МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ФРАНКА

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

Кваліфікаційна наукова
праця на правах рукопису

РОКИЦЬКИЙ ІГОР ВОЛОДИМИРОВИЧ

УДК 579.25; 577.21

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

03.00.22 – молекулярна генетика

Подається на здобуття наукового ступеня кандидата біологічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ І. В. Рокицький

Науковий керівник Осташ Богдан Омелянович, доктор біологічних наук, професор

Львів–2018

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. – Львівський національний університет імені Івана Франка, Львів. – Державна установа «Інститут харчової біотехнології та геноміки Національної академії наук України», Київ, 2018.

Закономірності вживання кодонів в геномах бактерій є питанням важливим та недостатньо вивченим. Існує багато теорій щодо механізмів виникнення зміщення в переважному вживанні кодонів (ПВК) та пояснень їхньої ролі в певних процесах. Деякі теорії навіть суперечать одна одній. У цій роботі для вивчення проблеми кодонного складу використано геноми бактерій роду *Streptomyces*. Їхні геноми досі не вивчались у цьому керунку, хоча містять виразні відмінності 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 GTR модель. Вона використовує різні частоти нуклеотидів (4 параметра), і різні частоти замін між нуклеотидами (6 параметрів).

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 має вплив на топологію філогенетичного дерева. Це підкреслює важливість пошуку оптимальних моделей заміщення.

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 кожної із 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. Дослідивши і узагальнивши дані з 50 стрептоміцетних геномів, 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.

“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: Expectation-maximization) 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 моделей для різних ортологічних груп генів з різних стрептоміцетів. Застосувавши цей підхід до стрептоміцетних генів, ми спостерігали відмінність в патернах для білків з різною функцією. Ці відмінності корелюють із фізико-хімічними властивостями білка і його належністю до первинного чи вторинного метаболізму. Також, відрізняється частота синонімічних та несинонімічних заміщень в генах різних білків.

Наявність природного екстремального ПВК 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*). Результати розрахунків для елонгації лейцинових кодонів за допомогою близько-споріднених тРНК показали, 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 ККГ, і все ж він може транслюватися не менш точно, ніж популярні кодони.

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, шоб 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 у вектор 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.

В результаті ми отримали плазміду pRV3 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* J1074 (SAM2).

Очікується, що репортерна система дасть 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*, що кодує $tRNA^{Leu}_{UAA}$. Як приклад, нами отримано перші докази, що вказують на містрансляцію TTA-вмісних генів у *S. albus*.

Ключові слова: геноміка, кодонний склад, ПВК, еволюційні моделі, репортерна система, актинобактерії, *Streptomyces*.

Список публікацій за темою дисертаційної роботи

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 тРНК, 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 ТТА 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.

SUMMARY

Rokytskyy I.V. **Bioinformatic and experimental models of codon composition of *Streptomyces* genomes. - Qualifying scientific work on the rights of manuscripts.**

Thesis for a candidate degree in biological sciences in specialty 03.00.22 molecular genetics. - Ivan Franko National University of Lviv, Lviv. – Institute of Food Biotechnology and Genomics, National Academy of Sciences of Ukraine, Kiyv, 2018.

The patterns of codon use in the genomes of bacteria are an important and insufficiently studied topic. There are many theories about mechanisms of occurrence of codon bias and explanations of its role in certain processes, some of which are contradictory. In this paper, genomes of bacteria of the genus *Streptomyces* are used to study the codon composition problem. Their genomes have not yet been studied from this viewpoint, although they possess distinct codon biases and should therefore be a good model for such studies. The industrial value of streptomycetes also lends some interest to investigations of the codon composition.

We selected three genes *Streptomyces coelicolor* (*sco*), protein product of which are significantly different in their cellular function. The *sco1728* gene encodes YtrA type transcriptional factor from the GntR family. This gene and protein represent a family of transcriptional factors that interact with DNA. Product of the gene *sco2706* - glycosyltransferase, similar to a number of enzymes involved in biosynthesis of lipopolysaccharides; this gene contains TTA codon and is therefore subject to control by *bldA* tRNA. The *sco2706* gene represents enzymes. Finally, *sco3894* represents transmembrane proteins as it encodes a probable flippase (exporter) of the lipid-containing precursors of peptidoglycan in bacteria. For each of the selected genes a set of orthologous amino acid and

nucleotide sequences has been generated (Kuzniar 2008). We used these data to deduce optimal models of evolution.

The transcriptional regulator (*sco1728*) and the glycosyltransferase (*sco2706*) are characterized by a K3Pu model that takes into account three parameters in the evolution of the sequences: one parameter for describing the rate of transitions and two parameters for the rate of transversions. The group of transmembrane protein orthologs (*sco3894*) fits the GTR model. It uses different frequency of nucleotides (4 parameters), and describes nucleotides substitutions with different frequencies (6 parameters).

The selection of optimal models for the evolution of amino acid sequences has shown that each group of proteins has its unique optimal evolutionary model. The evolution of Sco1728 orthologs (transcription regulator of the GntR family) is best described by the JTT matrix. The evolution of the group of glycosyltransferase protein orthologs of Sco2706 is most accurately described by the WAG matrix. LG matrix was an optimal model for the transmembrane protein set exemplified by Sco3894. It was found that the selection of the matrix had an effect on the topology of the phylogenetic tree. This emphasizes the importance of finding the best substitution models.

A large array of genomic data is now available, which allows us to investigate the "horizontal" and "vertical" patterns of codon usage in streptomycetes. Certain codons are adjacent to each other with a probability lower or higher than null expectation, namely that these codon pairs were formed randomly (according to background frequencies of nucleotide usage in genomes). The expected and observed frequencies of the "horizontal" use of the codon pairs can be determined and estimated statistically. The Anaconda software (Moura 2007) can analyze the sequenced and annotated genome and calculate the number of each of the possible code pairs. Positive context means that dicodon occurs at a frequency exceeding two standard deviations from the mean random value (inferred from nucleotide frequenciens) in normal distribution. A pair of codons

encountered at a frequency less than a statistical incident will respectively have a "negative" context. We investigated and summarized data from 50 streptomycetic genomes and identified nine positive dicodone associations: UAU-CUG, CUG-CGC, GUA-CGG, GAU-CCG, CUC-ACC, CUC-GCC, CUC-GGC, GAA-CUC, GAA-CUG. Two negative associations were also found: CUC-CUG, CUC-GAG.

We investigated the "vertical" codon substitution by comparing the selected gene and its orthologs from a number of *Streptomyces*. This will allow us 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. Therefore, one needs simple and affordable codon substitution analysis tools, including web-directed applications. We turned our attention to DART software package, including the Xrate program (Klosterman et al. 2008). This program allows one to build codon models using the Expectation-maximization (EM) algorithm to train the original model M0 (Holmes and Rubin 2002). We created the Xrate-based web application. Using the application, "bubble" graphs of codon models can be computed for different orthogous groups of genes from different streptomycetes. Applying this approach to streptomycetes genes, we observed a difference in codon substitution patterns for proteins of different orthology and phylogeny. These differences correlate with the physical and chemical properties of the protein and its relation to the primary or secondary metabolism. Also, the frequencies of synonymous and nonsynonymous substitutions in the genes of different protein families is different.

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. Consequently, we can speak of a certain regulatory function of the *bldA* gene mediated by the TTA codon (e.g., its rare and nonrandom placement in *Streptomyces* genes). It is assumed that the translation of rare codons should occur

with high accuracy; otherwise it would not function as a genetic switch. On the other hand, a number of studies (Clarke and Clark 2008; Doma and Parker 2007) indicate that rare codons are more often mistranslated than popular ones (those used frequently and in highly expressed genes). Consequently, we run into a contradiction: TTA should be rare to control only certain genes; it is decoded by only one tRNA, yet it has to remain accurate to prevent leakiness of regulatory action. We conducted a series of studies *in silico* for a more detailed study of the decoding features of this codon. In particular, a set of tRNA genes of six streptomycetes genomes was identified and analyzed: *S. coelicolor* M145, *S. albus* J1074, *S. ghanaensis* ATCC14672, *S. clavuligerus* ATCC27076, *S. venezuelae* ATCC14115 and *S. lividans* TK24. In the obtained samples, we counted the number of copies of the tRNA genes as parameter, which correlated with the concentration of tRNA in the cell (dos Reis 2004). Errors in translation can be considered as a competition between acceptor and non-acceptor tRNAs for codon recognition. Consequently, the probability of translation depends on the ratio of concentration of focal tRNA in the cell to the total concentration of closely related tRNAs (one-nucleotide difference from focal tRNA; possesses the highest probability of suppressing the absence of focal tRNA). We applied Shah's mathematical model (Shah and Gilchrist 2010), which determines the accuracy of translation by the ratio of related (focal) and closely-related tRNA (tF/tN). 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. Consequently, the codon can be rare, it can meet low GCN criterium, and in the same time it can be translated with no less accuracy than the popular codons (for which high concentration of tRNA exists).

Our study was also focused on studies of the factors that modulate the expression of the rare leucine codon TTA in streptomycetes. 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 net 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 us to 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. With all aforementioned conditions in mind, 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 *Streptomyces coelicolor*, whose translation product can hydrolyze a colorless lactose analogue X-Gal to form a deep-colored compound, Indigo Blue (Shuman and Silhavy 2003). The second element of the reporter system is the strain *Streptomyces albus* J1074, which is naturally devoid of all X-Gal hydrolyzing activity (King and Chater 1986).

We tested Sco3479's reporter ability by constructing a series of plasmids based on the pTES vector: pOOB109, pOOB110 and pOOB114 (the latter is a non-functional *sco3479* gene containing a stop codon at the beginning of the open reading frame). 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 is cloned 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 (Horbal et al., 2014). The cumate added to the environment will bind to the CymR repressor, leading to the release

of the *cmt* operator. Removing repression from the operator will trigger 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. We have confirmed the correctness of the nucleotide sequence of the *sco3479* genes in the plasmids pRV3 and pRV4 by sequencing. Plasmids pRV3 and pRV4 were transfered to *S. albus* J1074 (SAM2).

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 performed experiments that confirmed the proper functioning of the created reporter system in the wild type *S. albus* and its derivative deficient in the *bldA* gene encoding $tRNA^{Leu}_{UAA}$. As an example, we obtained the first evidence indicating the mistranslation of TTA-containing genes in *S. albus*.

