Успехи биологической химии, т. 63, 2023, с. 349–378

ВЛИЯНИЕ НУКЛЕОТИДНОГО СОСТАВА ГЕНА И ЕГО РЕГУЛЯТОРНЫХ ЭЛЕМЕНТОВ НА ЭФФЕКТИВНОСТЬ ЭКСПРЕССИИ БЕЛКА В *ESCHERICHIA COLI*

А. И. ЗАБОЛОТСКИЙ,
С. В. КОЗЛОВСКИЙ, А. Г. КАТРУХА

Московский государственный университет имени М.В.Ломоносова, Москва

I. ВВЕДЕНИЕ

Рекомбинантные белки широко используются в биомедицинской, пищевой и других биотехнологических сферах. Среди множества организмов, используемых для экспрессии негликозилированных рекомбинантных белков, *Escherichia coli* остается наиболее популярным и активно применяемым [1, 2]. Этому способствует множество факторов: достаточное глубокое понимание метаболических процессов *E. coli*, высокая скорость роста, сравнительная дешевизна и доступность сред для культивирования, широкие возможности для масштабирования, а также доступность огромного набора штаммов, экспрессионных векторов и инструментов генной инженерии [1, 2].

Увеличение уровня экспрессии является одной из ключевых задач как при масштабном производстве, так и исследовании белков, в

Принятые сокращения: РНКП – ДНК-зависимая РНК-полимераза; GFP – зеленый флуоресцирующий белок (*от англ.* Green Fluorescent Protein); НТО – нетранслируемая область; РИТ – регион инициации трансляции; КП – кодирующая последовательность; ШД – Шайна — Дальгарно (последовательность); БЦДГ – бисцистронный дизайн гена; RF – фактор высвобождения (*от англ.* Release Factor); CAI – индекс адаптации кодонов (*от англ.* Codon Adaptation Index); nTE – нормализованная трансляционная эффективность (*от англ*. Normalized Translational Efficiency); tAI – индекс адаптации тРНК (*от англ.* tRNA Adaptation Index).

Адрес для корреспонденции: zabolotscky.artur@yandex.ru

связи с чем большие усилия направлены на разработку методов оптимизации данного процесса. Такие подходы как выбор оптимальных штаммов и экспрессионных векторов, подбор условий экспрессии (выбор питательной среды, условий роста и индукции культуры) уже широко используются на практике для увеличения уровня экспрессии белков в *E. coli* [3]. Однако зачастую проблема низкого выхода белка связана с нуклеотидной последовательностью соответствующего гена и его регуляторных участков.

Как кодирующие, так и некодирующие последовательности гена содержат элементы, которые влияют на правильность фолдинга и величину выхода белка на всех стадиях экспрессии. Влияние многих составляющих элементов контроля экспрессии взаимосвязано, что затрудняет полное понимание механизма их действия, как по отдельности, так и в комбинации.

Тем не менее, за многие годы исследований был достигнут заметный прогресс в понимании влияния нуклеотидной последовательности на продукцию белка. Данный обзор посвящён описанию современных представлений о влиянии нуклеотидного состава гена и его регуляторных последовательностей на уровень продукции белка на разных стадиях экспрессии в *E. coli.*

II. ВЛИЯНИЕ НУКЛЕОТИДНОГО СОСТАВА ГЕНА НА ЭФФЕКТИВНОСТЬ ЭКСПРЕССИИ

СТАДИЯ ТРАНСКРИПЦИИ

Промотор и прилежащий к нему регион

Первым этапом в синтезе белка является транскрипция гена, осуществляемая ДНК-зависимой РНК-полимеразой (РНКП). Скорость инициации и эффективность синтеза мРНК опосредуется последовательностью промотора. Промотор представляет собой область ДНК перед геном, где соответствующие белки (такие как РНК-полимераза и факторы транскрипции) связываются, чтобы инициировать транскрипцию. При рекомбинантной экспрессии используют два основных типа промоторов. Индуцибельные промоторы представляют собой регулируемые промоторы, которые становятся активными в клетке только в ответ на специфический стимул. Конститутивные промоторы представляют собой нерегулируемые промоторы, которые активны в клетке при любых обстоятельствах. Сила промотора наиболее часто экспериментально определяется через относительный уровень мРНК или белка-репортера (наиболее часто зеленого флуоресцентного белка – GFP), производимых при экспрессии с данным промотором

[4, 5]. Существуют большие базы с собранными последовательностями промоторов и их экспериментально определенной силой, такие как: BIOFAB (International Open Facility Advancing Biotechnology) (http://parts.igem.org/Collections/BioFAB), Anderson promoter library (http://parts.igem.org/Promoters/Catalog/Anderson) и другие.

Для получения промоторов с наибольшей эффективностью на практике чаще всего используют метод, заключающийся в создании библиотек с рандомизованными последовательностями промотора с дальнейшим анализом уровня экспрессии репортерных генов с полученными вариантами промоторов. Например, рандомизация последовательностей –10/–35 конститутивного промотора Ptrc [6] или последовательностей –17/+3 индуцибельного промотора T7 [7, 8] Хотя большинство ключевых принципов инициации транскрипции известны, модели для прогнозирования силы промотора по нуклеотидной последовательности все еще находятся в стадии разработки. Несмотря на это экспериментальные данные показывают, что использование сильного промотора с разными последовательностями почти всегда приводит к воспроизводимому увеличению уровня транскрипции и следовательно экспрессии белка [6].

Скорость РНКП

Еще одним показателем, влияющим на скорость транскрипции, является «содержание водородных связей» или процентное содержание GC среди нуклеотидов (GC%) от области инициации транскрипции до ~15 кодона кодирующей последовательности. Данный параметр влияет на энергию, затрачиваемую РНК-полимеразой на плавление ДНК и, следовательно, на скорость первых этапов транскрипции. Показано, что кодирующие последовательности с более высоким уровнем транскрипции в *E. coli* имеют более низкое значение GC% в начале гена, чего не было обнаружено в эукариотах[9].

СТАБИЛЬНОСТЬ И ТОКСИЧНОСТЬ МРНК

Токсичность мРНК

Нередко экспрессия гетерологичных белков вызывает снижение темпа роста клетки, что обычно связано с токсичностью экспрессируемого белка или метаболической нагрузкой [10]. Однако у *E. coli* был обнаружен опосредованный мРНК эффект, в рамках которого некоторые специфические гетерологичные последовательности мРНК оказываются токсичными для бактериальных клеток. Так, при экспрессии синонимичных вариантов гена GFP в *E. coli* были обнаружены последовательности, транскрипция которых ингибировала

рост клеток вне зависимости от прохождения трансляции гена [11]. Механизм данного эффекта только изучается и может быть связан с токсическим эффектом, вызываемым специфическими вторичными структурами мРНК.

Деградация мРНК

Сопряжение транскрипции и трансляции в бактериях обеспечивает покрытие мРНК рибосомами и различными белковыми факторами со стадии синтеза мРНК, что обеспечивает некоторую степень защиты от действия РНКаз. Высокая плотность рибосом на активно транслируемых мРНК может эффективно препятствовать эндорибонуклеазной активности [12]. Несмотря на это, сравнение стабильности бактериальных и эукариотических мРНК показывает, что бактериальные мРНК существуют сравнительно недолго. Время полужизни большинства бактериальных мРНК колеблется от 40 секунд до 60 минут, тогда как для некоторых эукариотических мРНК этот параметр может достигать нескольких дней [13]. Это зачастую связывают с отсутствием конкретных механизмов и белков, обеспечивающих защиту мРНК от деградации у прокариот, например таких как поли-А связывающий белок у эукариот [14]. Особенно уязвимыми участками полностью незащищёнными рибосомами являются 5'- и 3'-нетранслируемые области мРНК (3'-/5'-НТО мРНК), которые больше всего подвержены эндонуклеазной активности и в *E. coli* определяют стабильность мРНК..

Так, в случае мРНК ompA (фрагмент мРНК белка внешней мембраны, outer membrane protein – OmpA) ее 5'-НТО, образующая вторичную структуру, служит элементом стабильности, увеличивающим период полужизни данной мРНК в 4 раза [15, 16]. Предполагается, что присутствие высокоструктурированных 5'-областей ингибирует связывание эндорибонуклеаз (например, РНКазы Е) с 5'-концевыми участками мРНК, не защищенными белками [17]. Однако выделить влияние стабильности 5'-концевой области мРНК на эффективность последующей экспрессии достаточно сложно, поскольку нуклеотидный состав в данной области также играет ключевую роль в чрезвычайно важном процессе инициации трансляции [18–20], для которого вторичные структуры являются крайне нежелательными.

Изменения в нуклеотидной последовательности 3'-НТО также могут приводить к увеличению периода полужизни мРНК и к увеличению уровня экспрессии белка. Так, показано, что замена и/или

укорачивание 3'-конца мРНК увеличивает стабильность последней и количество производимого белка [21]. В основе этого эффекта могут лежать (1) уменьшение образования на 3'-конце мРНК внутренних вторичных структур, способных подвергаться воздействию РНКаз, специфичных к двуцепочечным РНК [22], (2) уменьшение взаимодействия между 3'-и 5'-концами, приводящего к деградации мРНК под воздействием РНКаз [23], (3) наличие особых мотивов (AU-богатые мотивы [24], сайты РНКаз, сайты связывания sRNA и др. [22]).

