НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ДЕРЖАВНА УСТАНОВА «ІНСТИТУТ ХАРЧОВОЇ БІОТЕХНОЛОГІЇ ТА ГЕНОМІКИ НАН УКРАЇНИ»

РОКИЦЬКИЙ ІГОР ВОЛОДИМИРОВИЧ

УДК 579.25; 577.21

БІОІНФОРМАТИЧНІ ПІДХОДИ ТА РЕПОРТЕРНА СИСТЕМА ДЛЯ ДОСЛІДЖЕННЯ ОСОБЛИВОСТЕЙ ВЖИВАННЯ КОДОНІВ У ГЕНОМАХ СТРЕПТОМІЦЕТІВ

03.00.22 – молекулярна генетика

АВТОРЕФЕРАТ
дисертації на здобуття наукового ступеня
кандидата біологічних наук

Київ – 2019

Дисертацією є рукопис
Роботу виконано на кафедрі генетики та біотехнології Львівського національного університету імені Івана Франка

Науковий керівник доктор біологічних наук, професор
Осташ Богдан Омелянович,
Львівський національний університет імені Івана Франка,
провідний науковий співробітник науково-дослідної частини

Офіційні опоненти: доктор біологічних наук, професор
Сиволоб Андрій Володимирович,
ННЦ «Інститут біології та медицини» Київський національний університет імені Тараса Шевченка,
професор кафедри загальної та медичної генетики

кандидат біологічних наук, старший науковий співробітник
Карпов Павло Андрійович,
ДУ «Інститут харчової біотехнології та геноміки НАН України», завідувач лабораторії біоінформатики та структурної біології

Захист дисертації відбудеться «22» квітня 2019 р. о 13^{00} годині на засіданні спеціалізованої вченої ради Д 26.254.01 в ДУ «Інститут харчової біотехнології та геноміки НАН України» за адресою: 04123, м. Київ, вул. Осиповського, 2а. Факс: (044) 434 3777. Адреса електронної пошти: d26.254.01@ukr.net.

З дисертацією можна ознайомитись у бібліотеці ДУ «Інститут Харчової біотехнології та геноміки НАН України» за адресою: 04123, м. Київ, вул. Осиповського, 2а.

Автореферат розіслано 22 березня 2019 року.

Вчений секретар
спеціалізованої вченої ради,
кандидат біологічних наук, доцент Н.Л. Пастухова

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Обґрунтування вибору теми дослідження. Станом на січень 2018 р, встановлено нуклеотидну послідовність (секвеновано) більше 130 тисяч геномів різних видів бактерій. Зі зниженням часових та фінансових витрат та розвитком нових методів секвенування, кількість геномної інформації продовжує зростати експоненційно. Аналіз такої кількості інформації про нуклеотидні та амінокислотні послідовності вимагає застосування математичних моделей – узагальненого та інколи спрощеного відображення масиву даних. Наприклад, дослідження еволюції генетичних послідовностей здійснюють з використанням моделей заміщення нуклеотидів чи амінокислот. Їх застосовують для реконструкції філогенетичних зв'язків, пошуку та класифікації гомологічних послідовностей тощо. Зокрема, ці моделі застосовують в епідеміології, природоохоронній діяльності та криміналістиці (Bernard et al. 2007).

Сьогодні широкого застосування набувають кодонні моделі (Yang et al. 2000), що комбінують інформацію про нуклеотидні заміщення в кодувальній послідовності та відповідні заміщення амінокислот. Це дає змогу відслідковувати синонімічні заміщення, які не впливають на послідовність білка та є "невидимими", в моделях амінокислотних заміщень. Розуміння причин нерівномірного вживання синонімічних кодонів (переважне вживання кодонів, ПВК) дасть змогу розширити уявлення про механізми молекулярної еволюції. У свою чергу, таке розуміння удосконалить нові методи аналізу геномів.

Для вивчення механізмів вживання кодонів у геномах складно переоцінити вдалий вибір модельного організму. Зокрема, зручно вивчати ці механізми на геномах, що мають яскраво виражені зміщення у вживанні кодонів, які корелюють зі здатністю відповідного організму експресувати певну просту (таку, що легко виявляти) ознаку. Бактерії роду *Streptomyces* (стрептоміцети) мають такі властивості. Їхні GC-багаті геноми аномально збіднені за лейциновим кодоном TTA, і останній зустрічається лише в життєво неважливих генах вторинного метаболізму та морфогенезу. Природа такого ПВК маловивчена, що робить цих представників ідеальним об'єктом для побудови експериментальних та математичних моделей. Крім того, порушення експресії TTA-вмісних генів має простий фенотиповий вияв, як-от порушення спорулювання чи блокування синтезу забарвлених сполук. Бактерії роду *Streptomyces* широко використовують для продукування медично цінних вторинних метаболітів – гербіцидів, імуносупресантів, антибіотиків. Вивчення закономірностей кодонного складу геномів стрептоміцетів дасть змогу оптимізувати експресію промислово важливих генів у цих бактеріях.

Зв'язок роботи з науковими програмами, планами, темами. Дисертацію виконано у науково-дослідній лабораторії генетики, селекції та генетичної інженерії продуцентів антибіотиків (НДЛ-42) при кафедрі генетики та біотехнології Львівського національного університету імені Івана Франка. Роботу виконано у межах бюджетної теми БГ-41Нр "Універсальний генетичний механізм контролю продукції біологічно активних речовин стрептоміцетами." (№ державної реєстрації 0116U008070, 2016-2018 рр.).

Мета та завдання дослідження. Мета дисертаційної роботи – з'ясувати особливості кодонного складу геномів *Streptomyces* та його еволюції, розробити новий біоінформатичний інструмент та репортерну систему як нові знаряддя для вивчення трансляційної регуляції генів вторинного метаболізму. Для досягнення мети поставлено такі *завдання*:

1. Сформувати референтні вибірки ортологічних і функціонально споріднених генів *Streptomyces*, встановити для них оптимальні механістичні моделі нуклеотидних та амінокислотних заміщень;
2. Визначити особливості контекстного вживання кодонів на основі великого масиву геномів стрептоміцетів;
3. Створити веб-сервіс для візуалізації кодонних заміщень у великих масивах даних і на його основі визначити особливості кодонних заміщень для різних функціональних класів генів;
4. Описати можливі варіанти трансляції та містрансляції рідкісного лейцинового кодона TTA у стрептоміцетів, виходячи з різних підходів *in silico* до оцінки концентрації тРНК в клітині;
5. Сконструювати репортерну систему для виявлення та вивчення факторів, що модулюють експресію рідкісного лейцинового кодона TTA у стрептоміцетів.

Об'єктом дослідження є генетичні закономірності вживання та заміщення кодонів у геномах стрептоміцетів.

Предмет дослідження – гени тРНК та білок-кодувальні послідовності у геномах стрептоміцетів.

Методи дослідження: мікробіологічні (культивування штамів бактерій *in vitro*), генетичні (отримання та вивчення мутацій, генетична трансформація клітин *Escherichia coli*, кон'югаційні схрещування між *E. coli* та актиноміцетами), генно-інженерні (виділення та аналіз сумарної та плазмідної ДНК, конструювання рекомбінантних молекул ДНК, гель-електрофорез ДНК, полімеразна ланцюгова реакція), біоінформатичні (комп'ютерний аналіз нуклеотидних та амінокислотних послідовностей, філогенетичний аналіз, аналіз баз даних, передбачення структури та функції білків), комп'ютерні (створення онлайн-застосунків на мові програмування Python).

Наукова новизна отриманих результатів. Вперше сконструйовано кумат-регульовану β-галактозидазну репортерну систему на основі штаму *Streptomyces albus* для вивчення особливостей трансляції рідкісних кодонів (зокрема TTA) у генах біосинтезу антибіотиків. За допомогою цієї системи вперше продемонстровано блокування трансляції кодона TTA у мутанті *S. albus* за геном лейцил-тРНК *bldA*. Описано оптимальні моделі кодонних заміщень для функціонально різних груп генів актинобактерій. Вперше створено онлайн-сервіс для візуалізації кодонних заміщень у великих масивах даних на основі "бульбашкових" діаграм.