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

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

а/к – амінокислота (-и, -ний)

ККГ — кількість копій генів

ПЛР – полімеразна ланцюгова реакція

п.н. – пар нуклеотидів

т.п.н. – тисяч пар нуклеотидів

Am^r – стійкий (стійкість) до аміноглікозидного антибіотика апраміцину

BLAST (basic local alignment sequence tool) – онлайн-метод локального порівняння послідовностей

ЕМ-алгоритм – алгоритм максимізації очікування

ORF (open reading frame) – відкрита рамка зчитування

tF – концентрація акцепторної тРНК в клітині

tN – концентрація близькоспорідненої тРНК в клітині

X-Gal – 5-Br-4-Cl-3-індоліл-β-D-галактопіранозид

ВСТУП

Актуальність теми. 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 організма 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,* 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 й 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 беспосередньої участі у виконанні експериментів. Планування експериментів, аналіз та обговорення отриманих результатів проведені спільно з науковим керівником, д-ром біол. наук Б.О. Осташем, проф. В.О. Федоренком, канд. біол. наук 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 найменування). 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.

РОЗДІЛ 1. ОГЛЯД ЛІТЕРАТУРИ

1.1. Генетичний код та еволюція кодувальних послідовностей – засади

Геноми бактерій представлено здебільшого однією молекулою ДНК (хромосома), крім того можуть бути позахромосомні елементи – плазміди, фаги тощо. Елементарною структурною одиницею ДНК є нуклеотид, що складається з залишку фосфорної кислоти, дезоксирибози та однієї з 4х азотистих основ. Пуринові основи аденін (A) і гуанін (G), складаються з двох гетероциклів, тоді як піримідинові основи цитозин (C) і тимін (T) – одного. Два ланцюги ДНК з'єднані в подвійну спіраль за допомогою водневих зв'язків між пуриновими та піримідиновими основами, принцип комплементарності (Chargaff 1951).

Геном — це сукупність всієї спадкової генетичної інформації організму, тобто всіх генів, некодувальних послідовностей ДНК та позахромосомного генетичного матеріалу. Реалізація спадкової інформації, записаної в геномі, відбувається у результаті двох послідовних клітинних процесів — транскрипції та трансляції. В процесі транскрипції інформація з окремих генів “переписується” шляхом синтезу одноланцюгової молекули РНК на матриці ДНК. Транскрипція забезпечує необхідну кількість інформації (РНК) про ген, щоб швидко збільшити концентрацію його кінцевого продукту (зазвичай білка) в клітині. Аналіз транскриптому дозволяє виокремити гени, експресія яких необхідна на тому чи іншому етапі життєвого циклу клітини. Утворення поліпептидних продуктів генів відбувається за допомогою рибосом (і на основі мРНК) в процесі трансляції. Кожен з цих процесів має низку складних регуляторних механізмів, які в першу чергу забезпечують точність інтерпретації генетичної інформації.

Під час транскрипції, частини геному транскрибуються в одноланцюгову молекулу РНК. Деякі транскрипти можуть прямо виконувати

функцію (тРНК, рРНК, нкРНК). мРНК слугують матрицею для синтезу білків на рибосомах – кожні три послідовні нуклеотиди в ланцюзі забезпечують точне розпізнавання та комплементарну взаємодію з антикодоном аміноацильованої тРНК. Ці триплети називаються кодонами, а відповідність кодонів амінокислотам називається генетичним кодом, який майже універсальний для всіх відомих форм життя на Землі (Knight et al. 1999). Із $4^3 = 64$ можливих кодонів 61 визначає амінокислотний залишок, решта три – стоп-кодони, що визначають кінець кодувальної послідовності (на цих кодонах рибосома припиняє трансляцію). Генетичний код зазвичай репрезентують у вигляді таблиці, як показано на рис. 1.1

1й нуклеотид	2й нуклеотид: U	2й нуклеотид: C	2й нуклеотид: A	2й нуклеотид: G
U	UUU, UUC – Phe UUA, UUG – Leu	UCU, UCC, UCA, UCG – Ser	UAU, UAC – Tyr UAA, UAG – Stop	UGU, UGC – Cys UGA – Stop UGG – Trp
C	CUU, CUC, CUA, CUG – Leu	CCU, CCC, CCA, CCG – Pro	CAU, CAC – His CAA, CAG – Gln	CGU, CGC, CGA, CGG – Arg
A	AUU, AUC, AUA – Ile AUG – Met	ACU, ACC, ACA, ACG – Thr	AAU, AAC – Asn AAA, AAG – Lys	AGU, AGC – Ser AGA, AGG – Arg
G	GUU, GUC, GUA, GUG – Val	GCU, GCC, GCA, GCG – Ala	GAU, GAC – Asp GAA, GAG – Glu	GGU, GGC, GGA, GGG – Gly

Рис. 1.1 Таблиця генетичного коду (Сиволоб 2008).

Генетичний код виродженій, оскільки 61 кодон кодує тільки 20 амінокислот. Таким чином, більшість амінокислот зашифровано більш ніж одним кодоном. Кодони, які кодують одну амінокислоту, отримали назву синонімічних. Відповідно, несинонімічні кодони визначають різні амінокислотні залишки.

Геном – динамічна система нуклеотидних послідовностей, яка постійно зазнає змін. Зміни, що закріплюються (успадковуються) – це мутації. Що є джерелом мутацій? По-перше геном має реплікуватися (подвоїтися) перед кожним поділом клітини, і процес реплікації не є абсолютно безпомилковим. Спонтанний та природний індукований мутагенез, помилки асоційовані з транскрипцією та субоптимальними процесами репарації є іншими джерелами мутацій. Виходячи з характеру нуклеотидних замін у кодувальній послідовності можна розрізнити синонімічні та несинонімічні заміщення кодонів.

Мутації можуть призводити до зміни пристосованості живої системи та зазнавати дії факторів природного добору, або бути еволюційно нейтральними. Ці та низка інших (неселективних) факторів визначають долю мутації – її еволюційне закріплення, поширення чи втрату. Крім цього еволюційні події відбуваються в часі. Чим більше часу минуло із моменту розділення незалежних еволюційних ліній, тим більше генетичних відмінностей буде між ними. Аналізуючи відмінності між генетичними послідовностями різних організмів за допомогою різноманітних методів, можна реконструювати їх еволюційну історію, зв’язки між ними.

Еволюційна відстань між двома послідовностями прямо пропорційна часу дивергенції, що виражається зростанням кількості та якості нуклеотидних або амінокислотних заміщень між ними. Тому з плином часу відстань між будь-якими двома гомологічними (такими, що мають спільного еволюційного предка) послідовностями (виміряна у кількостях нт/ак заміщень) збільшується. Математично наші уявлення про особливості нт/ак

заміщень у певній вибірці послідовностей можна описати еволюційними моделями. Для моделей, що розглядають еволюцію нуклеотидних послідовностей, нуклеотидні заміни є результатом мутації та її наступного добору, закріплення чи вилучення. Таким чином, нуклеотидні заміни в таких моделях розглядають, як наслідок комплексного процесу.

1.2. Моделі еволюції генетичних послідовностей

Розрізняють параметричні та емпіричні моделі еволюції генетичних послідовностей – на основі різних підходів до їхнього обчислення. У параметричному (синонім: механістичному) підході, імовірності заміщень вираховують з низки параметрів, які або визначаються на основі теоретичних міркувань, або вираховуються безпосередньо з даних. Емпіричні моделі часто мають більше параметрів, які обчислюються як статистичні показники великого масиву даних. Ці показники перетворюють у певну систему рахунків, які далі використовуються для аналізу нових вибірок генетичних послідовностей.

1.2.1. Параметричні моделі еволюції нуклеотидних послідовностей. Параметричні моделі найчастіше використовуються для опису еволюції нуклеотидних послідовностей, де відносно невелика кількість можливих параметрів. Найпростіша з таких моделей - однопараметрична модель Джакса і Кентора JC69 (Jukes and Cantor 1969). В основі моделі лежить припущення, що заміна одного нуклеотиду на будь-який з трьох інших за одиницю часу відбувається з імовірністю (частотою, швидкістю) α, однаковою для всіх можливих варіантів. Для розкриття логіки, що використовується під час розробки еволюційних моделей загалом, розгляньмо модель Джакса і Кентора докладніше.

Розглянемо одну позицію нуклеотида в певній послідовності. Нехай у момент часу 0 в позиції знаходиться нуклеотид A, а імовірність знаходження A в цій позиції в момент часу 0 дорівнює 1. Оскільки A може замінитися на один із трьох інших нуклеотидів – T, C або G – за одиницю часу із імовірністю α, тоді через одиницю часу, у момент часу 1, A буде замінений на інший нуклеотид із імовірністю 3α або залишиться в цій позиції із імовірністю:

$$P_{A(1)}=1-3\alpha \quad . \qquad (1.1)$$

Таким чином, за першу одиницю часу можливі дві різні події у розглянутій позиції: Нуклеотид A залишиться незмінним із імовірністю 1 - 3α , або заміниться на інший нуклеотид (з трьох можливих) із імовірністю 3α. Проте, у момент часу 2, потрібно врахувати третій можливу подію. Якщо за першу одиницю часу A був замінений на інший нуклеотид, то за другу одиницю часу в розглянутій позиції могла відбутися зворотна заміна (реверсія) – нуклеотид міг знову замінитися на A. Імовірність того, що нуклеотид A залишатиметься в одній позиції із врахуванням усіх трьох можливих подій в момент часу 2 становитиме:

$$P_{A(2)}=(1-3\alpha)P_{A(1)}+\alpha(1-P_{A(1)}) \quad , \qquad (1.2)$$

а підставивши значення з рівняння 1.1 матимемо:

$$P_{A(2)}=(1-3\alpha)2+4\alpha \quad , \qquad (1.3)$$

$$P_{A(2)}=1-\alpha(2+9\alpha) \quad . \qquad (1.4)$$

Узагальнюючи вищесказане, розрахунок імовірності знаходження A у розглянутій позиції у будь-який момент часу t + 1 буде визначатися виходячи із імовірності знаходження A у розглянутій позиції у попередній момент часу t:

$$P_{A(t+1)}=(1-3\alpha)P_{A(t)}+\alpha(1-P_{A(t)}) \quad . \qquad (1.5)$$

Зміну імовірності перебування А у певній позиції в нуклеотидній послідовності у будь-який момент часу t, можна виразити так:

$$\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)} = -3\alpha P_{A(t)} + \alpha\left(1 - \right) \quad , \qquad (1.6)$$

$$\Delta P_{A(t)} = \alpha\left(1 - 4P_{A(t)}\right) \quad . \qquad (1.7)$$

Ми розглядали час як дискретну величину, але з точки зору неперервного процесу ΔP(t) виступатиме як темп змін за час t. Відповідно імовірність знаходження А в позиції можна розрахувати за допомогою:

$$\frac{dP_A(t)}{dt} = \alpha\left(1 - 4P_{A(t)}\right) \quad . \qquad (1.8)$$

Розв'язок отриманого диференційного рівняння:

$$P_{A(t)} = \frac{1}{4} + \left(P_{A(0)} - \frac{1}{4}\right)e^{-4at} \quad . \qquad (1.9)$$