СТАДИЯ ТРАНСЛЯЦИИ

Инициация трансляции

Эффективность связывания рибосом. В *E. coli* инициация трансляции в наибольшей степени определяет скорость прохождения трансляции и выход конечного белка, поэтому эта стадия является наиболее важной мишенью при оптимизации экспрессии белка [25, 26]. Скорость инициации трансляции зависит от множества факторов, одним из которых является аффинность гибридизации 16S рРНК, в составе 30S субъединицы рибосомы, с последовательностью Шайна-Дальгарно (ШД), находящейся в регионе инициации трансляции (РИТ) мРНК. Было показано, что энергия связывания между последовательностями 16S рРНК и ШД хорошо коррелирует с наблюдаемым уровнем экспрессии белка [6, 27, 28].

Несмотря на хорошую корреляцию все попытки разработки универсальных ШД, работающих с любыми кодирующими последовательностями (КП), не давали желаемого результата. Стандартная последовательность ШД, которая хорошо инициирует трансляцию с одной последовательности, может практически не работать с другой, требуя разработки оптимальной ШД для каждой КП [6, 27, 29]. Данный факт скорее всего связан с тем, что нуклеотидный состав в области ШД также влияет на формирование вторичных структур мРНК в области РИТ, которые, в свою очередь, влияют на уровень экспрессии. Таким образом, замена последовательности ШД может привести к образованию вторичных структур на 5' конце мРНК, уменьшающих эффективность инициации и уровень экспрессии. Поэтому оптимизация последовательности ШД проводится индивидуально для отдельного гена.

Примером оптимизации ШД без нарушения вторичной структуры 5'-конца мРНК, считается использование бицистронного дизайна гена (БЦДГ) [6]. В основе принципа БЦДГ лежит изоляция и сопряжение инициации трансляции двух слитых КП. При этом трансляция нижележащей КП, как полагают, является результатом повторной

инициации трансляции рибосомами [30, 31]. Оптимальная структура мРНК в области РИТ обеспечивается ШД-последовательностью и короткой кодирующей последовательностью первой рамки, изолируя влияние структур на нижележащую ШД, позволяя подобрать её оптимальный вариант. Данный изолирующий эффект в теории упрощает разработку стандартизированных ШД, эффективно инициирующих трансляцию с любыми КП [6, 27, 29].

Вторичная структура мРНК в регионе инициации трансляции. Многочисленные исследования показывают, что формирование прочных вторичных структур в 5'-концевой области мРНК (регион инициации трансляции – 5'-НТО мРНК + 5'-концевой участок КП мРНК) оказывает наиболее существенное влияние на уровень экспрессии белка в *E. coli* [18–20, 32–34]. Данный эффект часто объясняют влиянием вторичных структур РНК на эффективность связывания рибосом и инициацию трансляции (Рис. 1) [19], которая считается ключевой стадией, определяющей скорость процесса в целом [35]. Показано, что формирование прочных шпилек в 5'-концевых областях мРНК снижает эффективность трансляции в сотни раз [36].

Данная теория подтверждается демонстрацией существования вторичных структур мРНК *in vivo* при помощи веществ, избирательно реагирующих с неспаренными основаниями РНК, таких как зонды SHAPE [37] или DMS [38]. После обработки клетки данными веществами производят секвенирование мРНК, что позволяет картировать модификации и выявить неструктурированные и структурированные области мРНК. Одно из таких исследований в *E. coli* продемонстрировало, что эффективность трансляции генов в значительной степени определяется отсутствием вторичных структур мРНК в области РИТ, ограничивающих доступность ШД для рибосом [39, 40].

Элонгация трансляции

«Быстрые» и «медленные» кодоны. На данный момент не известно ни одного организма, клетки которого содержали бы полный набор тРНК с антикодонами, комплементарными каждому из 61-го кодирующего триплета. Так, *E. coli* имеет 39 тРНК с различными антикодонами (Рис. 2) [41]. Трансляция синонимичных кодонов с помощью одной тРНК происходит посредством кодон-антикодонных взаимодействий, при которых первые два основания кодона спариваются с антикодоном путем стандартных Уотсон-Криковских взаимодействий (A/U, G/C), а в третьем положении кодона допускается взаимодействие G/U, A/I, и т.д. благодаря «вобблингу» (wobbling, колебание). Однако сродство, по которому синонимичные

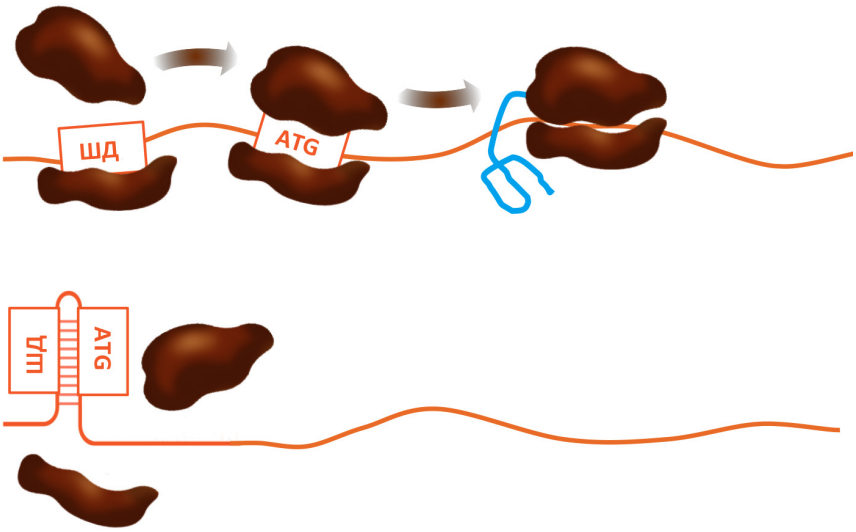

Рис. 1. Схематическое представление механизма влияния вторичных структур в области РИТ на эффективность инициации трансляции.

Рибосома легко связывается с ШД последовательностью на мРНК без вторичной структуры в 5'-концевой области (сверху), затруднения связывания рибосомы с ШД последовательностью заключённой во вторичные структуры мРНК (снизу).

кодоны распознаются одной тРНК, не одинаково. Например, тРНК типа 5′–GNN–3′ имеют более высокую аффинность связывания с кодонами 5′–NNC–3′, чем с кодонами 5′–NNU–3′ [42, 43]. Еще одним важным фактором при прочтении синонимичных кодонов в теории «вобблинга» являются модификации тРНК. У *E. coli* очень мало кодонов, считываемых тРНК, не претерпевшей модификаций [44]. Так, например модификация «колеблющегося» основания U – cmo^5U, была обнаружена у тРНК в семействах кодонов Ala, Leu, Pro, Ser, Thr и Val [45]. Данная модификация позволяет соответствующей тРНК распознавать кодоны мРНК оканчивающиеся на U, A, G и C. Изначально считалось, что данное модифицированное основание имеет одинаковый уровень сродства к любому партнеру, однако в дальнейшем исследования показали, что уровень сродства с кодонами оканчивающимися на A и U выше, чем к G/C оканчивающимся кодонам [46], что также влияет на скорость прочтения данного кодона.

Таким образом кинетика трансляции синонимичных кодонов зависит от множества факторов, таких как: (1) доступность необхо-

димой аминоацилированной тРНК, зависящей от общего содержания копий генов данной тРНК и уровня их экспрессии [47, 48] (2) наличие и уровень модификаций тРНК [44, 49] (3) конкуренция различных тРНК за кодон, выражающаяся в их относительном сродстве и силе связывания [50].

В целом «медленным» можно назвать кодон, который приводит к замедлению элонгации трансляции при его прочтении рибосомой, вне зависимости от причины данного явления. За многие годы исследований неоднократно предпринимались попытки количественной оценки вклада различных кодонов в скорость и уровень экспрессии. Самым простым таким показателем является индекс CAI (codon adaptation index, индекс адаптации кодонов), который является оценкой представленности различных кодонов в исследуемом гене по отношению к кодонному составу в наборе стандартных генов с высоким уровнем экспрессии [51]. Данный индекс достаточно часто используется в алгоритмах оптимизации кодонного состава при гетерологичной экспрессии белков в *E. coli* [52], однако он является крайне несовершенным в силу ряда причин: (1) в мРНК выделяют отдельные регионы, имеющие свои закономерности кодонного предпочтения, зачастую не связанные с кодонным предпочтением внутри целого организма. Например, 5'-конец КП генов прокариот содержит участок из ~15 кодонов, которые зачастую являются «медленными», что важно для правильного прохождения первоначальных этапов элонгации трансляции; (2) отсутствуют критерии выбора стандартных высокоэкспрессируемых генов из-за наличия большого количество факторов влияющих на данный показатель; (3) возможны различия между экспрессионными аппаратами и особенностями кодонного предпочтения организма для рекомбинантной экспрессии и организма донора гена.