Особистий внесок здобувача. Результати, викладені у дисертації, автор отримав особисто або за безпосередньої участі у виконанні експериментів. Планування експериментів, аналіз та обговорення отриманих результатів проведені

спільно з науковим керівником, д.б.н. Б.О. Осташем, проф. В.О. Федоренком, к.б.н. М.В. Рабик, аспірантами кафедри генетики та біотехнології ЛНУ ім. І. Франка О. Кошлою та О. Ющуком, з якими автор має спільні публікації.

Апробація результатів дисертації. Результати досліджень репрезентовані на XI-XII Міжнародних конференціях "Молодь і поступ в біології" (Львів, Україна, 2016-2017); на звітних наукових конференціях Львівського національного університету імені І. Франка (2015-2017); міжнародних конференціях "Bacterial Networks" (9-14 вересня, Сан-Феліу-де-Гішульс, Іспанія), "Integrative Biology and Medicine" (2-7 жовтня 2017 р., Київ).

Структура та обсяг дисертації. Дисертація складається з вступу, огляду літератури, матеріалів і методів, результатів досліджень, обговорення результатів досліджень, висновків, списку використаних джерел (102 найменувань) і трьох додатків. Роботу викладено на 129 сторінках машинописного тексту і проілюстровано 39 рисунками та 6 таблицями, а також наведено 26 математичних формул.

Практичне значення отриманих результатів. Полягає у можливості їхнього використання для дослідження генетичних механізмів контролю продукції біологічно активних речовин стрептоміцетами. Отримані результати, сконструйовані штами *E. coli*, стрептоміцетів і рекомбінантні молекули ДНК використовують у навчальному процесі на кафедрі генетики та біотехнології Львівського національного університету імені Івана Франка згідно положення про структурний підрозділ "Колекція культур мікроорганізмів — продуцентів антибіотиків", затвердженого рішенням Вченої ради Львівського національного університету імені Івана Франка (протокол №15/2 від 24.02.2016 р.).

Публікації. Результати дисертації опубліковано в шести статтях у фахових наукових журналах та тезах трьох доповідей на конференціях.

ОСНОВНИЙ ЗМІСТ РОБОТИ

ОГЛЯД ЛІТЕРАТУРИ

В огляді літератури розглянуто сучасні дані щодо розроблених моделей еволюції генетичних послідовностей – нуклеотидних, амінокислотних, та кодонних. Описано вивчені особливості структури генетичного коду стрептоміцетів, зокрема феномен рідкісного лейцинового кодону TTA. Також, висвітлено перспективи використання стрептоміцетів як модельного організму для дослідження особливостей вживання кодонів у геномах.

МАТЕРІАЛИ ТА МЕТОДИ ДОСЛІДЖЕНЬ

У роботі використано штами *Escherichia coli* GB2005, WM6026, а також штами *Streptomyces albus* J1074, SAM2, OK3 та їхні похідні, отримані у цій роботі. В конструюванні кодон-репортерної системи було використано плазміди pGCymRP21, pTES, pRV3, pRV4, pOOB109, pOOB114. Для вирощування штамів актиноміцетів дикого типу і тих, що містять лейцин-специфічні кодонні репортери, використовували рідкі та агаризовані живильні середовища TSB, LB, BC.

Трансформацію та електропорацію клітин *E. coli* здійснювали згідно стандартних методик (Green та Sambrook 2012). Кон'югаційні схрещування

Escherichia-Streptomyces виконували, як описано (Ha et al. 2008; Kaiser 2000). Виділення та аналіз сумарної та плазмідної ДНК, ферментативну обробку ДНК, ампліфікацію генів за допомогою ПЛР та їх субклонування здійснювали за стандартними методиками (Green та Sambrook 2012).

Для біоінформатичних аналізів використовували BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), IQ-Tree Web Service (Trifinopoulos et al. 2016), phylogeny.fr (Dereeper et al. 2008), PhyloPhLan (Segata et al. 2013).

РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ ТА ЇХНЄ ОБГОВОРЕННЯ

Дослідження закономірностей вживання кодонів у геномах стрептоміцетів. Закономірності вживання кодонів у геномах бактерій є питанням важливим та недостатньо вивченим. Існує багато теорій щодо механізмів виникнення ПВК та пояснень їхньої ролі в певних процесах, але частина з них має суперечливий характер. У цій роботі для вивчення проблеми кодонного складу використано геноми бактерій роду *Streptomyces*. Їхні геноми досі не вивчались у цьому керунку, хоча містять виразні зміщення в частотах вживання кодонів і тому мають бути доброю моделлю для таких досліджень. Промислова цінність стрептоміцетів також робить дослідження кодонного складу цікавим у прикладному вимірі.

Відібрано три гени *Streptomyces coelicolor* (*sco*), продукти яких суттєво відрізняються за своєю функцією в клітині. Ген *sco1728* кодує транскрипційний фактор YtrA-типу з родини GntR. Цей ген і білок репрезентує родину транскрипційних факторів, які взаємодіють з ДНК. Продукт гена *sco2706* – глікозилтрансфераза, схожа до низки інших ферментів задіяних в процесі біосинтезу ліпополісахаридів; цей ген містить кодон TTA і відтак підлягає контролю з боку тРНК *bldA*. Ген *sco2706* репрезентує ензими. Нарешті, обрано ген *sco3894*, що репрезентує трансмембранні білки, оскільки кодує імовірну фліппазу (експортер) ліпід-вмісних попередників пептидоглікану бактерій. Для кожного з обраних генів створено вибірку ортологічних амінокислотних та нуклеотидних послідовностей (Kuzniar 2008). Зібрані дані ми використали для визначення оптимальної моделі еволюції.

Ортологам транскрипційного регулятора (*sco1728*) і глікозилтрансферази (*sco2706*) притаманна модель K3Pu, що враховує три параметри в еволюції послідовностей: один параметр для опису швидкості транзицій та два параметри — для швидкості трансверсій. Групі ортологічних генів трансмембранного білка (*sco3894*) відповідає модель GTR. Вона використовує різні частоти нуклеотидів (4 параметри), і різні частоти замін між нуклеотидами (6 параметрів). Дані, отримані на одній групі генів, далі підтверджено при аналізі більших вибірок (див. дисертацію).

Підбір оптимальних моделей еволюції амінокислотних послідовностей показав, що кожній групі білків відповідає своя еволюційна модель. Еволюцію ортологів Sco1728 (транскрипційний регулятор родини GntR) найкраще описує матриця JTT. Еволюція групи ортологів білка глікозилтрансферази Sco2706 найточніше описується матрицею WAG. Оптимальною моделлю для ортологів трансмембранного білка Sco3894 є матриця LG. Виявлено, що вибір матриці

заміщення має вплив на топологію філогенетичного дерева (рис. 1). Це підкреслює важливість пошуку оптимальних моделей заміщення.