Це рівняння дійсне для будь-яких початкових умов. Наприклад, в момент часу 0, коли А зберігається (відсутність замін), $P_{A(0)} = 1$, тоді імовірність, що нуклеотид залишиться в позиції через деякий час t становитиме (також і для інших нуклеотидів, оскільки модель припускає однакову частоту всіх заміщень):

$$P_{AA(t)} = P_{CC(t)} = P_{GG(t)} = P_{TT(t)} = \frac{1}{4} + \frac{3}{4}e^{-4} \quad . \qquad (1.10)$$

Якщо ж початковою умовою обрати відсутність А в розглянутій позиції, $P_{A(0)} = 0$, тобто в цій позиції був один з інших нуклеотидів, також врахуємо, що згідно моделі частота замін нуклеотидів C/A, G/A та T/A однакова, тоді імовірність зустріти А в цій позиції через час t становитиме:

$$P_{CA(t)} = P_{GA(t)} = P_{TA(t)} = \frac{1}{4} - \frac{1}{4}e^{-4at} \quad . \qquad (1.11)$$

Таким чином, можна розглядати спільну швидкість заміни чи незаміни нуклеотиду *i* на нуклеотид *j* в будь-який момент часу t, $P_{ii(t)}$ та $P_{ij(t)}$ відповідно:

$$P_{ii(t)} = \frac{1}{4} + \frac{3}{4}e^{-4at} \quad , \qquad (1.12)$$

$$P_{ij(t)} = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} . \qquad (1.13)$$

Це дозволяє зробити наступний висновок. Якщо в момент часу 0 в розглянутій позиції знаходився нуклеотид *i*, то із часом від нуля до нескінченності імовірність збереження нуклеотиду *i* в цій позиції нелінійно знижується від 1 до 0,25. Відповідно нелінійно зростає швидкість заміни *i* на нуклеотид *j* від 0 до 0,25. Це значення дорівнює імовірності зустрічі нуклеотиду *i* в позиції, де першочергово був інший нуклеотид, що також зростає від 0 до 0,25 (рис. 1.2).

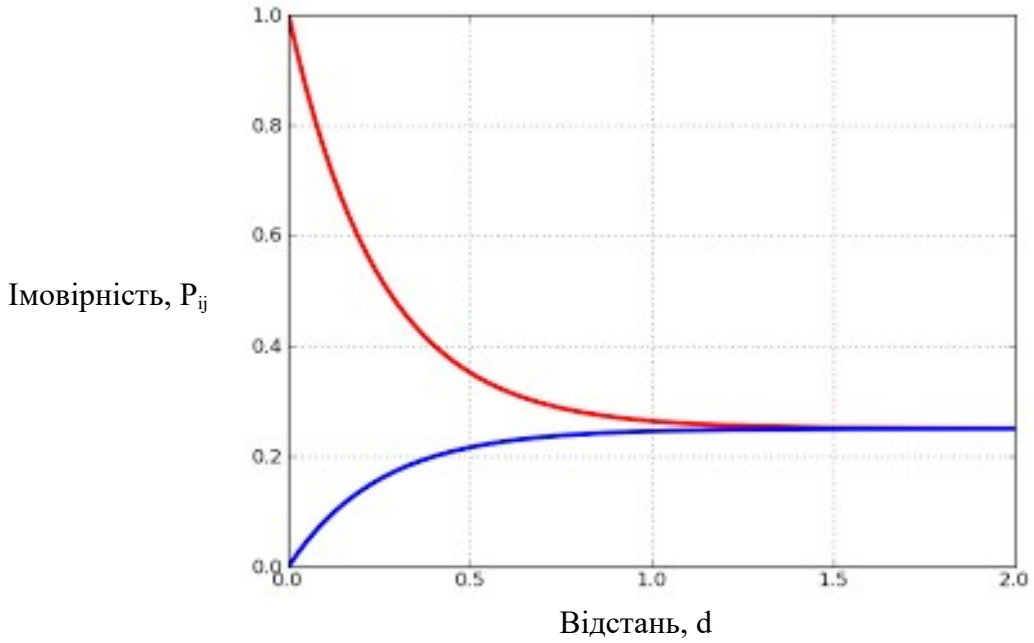


Рис. 1.2 Імовірність P_{ij} зміни початкового стану *i* до кінцевого стану *j* в залежності від еволюційної відстані *d* між двома послідовностями згідно моделі JC69. Червона крива: імовірність того, що *i* залишиться незмінним. Синя крива: імовірність того, що *i* заміниться на *j* (P_{ij}). Зі зростанням еволюційної відстані P_{ij} сягає плато (0,25, або 25%).

Отже, із часом імовірність знаходження нуклеотиду *i* в розглянутій позиції наближається до 0,25 незалежно від того, який нуклеотид знаходився в цій позиції початково. Це справедливо для будь-якого із чотирьох нуклеотидів, згідно із моделлю Джакса – Кентора, а також із плином часу імовірність зустрічності кожного в кожній позиції розглянутої послідовності тяжіє до 0,25. Відповідно вміст кожного із нуклеотидів у всій послідовності прямує до 25%. Таким чином, модель описує тенденцію до стану рівноваги між вмістом усіх чотирьох нуклеотидів у послідовності – або до стаціонарного стану.

Виходячи з розглянутої однопараметричної моделі Джакса і Кентора (JC69), для двох послідовностей, що зазнали дивергенцію в момент часу t = 0, імовірність того, що в момент часу t вони мають в певній позиції нуклеотид A становитиме $P_{AA(t)}^2$. Оскільки в нас відсутня інформація про вихідну послідовність, потрібно врахувати всі можливі сценарії, що призвели до появи однакового нуклеотиду (в нашому випадку A) в одній позиції в двох послідовностях:

$$I(t) = P_{AA(t)}^2 + P_{CA(t)}^2 + P_{GA(t)}^2 + P_{TA(t)}^2 \quad , \tag{1.14}$$

або, виходячи з раніше отриманих виразів,

$$I(t) = \frac{1}{4} + \left(\frac{3}{4}\right)^{-8\alpha t} \quad . \tag{1.15}$$

Звісно, що аналогічні міркування справедливі для всіх позицій аналізованих послідовностей. В еволюційному аналізі відмінності між послідовностями можна представити в розрахунку на довжину послідовностей як пропорцію різних нуклеотидів — *p*-відстань, що розраховується:

$$p = \frac{D}{L} \quad , \tag{1.16}$$

де D — кількість відмінних нуклеотидних позицій (сайтів), L - загальна кількість нуклеотидних сайтів, за якими порівнюються дві послідовності.

При обчислені *p*-відстані ділянки, попарного вирівнювання з прогалинами (результат вставок, делецій) не розглядаються.

Таким чином, імовірність відмінностей між двома послідовностями в розрахунку на позицію (сайт) в момент часу t дорівнює:

$$P = 1 - I(t) = \frac{3}{4} + \left(1 - e^{-8\alpha t}\right), \tag{1.17}$$

$$8\alpha t = -\ln\left(1 - \frac{4}{3}p\right). \tag{1.18}$$

Час дивергенції між двома послідовностями зазвичай не відомо і ми не можемо оцінити α. Тому прийнято оцінювати число замін на позицію (К) з моменту дивергенції двох послідовностей:

$$K = \frac{-3}{4}\ln\left(1 - \frac{4}{3}p\right). \tag{1.19}$$

Як бачимо, для розрахунку дистанції Джакса – Кентора між двома послідовностями необхідно знати лише *p*-дистанцію між ними. Також необхідно зазначити, що дистанція Джакса – Кентора може бути розрахована тільки при *p*-дистанції менше 0,75, оскільки при p > 0,75 аргумент логарифму набуває негативного значення.

Моделі нуклеотидних заміщень – це матриці 4×4, що містять 16 станів. Наприклад, модель Джакса-Кентора JC69 буде мати вигляд, як наведено на рис. 1.3.

У наступні три десятиліття, модель JC69 було багаторазово доповнено та розширено. Це привело до появи низки нових моделей, щоб дозволити різні імовірності відношення транзицій/трансверсій, запропонована Кімурою – модель K2P (Kimura 1980; Kimura 1981), для оцінки різних рівноважних частот (Felsenstein 1981; Hasegawa, Kishino and Yano 1985), а також моделювати різні значення для пуринових та піримідинових заміщень (Tamura 1992; Tamura and Nei 1993).

а *б*

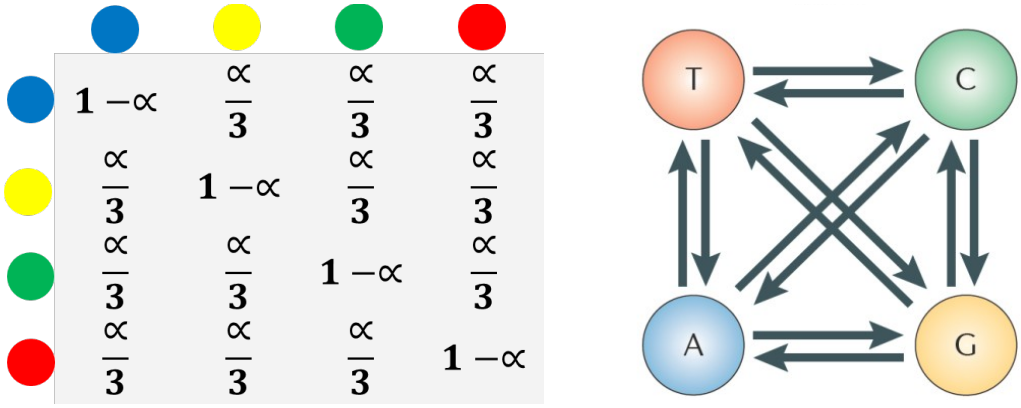


Рис. 1.3 Презентація моделі еволюції нуклеотидних послідовностей JC69 у вигляді матриці швидкостей заміщення (***а***) і графічної схеми (***б***), де використання кіл різного кольору означають 4 нуклеотидні основи. Стрілки однакового розміру, вказують на однакову частоту нуклеотидів і швидкість їхнього заміщення.