Для более точной оценки скорости прочтения кодона был предложен критерий tAI (tRNA adaptation index, индекс адаптации тРНК) [53]. Этот показатель учитывает количество копий генов тРНК для данного кодона в клетке хозяина (которое, как предполагается, коррелирует с содержанием тРНК в клетках), а также учитывает эффективность связывания кодонов с антикодонами, связанную с правилами колебания Крика. Данный показатель является более точным, однако не учитывает изменение концентраций тРНК, в том числе аминоацил- и модифицированных тРНК, в зависимости от разных условий. Например, у бактерий было обнаружено, что уровни аминоацилирования и модификаций различных тРНК, распознающих синонимичные кодоны, могут сильно колебаться в ответ на аминокислотное голодание, в зависимости от стадии деления и т.д. [54,

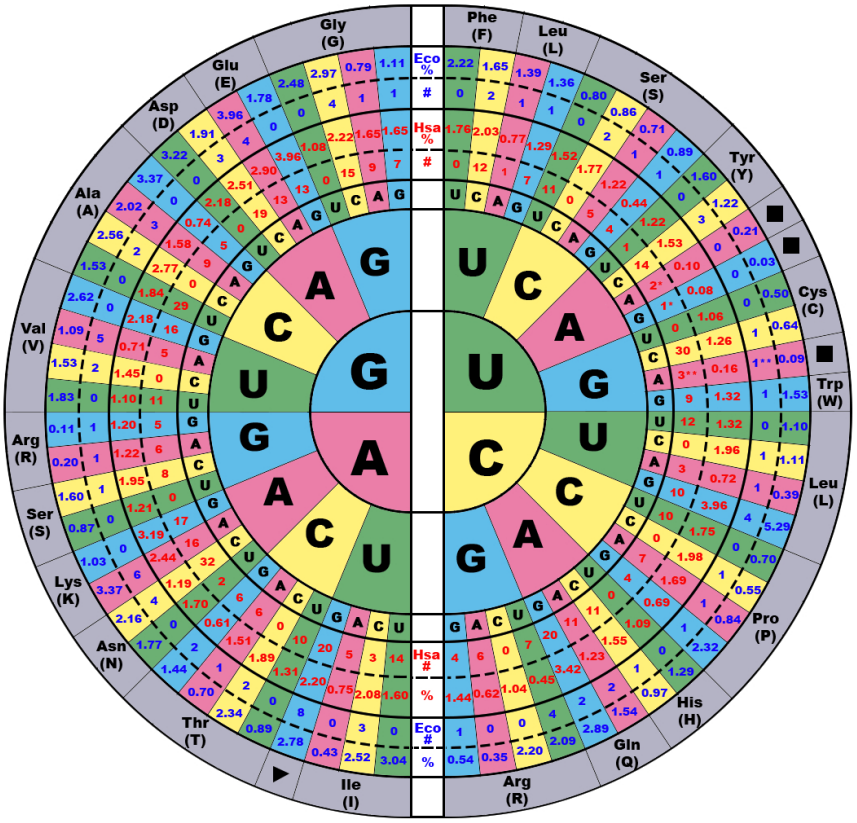

Рис. 2. Частота кодонов (%) во всех кодирующих последовательностях и количество копий генов тРНК (#) для *Escherichia coli* K12 (*Eco*, синий) и Homo sapiens (Has, красный) (информация взята из http://gtrnadb.ucsc.edu).

* Супрессорные-тРНК для стоп кодонов

** тРНК для селеноцистиина и супрессорные-тРНК для TGA стоп-кодона.

55]. Наиболее современным и информативным показателем, который учитывает не только общее количество тРНК, но и конкуренцию между всеми рибосомами за тРНК, соответствующую группе кодонов, является индекс нормализованной эффективности трансляции (nTE, normalized translational efficiency) [56]. Значение параметра nTE для кодона является отношением tAI к значению частоты транслирования данного кодона в организме.

Хотя рассчитанные значения скорости элонгации трансляции, полученные на основе параметров tAI и nTE, достаточно хорошо

коррелируют с экспериментальными данными, дальнейшее развитие методов анализа кодонного состава генов позволит давать еще более точные предсказания. Фактически наиболее важным для расчёта скорости прочтения кодона является значение концентрации зрелых (аминоацилированных и модифицированных) тРНК, готовых к доставке аминокислот к сайту трансляции. Однако если уровень аминоацилирования можно, предположительно, оценить количеством генов определенной тРНК (как это делают для параметров nTE и tAI), то с степень модифицированности тРНК охарактеризовать достаточно сложно. Механизмы, отвечающие за модификации тРНК в *E. coli*, на данный момент плохо изучены, однако исследователи считают, что их роль в регуляции скорости трансляции существенна [49].

Трансляционный скат. Еще одной особенностью трансляции *E. coli* является предпочтение более «медленных» кодонов в 5'-концевом участке КП мРНК, называемое трансляционным скатом [56, 57]. Анализ распределения «медленных» кодонов внутри генов *E. coli* выявил наличие в высокоэкспрессируемых генах участка, следующего за стартовым кодоном, включающего ~10–15 относительно редких/медленно транслируемых кодонов [56]. Данный факт также подтверждается результатами экспериментов, основанных на методе профилирования рибосом [32, 57, 58]. В основе метода лежит наблюдение, что транслирующая рибосома защищает от нуклеазной активности участок мРНК, на котором она расположена [12]. После ингибирования трансляции в клетках с помощью ингибиторов элонгации производится обработка РНКазами с последующим секвенированием защищенных рибосомами фрагментов РНК, что обеспечивает «запись» положения рибосомы в момент, когда трансляция была остановлена. Эксперименты показывают повышенное содержание рибосом на 5'-концевом участке КП мРНК, что указывает на относительно низкий темп трансляции данной области [32, 57, 58].

Предполагается, что сравнительно медленное начало процесса элонгации равномерно распределяет рибосомы, что снижает вероятность образования «заторов» рибосом во время элонгации высокоэкспрессируемых белков с относительно высокой плотностью рибосом на мРНК (Рис. 3) [57, 58]. Однако данный эффект также может объясняться тем, что эволюционный отбор против вторичных структур мРНК на 5'-конце для облегчения инициации трансляции высокоэкспрессируемых генов более важен, чем давление отбора в сторону хорошо транслируемых кодонов [19]. В этом случае скорость движения рибосом после инициации трансляции должна быть сбалансирована с необходимостью отсутствия вторичных структур на 5' конце мРНК.

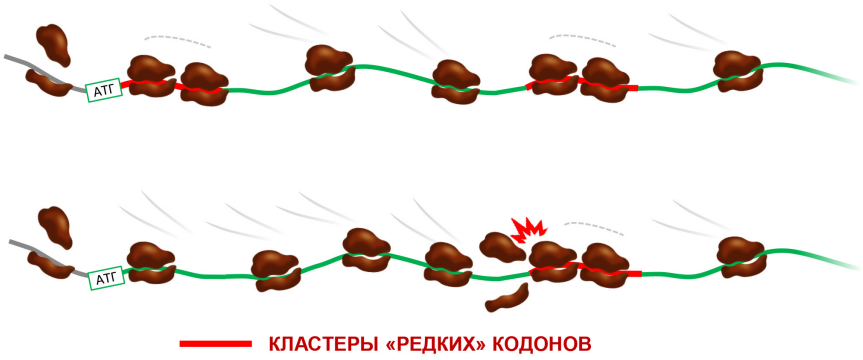

Рис. 3. Предположительный эффект трансляционного ската на процесс трансляции.

Зеленый – кластеры «частых» кодонов на мРНК; красный – кластеры «редких» кодонов.

Правильное распределение редких кодонов в регионе после стартового кодона (вверху) способствует равномерному распределению рибосом по мРНК. При замене кодонов и/или экспрессирующего хозяина (внизу), приводящих к изменению скорости движения рибосом в регионе 5' конца мРНК, возможно неравномерное распределение рибосом, приводящее к заторам и преждевременной терминации.

Кодонное группирование. Распределение синонимичных кодонов в пределах открытой рамки считывания имеет свои закономерности: вместо случайного распределения существует кодонное группирование [59]. Исследования демонстрируют, что взаимное расположение идентичных и некоторых изоакцепторных (считываемых одной и той же тРНК) кодонов в непосредственной близости друг к другу, как правило, выгодно для процесса трансляции.

Данный эффект предположительно связан с тем, что различные модификации «колеблющихся» оснований тРНК влияют не только специфичность, но и на сродство/эффективность молекул тРНК в распознавании различных кодонов, что в теории может способствовать развитию синонимичных паттернов группирования кодонов у бактерий под давлением эволюционного отбора. Только идентичные пары кодонов и неидентичные пары, в которых два кодона распознающихся с одинаковой (или близкой) высокой аффинностью одной и той же модифицированной тРНК, будут благоприятствовать процессу трансляции и накапливаться в бактериальных геномах [59].

ШД-подобные последовательности. ШД-подобные последовательности представляют из себя элементы, комплементарные или частично комплементарные анти-ШД последовательности 16S рРНК рибосомы. Исследования показывают, что у бактерий присутствие ШД-подобных последовательностей в КП может приводить к замедлению процесса элонгации трансляции и значительному снижению продукции белка [60]. Обнаружено, что у большинства прокариот, включая *E. coli,* ШД-подобные мотивы подвергались негативному отбору на протяжении эволюции [61, 62].

Пара кодонов AGG–AGG (Arg–Arg), является одной из самых медленно транслируемых *in vivo* предположительно из-за значительной аффинности к фрагменту 16S рРНК комплиментарному ШД последовательности. Данное предположение подтверждается тем, что даже увеличение пула соответствующих $тРНК^{ARG}_{AGG}$ за счет введения мультикопийной плазмиды с геном этой тРНК (argU), не увеличивает скорость трансляции данной пары [63, 64]. Помещение последовательности AGG-AGG на 5'-концевом участке КП в значительной мере снижает уровень экспрессии. Негативный эффект выражен тем сильнее, чем ближе к стартовому кодону находится данная последовательность [65].

Помимо уменьшения уровня экспрессии, аналогичный тандем кодонов AGG–AGA может привести не только к замедлению, но и к преждевременной терминации трансляции с образованием укороченного белка [64]. Таким образом следует избегать данных последовательностей при рекомбинантной экспрессии генов в *E. coli.*

Рибосомные столкновения и рибосомные заторы. В клетках *E. coli* одну молекулу мРНК одновременно транслируют сразу несколько рибосом, при этом образуя так называемые полисомы (Рис. 4). Образование полисом может увеличивать эффективность трансляции, путём защиты мРНК от деградации и продлением времени ее существования в качестве матрицы для трансляции [12]. Также рибосомы за счет своей хеликазной активности [66] могут дестабилизировать вторичные структуры мРНК, влияя на доступность сайта связывания рибосомы для последующих стадий инициации трансляции [67, 68].