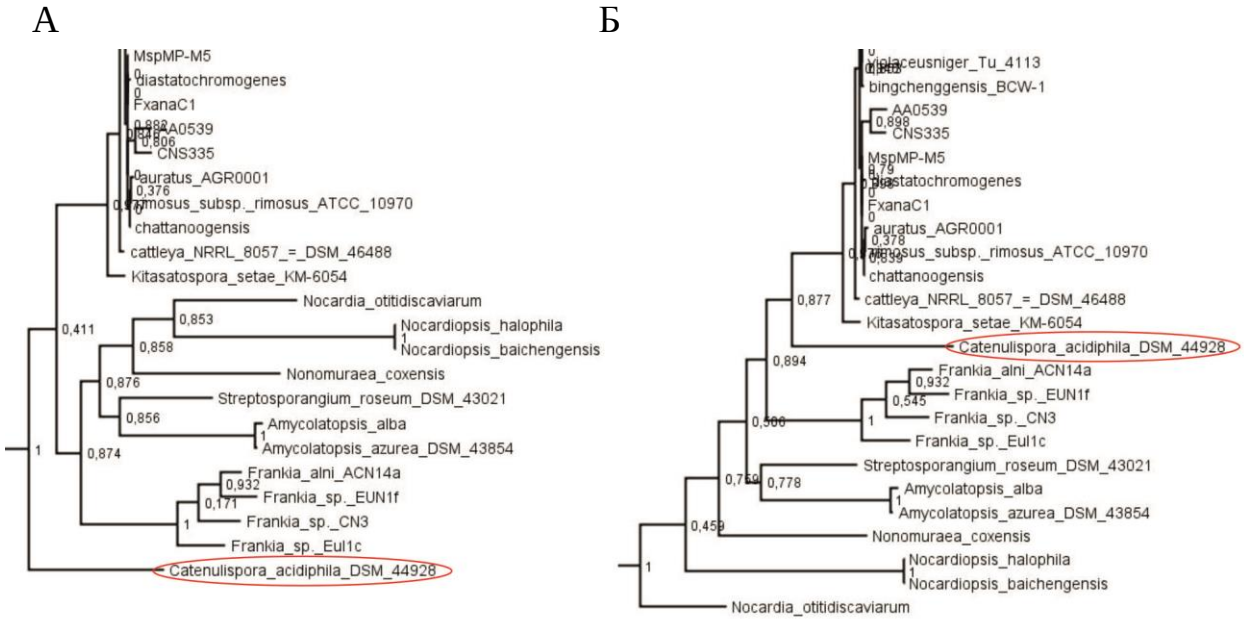

Рис. 1. Філогенетичні дерева на основі вирівнювання набору ортологічних білків родини AdpA, побудовані з використанням матриць JTT (А) та WAG (Б). Червоним овалом відмічено кладу родини *Catenulispora acidiphila*, яка займає відмінну позицію в двох деревах. Цифри на нодах дерева – коефіцієнт надійності топології (1 = 100% імовірність, що усі гілки клади будуть разом на філогенетичному дереві, незалежно від параметрів реконструкції).

Нині доступний великий масив геномних даних, який дає змогу досліджувати “горизонтальні” та “вертикальні” закономірності вживання кодонів у стрептоміцетів. Певні кодони знаходяться поруч один з одним з імовірністю, меншою або більшою, ніж очікується, якщо б ці кодонні пари формувалися випадково (відповідно до фонових частот вживання нуклеотидів у геномах). Очікувані й фактичні частоти “горизонтального” вживання пар кодонів можна обчислити й оцінити статистично. Спеціалізоване програмне забезпечення Anaconda (Moura 2007) дає змогу проаналізувати секвенований й анотований геном та підрахувати кількість усіх можливих кодонних пар. Позитивний контекст матимуть дикодони, частота зустрічності яких перевищує два стандартні відхилення від середнього значення у нормальному розподілі. Пари кодонів, які зустрічаються з частотою меншою ніж статистична випадковість, відповідно матимуть “негативний” контекст. Ми виконали такий аналіз для 50 найкраще досліджених та анотованих геномів *Streptomyces*. Узагальнивши дані, ми виявили дев’ять позитивних дикодонних асоціацій: UAU-CUG, CUG-CGC, GUA-CGG, GAU-CCG, CUC-ACC, CUC-GCC, CUC-GGC, GAA-CUC, GAA-CUG. Також виявлено дві негативні асоціації: CUC-CUG, CUC-GAG (рис 2.).

Відповідно, отримані результати можна узагальнити у наступні паттерни: «позитивні» — KAU-CYG (A та D рис. 2), GWA-CKS (C та F рис. 2), CUS-VSC (B та E рис. 2), та «негативний» — CUC-SWG (E рис. 2). Цікаво, що частина цих

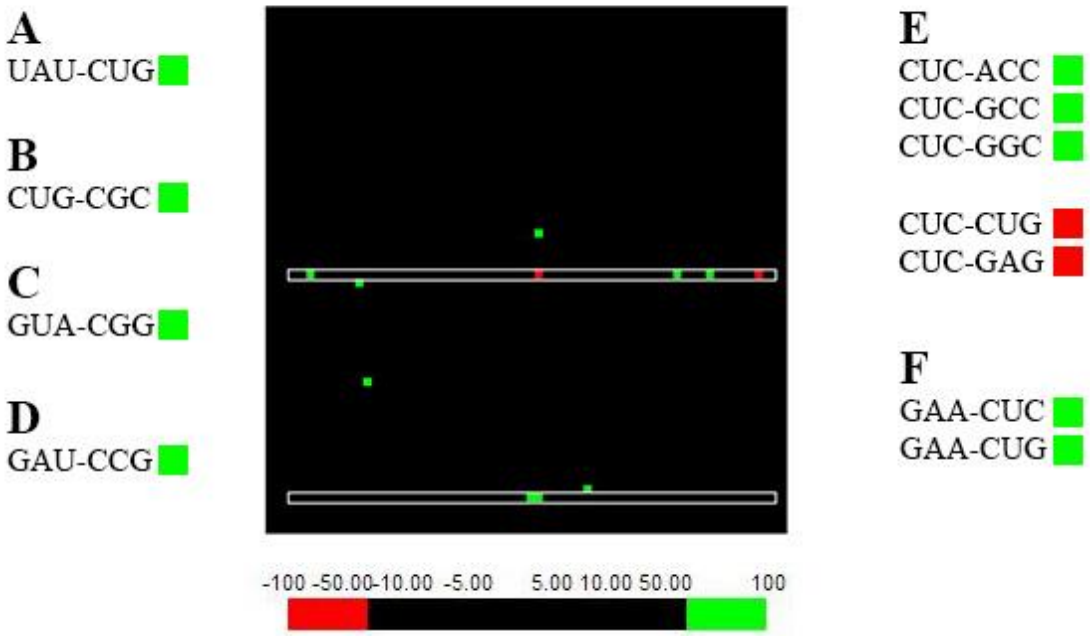

Рис. 2. Усереднена теплова карта 50 геномів *Streptomyces*. Для ідентифікації конкретних контекстів дикодонів карта була відфільтрована для відображення кодонів, кількість яких перевищує 50. Червоний −«негативний» контекст та зелений - «позитивний» контекст. Всі інші випадки (відсутність контексту) забарвлено в чорний колір. A, B, C, D − одиничні пікселі (з верху до низу), E і F−виділені рядки карти, відповідно.

паттернів описує залежність стосовно лейцинових кодонів — CUS-VSC та CUC-SWG. Також, виявлені залежності переважно описують асоціацію між U/T у першому кодоні, та G/C – в другому. Такі патерни показують, що в дикодонах не можна виділити однозначний вплив нуклеотида в певній позиції. Втім, наявність патернів свідчить, що взаємне розташування нуклеотидів вздовж послідовності відіграє певну роль у загальній картині вживання кодонів, а контекстне вживання кодонів слід враховувати поряд із ПВК.

Стосовно кодона TTA нам не вдалося виявити жодного відхилення у вживанні від випадкового (статистичного очікуваного; див. вище). Це може означати, що кодон TTA справді не має контекстних особливостей вживання. Однак, варто мати на увазі й обмеження нашого підходу. Зокрема, кодон TTA складає найменшу групу, і в таких групах складно виявляти значні відхилення. У цьому дослідженні ми визначали залежність вживання кодона від сусіднього, в той час як може існувати контекст від кількох сусідніх кодонів (Chevance et al. 2017).

“Вертикальні” кодонні заміщення ми досліджували, порівнюючи обраний ген та його ортологи з низки видів *Streptomyces*. Це дозволить вивчити як еволюціонував геном стрептоміцетів на кодонному рівні. Для цього необхідно змоделювати та візуалізувати закономірності кодонних заміщень у масиві генетичних послідовностей. Відтак необхідно мати прості й доступні інструменти аналізу кодонних заміщень, зокрема веб-спрямовані застосунки. Ми зупинилися на програмному пакеті DART, зокрема програмі Xrate (Klosterman et al. 2008). Ця програма дає змогу будувати кодонні моделі, застосувавши алгоритм Очікування-Максимізації (EM) для тренування вихідної моделі M0 (Holmes and Rubin 2002). Ми створили веб-застосунок обчислення кодонних моделей на основі Xrate. Використовуючи застосунок, ми отримуємо “бульбашкові” графіки кодонних заміщень для різних ортологічних груп генів (рис. 3).