Для прикладу, розглянемо використання моделей JC69 та K2P для визначення еволюційної відстані між послідовностями 1 та 2:

1: A C C T G T A A T C

2: A C G T G C G A T C

* * *

Ці послідовності довжиною в десять нуклеотидів відрізняються в трьох позиціях, що відмічено символом *. Згідно формули 1.16, p-відстань становитиме 0,3. Можна обчислити частоту транзицій та трансверсій, значення яких необхідні для моделі Кімури і які становитимуть 0,2 (T|C, A|G) та 0,1 (C|G) відповідно. Тоді еволюційна відстань згідно моделі JC69 (ф-ла 19) становить 0,383, а згідно моделі K2P – 0,402. Отже, маємо відмінні результати для однакових даних при використанні різних моделей. Такі відмінності матимуть вплив на подальший аналіз послідовностей, тому необхідно використовувати ті еволюційні моделі, які найбільше підходять для послідовностей, що вивчають.

Параметричні моделі гнучкі, їх можна адаптувати конкретного масиву даних, наприклад, змінюючи відношення швидкостей транзицій/трансверсій для різних груп генів (Yang 1999). Загалом, чим більше параметрів, які можна обчислити з даних, тим краще модель буде описувати ці дані. Проте, є ряд недоліків при використанні занадто багатьох параметрів. Перш за все, це наявний обсяг даних, що використовують для аналізу: малі вибірки ведуть до відхилень, похибок в обчислені параметрів, зокрема, якщо є низка параметрів, що компенсують один одного. Існує також ризик, що через велику кількість параметрів, модель більше не описуватиме загальний процес, а лише певну вибірку.

1.2.2. Параметризовані моделі еволюції кодонних послідовностей. Моделі наведені в цьому розділі, визначаються матрицею швидкостей Q, яка описує миттєві швидкості заміщення між 61 змістовним кодоном і містить параметри, які обчислюються відповідно до вибірки досліджуваних даних.

Перші параметризовані моделі, що описують еволюцію на кодонному рівні були запропоновані незалежно Голдманом та Янгом (1994) та Мюзом і Гаутом (1994). Для аналізу еволюційних процесів у кодонних послідовностях довгий час використовували параметризовані моделі. Перш за все це пояснюється необхідністю великої кількості даних для створення емпіричних моделей, яких у 1990х роках не було. Лише в 2005 році описана перша емпірична кодонна модель (Schneider, Cannarozzi and Gonnet 2005).

Мюз і Гаут представили свою модель як метод оцінки швидкостей синонімічних та несинонімічних заміщень. Ними запропоновано модель засновану на марковському процесі (імовірність стану N залежить лише від x попередніх станів, а не від усієї історії системи), орієнтовану на пошук максимально вірогідного еволюційного шляху.

Модель використовує шість параметрів, α та β для синонімічних і несинонімічних коефіцієнтів заміщення, а також чотири нуклеотидні частоти

π_n. Враховуються тільки випадки, коли кодони *i* та *j* відрізняються одним нуклеотидом:

$$Q_{i,j}=\begin{cases} \alpha\pi_n, & \text{для синонімічних з} \\ \beta\pi n, & \text{для несинонімічних} \\ 0, & \text{заміщення більше,ніж одно} \end{cases} . \qquad (1.20)$$

Модель описує матрицю 61 × 61 змістовних кодонів. Рівноважна частота кодону, що складається з нуклеотидів *I*, *J* і *K*, може бути отримана, як:

$$\pi_i \pi_j \pi_k / (1 - \prod stop) , \qquad (1.21)$$

де Π_{stop} є сума добутків $\pi_i\pi_j\pi_k$ для стоп-кодонів.

Матрицю Q можна нормалізувати, два параметра α та β може бути замінено одним параметром (співвідношенням синонімічних і несинонімічних швидкостей заміщення), як це було зроблено в пізніших моделях.

Моделі Мюза і Гаута не вистачало деяких бажаних властивостей, які були передбачені моделлю Голдмана і Янга (1994), тому вона не знайшла широкого застосування. Голдман і Янг представили свою модель в основному як модель Маркова для філогенетичних реконструкцій на основі кодонних замін. Також автори виявили, що модель може бути використана для отримання оцінки максимальної вірогідності швидкості коефіцієнта транзиції/трансверсії та синонімічних/несинонімічних заміщень.

Модель Голдмана і Янга — це матриця 61 × 61, результат процесу Маркова з низкою параметрів. Параметр *κ* описує імовірність транзицій. Схожість між різними амінокислотами вираховується згідно фізико-хімічних відстаней між амінокислотами за Ґрантхемом (1974), як параметр $d_{aai,aaj}$, де aa_i – амінокислота що кодується кодоном *i*. Також є параметр *V*, який представляє мінливість гена – його тенденцію зазнати несинонімічне заміщення.

Модель Голдмана і Янга використовувалася для визначення впливу селективних сил в межах родів (Yang 1998), специфічних сайтах (Nielsen and Yang 1998), та сайтах і родах одночасно (Yang and Nielsen 2002). Відповідно модель 1994 року дещо змінили для більш наглядного відображення синонімічних та несинонімічних заміщенн, шляхом введення відношення *dN*/*dS*, як параметра ω. Матриця швидкостей заміщень використана в цих роботах:

$$Q_{i,j}=\begin{cases}\pi_i & \text{, якщо синонімічні транс}\\ \kappa\pi_i & \text{, якщо синонімічні транз}\\ \omega\pi_i & \text{, якщо несинонімічні тра}\\ \omega\kappa\pi_i & \text{, якщо несинонімічні тра}\\ 0 & \text{, якщо більше однієї замі}\end{cases} . \qquad (1.22)$$

Ці методи створено для порівняння двох або більше гіпотез про дію сил природного добору. Для визначення впливу останніх різним гілкам (видам) присвоювалося довільне значення ω. Параметри зазнавали оптимізації за допомогою побудови дерев, а про правдоподібність отриманих значень судили за порівнянням з нульовою гіпотезою, де всі гілки мають однакове значення $\omega \leq 1$. За значенням параметру відношення несинонімічних до синонімічних заміцень можна говорити про напрям дії добору: нейтральний – $\omega = 1$; позитивний (або рушійний у вітчизняній літературі) – $\omega > 1$; негативний (або стабілізуючий) – $\omega < 1$.

Моделі Янга та співробітників широко використовуються і нині є найпопулярнішими кодонними моделями, оскільки враховують основні аспекти еволюції ДНК двома параметрами κ і ω. Важлива причина їхнього успіху – інтеграція в пакет програмного забезпечення paml (Yang 1997).

1.3. Емпіричні моделі

1.3.1. Емпіричні моделі еволюції амінокислотних послідовностей. Головна відмінність еволюційних моделей полягає в тому, чи параметри обчислюють кожного разу для вибірки даних, чи оцінюють один раз, на великому наборі даних. Параметричні моделі описують всі заміни в залежності від ряду параметрів, які оцінюють (підбирають) для кожного набору даних. Це має таку перевагу, що модель може бути пристосована до особливостей конкретної вибірки. Емпіричні моделі створюються шляхом оцінки багатьох параметрів з великого набору даних. Ці параметри потім фіксуються і будуть використовуватися для кожного нового набору даних. Перевага таких моделей в тому, що ці параметри можна обчислити точніше (на основі великого і/або різноманітного масиву), і такі моделі потенційно можуть мати універсальне застосування.

На відміну від параметричних моделей заміщення в ДНК, еволюційні моделі для амінокислотних послідовностей, як правило, виведені емпірично (Whelan and Goldman 2001). Перші моделі еволюції амінокислотних послідовностей, PAM (Dayhoff, Schwartz and Orcutt 1978) та JTT (Jones, Taylor and Thornton 1992), обчислено різними методами підрахунку заміщень в вирівнюваннях близькоспоріднених послідовностей і подані як матриці заміщення амінокислот розміром 20x20. Зовсім недавно, емпіричні моделі амінокислотних заміщень були обчислені з використанням методів максимальної вірогідності (Kosiol and Goldman 2005). Цей підхід вніс правки в низку методологічних недоліків з використання рахунків (Adachi and Hasegawa 1996; Whelan and Goldman 2001; Le and Gascuel 2008; Gonnet, Cohen and Benner 1993).

Слід зазначити важливі припущення, на яких ґрунтуються всі описані моделі. Всі сайти (ділянки) в межах множинного вирівнювання послідовностей еволюціонують незалежно один від одного, але є

учасниками одного еволюційного процесу, а відповідно мають однакову еволюційну швидкість. Також, всі сайти вирівнювання пов'язані з одним філогенетичним деревом (Halpern and Bruno 1998).

Однією із найпопулярніших серій матриць амінокислотних заміщень стали BLOSUM (**BLO**cks **SU**bstitution **M**atrix) (Henikoff and Henikoff 1992). Ця серія матриць базується на логарифмі співвідношення так званих цільових до фонових частот заміщення амінокислот. Тут цільовим частотам відповідають імовірності зустріти певну пару амінокислотних залишків у вирівнюванні гомологічних білків, а фонові частоти – це імовірність зустріти таку пару при вирівнюванні випадкових (негомологічних) білків. Фонові частоти фактично дорівнюють частотам амінокислотних залишків у досліджуваному протеомі. Матрицю рахунків виводять безпосередньо із вирівнювань, не спираючись на філогенетичне дерево, чи певну еволюційну модель.

Для побудови BLOSUM-матриць використано базу Blocks – це множинні вирівнювання фрагментів білків. Єдиним критерієм відбору білків до BLOCKS – їхня здатність до досконалого вирівнювання з іншими білками, незалежно від філогенетичної спорідненості.