В то же время слишком высокая загрузка мРНК в сочетании с участками быстрой трансляции и/или замедлением прохождения рибосом из-за наличия медленно транслируемых кодонов, может привести к рибосомным столкновениям и рибосомным заторам, которые в итоге снижают эффективность трансляции за счет замедления или полной терминации трансляции [69]. В частности, столкновения

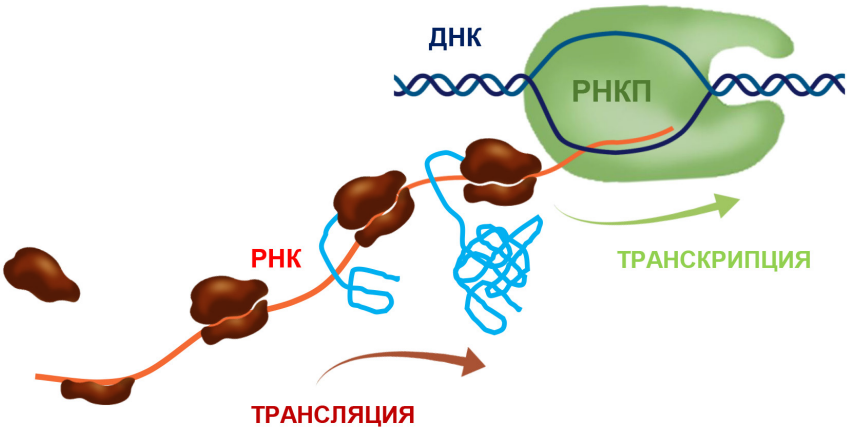

Рис. 4. Схематическое представление одновременного прохождения транскрипции/трансляции в бактериях и образования полисом.

могут либо стимулировать спонтанную диссоциацию застрявших рибосом, либо запускать пути контроля застрявших рибосом, приводящих к диссоциации рибосомы и деградации мРНК [70, 71].

Кроме того, в *E. coli* существует механизм «высвобождения» застопорившихся рибосом за счет распознавания заблокированных субъединиц 50S и последующей протеолитической деградации частично сформированной полипептидной цепи [72]. Таким образом, в бактериях имеются механизмы котрансляционной деградации как мРНК, так и растущего пептида, способные влиять на эффективность трансляции.

На оптимальное распределение рибосом на мРНК может влиять не только общий кодонный состав, но и, как было сказано ранее, наличие «трансляционного ската». Медленная элонгация на ранних этапах трансляции может влиять на правильность дальнейшего равномерного распределения рибосом по всей мРНК, тем самым предотвращая рибосомные столкновения и заторы [57].

Котрансляционный фолдинг белков. Исследования показывают, что на определенных этапах трансляции замедление элонгации зачастую является критическим для правильного фолдинга растущей полипептидной цепи [73].

In vivo сворачивание белка начинается совместно с трансляцией, когда формирующийся лидерный пептид выходит из туннеля рибосомы. Вариации в скорости локальной трансляции могут способствовать локальному сворачиванию белка, позволяя упорядочен-

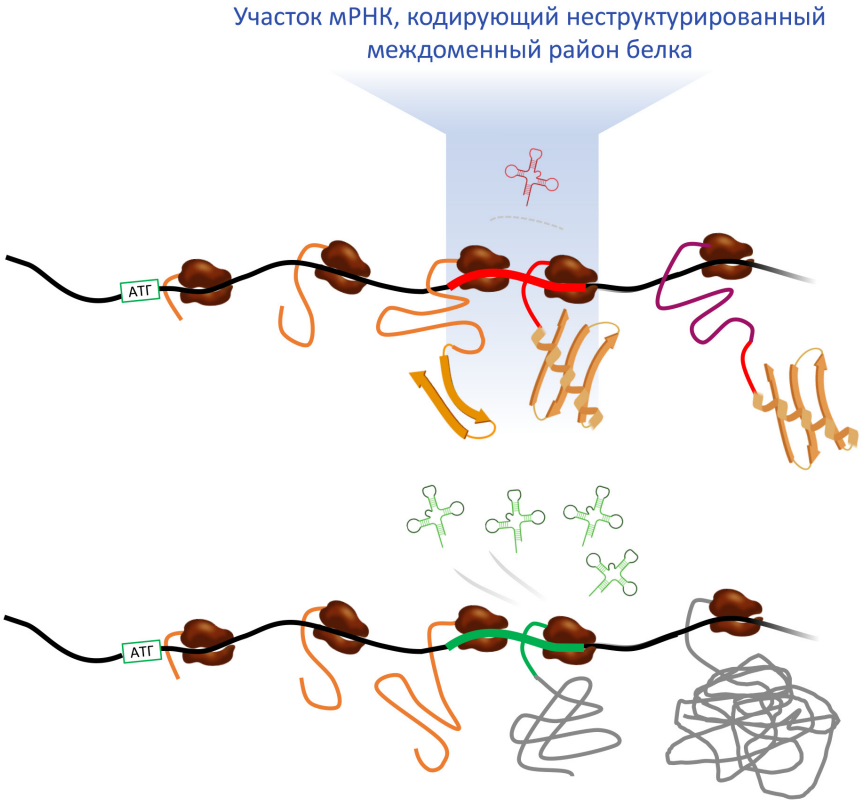

Рис. 5. Влияние изменения скорости прочтения мРНК, транслирующей неструктурированный междоменный участок, на правильный фолдинг всей белковой молекулы. При прохождении участка, кодирующего неструктурированный междоменный район белка, рибосомы могут замедляться, что дает время образовавшейся полипептидной цепи правильно свернуться (сверху). При смене экспрессионного хозяина (изменение пула тРНК) или при замене кодонов при неправильной оптимизации данный регион может проходится рибосомами быстрее, приводя к неправильному фолдингу конечного белка (снизу).

ное и последовательное структурирование возникающих доменов полипептидных цепей, выходящих из рибосомы [74, 75]. Снижение скорости трансляции увеличивает время, необходимое синтезированной полипептидной цепи для правильного сворачивания и формирования структурного домена еще до появления большего количества аминокислотных остатков других доменов (Рис. 5). В альтернативном случае ускорение трансляции может позволить

целому домену появиться согласованным образом, не приводя к образованию неправильных и неполноценных структур [76, 77].

Более того, существуют данные о наличии важных для фолдинга редких кодонов не только в неструктурированных междоменных областях, но и внутри структурированных доменов. Это позволяет предположить, что замедление трансляции может иметь особое значение для сворачивания более мелких структурных субэлементов [78].

Экспериментально было продемонстрировано, что замена редких кодонов синонимичными «быстрыми» кодонами может привести к неправильному фолдингу, который приводит к агрегации (образованию телец включения) или деградации белка [79, 80], или же к появлению вариантов белка с измененными функциональными параметрами. Так, синонимичные мутации могут влиять даже на субстратную специфичность ферментов [81], а также на профили фосфорилирования и активность белков [82].

Также, следует отметить, что некоторые структурно и/или функционально важные области (например, каталитические центры) могут кодироваться частыми кодонами не из-за кинетики сворачивания, а в связи с более точной трансляцией частых кодонов, понижая вероятность ошибки в данной области [83].

Вторичная структура 3' конца мРНК. В 3'-концевых областях мРНК *E. coli* также обнаруживается сниженное содержание GC-пар [33, 84]. Причиной этого предположительно является отбор против прочных вторичных структур РНК, формирование которых на 3'-конце мРНК может влиять на нормальную терминацию трансляции [33, 84–86]. Свидетельством этого является то, что сниженный GC состав на 3' конце, хоть и слабо, но коррелирует с увеличением экспрессии белка в *E. coli* [33]. Интересными также являются различия в паттернах использования кодонов между эукариотами и прокариотами. Предпочтение на 3'-конце КП гена к кодонам, кончающимся на A/T у бактерий более выраженно чем у эукариот. [87].

Выбор стоп-кодона. В *E. coli* отсутствуют тРНК, способные прочитывать три стоп-кодона UAA, UAG и UGA. Вместо этого стоп-кодон UAG декодируется фактором освобождения 1 (release factor 1, RF1), UGA–RF2, а UAA распознается как RF1, так и RF2. Процесс терминации трансляции на всех трех стоп-кодонах стимулируется RF3 [88, 89].

Наиболее часто используемым стоп-кодоном *E. coli* является UAA, а самым редко используемым – UAG [84, 90]. Основание, следующее за стоп-кодоном, может также быть важным элементом сигнала терминации трансляции. При этом эффективность терминации значительно

варьирует в зависимости от стоп-кодона и четвертого основания, в пределах от 80% UAAU, который на ряду с UAAG используется большинством высокоэкспрессируемых генов, до самого низкоэффективного UGAC–7% [91]. Данный факт наиболее часто связывают с вкладом данного основания в эффективность связывания фактора высвобождения 3 (RF3). Некоторые данные также указывают на вероятность того, что большее количество оснований (+4–+10) после стоп кодона могут вносить вклад в эффективность терминации также через взаимодействие с RF3 [92, 93], однако точного доказательства еще не получено.

Различие в эффективности терминации различных стоп-сигналов как предполагается, может быть связано со скоростью и эффективностью считывания терминационного сигнала [94]. Данный показатель в свою очередь зависит от концентрации факторов высвобождения для каждого стоп-кодона, аффинности их связывания, а также с особенностями скорости их рециклирования [90]. А поскольку кодон UAA считывается обоими факторами высвобождения, данные показатели меньше влияют на эффективность его прочтения. Этот факт подтверждается данными о том, что при повышении экспрессии RF3 рост эффективности декодирования сильных стоп-сигналов UAAU и UAAG выражен сильнее, нежели в случае более слабых [95]. Эффективность UAA также была показана в экспериментах с делецией гена RF3, при которой он обладал наибольшей эффективностью и точностью по сравнению с UGA, в случае которого делеция RF3 приводила к увеличению случаев перекодирования или скоплений рибосом возле стоп-кодона [96].