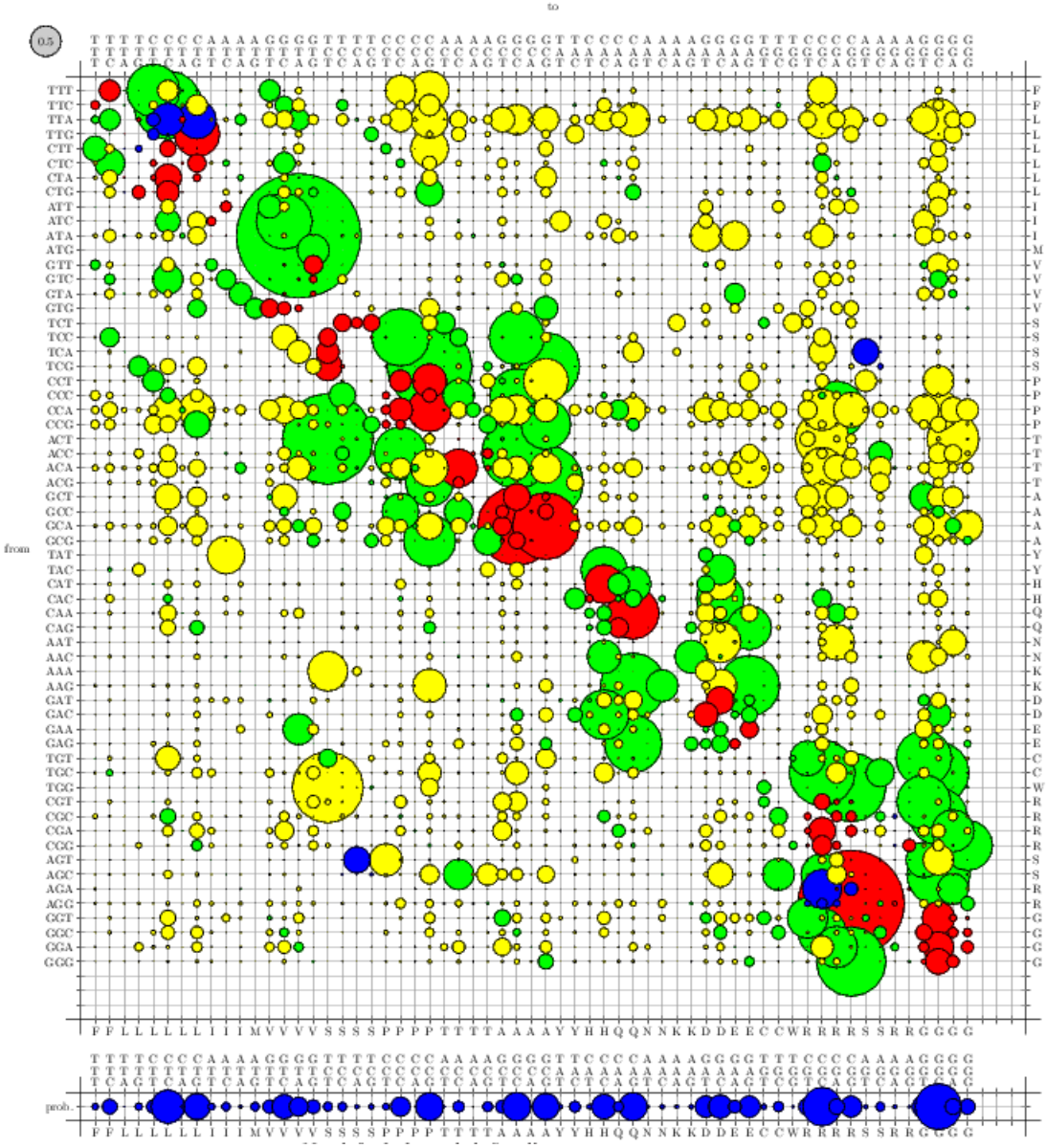

Рис. 3. Заміщення кодонів в генах-ортологах транскрипційного фактора (*sco1728*). Зелений — несинонімічні заміщення кодонів з різницею в 1 нк. Червоний — синонімічні заміщення з різницею в 1 нк. Жовтий — несинонімічні заміщення з

різницею більше ніж 1 нк. Синій — синонімічні заміщення з різницею більше ніж 1 нк. Шкала знизу показує сумарну імовірність заміщення кодонів.

Діаметр бульбашки і її колір вказує на частоту і тип заміщення, відповідно. Різні класи ортологічних генів ведуть до якісно відмінних бульбашкових діаграм (рис. 4).

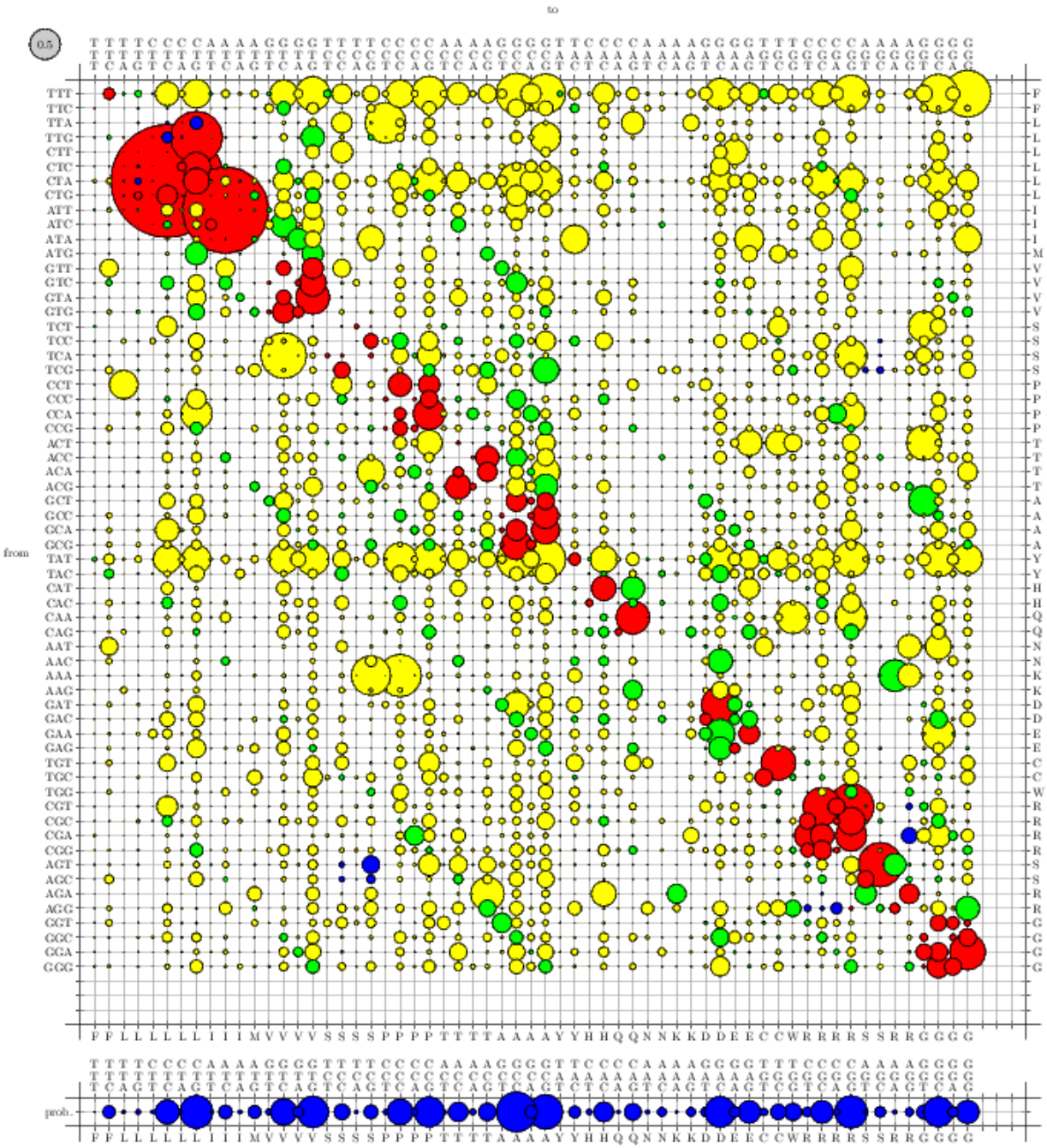

Рис. 4. Графік частоти заміщень кодонів в ортологічних генах глікозилтрансферази (*sco2706*). Кольоровий код – див. рис. 3.