Рахунок заміщення $S_{i,j}$ амінокислоти *i* амінокислотою *j* в матриці BLOSUM розраховується так:

$$BLOSUM_{i,j} = S_{i,j} = 2\log_2 \frac{q_{i,j}}{e_{i,i}}, \qquad (1.23)$$

де $q_{i,j}$ нормалізована частота зустрічності заміни *i* на *j* (сума всіх частот для всіх амінокислот рівна одиниці) в межах вибірки даних, а $e_{i,j}$ – частота очікування появи заміни *i* на *j*. Тут в чисельнику маємо цільову частоту , в знаменнику – фонову. Значення заокруглюють і записують в матрицю. Процес еволюції вважають оборотним (швидкість/імовірність заміщення, наприклад, Ala→Pro = Pro→Ala), тому матриця симетрична і подається лише діагональ та одна половина (рис. 1.4).

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Рис. 1.4 BLOSUM62 – матриця рахунків заміщень амінокислот. Номер матриці (62) означає, що амінокислотні послідовності, які мали більше 62% ідентичності, групували разом і розглядали як одну послідовність (тому вони не робили незалежного внеску в обчислення рахунків матриці).

BLOSUM-матриці використовують для оцінки попарних вирівнювань амінокислотних послідовностей. Тут аналізують пару амінокислотних залишків в одній позиції вирівнювання. Виходячи із значень в матриці оцінок можна судити про біологічний зміст цієї пари амінокислот. Тобто оцінюють, чи таке заміщення широко зустрічається в природі (додатній рахунок), чи це рідкісне явище (від'ємний). Для прикладу, заміщення метіоніну лейцином зустрічається часто (Met/Leu: 2), тоді як триптофан заміщується проліном вкрай рідко, нижче за фонову , випадкову зустрічність цих амінокислот, і тому значення такого заміщення буде від’ємне (Trp/Pro: 4). Найвищий рахунок має збереження амінокислоти – позиції в яких не відбулося заміщення (Trp/Trp: 11). Матриці BLOSUM лежать в основі найпопулярнішого пакету програм попарного вирівнювання BLAST. Однак, оскільки послідовності у базі BLOCKS (джерело вихідних даних для

BLOSUM) не мають еволюційної спорідненості, то BLOSUM не використовуються для філогенетичних реконструкцій.

1.3.2. Емпіричні моделі еволюції кодонних послідовностей. Моделі кодонних послідовностей запропоновані до 2005 року були параметричними і охоплювали лише деякі аспекти еволюції ДНК. До моделей не включені заміщення амінокислот, спричинені зміною більш ніж одного нуклеотиду в кодоні. Тому, щоб вивчити повну картину еволюції на рівні кодонів, була створена емпірична модель кодонних заміщень з великого набору вирівнювання кодувальних послідовностей (Schneider, Cannarozzi and Gonnet 2005). Для емпіричних моделей, повна матриця заміщень оцінюється один раз із вирівнювання великого набору генетичних послідовностей.

Для виведення моделі Шнайдер та співробітники використали низку доступних геномів для яких були дані про наявність ортологічних генів. Кодонне вирівнювання послідовностей здійснили керуючись вирівнюванням амінокислотних послідовностей та подальшою заміною амінокислотних залишків на кодони з відповідних нуклеотидних послідовностей. Методом підрахунку заміщень кодонів виведено матрицю оцінок $S_{i,j}$, як і у випадку з емпіричними матрицями амінокислотних заміщень:

$$S_{i,j}=10\,lg\frac{\pi_j M_{i,j}(t)}{\pi_i \pi_j}, \qquad (1.24)$$

де $M_{i,j}$ –відповідає імовірності заміни кодону *i* кодоном *j* в відповідний момент часу *t*, а π_c– рівноважна частота кодону *c* в межах вирівнювання.

Через відсутність параметрів, основне застосування такої моделі – це обчислення рахунків при кодонних вирівнюваннях ДНК таксономічно обмеженої групи послідовностей. Такі моделі мають кращі показники якості вирівнювання послідовностей ніж аналогічні для амінокислотних послідовностей, особливо на незначних еволюційних відстанях.

1.4. Особливості вживання кодонів

Низка еволюційних сил формують первинну структуру кодувальних компонентів геному. Це, зокрема, дуплікація генів, хромосомні перегрупування, рекомбінація ДНК, делеції і вставки, транспозиції мобільних елементів, нуклеотидні поліморфізми, нуклеотидні повтори тощо. Також, виродженість генетичного коду має важливе значення для еволюції первинної структури генів, оскільки це забезпечує велику кількість варіантів для створення послідовності відкритої рамки зчитування (ORF) будь-якого конкретного білка. Наприклад, пентапептид “Val-Leu-Pro-Ile-Ile” згідно генетичного коду (див. рис. 1.1) можна закодувати 864 різними нуклеотидними послідовностями. Чим довший поліпептидний ланцюг, тим більшою буде кількість можливих кодувальних послідовностей. Однак еволюційні сили накладають обмеження на вживання кодонів, зокрема синонімічних. Відповідно використання синонімічних кодонів в генах невипадкове, що вказує на існування механізмів, які обмежують “ступінь свободи” кодувальної послідовності. Правила, якими керують вживанням синонімічних кодонів у ORF лише частково зрозумілі (Plotkin and Kudla 2011). Втім, очевидно, що особливості вживання кодонів відображають дію двох основних еволюційних сил: селективних, що змінюють ефективність декодування мРНК, і мутаційних (випадкових) – що однаково діють як на (білок-)кодувальну, так і на некодувальну частини генома.

Перевага у частоті вживання одних синонімічніх кодонів над іншими – переважне вживання кодонів або ПВК – один з проявів залежностей між вживанням різних кодонів. ПВК помітно на різних рівнях організації генетичного матеріалу. В різних видів вживання синонімічних кодонів корелює із GC-складом відповідних геномів. Наприклад, в геномі *Lactobacillus acidophilus* (GC-склад генома ≈ 50%) два лізинові кодони (AAG та AAA) зустрічається практично з однаковою частотою, тоді як в геномі

Streptomyces venezuelae (GC-склад – 72 %) домінує AAG (98% усіх лізинових кодонів). У межах одного генома (на міжгенному рівні) різні гени збагачені різними синонімічними кодонами. Це може виступати як прояв механізму по регуляції експресії генів. Зміна концентрації в клітині акцепторних тРНК для різних синонімічних кодонів відповідатиме змінам в рівні експресії генів багатих на відповідні синонімічні кодони (Guthrie and Chater 1990). В межах одного гена (на внутрішньогенному рівні) також можна спостерігати нерівномірне вживання синонімічних кодонів. Наприклад, в про- та евкаріотів перші 100-120 п.н. кодувальної послідовності гена часто містять непропорційно багато “непопулярних” кодонів (яким відповідає низька концентрація відповідних тРНК), а послідовність ближче до стоп-кодона збагачено на їхні синонімічні, “популярні”, версії. Одні дослідники вважають, що це допомагає уникнути зіткнення чи зупинки рибосом при ініціації трансляції; інші наводять докази на користь важливості непопулярних кодонів для формування стабільної мРНК (Shah and Gilchrist 2011).

На рис. 1.5 наведено приклад нерівномірного вживання кодонів в межах однієї рамки зчитування, а саме теорія “рампи”. В низці робіт показано, що в межах гена синонімічні кодони, для яких є низька концентрація акцепторних тРНК в цитоплазмі, розташовані ближче до 5' кінця послідовності (перші 100-120 п.н.). Теорія “рампи” пояснює цей феномен як механізм, який сповільнює проходження рибосомами цієї ділянки в мРНК, що далі допомагає уникнути зіткнення рибосом на ініціаторному відтинку транс крипта (Tuller et al. 2010).

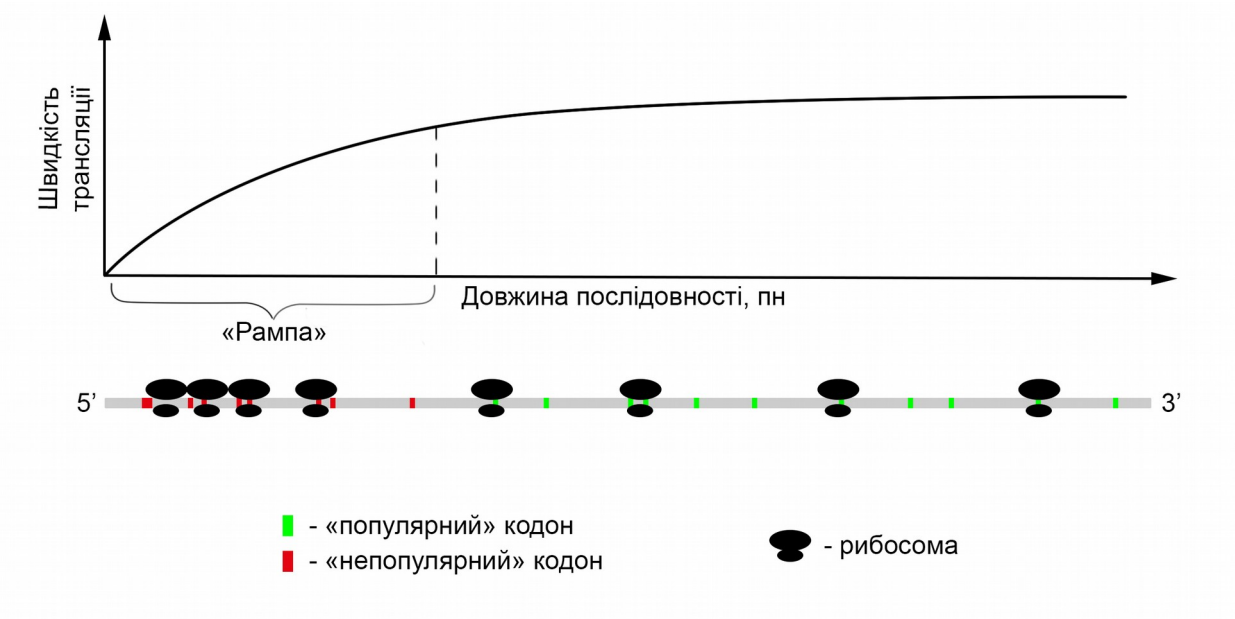


Рис. 1.5 “Рампа” утворена скупченням рідкісних синонімічних кодонів на початку мРНК – гальмує проходження ініціаторної ділянки мРНК, і так рівномірно розподіляє рибосоми по всій довжині мРНК.

