обратную трансляцию потенциально неоднозначной, поскольку большинство аминокислот кодируются разными кодонами. Общий подход к преодолению этой трудности основан на имитации использования кодонов в пределах вида [1].

Оптимальное решение данной задачи будет актуально для лабораторий, занимающихся белковой и генной инженерией. В лабораториях подобного типа имеется потребность во встраивании генетической информации в тест-систему для экспрессии или полноценный организм.

К процессу кодирования полипептида можно подойти разными биоинформатическими способами. Существует различное программное обеспечение, работающее по разным алгоритмам. К числу подходов в создание алгоритмов можно выделить имитацию кодонов, которые представляют собой наиболее вероятный невырожденный участок, для генерации ДНК. Альтернативные подходы используют алгоритмы скрытых марковских моделей (СММ). СММ называют статистическую модель, имитирующую работу процесса похожего на марковский с неизвестными параметрами, и задачей ставится разгадывание данных параметров на основе наблюдаемых. Полученные параметры могут быть использованы в дальнейшем анализе, например, для распознавания образов. СММ может быть рассмотрена как простейшая байесовская сеть

доверия. Данная статистическая модель нашла широкое использование в решение разного рода биоинформатических задач.

Программным обеспечением, в основе которого лежит использование СММ можно назвать EasyBack. Принцип его работы заключается не в имитации использования кодонов в пределах целевого вида, а в использовании критериев сходства последовательностей. Модель обучается с использованием набора белков с известными кодирующими последовательностями кДНК, сконструированных из входного белка путем запроса баз данных NCBI с помощью BLAST. В отличие от существующего программного обеспечения, предлагаемый метод позволяет оценить качество прогнозирования. К недостаткам подобных программ можно отнести их формальную устарелость, трудности в нахождение подобного обеспечения и их использование. В данный момент популярность набирают программные обеспечения, позволяющие пользоваться вычислениями не локально, а удаленно. К тому же для пользователей, работающих в области вычислительной биологии, снижается порог вхождения ввиду уже готовых алгоритмов.

В данной статье будет описан опыт использования метода обратной трансляции на примере белка *Glycine max*(сои) WIN, с внесенным однонуклеотидным полиморфизмом и перевода его в синтезированную ДНК матрицу. Проведение обратной трансляции проходили на веб-ресурсе «Sequence Manipulation Suite» [2].

Были выставлены требование к проведенным вычислениям, результат трансляции синтезированной нуклеотидной цепи в белок, должна быть 100% и соответствовать точности сохранения сходства нуклеотидов в 70% и выше.

Материалы и методы исследований

Полипептидная последовательность белка WIN была взята с базы данных Uniprot. По методу гомологичного сходства был предсказан

однонуклеотидный полиморфизм. Который после был индуцирован в полипептидную цепь, через редактирование исходного документа fasta-формата и замены одной аминокислоты на другую. В дальнейшем в качестве методов контроля был проведен алгоритм обратной трансляции нативной полипептидной цепи. Обратная трансляция осуществлялась в виртуальной лаборатории для работы с первичными последовательностями SMS. Алгоритм, лежащий в основе обратного транслятора, описывается как имитационный подбор часто встречающихся кодонов у искомого таксона. Было использовано два типа подобранных кодонов. Первый тип являлся стандартным и предлагался самой программой, его таксономический тип соответствовал *E.Coli*. Для второго типа были подобраны соответствующие триплеты для *Glycine max*. Подбор имитационных кодонов проходил на Codon Usage Database [3]. Полученный результат в виде последовательности нуклеотидов был сверен с CDS (белок-кодирующая область) у референсной ДНК. Информация о первичной структуре референсной ДНК и аннотация к ней была взята с базы данных NCBI(Национальный центр биотехнологической информации) [4]. Сходство новосинтезированной последовательности ДНК и референсной производилось через стандартную методику локального выравнивания на сервисе Emboss [5]. Оценка сохранности смысла новых кодонов проводилось в той же виртуальной лаборатории, где и производилась обратная трансляция. Оценка результатов выравнивания проводилось в самом программном обеспечении с использованием алгоритма Нидлмана-Вунша.