Застосувавши цей підхід до стрептоміцетних генів, ми спостерігали відмінність в паттернах для білків з різною функцією. Зокрема, представлені транскрипційний фактор — *sco1728* та фермент глікозилтрансфераза — *sco2706* (рис. 3, 4, див. також дисертацію, де наведено більше прикладів). Ці відмінності корелюють із фізико-хімічними властивостями білка (цитозольні чи мембранні) і його належністю до

первинного чи вторинного метаболізму. Також, відрізняється частота синонімічних та несинонімічних заміщень в генах різних білків.

Підсумовуючи отримані результати натренованих моделей кодоних заміщень в межах роду *Streptomyces*, ми можемо виділити наступні закономірності еволюції для різних груп кодувальних послідовностей. У всіх трьох групах спостерігається заміщення кодонів, які багаті на A/T пари, на кодони з G/C парами, що суперечить тенденції до збагачення геномів бактерій A/T парами. Можемо зробити наступне припущення: високий GC-склад геномної ДНК стрептоміцетів – це наслідок кодоних заміщень під дією еволюційних сил. Тобто, існує форма рушійного добору на кодонному рівні, яка протидіє мутаційним процесам, що збагачують геном на AT-пари (Hershberg and Petrov 2010).

Виникає питання про селективні переваги у використанні GC-багатих кодонів. У стрептоміцетів велика кількість довгих генів (більше 3000 п.н.), що кодують вторинні метаболіти. Переважання GC-пар в таких генах відповідатиме низькій частоті виникнення A/T-багатих стоп-кодонів (TAG, TGA та TAA). Чим вищий GC-склад геному, тим сприятливші умови до появи довгих генів. Добір на кодонному рівні також може бути спрямований на уникання зайвих сайтів зв'язування рибосоми (RBS). Так, імовірність появи A/T-багатих сайтів RBS в непризначеному місці менша в генах з переважанням GC- багатих кодонів. Це в свою чергу несе перевагу в еволюційному розумінні у вигляді нижчої частоти помилкових транскриптів, так мінімізуючи марнування ресурсів клітини (Li et al. 2012). Відповідно, можемо поставити під сумнів постулат, що зсув вживання кодонів – наслідок високого GC-складу стрептоміцетів. Не виключено, що високий GC-склад постав унаслідок природного добору GC-багатих кодонів. Таке тлумачення появи GC-багатих геномів стрептоміцетів не суперечить присутності мутаційного процесу в бік AT-пар. Ми припускаємо лише, що мутаційний процес і природний добір чинять тиск на кодувальну послідовність у протилежних керунках, що в сумі зумовлює збагачення кодувальних послідовностей GC-багатими кодонами.

Дослідження рідкісного лейцинового кодона TTA у стрептоміцетів. Наявність природного екстремального зміщення у вживанні лейцинового кодону TTA – одна із причин вибору стрептоміцетів як об'єкта досліджень. Цей кодон впізнається лейциновою тРНК, що кодується геном *bldA*. При делеції цього гена порушується морфогенез та вторинний метаболізм бактерії. Отже, можна говорити про певну регулювальну функцію гена *bldA*, опосередковану кодоном TTA. Таке припущення передбачає, що трансляція рідкісного кодона має відбуватися з високою точністю. З іншого боку, низка досліджень (Clarke and Clark 2008; Doma and Parker 2007) вказує на те, що рідкісні кодони зазвичай частіше містранслюються ніж популярні (ті, що вживаються в генах з високою частотою). Отже, TTA має бути рідкісним, аби контролювати лише певні гени, має транслюватися лише однією тРНК і водночас залишатися точним. Ми провели низку досліджень *in silico* щоб краще зрозуміти декодування цього кодону. Зокрема, визначили й проаналізували набір генів тРНК шести стрептоміцетних геномів: *S. coelicolor* M145, *S. albus* J1074, *S. ghanaensis* ATCC14672, *S. clavuligerus* ATCC27076, *S. venezuelae* ATCC14115 та *S. lividans* TK24. В отриманих вибірках ми підрахували кількість копій генів тРНК, як

показник, що корелює із концентрацією тРНК в клітині (dos Reis 2004). Появу помилок при трансляції можна розглядати як конкуренцію між акцепторними та неакцепторними тРНК за розпізнавання кодона. А отже, імовірність містрансляції залежить від відношення концентрації фокальної тРНК в клітині до сукупної концентрації близько-споріднених тРНК (відрізняються від акцепторної тРНК одним нуклеотидом, і тому найімовірніше вестимуть до містрансляції). Застосувано математичну модель Шаха (Shah and Gilchrist 2010), що визначає точність трансляції через співвідношення споріднених (фокальних) та близько-споріднених тРНК (*tF*/*tN*).

Кодон	S. coelicolor	S. albus	S. venezuelae	S. lividans	S. ghananensis	S. clavuligerus
UUA	0.006110	0.006110	0.006110	0.006110	0.006110	0.006110
UUG	0.013720	0.014980	0.012460	0.013720	0.013720	0.013720
CUA	0.007370	0.008630	0.009890	0.007370	0.008630	0.008630
CUG	0.018690	0.019950	0.018690	0.018690	0.018690	0.018690
CUC	0.013600	0.012400	0.012400	0.011080	0.012400	0.013660
CUU	-	-	-	-	-	-

Рис. 5. Теплова карта швидкостей елонгації лейцинових кодонів за допомогою близько-споріднених тРНК. Шкала від зеленого до червоного відповідає величині показника (від меншого до більшого, відповідно).

Результати розрахунків для елонгації лейцинових кодонів за допомогою близько-споріднених тРНК показали (рис. 5), що рідкісний кодон TTA має низькі показники швидкості декодування неакцепторними тРНК і відповідно має меншу імовірність містранслюватися ніж інші лейцинові кодони. Отже, кодон може бути рідкісним, йому може відповідати низька кількість копій генів (ККГ), і все ж він може транслюватися не менш точно, ніж популярні кодони (Rokytskyy et al. 2016). Такий результат узгоджується із регуляторною функцією цього кодона.

Далі наше дослідження стосувалося вивчення факторів, що модулюють експресію рідкісного лейцинового кодона TTA у стрептоміцетів. Кодон TTA зустрічається лише в генах вторинного метаболізму, що задіяні у складних каскадах реакцій. Фенотиповий прояв змін в таких генах — це результат впливу багатьох факторів. Важливо сконструювати систему з найкоротшим шляхом від кодону до фенотипу, що, в ідеалі, постає унаслідок експресії одного гена. Обраний ген має бути таким, щоб можна було легко виявити активність його білкового продукту і очистити останній. Це, в свою чергу, дасть змогу кількісно судити про вплив рідкісного кодону на трансляцію безпосередньо. Також, помістивши таку конструкцію в штам з делетованим геном *bldA*, матимемо змогу вивчити вплив різних генетичних та середовищних факторів на (міс)трансляцію TTA. Ми

сконструювали TTA-кодон-специфічну репортерну систему з можливістю якісного та кількісного аналізу активності репортерного білка. Система базується на гені β-галактозидази *sco3479* (*lacZ*$_{sc}$) зі штаму *Streptomyces coelicolor*, продукт якого, білок Sco3479, може розщеплювати безбарвний аналог лактози — X-Gal – з утворенням кольорової сполуки, індиго синього (Shuman and Silhavy 2003). Другим елементом репортерної системи є штам *Streptomyces albus* J1074, що не гідролізує X-Gal (King та Chater 1986).

Для перевірки репортерних властивостей Sco3479 сконструйовано низку плазмід на основі вектора pTES: pOOB109, pOOB110 та pOOB114 (остання – нефункціональний ген *sco3479* , що містить стоп-кодон на початку відкритої рамки зчитування). Плазміди надають штамові *S. albus* здатності розщеплювати X-Gal в середовищі, за рахунок експресії гена *lacZ*$_{sc}$, з утворенням синього продукту. Контрольний штам J1074-pOOB114, як очікувалось, залишався безбарвним. Ген *sco3479* містить низку Leu кодонів на початку гена, які можна замінити на TTA. Ми замінили кодон CTC в 8-ій позиції TTA кодоном та внесли шість гістидинових кодонів CAC перед стоп-кодоном. Ген клоновано за допомогою вектора pGCymRP21, який містить кумат-залежну систему контрольованої експресії клонованих генів. А саме, за відсутності індуктора, кумінової кислоти (кумату), транскрипція цільового гена пригнічується репресорним білком CymR (Horbal et al. 2014). Кумат, доданий в середовище, зв'язуватиметься з репресором CymR, що вестиме до вивільнення оператора *cmt*. Зняття репресії з оператора запустить експресію цільового гена.