Інше пояснення трансляційної рампи – вона забезпечує достатньо часу для формування правильної вторинної структури білка в процесі нарощування поліпептидного ланцюга. Не виключено наразі й таку версію, що це явище не має біологічної ролі, а є лише наслідком нейтральної еволюції.

Іншим аспектом вживання кодонів в межах однієї ORF виступає контекст послідовної пари кодонів (гексаплет, або дикодон). А саме, між останнім нуклеотидом першого кодону і першим нуклеотидом другого ($N_1N_2\mathbf{N_3}$-$\mathbf{N_1}N_2N_3$) спостерігається невипадкова асоціація, так утворюючи певний $\mathbf{N_3}$-$\mathbf{N_1}$ контекст (Moura et al. 2007). Контекстне вживання кодонів має відношення до декодувальної точності. Зокрема трансляційні механізми чутливі до природи кодонної пари, яка присутня в A- і P-сайтах рибосоми.

E.coli використовує дикодони в невипадкових патернах (Gutman and Hatfield 1989). Сильна кореляція між нуклеотидами в кодоні та хиткими позиціями сусідніх кодонів вказує на організацію кодонів в деякому

оптимальному контексті (Curran et al. 1995). Також показана сильна відмінність в контексті вживання кодонів в генах із сильною та слабкою експресією (Gouy 1987; Shpaer 1986; Yarus and Folley 1985). Контекстне вживання кодонів має деякий ефект на ПВК, хоч і слабший ніж вплив сил природного добору. Припускають, що це допомагає уникнути утворення небажаних елементів вторинної структури мРНК або додаткових сайтів зв'язування мРНК із рРНК (Bulmer 1988; Eyre-Walker and Bulmer 1993; Hartl, Moriyama and Sawyer 1994; Li et al. 2012). Усе сказане вище веде до важливого висновку. Ген, який складається з оптимальних синонімічних кодонів (яким відповідає висока концентрація тРНК), може бути неоптимальним з точки зору кодонних контекстів чи вторинної структури мРНК (Goodman et al. 2013; Napolitano et al. 2017; Chevance and Hughes 2017). Тому важливе цілісне розуміння усіх тонкощів організації й еволюції кодонної послідовності, а також ступеня впливу цих факторів на експресію і подальшу функцію білка.

1.5. Актиноміцети, як модельний організм для вивчення закономірностей вживання кодонів

1.5.1. Стрептоміцети. Рід *Streptomyces* відноситься до грам-позитивних, GC-багатих бактерій класу *Actinobacteria*, порядку *Streptomycetales*, родини *Streptomycetaceae*. Типовою властивістю геномної ДНК актинобактерій загалом є високий вміст GC від 72 до 75 % (Piepersberg 1993; Ventura et al. 2007; Barka et al. 2015). Цю особливість використовують в ідентифікації генів (Bibb, Findlay and Johnson 1984). *Streptomyces* є одним з небагатьох родів бактерій з лінійною хромосомою (Chen et al. 2002). Хромосома *Streptomyces coelicolor* A3(2), завдовжки 8,67 млн. пн. з вмістом GC 72,1%, за даними останньої анотації містить 7825 білок-кодувальних генів. Лінійні кінці хромосом *Streptomyces* складаються з термінальних інвертованих

повторів різних розмірів (24-500 т.п.н.). Вони зв'язані з білками з обидвох вільних 5' кінців. Ці білки, ймовірно, виступають в ролі праймерів для синтезу останнього фрагмента Оказакі відстаючого ланцюга при завершені двобічної реплікації ДНК, після ініціювання в типовому сайті OriC в центрі хромосоми (Musialowski et al. 1994; Chang, Kim and Cohen 1994; Chen et al. 2002; Redenbach et al. 1996; Wolanski et al. 2014).

Streptomyces є найбільшою бактерійною групою виробників біоактивних речовин, і забезпечують 70% від загального виробництва антибіотиків, які використовуються в медицині та ветеринарії. Промислове використання *Streptomyces* не обмежується їх вторинними метаболітами. Інші корисні сполуки – це гідролітичні ферменти або цілий мікроорганізм як такий. Наприклад, біофунгіциди для захисту коренів рослин проти грибкових інфекцій, містять висушені спори і міцелій штаму *Streptomyces griseoviridis*.

1.5.2. Роль рідкісного кодону TTA в біології стрептоміцетів. Характерною рисою стрептоміцетів, як зазначено вище, є продукція вторинних метаболітів, яка співпадає із початком морфологічної диференціації. Біосинтез антибіотиків вимагає узгодженої роботи великої кількості генів, які потребують складних регуляційних механізмів для оптимального використання клітинних ресурсів. У цьому процесі важливу роль відіграють регуляторні гени вищого рівня, що здійснюють глобальну регуляцію (зазвичай це транскрипційні фактори, що впливають як на вторинний метаболізм, так і на інші аспекти біології актиноміцетів, такі як морфологічна диференціація). Глобальними регуляторами можуть виступати гени, що контролюють процеси транскрипції, посттранскрипційної модифікації РНК, або трансляції. Маніпуляції із такими регуляторними генами дозволяють підвищувати рівні синтезу вторинних метаболітів або запускати експресію "мовчазних" кластерів генів продукції вторинних

метаболітів (тих, які не експресуються за умов стандартної ферментації) (Chater 1993; Chater 2006; Bibb 1996; Barka et al. 2015).

Найдокладніше вивченим механізмом трансляційної регуляції у актинобактерій є *bldA*-залежна регуляція. Ген *bldA* кодує тРНК$^{Leu}_{UAA}$ – єдина, що здатна ефективно декодувати лейциновий кодон TTA (Lawlor, Baylis and Chater 1987). Вперше мутацію у цьому гені описано у 1967 році (Hopwood 1967) у модельного стрептоміцета *Streptomyces coelicolor* A3(2), яку було названо S48 та згодом перейменовано на *bldA*. Мутанти за геном *bldA* нездатні формувати повітряний міцелій і не синтезують усіх чотирьох відомих для *S. coelicolor* (на той час) антибіотиків (Merrick 1976).

Показано, що експресія TTA-вмісних генів залежить від функціонування гена *bldA*. Це, а також запізніле (порівняно з іншими тРНК) нагромадження *bldA*-тРНК у стаціонарній фазі росту навели дослідників на думку про часову регуляцію експресії TTA-вмісних генів за рахунок обмеження концентрації зрілої аміноацильованої *bldA*-тРНК (Leskiw et al. 1991; Leskiw et al. 2005). Також, *bldA*-мутанти із подібними фенотипами було знайдено у різноманітних стрептоміцетів (Kwak, McCue and Kendrick 1996; Trepanier et al. 2002).

Дослідження особливостей вживання TTA-кодону у геномах родини стрептоміцетів проводилися різними методами та у різних масштабах (Chater and Chandra 2008). У *S. coelicolor* 145 хромосомних генів є TTA-вмісними. Це гени синтезу антибіотиків, морфогенезу, гідролізу рослинних і тваринних вуглеводних полімерів (целюлоза, хітин тощо), гени з невідомою функцією. Більшість з цих генів містять тільки один TTA-кодон, переважно на початку мРНК. Вважають, що відсутність достатніх кількостей зрілої тРНК$^{Leu}_{UAA}$ в експоненційній фазі росту обмежує трансляцію усіх UUA-вмісних транскриптів у цій фазі, і відтак це слугує механізмом,що запускає вторинний метаболізм і морфогенез саме в стаціонарній фазі. Однак, досліджено роль у первинному та вторинному метаболізмі *S. coelicolor* лише

декількох TTA-вмісних генів, функція решти таких генів невідома (Fernandez-Moreno et al. 1991; Li et al. 2007; White and Bibb 1997; Bibb 1996). Незрозуміло також, що саме визначає запізнілий час появи зрілої тРНК$^{Leu}_{UAA}$; відомо лише, що це не пов'язано із запізнілою транскрипцією РНК-попередника (Pettersson et al. 2011). Імовірно, загальмоване аміноацилювання цієї тРНК, або її посттранскрипційні модифікації є причиною особливої часової динаміки накопичення зрілої тРНК$^{Leu}_{UAA}$.

1.6. Перспективи

Як згадувалося вище, актинобактерії широко використовуються в промисловості як продуценти цінних вторинних метаболітів, зокрема антибіотиків. Це складні системи біохімічного синтезу, в яких задіяна велика кількість генів. Одні гени кодують ферменти задіяні в синтезі, інші — виконують регуляційну функцію. Одним з вищих механізмів регуляції виступає особливе вживання синонімічних кодонів в кодувальних послідовностях генів вторинного метаболізму. Цей механізм недостатньо вивчений та не до кінця зрозумілий, проте низка експериментальних результатів демонструють важливу роль ПВК та контекстів. Виходячи з цього, глибше розуміння особливостей та вивчення закономірностей кодонного складу цієї групи бактерій нестиме безпосереднє практичне застосування – зокрема для покращення промислових якостей штамів актинобактерій. Для прикладу, це може бути розробка нових технологій ферментації чи конструювання нових штамів-надпродуцентів з оптимізованим складом кодонів відповідних генів. Також, такі знання допоможуть краще оптимізувати кодувальні послідовності генів для гетерологічної експресії з метою отримання білка (Baltz 2010).

Для кращого розуміння інформації про вживання кодонів в геномі необхідні методи візуалізації даних, отриманих в процесі математичного

моделювання. Це має бути такий метод, що допоможе одночасно охопити як загальну картину, так і окремі особливості кодонного складу геному. Також важлива доступність цих методів для широкого кола науковців-нефахівців у галузі обчислювальної біології.

Наступним важливим аспектом вивчення особливостей кодонного складу актинобактерій є перевірка математичних гіпотез на експериментальній моделі. Наразі описано низку фенотипів різних актинобактерій пов'язаних зі змінами чи порушенями зчитування певного кодона. Коли організм не має здатності експресувати гени з одним із синонімічних кодонів — спостерігається зміна фенотипових ознак. Тим не менше, такі фенотипові зміни — складний критерій для вивчення вживання синонімічних кодонів, оскільки в формуванні фенотипової ознаки задіяно багато генів. Такі дані важко аналізувати з врахуванням всіх можливих факторів впливу на фенотип. Таким чином, постає завдання створення простої моделі з якнайменшою кількістю ланок від зміни синонімічного кодону до продукту (фенотипу), який можна виміряти якісно та кількісно.