Неэффективная терминация трансляции может вызывать формирование рибосомных заторов, что оказывает негативное влияние на элонгацию трансляции, когда несколько рибосом «собираются» на транслирующей мРНК перед стоп-кодоном [97]. Особенно большую роль данный фактор может играть в трансляции коротких генов. Поэтому именно в небольших, но высокоэкспрессируемых генах *E. coli* наиболее часто присутствует кодон UAA, а не менее эффективный стоп кодон UGA [98]. Кроме того, диссоциация факторов терминации важна для обновления пула свободных рибосом, что имеет решающее значение для быстрой трансляции и увеличения экспрессии белка [97].

III. ПРИМЕНЕНИЕ ЭФФЕКТОВ КОДОННОГО ПРЕДПОЧТЕНИЯ

УЛУЧШЕНИЕ ЭКСПРЕССИОННОГО ШТАММА

Как указывалось выше, скорость прочтения кодонов и соответственно трансляции зачастую связана с количеством тРНК, соответствующей определенным кодонам. Были предприняты попытки увеличить уровень экспрессии гетерологичных белков путем введения в экспрессионный штамм дополнительных копий генов тРНК, соответствующих редким, медленно транслируемым в *E. coli* кодонам. Например, группа штаммов *E. coli* Rosetta™ 2(DE3)pLysS содержит плазмиду pRARE с генами тРНК, распознающих следующие кодоны: AGG, AGA, AUA, CUA, CCC, GGA, а штамм BL21-CodonPlus – плазмиду pRIL, несущую гены тРНК для 4 редких для *E. coli* кодонов [99]. Однако данный метод решает лишь проблемы, связанные с медленным прохождением рибосом и полностью не учитывает особенности кодонного предпочтения различных частей гена.

КОДОННАЯ ОПТИМИЗАЦИЯ ГЕНОВ

Давление естественного отбор сформировало правила кодонного предпочтения для отдельных областей гена, и теперь основная задача состоит в том, чтобы понять данные правила для эффективной экспрессии белков в гетерологичных системах. По мере расширения нашего понимания этих механизмов и выявления новых закономерностей, оптимизация генов постепенно отходит от общепринятой концепции, согласно которой синтетические гены должны содержать как можно больше частых/быстро транслируемых кодонов для достижения высоких уровней экспрессии белка.

Стандартные конструкции для повышения экспрессии

Наиболее простым методом оптимизации уровня экспрессии могло бы являться использование стандартных регуляторных элементов (промотор, ШД, РИТ и др.), которые можно надежно и воспроизводимо использовать в комбинации с любыми генами для увеличения уровня их экспрессии [6]. Однако из-за неполного понимания всех факторов, определяющих эффективность экспрессии белка, разработка и применение подобных конструкций находятся лишь на начальных стадиях. Даже в пределах хорошо изученных организмов, таких как *E. coli*, кажущиеся простыми генетические конструкции ведут себя по-разному в разных условиях (разные экспрессионные конструкции, штаммы, среды и условия культивирования) [100].

Множество экспериментальных данных указывает на то, что именно 5'-НТО мРНК является областью, в большей мере определяющей эффективность трансляции в *E. coli*. Как указывалось выше, основным фактором в данном регионе мРНК, негативно влияющим на уровень экспрессии белка, принято считать образование вторичных структур мРНК. Считается, что данные структуры образуются не только внутри самой 5'-НТО мРНК, но и с участием нижележащих 10–15 кодонов области КП [6, 19, 20]. Данный факт усложняет создание стандартных конструкций, поскольку последовательность ШД, которая хорошо инициирует трансляцию для одной кодирующей последовательности, может вообще не функционировать с другой кодирующей последовательностью [27]. Тем не менее, были предприняты попытки создания стандартных конструкций при помощи изоляции этих двух областей. В одном исследовании, для устойчивой экспрессии гена использовались стандартизованные модули с 5'-НТО мРНК и N-концевыми белковыми фрагментами, впоследствии отщепляемыми для получения целевого белка [101]. Еще в одной работе были применены бисцистронные модули, которые оказались очень полезными для изоляции последовательностей, отвечающих за эффективную инициацию трансляции от ШД и кодирующей последовательности белка [6, 18].

Создание стандартных промоторов является не настолько сложной задачей, поскольку было показано, что их сила не слишком сильно варьирует при экспрессии различных генов и остаётся на предсказуемом уровне [18].

Компьютерная оптимизация генов

Наиболее часто используемым методом компьютерной оптимизации является адаптация кодонного состава гена к различным индексам использования кодонов выбранного экспрессирующего хозяина [52]. Основным примером является индекс наиболее часто используемых кодонов в высоко экспрессируемых эталонных генах хозяина – CAI (алгоритмы CAI calculator, CAIcal, CodonO, CodonW).

Некоторые академические и коммерческие алгоритмы также принимают во внимание дополнительные параметры, такие как содержание GC-пар и избегание определенных мотивов, таких как ШД-подобные последовательности, сайты РНКазы Е или повторов, образующих прочные вторичные структуры внутри КП. И лишь несколько алгоритмов дополнительно направлены на минимизацию вторичных структур в 5'-области мРНК, хотя данный показатель, безусловно, является одним из ключевых для инициации трансляции и конечного уровня экспрессии белка [102].

Другим методом оптимизации генов является кодонная гармонизация. При данном подходе кодоны в гене заменяются по принципу равенства их относительной скорости трансляции i) в клетке, являющейся источником данного гена ii) в клетке – гетерологичном экспрессионном хозяине [103]. Относительную скорость трансляции в данном методе выражают через частоту использования данных кодонов в организме [73, 82, 104]. Данный подход наиболее часто применяется для оптимизации фолдинга белка в *E. coli* [105, 106].

Даже учитывая весь прогресс, на данный момент мы не имеем полной картины влияния кодонного состава на экспрессию и фолдинг гетерологичных белков, что серьезно усложняет применение компьютерной оптимизации. Многие разработанные алгоритмы, учитывающие большое количество параметров (например Eugene, DNA-Tailor), оставляют за пользователем установку конкретных приоритетов в оптимизации, что на практике является трудно реализуемой задачей из-за сложности оценки веса каждой характеристики. Частичным решением данной проблемы может являться применение технологий машинного обучения, которые показывают себя как отличный инструмент работы с большими массивами связанных данных. Различные типы машинного обучения могут использоваться для создания надежных алгоритмов для улучшения последовательности синтетических генов. Примером применения технологий глубокого обучения является создание метода полноценной оптимизации генов [107]. В данной работе нейронная сеть обучалась кодонным ландшафтам на 4906 генах *E. coli* из базы данных NCBI. Результаты экспериментальной проверки показывают, что данный метод повышения экспрессии белка вполне эффективен и конкурентоспособен. На данный момент технологии машинного обучения в основном применяют для анализа кодонных предпочтений эукариот, таких как *Saccharomyces cerevisiae* [108–110].

Такие подходы могут быть перспективны для разработки более совершенных методов прогнозирования различных показателей, таких как уровень экспрессии и фолдинг. Однако для алгоритмов машинного обучения требуются большие массивы данных, которые должны быть однородны по своему составу для избегания создания неверных соотношений между параметрами внутри нейронной сети. Другой большой проблемой является то, что машинное обучение не обязательно ведет к более глубокому пониманию биологии и реальных процессов, а лишь хорошо стремится увеличить заданный показатель и на практике зачастую является «черным ящиком» [107].

Таким образом, улучшение продукции гетерологичных белков с помощью компьютерных алгоритмов оптимизации кодонов до сих пор остается методом проб и ошибок. Тестирование нескольких вариантов повышает вероятность успеха, но также увеличивает затраты на работу.

Методы оптимизации, основанные на создании библиотек

Альтернативным подходом к компьютерной оптимизации и стандартным конструкциям является получение библиотек синтетических плазмид с рандомизированными участками генов, требующими оптимизации, и дальнейшим скринингом полученных последовательностей по уровню экспрессии. Метод синтетических библиотек широко применяют для оптимизации промоторов [5, 6], последовательностей ШД, РИТ [6, 18] и региона 5'-НТО + КП мРНК [111, 112]. Наряду с методами машинного обучения рандомизированные плазмидные библиотеки в меньшей степени зависят от полноты понимания механизмов и особенностей кодонного предпочтения, поскольку позволяют одновременно анализировать множество вариантов последовательностей. Развитие методов синтетических библиотек и анализ данных, полученных из экспериментов данного типа, также способствует развитию новых компьютерных алгоритмов оптимизации.

Основным подходом при скрининге библиотек является слияние КП гена с репортерным белком. При этом уровень экспрессии или другие типы изменений детектируются по изменению интенсивности репортерного сигнала. В качестве репортерных белков могут применятся флуоресцентные белки, такие как mCherry, GFP или superfolderGFP – специальная форма GFP, разработанная для экспериментов по экспрессии слитых белков в *E. coli* [113]. Скрининг при данном выборе репортера происходит путём отбора отдельных колоний высеянных на чашку Петри [114] или помощи проточного сортинга [19] по интенсивности флуоресценции. Другим типом белков-репортёров являются факторы устойчивости к антибиотикам. Скрининг в данном случае происходит посредством высевания полученных клонов на чашки Петри с постепенным увеличением концентрации антибиотика, при этом клоны с наибольшим уровнем экспрессии слитого белка будут проявлять наибольшую выживаемость [111].