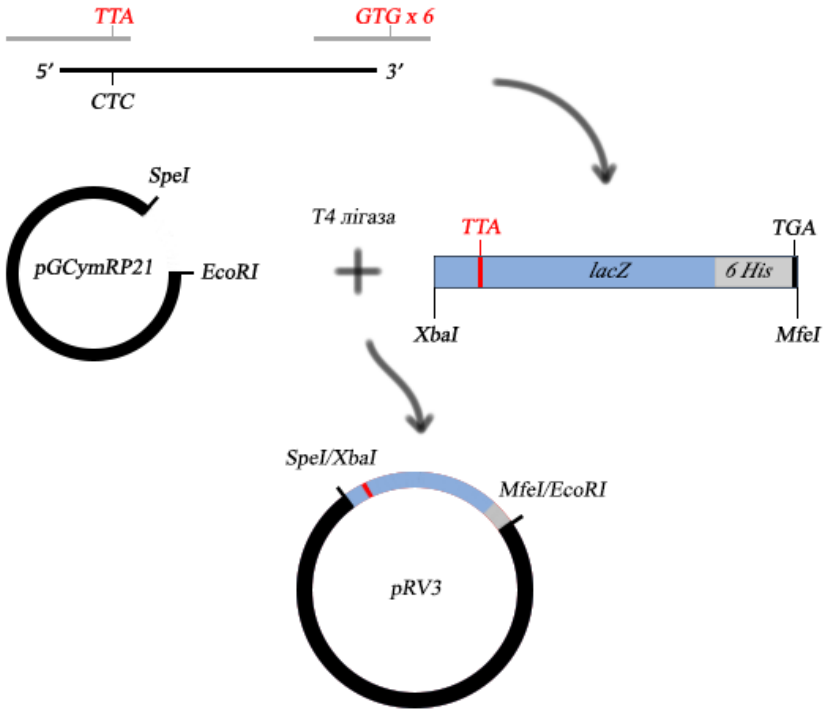

Рис. 6. Конструювання фрагменту *sco3479 (lacZ)* із TTA кодоном та 6His-тагом і його вбудовування у вектор pGCymRP21.

В результаті ми отримали плазміду pRV3 (рис. 6) з геном *sco3479*, що містить лейциновий TTA кодон та 6-His таг та плазміду pRV4 з геном *sco3479* дикого типу,

що має 6-His таг. Ми довели коректність нуклеотидної послідовності генів *sco3479* у плазмідах pRV3 й pRV4 за допомогою секвенування. Плазміди pRV3 та pRV4 перенесли в *S. albus* SAM2. Сконструйовані штами перевірили методом ПЛР та шляхом якісних реакцій на індикаторних середовищах (рис. 7).

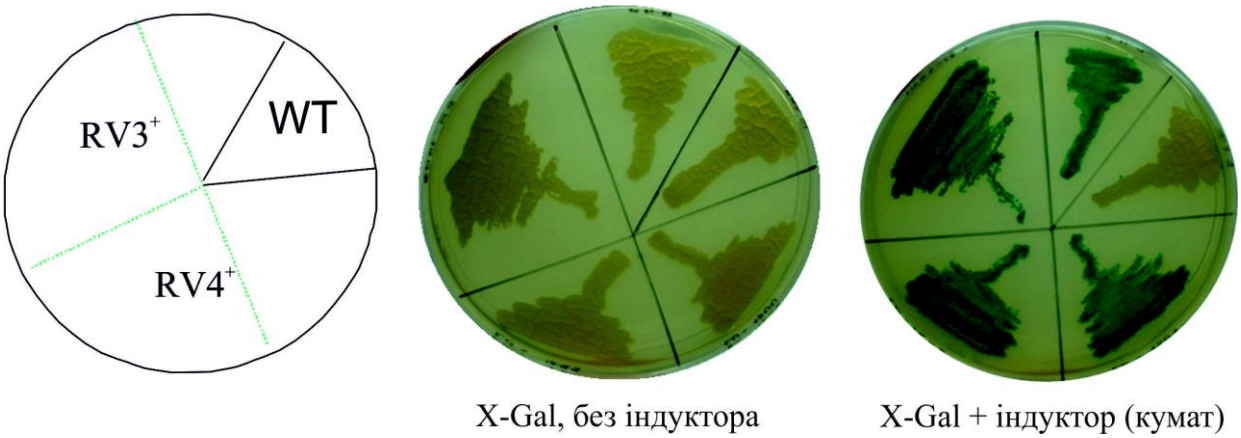

Рис. 7. Ріст *S. albus* J1074 із внесеними плазмідами pRV4 (RV$^+$), pRV3 (RV3$^+$) та дикого типу (WT), на X-Gal-вмісному середовищі з куматом або без нього.

Отримана репортерна система дає змогу виконати щонайменше якісну оцінку фенотипу – забарвлення міцелію свідчитиме про трансляцію чи містрансляцію. Відсутність синього кольору слугуватиме непрямим доказом блокування експресії репортера на рівні трансляції. Нами виконано експерименти, які підтвердили належне функціонування створеної репортерної системи у дикому типі *S. albus* та його похідному з делецією гена *bldA* — *S. albus* ОК3, що кодує $tRNA^{Leu}_{UAA}$. (рис. 8). Зокрема, показано, що відсутність гена *bldA* блокує експресію TTA-вмісного репортерногог гена *lacZ*, але не його CTC-вмісного варіанту. Отже, кодон TTA нездатний ефективно транслюватися за відсутності акцепторної тРНК.

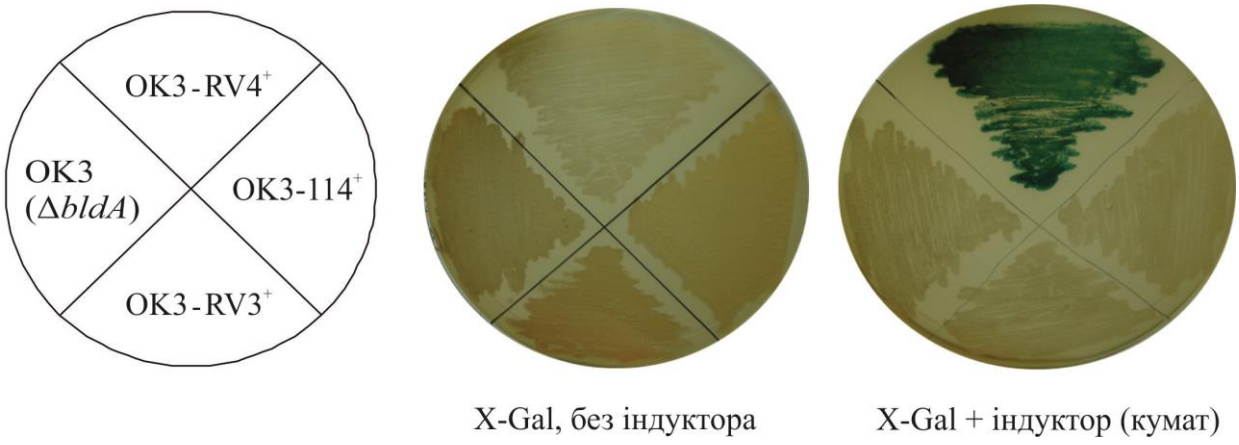

Рис. 8. Газони штамів *S. albus* , що містять плазміди pRV4 (OK3-RV4$^+$), pRV3 (OK3-RV3$^+$), pOOB114 (OK3-114$^+$) та дикого типу (OK3), на TSA з X-Gal (30 мМ), без (ліворуч) та із додаванням (праворуч) кумату.