РОЗДІЛ 2. МАТЕРІАЛИ ТА МЕТОДИ

2.1. Матеріали

2.1.1. Штами бактерій, плазміди. Штами бактерій та рекомбінантні молекули ДНК використані в роботі подано в табл. 2.1 і 2.2, відповідно.

Таблиця 2.1

Штами мікроорганізмів використані та отримані при виконанні роботи

Назва штаму	Генотип/Примітки	Посилання
1	**2**	**3**
	Escherichia coli	
GB2005	Штам для рутинного клонування ДНК; F⁻*mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K λ⁻*rps*L *nup*G *fhu*::IS2 *rec*ET (Ton^r)	Fu et. al. 2012, Yin et. al. 2015
WM6026	Штам для кон'югативного перенесення ДНК, підтримує *pir*-залежні реплікони, ауксотроф за діамінопімеліновою кислотою; *lacI*^{q} *rrnB3* Δ*lacZ*4787 *hsdR514* Δ*araBAD567* Δ*rhaBAD568 rph-1* *att*λ::pAE12 (Δ*ori*R6K-*cat*::Frt5) Δ*endA*::Frt *uidA*(ΔMluI)::*pir* *att*HK::pJK1006Δ(*ori*R6K-*cat*::Frt5 *trfA*::Frt)	Blodgett, Zhang та Metcalf 2005, Luo et al. 2013

Продовження таблиці 2.1

1	2	3
	Streptomyces albus	
J1074	Похідний *S. albus* G (*ilv-1 sal2*); не продукує забарвлених вторинних метаболітів	Chater and Wilde 1980; Menéndez et al. 2004; Zaburannyi et al. 2014
Δpse (Δ*pse*B4), або SAM2	Похідний *J1074;* безмаркерна делеція псевдо-*attB*$^{\varphi C31}$ сайту з genoma	Bilyk and Luzhetskyy 2014
ΔbldA, або ОК3	Похідний *J1074;* делеція гена Leu tRNA UAA - *bldA*	Koshla et al. 2017

Геномні послідовності використані в роботі наведено в базі даних GenBank

Таблиця 2.2

Перелік векторів, використаних у даній дисертаційній роботі

Назва	**Характеристика**	**Посилання**
1	**2**	**3**
pGCymRP21	Вектор експресії на основі інтегрази *int*-phiC31, містить сильний кумат-індуцибельний промотор *cmt-P21*; Am^r Sp^r Ap^r	Horbal, Fedorenko and Luzhetskyy 2014

Продовження таблиці 2.2

1	2	3
pTES	Вектор експресії на основі інтегрази *int*-phiC31, містить сильний промотор *ermEp*. Існує можливість вирізання послідовностей вектора по *loxP*-сайтах; Am^r	Herrmann et al. 2012
pRV3	Вектор pGCymRP21, що несе мутантну (TTA кодон в 8-ій позиції) версію гена *sco3479* під контролем промотора *cmt-P21*	Ця робота (пункт 3.5)
pRV4	Вектор pGCymRP21, що несе ген *sco3479* дикого типу під контролем промотора *cmt-P21*	Ця робота (пункт 3.5)
pOOB109	Вектор pTES що несе ген *sco3479* дикого типу під контролем тандему промоторів: частини *ermEp* та цілісного *SCO3479p*	Ця робота (пункт 3.5)
pOOB114	Вектор pTES що несе мутантну (стоп-кодон TGA в 19-ій позиції) версію гена *sco3479* під контролем промотора *ermEp*	Ця робота (пункт 3.5)

2.1.2. Праймери. Праймери, що використовувались у даній роботі представлені у таблиці 2.3.

Таблиця 2.3.

Перелік олігонуклеотидних послідовностей (праймерів), використаних у роботі

Назва	Послідовність нуклеотидів[1,2]	Опис
1	**2**	**3**
Sco3479_ 6Hisrp	5'-AAA AAG CTT CAA TTG TCA **GTG GTG GTG GTG GTG GTG**[2] CCC TAC ATG ACG GTC CTT G-3'	"зворотній" праймер для синтезу *sco3479*, що містить 6His-таг, та сайти рестрикції HindIII та MfeI (pRV3, pRV4)
LacZ_ XbaI_ TTA1	5'-AA ATC TAG ACG TGA CGT CCC ATG ACG TAA GGA ATC GAC GTG CTC CCC ATC ATC GGT GCT **TTA** CGC CCG TGG GC-3'	праймер для синтезу sco3479, що містить TTA-кодон та сайт рестрикції XbaI (pRV3)
Sco3479 upXbaI	5'-AA ATC TAG ACG TGA CGT CCC ATG ACG TAA G -3'	праймер для синтезу sco3479 із сайтом рестрикції XbaI (pRV4)
Sco3479 upBglII	5'-AA AAG ATC TGC ACG ACC TCA CGG TGC ATG-3'	праймер для синтезу sco3479 із сайтом рестрикції BglII (pOOB109, pOOB110, pOOB114)

Продовження таблиці 2.3

1	2	3
LacZ_ XbaI_ TGA2	5'-AA ATC TAG ACG TGA CGT CCC ATG ACG TAA GGA ATC GAC GTG CTC CCC ATC ATC GGT GC T CTC CGC CCG TGG GCG GAC CCT ACG ATG ACA TCC **TGA** CTC CGA CTC CCG ATG CGC GCG CG-3'	праймер для синтезу sco3479, що містить стоп-кодон (TGA) та сайт рестрикції XbaI (pOOB114)
Sco3479 rpMfeI BglII	5'-AA ACA ATT G AG ATC TCA CCC TAC ATG ACG GTC CTT G-3'	"зворотній" праймер для синтезу sco3479, що містить сайти рестрикції MfeI та BglII (pOOB109, pOOB110, pOOB114)

[1]сайти рестрикції позначено підкресленням
[2]жирним шрифтом виділено штучно уведені кодони (TTA, гексагістидиновий таґ), які не співпадає із послідовністю *SCO3479* дикого типу

2.1.3. Поживні середовища та умови культивування. Клітини бактерії *E. coli* вирощували при 37°С на багатому середовищі Лурія-Бертані (LB: 0,5 % дріжджовий екстракт, 1 % пептон, 1 % NaCl). Для селекції плазмідовмісних бактерій використовували апраміцин у кінцевій концентрації 100 мг/л (концентрація вихідного розчину 100 мг/мл). Для вирощування штамів актиноміцетів використовували повноцінні середовища. Схрещування *E. coli – Streptomyces* проводили на вівсяному середовищі ВС (вівсяне борошно – 50 г., агар – 20 г., вода водопровідна – до 1 л., pH 7.2) або СМ-агарі (соєве борошно – 20 г., маннітол – 20 г., агар – 22г, вода водопровідна - до 1 л., pH 7.2). Міцелій для виділення сумарної та плазмідної ДНК, вирощували у рідкому середовищі TSB (триптон – 17 г,

гідролізат сої – 10 г, NaCl – 5 г, K_2HPO_4 – 2,5 г, глюкоза – 2,5 г, вода дистильована – 1 л).

2.2. Основні молекулярно-генетичні методи

2.2.1. Полімеразна ланцюгова реакція (ПЛР). Готували реакційну суміш, до складу якої входили вода, буфер для ДНК-полімерази з $MgCl_2$, дезоксинуклеотидтрифосфати (дНТФ), праймери та фермент ДНК-полімераза. Суміш розділяли на кілька окремих проб, в кожну з яких додавали ДНК-матрицю.

Склад реакційної суміші

реагент	кінцева концентрація
стерильна деіонізована вода	-
10х PCR буфер	1х
2 мМ суміш дНТФ	0,2 мМ
праймер 1	0,1-1 мкл
праймер 2	0,1-1 мкл
полімераза	1,25 u/ 50мкл
матриця ДНК	10 пг- 1 мкг

Реакції проводилися на ампліфікаторах BioRad T100 й Thermo Pico24. Параметри реакції: початкова денатурація ДНК проводилася при 98°С протягом 30 с, потім 25 циклів з денатурацією при 98°С протягом 30 с, анілінгом праймерів при 55°С протягом 15 с та полімеризації при 72°С протягом 2 хв. В кінці проби інкубувалися при 72°С протягом 5 хв, а потім охолоджувалися до 4°С. Температура, при якій відбувалося відпалювання праймерів до матриці, визначалася з врахуванням довжини і нуклеотидного

складу праймерів, тривалість полімеризації визначалася в залежності від довжини фрагменту, який потрібно було ампліфікувати.

2.2.2. Рестрикційний аналіз ДНК. Метод базується на здатності ендонуклеаз рестрикції ІІ класу гідролізувати ДНК у певних сайтах (послідовностях нуклеотидів). У роботі використовували ферменти та відповідні інкубаційні буфери виробництва фірми “Thermo Scientific” (Литва). У реакційну суміш вносили 0,5-2 мкг аналізованої ДНК, розчиненої в ТЕ-буфері чи бідистильованій воді. До розчину ДНК додавали 2 мкл (1/10 об’єму) відповідного буферу і 1 U ферменту на 1мкг плазмідної ДНК, об’єм реакційної суміші доводили водою до 20 мкл. Суміш інкубували при 37°С протягом 1-2 год.

2.2.3. Електрофорез ДНК в агарозному гелі. Реактиви: агароза; ТАЕ буфер: 0,04 М *трис*-ацетат, 0,002 М ЕДТА; буфер для нанесення проб: 0,25 % бромфеноловий синій, 40 % сахароза, бромистий етидій 0,5 мкг/мл.

Агарозу вносили в ТАЕ буфер, нагрівали до повного розчинення та кип'ятили протягом 3 хв. Коли розчин охолоджувався до 50°С, додавали бромистий етидій. Охолоджений розчин заливали у кювету з гребінцем для формування лунок. Між дном кювети і гребінцем залишали шар агарози товщиною 0,5-1 мм. Після полімеризації гелю гребінець виймали, а кювету з гелем поміщали в електрофоретичну камеру з ТАЕ-буфером. У лунки вносили виділену ДНК, змішану з 6-кратним буфером для нанесення проб так, щоб об'єм буферу становив 1/6 об'єму кінцевої суміші. Електрофорез проводили за напруги 2-4 В/см протягом 30-40 хв. Фракції ДНК реєстрували, освітлюючи гель ультрафіолетовими променями. Одержану електрофореграму фотографували. Як ДНК-стандарт у всіх випадках використовували ДНК-маркер фірми “Thermo Scientific” SM 0311.