Некоторые исследователи считают, что слияние с репортером может искажать свойства целевого белка, такие как растворимость и уровень экспрессии, что может приводить к ложноположительным или ложноотрицательным результатам. Альтернативный метод, не

основанный на слиянии белков, недавно был разработан путем трансляционного связывания через БЦДГ интересующего белка с выбираемым репортером устойчивости к антибиотикам. Система TARSyn была продемонстрирована для высокопроизводительного отбора конструкций с оптимизированным 5'-концом мРНК, для экспрессии антител в *E. coli* [111].

Однако количество последовательностей, которые можно проанализировать без применения высокопроизводительных методов скрининга, является ограниченным. В среднем полная вырожденность без изменения аминокислотного состава 15 кодонов создает библиотеку размером ~2×10^7 вариантов, что невозможно проанализировать даже самыми современными методами с высокой пропускной способностью (высокопроизводительными методами). Поэтому большинство работ, использующих библиотечные методы оптимизации: (1) ограничиваются меньшим количеством вырожденных нуклеотидов либо (2) применяют методики скрининга, сокращающие количество вариантов (например TARSyn) [111]. Таким образом, несмотря на свои очевидные достоинства, данная методика наиболее эффективна для оптимизации отдельных коротких участков гена, таких как последовательность ШД, промотор, РИТ, 5'-НТО мРНК и др. Это связано как с ограниченной пропускной способностью метода не позволяющей анализировать огромные библиотеки, так и отсутствием простых методов генной инженерии для рандомизации длинных участков гена.

IV. ЗАКЛЮЧЕНИЕ

Разработка различных алгоритмов оптимизации по-прежнему остается сложной задачей, которая зачастую ограничивает применение синтетических генов в биотехнологических сферах из-за проблем с маленьким выходом или неправильным фолдингом белка. Большинство методов оптимизации генов используют устаревшие показатели, такие как CAI, уменьшение количества редких кодонов и т.д., не учитывая результаты, полученные в данной сфере за последние годы.

На данный момент очевидным фактом является то, что применение одного параметра для оптимизации всего гена, без учета кодонных предпочтений отдельных участков, не дает желаемого результата. Каждый участок гена – 5'/3'-НТО и КП мРНК, границы доменов и др. имеют разные паттерны нуклеотидного предпочтения и при оптимизации по-разному влияют на эффективность экспрессии и правильность фолдинга белка. Современные знания в какой-то мере

позволяют создавать конструкции с оптимизированными отдельными участками гена, как например компьютерная минимизация вторичной структуры 5'-конца мРНК для увеличения экспрессии [115] или оптимизация отдельных участков на границе доменов для повышения эффективности фолдинга белка [116, 117]. Однако из-за неполноты понимания механизмов данные методики все еще остаются предметом проб и ошибок. Поэтому создание оптимальных синтетических конструкций требует более глубокого понимания кодонного предпочтения как отдельных участков гена, так и в их комбинации.

Другим широко используемым подходом кодонной оптимизации является метод синтетических библиотек, в котором оптимизация достигается путем рандомизации регуляторных и/или кодирующих последовательностей генов с дальнейшим анализом уровня экспрессии по интенсивности сигнала репортерного белка. Основной проблемой данного метода является быстрый рост количества вариантов в библиотеке при увеличении рандомизируемой области. Даже анализ данных после рандомизации отдельного участка гена (например 5'-НТО + КП мРНК) является крайне нетривиальной задачей и на практике ограничивается ~ 300 тыс. вариантов [18]. Имеются несколько возможных вариантов преодоления данной проблемы. Первым является использование закономерностей, полученных из предыдущих библиотечных экспериментов для последующего ограничения рандомизируемых вариантов и снижения размера библиотеки. Полученные после нескольких таких итераций принципы построения библиотек будут в теории более адекватно описывать принципы кодонного предпочтения данного региона гена, что существенно облегчит анализ. В качестве альтернативы данному циклу можно использовать системы с репортерными белками, ограничивающими рост неоптимальных вариантов. Например, система TARSyn позволяет отбирать клоны с высокой экспрессией на основе устойчивости к антибиотикам [111].

Развитие и применение новых высокопроизводительных методик, таких как высокопроизводительное секвенирование, транскриптомные и протеомные методы анализа также позволяют получать более значимые и репрезентативные массивы данных, которые можно использовать для создания точных предсказывающих теорий и алгоритмов оптимизации генов. Для анализа полученных огромных массивов данных можно использовать широко развивающиеся алгоритмы машинного обучения [118]. Подходы машинного обучения могут помочь в более объективном выявлении неизвестных функций и

факторов, и могут быть полезными для разработки более совершенных алгоритмов прогнозирования уровня экспрессии белков.

В целом, как для понимания фундаментальных принципов кодонного предпочтения и экспрессии генов, так и для решения практических задач необходимо более глубокое изучение данной области. Одной из важных проблем является раскрытие модели скорости прочтения отдельных кодонов в нормальных условиях. Существующие на данные момент метрические показатели, такие как tAI и nTE, позволяют делать неплохие предсказания, однако все еще не включают учёт количества аминоацилированных и модифицированных аминоацил-тРНК. Многие организмы, включая *E. coli,* могут дополнительно изменять профиль модифицирования и аминоацилирования тРНК в различных условиях роста или стресса. В этом смысле «оптимальность» кодонов также может рассматриваться как динамический параметр: разные условия требуют разных «быстрых» или «медленных» кодонов для создания подходящей реакции на изменяющиеся условия [119]. Поэтому для более глубокого понимания требуется нормализация условий проведения экспериментов. Более точные модели предсказания скорости прочтения кодонов могут помочь решить вопросы скорости движения рибосом, трансляционного ската, трансляционных пауз и трансляционно-зависимого фолдинга.

Другой, не менее важной задачей, является точная модель сегментации гена и взаимосвязи кодонного состава его отдельных сегментов. Каждый сегмент гена имеет разные требования для оптимизации кодонов и выяснение взаимного влияния сегментов на кодонное предпочтение друг друга также крайне важно. При анализе параметра кодирующей последовательности, потенциально влияющего на экспрессию, например влияние вторичных структур мРНК на уровень экспрессии, правильным было бы минимизировать взаимодействие между отдельными сегментами, чтобы иметь возможность учитывать влияние каждого сегмента на данный параметр. Такой подход позволит более точно изучить кодонные предпочтения как отдельных сегментов, так и их совокупности.

Несмотря на впечатляющий прогресс, дальнейшее совершенствование и разработка новых экспериментальных и вычислительных подходов будет иметь важное значение для решения ключевых вопросов в данной тематике.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

ЛИТЕРАТУРА

1. Huang, C.J., Lin, H., & Yang, X. (2012). Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements. *Journal of Industrial Microbiology and Biotechnology*, **39**, 383–399.
2. Baeshen, M.N., Al-Hejin, A.M., Bora, R.S., Ahmed, M.M.M., Ramadan, H.A.I., Saini, K.S., Redwan, E.M. (2015). Production of biopharmaceuticals in E. Coli: Current scenario and future perspectives. *Journal of Microbiology and Biotechnology*. Korean Society for Microbiolog and Biotechnology, **25**, 953–962.
3. Packiam, K.A.R., Ramanan, R. N., Ooi, C.W., Krishnaswamy, L., & Tey, B.T. (2020, April 1). Stepwise optimization of recombinant protein production in Escherichia coli utilizing computational and experimental approaches. *Applied Microbiology and Biotechnology*, **104**, 3253–3266.
4. Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., … Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 14024–14029.
5. Alper, H., Fischer, C., Nevoigt, E., & Stephanopoulos, G. (2005). Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences*, **102**, 12678–12683.
6. Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.A., Endy, D. (2013). Precise and reliable gene expression via standard transcription and translation initiation elements. *Nature Methods*, **10**, 354–360.
7. Conrad, T., Plumbom, I., Alcobendas, M., Vidal, R., & Sauer, S. (2020). Maximizing transcription of nucleic acids with efficient T7 promoters. *Communications Biology*, **3**, 439.
8. Komura, R., Aoki, W., Motone, K., Satomura, A., & Ueda, M. (2018). High-throughput evaluation of T7 promoter variants using biased randomization and DNA barcoding. *PLoS ONE*, **13**, e0196905.
9. Villada, J.C., Duran, M.F., & Lee, P.K.H. (2020). Interplay between Position-Dependent Codon Usage Bias and Hydrogen Bonding at the 5′ End of ORFeomes. *mSystems*, **5**.
10. Gorochowski, T.E., Chelysheva, I., Eriksen, M., Nair, P., Pedersen, S., & Ignatova, Z. (2019). Absolute quantification of translational regulation and burden using combined sequencing approaches. *Molecular Systems Biology*, **15**, e8719.
11. Mittal, P., Brindle, J., Stephen, J., Plotkin, J.B., & Kudla, G. (2018). Codon usage influences fitness through RNA toxicity. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 8639–8644.
12. Dé Rique Braun, F., le Derout, J., & Ré Gnier, P. (1998). Ribosomes inhibit an RNase E cleavage which induces the decay of the rpsO mRNA of Escherichia coli. *The EMBO Journal,* **17**, 4790–4797.
13. Joel, J. (1993) Control of Messenger RNA Stability. Part I: Prokaryotes Part II: Eukaryotes: Elsevier. 495–517 p.
14. Kushner, S.R. (2004). mRNA decay in prokaryotes and eukaryotes: Different approaches to a similar problem. *IUBMB Life*, **56**, 585–594.
15. Emory, S.A., Bouvet, P., & Belasco, J.G. (1992). A 5'-terminal stem-loop structure can stabilize mRNA in Escherichia coli. *Genes & Development*, **6**, 135–148.
16. Arnold, T.E., Yu, J., & Belasco, J.G. (1998). mRNA stabilization by the ompA 59 untranslated region: Two protective elements hinder distinct pathways for mRNA degradation. *RNA*, **4**, 319–330.