Результаты

Информация о первичной структуре протеина сои WIN была взята из базы данных Uniprot, после чего над ним была произведена редакция, путем внедрения однонуклеотидного полиморфизма (рис. 1).

KPYSWRSKYGWTAFCGPVGPRGRDSCGKCLRVTNTGTGANTIVRIVDQCSNGGLDLDVGVFNRIDTD
GRGYQQGHLIVNYQFVDCGNELDLTKPLLSILDAP

KPYSWRSKYGWTAFCGPVGPRGRDSCGKCLRVTNTGTGANTIVRIVDQCSNGGLDLDVGVFNRIDTD
GRGYQQGHLIVNYQFVDCDNELDLTKPLLSILDAP

Рис.1 Сравнение аминокислотных цепей.
Сверху нативная полипептидная цепь, снизу с индуцированным полиморфизмом, красным обозначены аминокислоты с порядковым номером 86, где произошла замена с G на D

Обратной трансляции были подвергнуты: нативная цепь, в качестве метода контроля и цепь с индуцированным полиморфизмом. Была определена точность прогноза, путем сравнения сходства (Табл.1). Точность трансляции у всех кандидатов была равна 100%.

Таблица 1.
Точность прогноза

Тип сгенерированных кодонов	Точность обратной трансляции для нативной цепи	Точность обратной трансляции для цепей с однонуклеотидным полиморфизмом
E.Coli	76.4%	76.7%
Glycine max	75.1%	75.8%

Обсуждение

В результате проведенного алгоритма удалось получить нуклеотидные цепи, которые в обоих случаях сгенерированных кодонов попадали под точность свыше 70%. Интересным наблюдением является то, что имитация кодонов, таксономически близких к *E.Coli* дало более высокое сходство к референсной цепи, чем имитационные триплеты для *Glycine max.* Такой результат может объясняться не достаточной проработкой самой базы данных и малого количества аннотированной информации о генах для сои и наоборот более полную и комплексную

картину для кишечной палочки, которая является одним из самых популярных объектов исследования в области генетики и молекулярной биологии. Так же наблюдается динамика возрастания точности прогноза у обеих нуклеиновых кислот в случае индукции однонуклеотидных полиморфизмов. Прирост точности хоть и незначителен в этом случае, но статистически воспроизводим, что может объясняться искусственным редактированием самих триплетов в референсных генах, что снижает погрешность парного выравнивания. Исходя из анализа вышеперечисленных данных можно прийти к заключению, что в попытках совершить обратную трансляцию для белков недостаточно аннотированых организмов, допускается использование стандартных имитационных триплетов сгенерированных самой базой данных Codon Usage Database для сервиса SMS. Такой прогноз дает высокую точность предсказанных нуклеотидов, и погрешность не должна будет оказать влияние на уровень матричной экспрессии внутри самого генома. К тому же подобный сервер имеет низкий порог вхождение и постоянное техническое обслуживание, что делает его доступным для многих исследователей. Однако в работе не оценивалось сравнение программного обеспечения, работающих по принципу скрытых марковских моделей, чьи прогнозы ввиду более сложных методов обработки массива данных дают прогноз точнее. В дальнейшем планируется оптимизации для исследователей подобных алгоритмов, которые предоставят более простой способ работы с данными.

Список литературы

1. Richardson S, Wheelan S, Yarrington R, Boeke J: GeneDesign: rapid, automated design of multikilobase synthetic genes. Genome Res 2006, 16: 550–556. 10.1101/gr.4431306
2. https://www.bioinformatics.org/sms2/rev_trans.html
3. https://www.kazusa.or.jp/codon/

4. https://www.ncbi.nlm.nih.gov/gene/548085 - genomic database
5. https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=emboss_water-I20221109-063032-0834-27421528-p2m –web-tools for alignment

Penzin A.A.,
Timkin P.D.
All-Russian Research Institute of SOY, Russia, Blagoveshchensk

USING REVERSE TRANSLATION METHODS FOR PROTEIN ENGINEERING IN SILICO