Як приклад застосування нашої системи, нами отримано перші докази, що вказують на містрансляцію TTA-вмісних генів у *S. albus* (рис. 9). А саме, за тривалого інкубування з'являються синьо забарвлені колонії, виключно за умови індукції транскрипції гена *sco3479*. Низка контрольних експериментів виключили можливість того, що кумат стимулює деградацію хромогенного субстрату та інші альтернативні пояснення отриманих результатів.

а *б*

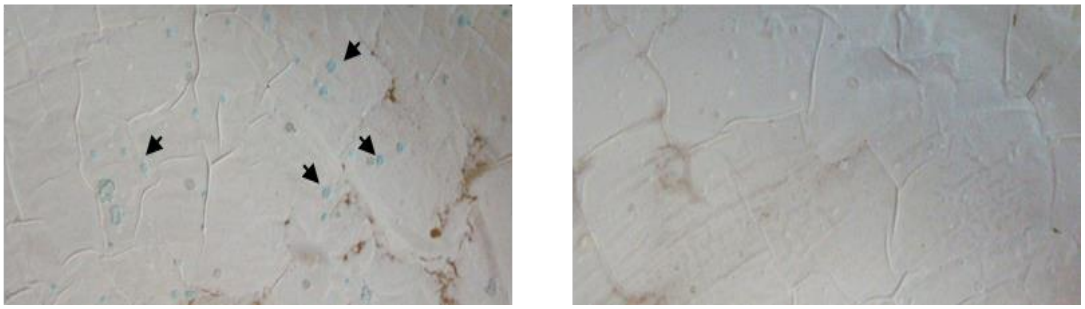

Рис. 9. Фотографія газонів штамів *S. albus* ΔbldA (ОКЗ) RV3^{+} на середовищі TSA з апраміцином, X-Gal (100 мМ) й куматом (*а*), або без нього (*б*). Стрілками позначено кілька репрезентативних колоній *S. albus*, що здатні розщеплювати хромогенний субстрат.

Отже, на основі гена β-галактозидази *sco3479* нами опрацьовано і протестовано просту репортерну систему, де порушення трансляції кодона TTA прямо ведуть до фенотипового вияву – зміни забарвлення міцелію. Усі отримані дані показують, що система функціонує як очікувано, і подальше її використання дасть змогу відповісти на низку питань про особливості трансляції кодона TTA у геномах стрептоміцетів.

ВИСНОВКИ

У результаті виконання дисертаційної роботи визначено оптимальні моделі заміщення нуклеотидних та амінокислотних залишків у масивах послідовностей, що походять з геномів *Streptomyces*; виявлено невипадкову асоціацію низки кодонів в цих геномах. Опрацьовано новий веб-орієнтований застосунок для візуалізації моделей кодонних заміщень. Продемонстровано новий підхід до біоінформатичного передбачення рівня містрансляції кодонів, а також нову експериментальну модель вивчення експресії рідкісного кодона TTA у стрептоміцетах.

1. Створено низку масивів ортологічних послідовностей, що відображають функціонально різні класи генів та білків з вторинного метаболізму бактерій роду *Streptomyces*. Це послідовності, що кодують один ензим, сім транскрипційних факторів та один мембранний білок.
2. Визначено оптимальні моделі заміщення для різних кодувальних послідовностей, а також їхніх продуктів трансляції. Показано, що групі функціонально-подібних послідовностей притаманні здебільшого одинакові моделі. В основному це моделі K3Pu та TVM для нуклеотидних

послідовностей стрептоміцетів. Для кожної із досліджених амінокислотних послідовностей притаманна своя оптимальна модель.

3. Кодувальним послідовностям з геномів *Streptomyces* притаманні дев'ять позитивних дикодонних асоціацій – таких, що зустрічаються з частотою вищою за випадкову: UAU-CUG, CUG-CGC, GUA-CGG, GAU-CCG, CUC-ACC, CUC-GCC, CUC-GGC, GAA-CUC, GAA-CUG. Також виявлено дві негативні асоціації: CUC-CUG, CUC-GAG. Спільною рисою усіх асоціацій є те, що вони характерні переважно лейциновим кодонам, із нуклеотидами A/T у центральній позиції кодона.
4. Створено веб-орієнтований застосунок візуалізації кодонних заміщень у масивах кодувальних послідовностей у вигляді "бульбашкового" графіка. Описано основні підходи до тлумачення такого графіка; виявлено тенденцію до спрямованого збагачення стрептоміцетних генів GC-багатими кодонами.
5. Створено та апробовано репортерну систему на основі гена β-галактозидази *sco3479* та штаму *Streptomyces albus* J1074, що дає змогу прямо вивчати вплив різних мутацій чи факторів середовища на експресію рідкісного кодона TTA.

ПЕРЕЛІК НАУКОВИХ ПРАЦЬ, ОПУБЛІКОВАНИХ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Статті

1. Gren, T., Ostash, B., Babiy, V., **Rokytskyy, I.** and Fedorenko, V. 2018. "Analysis of *Streptomyces coelicolor* M145 genes *sco4164* and *sco5854* encoding putative rhodaneses." *Folia Microbiol* 63(2):197-201 doi:10.1007/s12223-017-0551-6. *Особистий внесок здобувача – проведення філогенетичного аналізу, опис отриманих результатів аналізу, участь в обговорені результатів.*
2. Koshla, O., Lopatniuk, M., **Rokytskyy, I.**, Yushchuk, O., Dacyuk, Y., Fedorenko, V., Luzhetskyy, A. and Ostash, B. 2017. "Properties of *Streptomyces albus* J1074 mutant deficient in $tRNA^{Leu}_{UAA}$ gene *bldA*." *Arch Microbiol* 199(8):1175-1183. doi: 10.1007/s00203-017-1389-7. *Особистий внесок здобувача – конструювання плазмід для TTA-специфічної репортерної системи, опис методології, обговорення результатів.*
3. Rabyk, M., Yushchuk, O., **Rokytskyy, I.**, Anisimova, M. and Ostash, B. 2018. "Genomic Insights into Evolution of AdpA Family Master Regulators of Morphological Differentiation and Secondary Metabolism in *Streptomyces*." *Journal of Mol. Evol.* 86:166–178. doi:10.1007/s00239-018-9834-z. *Особистий внесок здобувача – філогенетичний аналіз різних родів актинобактерій, аналіз топології філогенетичних дерев, опис та обговорення отриманих результатів.*
4. **Rokytskyy, I.**, Koshla, O., Fedorenko, V. and Ostash, B. 2016. "Decoding options and accuracy of translation of developmentally regulated UUA codon in *Streptomyces*: bioinformatics analysis." *SpringerPlus* 5:982. doi:10.1186/s40064-016-2683-6. *Особистий внесок здобувача – підрахунок кількості копій генів, що кодують мРНК, швидкості елонгації та статистичний аналіз достовірності даних, опис методології, опис результатів та їх обговорення.*

5. **Rokytskyy, I.**, Kulaha, S., Mutenko, H., Rabyk, M. and Ostash, B. 2017. “Peculiarities of codon context and substitution within streptomycete genomes.” *Вісник Львів. Ун-ту. Сер.біол.* 75:66-74. *Особистий внесок здобувача – визначення контекстних залежностей кодонів в складі стрептоміцетних геномів, робота з програмою Anaconda та опис отриманих результатів.*
6. **Rokytskyy, I.** and Ostash, B. 2016. “Optimal models of nucleotide and aminoacid substitution for sequences derived from actinobacterial genera.” *Вісник Львів. Ун-ту. Сер.біол.* 72:75-81. *Особистий внесок здобувача – створення вибірок генетичних послідовностей стрептоміцетних генів та визначення оптимальних моделей еволюції, опис методології, опис результатів та їх обговорення.*

Тези

7. **Рокицький І.**, Кошла О. 19-21 квітня 2016 “Декодування та точність трансляції лейцинового кодону TTA у стрептоміцетів: аналіз *in silico” XII Міжна. Наук. Конф. "Молодь і поступ біології"* Львів, Україна. Тез.доп. – С. 134.
8. **Рокицький І.**, Кошла О. 25-27 квітня 2017 “Методи дослідження вживання кодонів в геномах *Streptomyces” XIII Міжна. Наук. Конф. "Молодь і поступ біології"* Львів, Україна. Тез.доп. – С. 116.
9. Oksana Koshla, **Ihor Rokytskyy**, Julia Sehin, Leif A. Kirsebom, Andriy Luzhetskyy and Bohdan Ostash. 9-14 september 2017 “Switch of the switch? Posttranscriptional tRNA modifications as regulators of Streptomyces biology” “Bacterial Networks” Sant Feliu de Guixols, Spain. Thesis – P. 59.