17. Baker, K.E., & Mackie, G.A. (2003). Ectopic RNase E sites promote bypass of 5'-end-dependent mRNA decay in Escherichia coli. *Molecular Microbiology,* **47**, 75–88.
18. Cambray, G., Guimaraes, J.C., & Arkin, A.P. (2018, November 1). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli. *Nature Biotechnology*, **36**, 1005–1015.
19. Goodman, D.B., Church, G.M., & Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. *Science*, **342**, 475–479.
20. Kudla, G., Murray, A.W., Tollervey, D., & Plotkin, J.B. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science*, **324**, 255–258.
21. Menendez-Gil, P., Caballero, C.J., Catalan-Moreno, A., Irurzun, N., Barrio-Hernandez, I., Caldelari, I., & Toledo-Arana, A. (2020). Differential evolution in 3′UTRs leads to specific gene expression in Staphylococcus. *Nucleic Acids Research*, **48**, 2544–2563.
22. Menendez-Gil, P., & Toledo-Arana, A. (2021). Bacterial 3′UTRs: A Useful Resource in Post-transcriptional Regulation. *Frontiers in Molecular Biosciences*, **7**, e617633.
23. Ruiz de los Mozos, I., Vergara-Irigaray, M., Segura, V., Villanueva, M., Bitarte, N., Saramago, M., Toledo-Arana, A. (2013). Base Pairing Interaction between 5′- and 3′-UTRs Controls icaR mRNA Translation in Staphylococcus aureus. *PLoS Genetics*, **9**, e1004001.
24. Zhao, J.P., Zhu, H., Guo, X.P., & Sun, Y.C. (2018). AU-rich long 3′ untranslated region regulates gene expression in bacteria. *Frontiers in Microbiology*, **9**, e3080.
25. McCarthy, J.E.G., & Gualerzi, C. (1990). Translational control of prokaryotic gene expression. *Trends in Genetics*, **6**, 78–85.
26. Laursen, B.S., Sørensen, H.P., Mortensen, K.K., & Sperling-Petersen, H.U. (2005). Initiation of Protein Synthesis in Bacteria. *Microbiology and Molecular Biology Reviews*, **69**, 101–123.
27. Salis, H.M., Mirsky, E.A., & Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, **27**, 946–950.
28. Barrick, D., Villanueba, K., Childs, J., Kalil, R., Schneider, T.D., Lawrence, C.E., Stormo, G.D. (1994). Quantitative analysis of ribosome binding sites in *E. coli*. *Nucleic Acids Research*, **22**, 1287-1295.
29. Nieuwkoop, T., Claassens, N.J., & van der Oost, J. (2019). Improved protein production and codon optimization analyses in Escherichia coli by bicistronic design. *Microbial Biotechnology*, **12**, 173–179.
30. Schoner, B.E., Belagaje, R.M., & Schoner, RG. (1986). Translation of a synthetic two-cistron mRNA in Escherichia coli (bovine growth hormone/human growth hormone/runaway replicon). *Biochemistry,* **83**, 8506–8510.
31. Makoff, A.J., & Smallwood, A.E. (1990). The use of two-cistron constructions in improving the expression of a heterologous gene in *E.coli*. *Nucleic Acids Research*, **18**, 1711–1718.
32. Boël, G., Letso, R., Neely, H., Price, W.N., Wong, K.H., Su, M., … Hunt, J.F. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature*, **529**, 358–363.
33. Allert, M., Cox, J.C., & Hellinga, H.W. (2010). Multifactorial Determinants of Protein Expression in Prokaryotic Open Reading Frames. *Journal of Molecular Biology*, **402**, 905–918.
34. del Campo, C., Bartholomäus, A., Fedyunin, I., & Ignatova, Z. (2015). Secondary Structure across the Bac-

terial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS Genetics*, **11**, e1005613.

35. Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.
36. Borujeni, A.E., Cetnar, D., Farasat, I., Smith, A., Lundgren, N., & Salis, H.M. (2017). Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Research*, **45**, 5437–5448.
37. Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E., & Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods*, **11**, 959–965.
38. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., & Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.
39. Mustoe, A.M., Busan, S., Rice, G.M., Hajdin, C.E., Peterson, B.K., Ruda, V.M., … Weeks, K.M. (2018). Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. *Cell*, **173**, 181–195.e18.
40. Kelsic, E.D., Chung, H., Cohen, N., Park, J., Wang, H.H., & Kishony, R. (2016). RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. *Cell Systems*, **3**, 563–571.e6.
41. Gouy, M., & Grantham, R. (1980). Polypeptide elongation and tRNA cycling in *Escherichia coli*: A dynamic approach. *FEBS Letters*, **115**, 151–155.
42. Crick, F.H.C. (1966). Codon-anticodon pairing: The wobble hypothesis. *Journal of Molecular Biology*, **19**, 548–555.
43. Söll, D., Jones, D.S., Ohtsuka, E., Faulkner, R.D., Lohrmann, R., Hayatsu, H., … Bock, R.M. (1966). Specificity of sRNA for recognition of codons as studied by the ribosomal binding technique. *Journal of Molecular Biology*, **19**, 556–573.
44. Agris, P.F., Vendeix, F.A.P., & Graham, W. D. (2007). tRNA's Wobble Decoding of the Genome: 40 Years of Modification. *Journal of Molecular Biology*. Academic Press, **366**, 1–13.
45. Kothe, U., & Rodnina, M.V. (2007). Codon Reading by tRNAAla with Modified Uridine in the Wobble Position. *Molecular Cell*, **25**, 167–174.
46. Ran, W., & Higgs, P.G. (2010). The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Molecular Biology and Evolution*, **27**, 2129–2140.
47. Dykeman, E.C. (2020). A stochastic model for simulating ribosome kinetics in vivo. *PLoS computational biology*, **16**, e1007618.
48. Vieira, J.P., Racle, J., & Hatzimanikatis, V. (2016). Analysis of Translation Elongation Dynamics in the Context of an Escherichia coli Cell. *Biophysical Journal*, **110**, 2120–2131.
49. de Crécy-Lagard, V., & Jaroch, M. (2021). Functions of Bacterial tRNA Modifications: From Ubiquity to Diversity. *Trends in Microbiology*, **29**, 41–53.
50. Gromadski, K.B., Daviter, T., & Rodnina, M.V. (2006). A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity. *Molecular Cell*, **21**, 369–377.
51. Sharp, P.M., & Li, W.H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research,* **15** 1281–1295.
52. Parret, A.H., Besir, H., & Meijers, R. (2016, June 1). Critical reflections on

synthetic gene design for recombinant protein expression. *Current Opinion in Structural Biology*, **38**, 155–162.

53. dos Reis, M., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Research*, **32**, 5036–5044.
54. Elf, J. (2003). Selective Charging of tRNA Isoacceptors Explains Patterns of Codon Usage. *Science*, **300**, 1718–1722.
55. Dittmar, K.A., Sørensen, M.A., Elf, J., Ehrenberg, M., & Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. *EMBO Reports*, **6**, 151–157.
56. Pechmann, S., & Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural and Molecular Biology*, **20**, 237–243.
57. Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.
58. Tuller, T., & Zur, H. (2015). Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Research*, **43**, 13–28.
59. Shao, Z.Q., Zhang, Y.M., Feng, X.Y., Wang, B., & Chen, J.Q. (2012). Synonymous codon ordering: A subtle but prevalent strategy of bacteria to improve translational efficiency. *PLoS ONE*, **7**, e33547.
60. Li, G.W., Oh, E., & Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, **484**, 538–541.
61. Diwan, G.D., & Agashe, D. (2016). The frequency of internal shine-dalgarno-like motifs in prokaryotes. *Genome Biology and Evolution*, **8**, 1722–1733.
62. Hockenberry, A.J., Jewett, M.C., Amaral, L.A.N., & Wilke, C.O. (2018). Within-Gene Shine–Dalgarno Sequences Are Not Selected for Function. *Molecular Biology and Evolution*, **35**, 2487–2498.
63. Chevance, F.F.V., le Guyon, S., & Hughes, K.T. (2014). The Effects of Codon Context on In Vivo Translation Speed. *PLoS Genetics*, **10**, e1004392.
64. Correddu, D., Montaño López, J. de J., Angermayr, S.A., Middleditch, M.J., Payne, L.S., & Leung, I.K.H. (2020). Effect of consecutive rare codons on the recombinant production of human proteins in Escherichia coli. *IUBMB Life*, **72**, 266–274.
65. Osterman, I.A., Chervontseva, Z.S., Evfratov, S.A., Sorokina, A.V., Rodin, V.A., Rubtsova, M.P., Sergiev, P.V. (2020). Translation at first sight: The influence of leading codons. *Nucleic Acids Research*, **48**, 6931–6942.
66. Takyar, S., Hickerson, R.P., & Noller, H.F. (2005). mRNA helicase activity of the ribosome. *Cell*, **120**, 49–58.
67. Andreeva, I., Belardinelli, R., & Rodnina, M.V. (2018). Translation initiation in bacterial polysomes through ribosome loading on a standby site on a highly translated mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 4411–4416.
68. Burkhardt, D.H., Rouskin, S., Zhang, Y., Li, G.-W., Weissman, J. S., & Gross, C.A. (2017). Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *eLife*, **6**, e22037.
69. Keiler, K.C. (2015). Mechanisms of ribosome rescue in bacteria. *Nature Reviews Microbiology*, **13**, 285–297.
70. Janssen, B.D., & Hayes, C.S. (2012). The tmRNA ribosome-rescue system. In *Advances in Protein Chemistry and Structural Biology*, **86**, 151–191.
71. Moore, S.D., & Sauer, R.T. (2007). The tmRNA system for translational surveillance and ribosome rescue.

Annual Review of Biochemistry, 76, 101–124.

72. Lytvynenko, I., Paternoga, H., Thrun, A., Balke, A., Müller, T.A., Chiang, C.H., … Joazeiro, C.A.P. (2019). Alanine Tails Signal Proteolysis in Bacterial Ribosome-Associated Quality Control. *Cell*, **178**, 76–v90. e22.
73. Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., … Komar, A.A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Molecular Cell*, **61**, 341–351.
74. Waudby, C.A., Launay, H., Cabrita, L.D., & Christodoulou, J. (2013). Protein folding on the ribosome studied using NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, **74**, 57–75
75. Gloge, F., Becker, A.H., Kramer, G., & Bukau, B. (2014, February). Co-translational mechanisms of protein maturation. *Current Opinion in Structural Biology*, **24**, 24–33.
76. Jacobson, G.N., & Clark, P.L. (2016, June 1). Quality over quantity: Optimizing co-translational protein folding with non-'optimal' synonymous codons. *Current Opinion in Structural Biology*, **38**, 102–110.
77. Sander, I.M., Chaney, J.L., & Clark, P.L. (2014). Expanding anfinsen's principle: Contributions of synonymous codon selection to rational protein design. *Journal of the American Chemical Society*, **136**, 858–861.
78. Chaney, J.L., Steele, A., Carmichael, R., Rodriguez, A., Specht, A.T., Ngo, K., Clark, P.L. (2017). Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Computational Biology*, **13**, e1005531.
79. Spencer, P.S., Siller, E., Anderson, J.F., & Barral, J.M. (2012). Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *Journal of Molecular Biology*, **422**, 328–335.
80. Zhang, G., Hubalewska, M., & Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural and Molecular Biology*, **16**, 274–280.
81. Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., & Gottesman, M.M. (2007). A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science*, **315**, 525–528.
82. Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J.M., Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, **494**, 111–115.
83. Drummond, D.A., & Wilke, C.O. (2008). Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, **134**, 341–352.
84. Eyre-Walker, A. (1996). The Close Proximity of Escherichia coli Genes: Consequences for Stop Codon and Synonymous Codon Use. *Journal of Molecular Evolution*, 42, 73–78.
85. Katz, L., & Burge, C. B. (2003). Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Research*, **13**, 2042–2051.
86. Rocha, E.P.C., Danchin, A., & Viari, A. (1999). Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Research*, 27, 3567–3576.
87. Zahdeh, F., & Carmel, L. (2019). Nucleotide composition affects codon usage toward the 3'-end. *PLoS ONE*, **14**, e0225633.
88. Capecchi, M.R. (1967). Polypeptide chain termination in vitro: isolation of a release factor. *Proceedings of the National Academy of Sciences*, **58**, 1144–1151.

89. Scolnick, E.M., & Caskey, C.T. (1969) Peptide chain termination, V. The role of release factors in mRNA terminator codon recognition. *Proceedings of the National Academy of Sciences*, **64**, 1235–1241.
90. Korkmaz, G., Holm, M., Wiens, T., & Sanyal, S. (2014). Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *Journal of Biological Chemistry*, **289**, 30334–30342.
91. Poole, E.S., Brown, C.M., & Tate, W.P. (1995). The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. *EMBO Journal*, **14**, 151–158.
92. Tate, W.P., & Mannering, S.A. (1996). Three, four or more: the translational stop signal at length. *Molecular Microbiology*, **21**, 213–219.
93. Namy, O., Hatin, I., & Rousset, J.-P. (2001). Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO reports*, **2**, 2001.
94. Sharp, P.M., & Bulmer, M. (1988). Selective differences among translation termination codons. *Gene*, **63**, 141–145.
95. Crawford, D.J.G., Ito, K., Nakamura, Y., & Tate, W.P. (1999). Indirect regulation of translational termination efficiency at highly expressed genes and recoding sites by the factor recycling function of Escherichia coli release factor RF3. *EMBO Journal*, **18**, 727–732.
96. Baggett, N.E., Zhang, Y., & Gross, C.A. (2017). Global analysis of translation termination in E. coli. *PLoS Genetics*, **13**, e1006676.
97. Pavlov, M.Yu., Freistroffer, D.V., Dincbas, V., MacDougall, J., Buckingham, R.H., & Ehrenberg, M. (1998). A direct estimation of the context effect on the efficiency of termination. *Journal of Molecular Biology*, **284**, 579–590.
98. Jin, H. (2002). Cis control of gene expression in *E.coli* by ribosome queuing at an inefficient translational stop signal. *The EMBO Journal*, **21**, 4357–4367.
99. Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in Biotechnology*, **22**, 346–353.
100. Kittleson, J.T., Wu, G.C., & Anderson, J.C. (2012, August). Successes and failures in modular genetic engineering. *Current Opinion in Chemical Biology*, **16**, 329–336.
101. Ki, M.R., & Pack, S.P. (2020, March 1). Fusion tags to enhance heterologous protein expression. *Applied Microbiology and Biotechnology*, **104**, 2411–2425.
102. Gould, N., Hendy, O., & Papamichail, D. (2014). Computational tools and algorithms for designing customized synthetic genes. *Frontiers in Bioengineering and Biotechnology*, **2**, e00041.
103. Tian, J., Yan, Y., Yue, Q., Liu, X., Chu, X., Wu, N., & Fan, Y. (2017). Predicting synonymous codon usage and optimizing the heterologous gene for expression in E. coli. *Scientific Reports*, **7**, e9926.
104. Rodriguez, A., Wright, G., Emrich, S., & Clark, P.L. (2018). %MinMax: A versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding. *Protein Science*, **27**, 356–362.
105. Angov, E. (2011). Codon usage: Nature's roadmap to expression and folding of proteins. *Biotechnology Journal*, **6**, 650–659.
106. Hillier, C.J., Ware, L. A., Barbosa, A., Angov, E., Lyon, J.A., Heppner, D.G., & Lanar, D.E. (2005). Process development and analysis of liver-stage antigen 1, a preerythrocyte-stage protein-based vaccine for Plasmodium falciparum. *Infection and Immunity*, **73**, 2109–2115.
107. Fu, H., Liang, Y., Zhong, X., Pan, Z.L., Huang, L., Zhang, H.L., … Liu, Z. (2020). Codon optimization with

deep learning to enhance protein expression. *Scientific Reports*, **10**, 17617.

108. Cuperus, J.T., Groves, B., Kuchina, A., Rosenberg, A.B., Jojic, N., Fields, S., & Seelig, G. (2017). Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Research*, **27**, 2015–2024.
109. Decoene, T., Peters, G., de Maeseneire, S.L., & de Mey, M. (2018). Toward Predictable 5′UTRs in Saccharomyces cerevisiae: Development of a yUTR Calculator. *ACS Synthetic Biology*, **7**, 622–634.
110. de Jongh, R.P.H., van Dijk, A.D. J., Julsing, M.K., Schaap, P.J., & de Ridder, D. (2020). Designing Eukaryotic Gene Expression Regulation Using Machine Learning. *Trends in Biotechnology*, **38**, 191–201.
111. Rennig, M., Martinez, V., Mirzadeh, K., Dunas, F., Röjsäter, B., Daley, D.O., & Nørholm, M.H.H. (2018). TARSyn: Tunable Antibiotic Resistance Devices Enabling Bacterial Synthetic Evolution and Protein Production. *ACS Synthetic Biology*, **7**, 432–442.
112. Mirzadeh, K., Martínez, V., Toddo, S., Guntur, S., Herrgård, M.J., Elofsson, A., Daley, D.O. (2015). Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector-Coding Sequence Junction. *ACS Synthetic Biology*, **4**, 959–965.
113. Pédelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C., & Waldo, G.S. (2006). Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnology*, **24**, 79–88.
114. Mirzadeh, K., Shilling, P.J., Elfageih, R., Cumming, A.J., Cui, H.L., Rennig, M., Daley, D.O. (2020). Increased production of periplasmic proteins in Escherichia coli by directed evolution of the translation initiation region. *Microbial Cell Factories*, **19**, e85.
115. Care, S., Bignon, C., Pelissier, M.C., Blanc, E., Canard, B., & Coutard, B. (2008). The translation of recombinant proteins in E. coli can be improved by in silico generating and screening random libraries of a –70/+96 mRNA region with respect to the translation initiation codon. *Nucleic Acids Research*, **36**, e6.
116. Hess, A.-K., Saffert, P., Liebeton, K., & Ignatova, Z. (2015). Mathematisch-Naturwissenschaftliche Fakultät Optimization of translation profiles enhances protein expression and solubility Optimization of Translation Profiles Enhances Protein Expression and Solubility. *PLOS ONE*, **10**, e127039.
117. Zhong, C., Wei, P., & Zhang, Y.H.P. (2017). Enhancing functional expression of codon-optimized heterologous enzymes in Escherichia coli BL21(DE3) by selective introduction of synonymous rare codons. *Biotechnology and Bioengineering*, **114**, 1054–1064.
118. Zrimec, J., Buric, F., Kokina, M., Garcia, V., & Zelezniak, A. (2021, June 10). Learning the Regulatory Code of Gene Expression. *Frontiers in Molecular Biosciences*, **8**, e673363.
119. Hanson, G., & Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology*, **19**, 20–30.