АНОТАЦІЯ

Рокицький І.В. Біоінформатичні підходи та репортерна система для дослідження особливостей вживання кодонів у геномах стрептоміцетів. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата біологічних наук за спеціальністю 03.00.22 – молекулярна генетика – ДУ «Інститут харчової біотехнології та геноміки НАН України», Київ, 2019.

У дисертації розглянуто, які оптимальні моделі еволюції характерні для різних груп білків, та питання, що стосуються використання кодонів у геномах стрептоміцетів. Це, в свою чергу, включає вивчення контекстної залежності у вживання кодонів та моделі еволюційного заміщення кодонів. Щодо останнього, запропоновано новий веб-сервіс для візуалізації кодонних заміщень — cTools (http://biotools.online). Також, обчислено точність декодування рідкісного у стрептоміцетів лейцинового кодона TTA. У ході виконання дисертації також створено TTA-специфічну кумат-індуцибельну репортерну системи на основі гена β-галактозидази — *sco3479* (*lacZ*$_{sc}$) та наведено приклади використання такої системи для вивчення експресії TTA-вмісних генів.

Ключові слова: геноміка, кодонний склад, переважне вживання кодонів, еволюційні моделі, репортерна система, актинобактерії, *Streptomyces*.

АННОТАЦИЯ

Рокицький И.В. Биоинформатические подходы и репортерная система для исследования особенностей употребления кодонов в геномах стрептомицетов. - Квалификационная научная работа на правах рукописи.

Диссертация на соискание научной степени кандидата биологических наук по специальности 03.00.22 – молекулярная генетика – ГУ «Институт пищевой биотехнологии и геномики НАН Украины», Киев, 2019.

В диссертации рассмотрено, какие оптимальные модели эволюции характерны для разных групп белков, и вопросы, касающиеся использования кодонов в геномах стрептомицетов. Это, в свою очередь, включает изучение контекстной зависимости в употреблении кодонов и модели эволюционного замещения кодонов. Что касается последнего, предложен новый веб-сервис для визуализации кодонних замещений - cTools (http://biotools.online). Также, вычислена точность декодирования редкого у стрептомицетов лейцинового кодона TTA.

Также, в диссертации описано конструирование TTA-специфической кумат-индуцибельной репортерной системы на основе гена β-галактозидазы - *sco3479* (*lacZ*$_{sc}$) и пример использования такой системы для изучения експерсии.

Ключевые слова: геномика, кодонний состав, эволюционные модели, репортерного система, актинобактерии Streptomyces.

SUMMARY

Rokytskyy I. Bioinformatics approaches and a reporter system for studying the peculiarities of codon usage in streptomyces genomes. – Qualifying scientific work, manuscript.

Thesis for the degree of Candidate of Biological Sciences on a specialty 03.00.22 – molecular genetics. – Institute of Food Biotechnology and Genomics of the National Academy of Sciences of Ukraine, Kyiv, 2019.

The patterns of codon usage in the genomes of bacteria are an important and insufficiently studied topic. In this dissertation, genomes of bacteria of the genus *Streptomyces* are used to study the codon composition problem.

First of all, optimal evolutionary models of genetic sequences were considered and the "horizontal" and "vertical" patterns of codon usage in streptomycetes were investigated. Certain codons are adjacent to each other with a probability lower or higher than expectations. Frequencies of the "horizontal" usage of the codon pairs can be determined and estimated statistically and presented as codon contexts. Positive context means that dicodon occurs at a frequency exceeding two standard deviations from the mean random value in normal distribution. A pair of codons encountered at a frequency less than a statistical incident will respectively have a "negative" context. Investigated and summarized data from 50 streptomycetic genomes shows nine positive dicodone associations: UAU-CUG, CUG-CGC, GUA-CGG, GAU-CCG, CUC-ACC, CUC-GCC, CUC-GGC, GAA-CUC, GAA-CUG, and two negative associations: CUC-CUG, CUC-GAG. The "vertical" codon substitutions was investigated by comparing the selected gene

and its orthologs from a number of *Streptomyces*. This will allow to study how the streptomycetes genome has evolved at the codon level. To do this, it is necessary to simulate and visualize the codon substitution patterns in an array of genetic sequences. In that purpose a brand new web service is offered - cTools (http://biotools.online). It can visualize codon substitutions in form of bubble plot.

The presence of a natural extreme bias in the usage of the leucine codon TTA is one of the initial reasons for us to choose *Streptomyces* as an object of research. This codon is recognized by the leucine tRNA encoded by *bldA* gene. The *bldA* deficiency leads to morphogenesis and secondary metabolism defects. It is assumed that the translation of rare codons should occur with high accuracy; otherwise it would not function as a genetic switch. Studies of the decoding features of TTA codon are based on correlation between number of copies of the tRNA genes and concentration of tRNA in the cell. Errors in translation can be considered as a competition between acceptor and non-acceptor tRNAs for codon recognition. The probability of translation depends on the ratio of concentration of focal tRNA in the cell to the total concentration of closely related tRNAs. The results of calculations for the elongation of leucine codons with the help of closely related tRNAs showed that the rare codon TTA has low rates of decoding with the near-cognate tRNA and, subsequently, has less chance of being mistranslated than other leucine codons.

Codon TTA is found only in genes responsible for secondary metabolism, which are involved in complex cascades of reactions. The phenotypic extent of changes caused by such genes is a result of interplay of numerous factors. It is important to design a system with the shortest path from the codon to the phenotype, which, ideally, arises as a result of the expression of one reporter gene. The reporter should be easily detected by the activity of its protein product. This, in turn, will allow directly quantify the impact of the rare codon on translation. Also, by placing such a construct in a strain with a deleted *bldA* gene, we will be able to study the effects of various genetic and environmental factors on the translation of TTA. We constructed a TTA-codon-specific reporter system with the possibility of qualitative and quantitative analysis of the activity of reporter protein. The system is based on the β-galactosidase gene *sco3479* (*lacZ*$_{sc}$) from strain *S. coelicolor*, whose translation product can hydrolyze a colorless lactose analogue X-Gal to form a deep-colored compound, Indigo Blue. The second element of the reporter is the strain *Streptomyces albus* J1074, which is naturally devoid of all X-Gal hydrolyzing activity.

We tested *sco3479*'s reporter ability by constructing a series of plasmids based on the pTES vector: pOOB109, pOOB110 and pOOB114. Plasmids conferred *S. albus* colonies to the ability to convert X-Gal into colored product. The control strain J1074-pOOB114, as expected, remained colorless. The *sco3479* gene contains a series of Leu codons at the beginning of the gene that can be replaced by TTA. We replaced the CTC codon in the 8th position of the TTA codon and insert six histidine codons CAC before the stop codon. The gene was inserted into a pGCymRP21 vector containing a cumate-dependent system of controlled expression of cloned genes. Namely, in the absence of the inductor, cumic acid (cumate), the transcription of the target gene is suppressed by the repressor protein CymR. The cumate added to the environment will bind to the CymR repressor, leading to the release of the *cmt* operator and the expression of the target gene.

As a result, we generated the plasmid pRV3 with the *sco3479* gene containing the leucine TTA codon and the 6-His tag and the pRV4 plasmid with the wild type *sco3479* gene having 6-His tag. It is expected that the reporter system will enable at least a qualitative assessment of the phenotype – the color of the mycelium will indicate either translation or mistranslation. The absence of a blue color will serve as an indirect proof of blocking the reporter's expression at the translational level. We have validated the proper function of the described reporter system in the wild type *S. albus* and its *bldA* –minus derivative. As an example, we obtained the first evidence indicating the mistranslation of TTA-containing genes in *S. albus*.

Key words: genomics, codon usage, codon bias, evolutionary models, reporter system, actinobacteria, *Streptomyces.*