2.2.4. Елюювання фрагментів ДНК з гелю. Мікропробірку типу Eppendorf об’ємом 1,5 мл зважували і записували його вагу. Використовуючи чистий скальпель, вирізали якомога менший фрагмент гелю, що містить

потрібний зразок ДНК і поміщали в пробірку, зважували і розраховували вагу агарозного зрізу. Додавали 10 мкл зв'язуючого буферу (тип 2) на кожні 10 мг агарозного фрагменту, інкубували на водяній бані при 55°С, перемішуючи, поки агароза повністю не розплавиться. Отриману суміш об'ємом 600 мкл, яка містила ДНК, переносили у колонку (GFX MicroSpin column), центрифугували при 14000 об./хв протягом 1 хв на центрифузі. Повторювали, доки вся проба не була оброблена таким чином. Потім додавали 500 мкл буферу для промивання (тип 1), центрифугували за таких самих умовах. Додавали 10-20 мкл буферу для елюції (тип 4), інкубували при кімнатній температурі 3-5 хв, центрифугували при 14000 об./хв протягом 2 хв.

2.2.5. Виділення плазмідної ДНК з клітин *E. coli.* Реактиви: Розчин I: 50 мМ глюкоза, 25 мМ *трис*-HCl (pH 8,0), 10 мМ ЕДТА (pH8,0). Розчин II: 0,2 М NaOH, 1 % SDS, 10мМ *трис*-HCl (pH 8,0), 1 мМ ЕДТА (pH8,0); 5М Калію ацетат; 3М Натрію ацетат; 7,5М Амонію ацетат; фенол (pH 8,0); ізопропанол; 96 % розчин етилового спирту; 70 % розчин етилового спирту; ТЕ-буфер: 10 мМ *трис*-HCl (pH 8,0), 1 мМ ЕДТА (pH8,0)

Бактерійні трансформанти нарощували в 10 мл селективного середовища протягом ночі. Біомасу осаджували центрифугуванням при 4000 об./хв протягом 15 хв на центрифузі марки Eppendorf 5804R. Клітини промивали розчином I без лізоциму і центрифугували при 4000 об./хв протягом 12 хв. Супернатант зливали, а осад ресуспендували та додавали 4 мл розчину II, перемішували, щоб суспензія стала в'язкою та витримували 5 хв при 0° С. Потім додавали 3 мл 5 М калію ацетату, перемішували і залишали на холоді. Через 15 хв суміш центрифугували при 9000 об./хв протягом 15 хв, супернатант переносили в чисту пробірку та додавали 1/10 об'єму натрію ацетату та 0,6 об'єму ізопропанолу. ДНК осаджували центрифугуванням при 9000 об./хв протягом 15 хв. Осад розчиняли в 300 мкл ТЕ-буферу. Потім додавали 2 об'єми амонію ацетату та інкубували

протягом 15 хв при кімнатній температурі. Центрифугували при 5000 об./хв протягом 15 хв на центрифузі марки Eppendorf 5417C. Надосадову рідину відбирали у чистий еппендорф і додавали 1/10 об'єму натрію ацетату та 0,6 об'єму ізопропанолу та інкубували при кімнатній температурі 10 хв. Після центрифугування при 12000 об./хв 10 хв, супернатант зливали, а осад підсушували і розчиняли в 200 мкл ТЕ-буферу. Для остаточної депротеїнізації до розчину додавали 200 мкл фенолу, насиченого буфером, центрифугували при 12000 об./хв і відбирали верхню (водну) фазу. Додавали 1/10 об'єму натрію ацетату, 2 об'єми етанолу. Потім центрифугували 10 хв при 12000 об./хв. Осад (плазмідна ДНК) промивали 70 % етанолом, підсушували і розчиняли в стерильному ТЕ-буфері.

2.2.6. Лігування фрагментів ДНК. Готували 10 мкл суміші фрагменту та розщепленого вектора (400-800 нг) у бідистильованій воді; реакційна суміш містила рівну чи вищу (до 3 раз) концентрацію фрагменту по відношенню до концентрації вектора.

Додавали наступні компоненти до суміші:

10-кратний буфер для лігування – 0,5 мкл,

Т4 ДНК-лігаза – 2-4 одиниці,

бідистильована H_2O – до 5 мкл.

Інкубували суміш при кімнатній температурі протягом 3 год.

2.2.7. Електротрансформація клітин *E. coli.* Свіжу колонію бактерійних клітин інокулювали в 2 мл рідкого середовища LB та вирощували в умовах аерації при 37°C протягом ночі. 1 мл „нічної" культури переносили в 100 мл рідкого середовища LB та підрощували до OD_{600} = 0,5-1,0 (кювета 1 см), що відповідає 10^{10} кл/мл. Клітини осаджували 15 хв при 4000 об./хв, 4°C на центрифузі і промивали в 400 мл стерильної холодної бідистильованої води. Клітини ресуспендували в 200-300 мкл холодного стерильного 10 % гліцеролу і переносили в стерильні епендорфи по 40 мкл для довготривалого зберігання при –70°C.

До 40 мкл суспензії клітин додавали 1-4 нг плазмідної ДНК і вносили суміш на дно охолодженої 2-мм кювети; здійснювали електротрансформацію (12,25 кВ/см, 50 мкФ, 129 Ом) тривалістю 5-6 мс. До кожного зразка додавали по 1 мл рідкого середовища LB, трансформовані клітини інкубували протягом 50 хв при 37°С. Отриману суспензію висівали на чашки зі селективним середовищем, яке містило антибіотик, та інкубували від 12 до 16 годин при 37°С.

2.2.8. Кон'югація *E. coli* – *Streptomyces*. Міжродову кон'югацію проводили за модифікованою методикою Мазодієра. Вирощували нічну культуру донорного штаму *E. coli* WM6026, що містила відповідну плазміду, до середини логарифмічної фази росту, клітини осаджували при 12 тис. об/хв та ресуспендували в 100 мкл стерильної дистильованої води. Суспензію спор штама-реципієнта (концентрація спор – 10^7 кл/мл) піддавали тепловому шоку при 50 °С протягом 10 хв. Спори актиноміцета та клітини штаму-донора *E. coli* змішували у співвідношенні 1:10 та висівали на чашки з підсушеним вівсяним середовищем. Інкубували 12–20 год та заливали розчином антибіотика, стійкість до якого визначається маркерним геном плазміди, і налідиксової кислоти. Інкубували чашки 3–6 діб, після чого підраховували та аналізували транскон'югантів.

2.2.9. Біоінформатичний аналіз. Нуклеотидні та амінокислотні послідовності, що використовувалися в роботі отримували з різних баз даних: сервера *Streptomyces* (*http://strepdb.streptomyces.org.uk/*), база даних генів та геномів, Кіото – KEGG (*http://www.genome.jp/kegg/*).Пошук в базах даних нуклеотидних і амінокислотних послідовностей проводили із використанням програми BLAST, що знаходиться на сервері Національного інституту біотехнологічної інформації, США (*http://www.ncbi.nlm.nih.gov*).

Множинне вирівнювання послідовностей здійснювали низкою алгоритмів, зокрема: MAFFT (Katoh and Toh 2008), Clustal Omega (Sievers et

al. 2011), MUSCLE (Edgar 2004) (*http://www.ebi.ac.uk/Tools/msa/*) та ProbCons (*http://probcons.stanford.edu/* - Do et al. 2005).

Для підбору моделей еволюції використовували веб сервіс IQ-Tree Web Service (http://iqtree.cibiv.univie.ac.at/; Trifinopoulos et al. 2016).

Філогенетичні дерева будували за допомогою програмних пакетів phylogeny.fr (Dereeper et al. 2008) та PhyloPhLan (Segata et al. 2013).

РОЗДІЛ 3. РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ

Як випливає з огляду літератури, закономірності вживання кодонів в геномах бактерій є питанням важливим та недостатньо вивченим. Існує багато теорій щодо механізмів виникнення змішень в переважному вживанні кодонів (ПВК) та пояснень їхньої ролі в певних процесах, частина з яких 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 ПВК і тому мають бути доброю моделлю для таких досліджень. Промислова цінність стрептоміцетів також робить дослідження кодонного складу цікавим у прикладному вимірі.

Виходячи із сучасного стану досліджуваної проблеми (див. Огляд літератури), ми окреслили низку завдань цієї роботи. Їхнє вирішення має привести до кращого розуміння механізмів виникнення і значення нерівномірного вживання кодонів у біології стрептоміцетів. Зокрема, для початку потрібно встановити чи є якісь характерні особливості заміщення (нуклеотидів/амінокислот) на добре вивчених нуклеотидному та амінокислотному рівнях генетичної інформації. Чи можна говорити про родову незалежність цих заміщень, чи всі закономірності є наслідком еволюційних процесів? Дослідження різних за функцією груп білків допоможе встановити оптимальні еволюційні моделі нуклеотидних та амінокислотних заміщень. Також, важливо дослідити особливості контекстного вживання кодонів у геномах стрептоміцетів.

Нині доступний великий масив геномних даних, який дає змогу будувати емпіричні моделі кодонних заміщень для різних класів генів стрептоміцетів. Але постає проблема у візуалізації таких моделей та їхньому тлумаченні. Відтак необхідно мати прості й доступні інструменти аналізу кодонних заміщень, зокрема веб-спрямований програмний інструментарій.

Іншим аргументом у виборі стрептоміцетів для дослідження закономірності вживання кодонів є наявність природного екстремального ПВК у вживанні лейцинового кодону TTA. Потрібно провести низку досліджень *in silico* для детальнішого вивчення особливостей декодування цього кодону та можливих варіантів виникнення такого явища. Наступний крок – виявити та вивчити фактори, що модулюють експресію рідкісного лейцинового кодона TTA у стрептоміцетів. Для цього необхідно сконструювати TTA-кодон-специфічну репортерну систему з можливістю якісного та кількісного аналізу активності репортерного білка. На вищеперелічених завданнях і було зосереджено нашу роботу.

3.1. Визначення оптимальних моделей нуклеотидних та амінокислотних заміщень у *Streptomyces*

Множинне вирівнювання послідовностей – важливий метод порівняльного аналізу генетичних послідовностей, і один з перших етапів філогенетичної реконструкції. Існує низка різних біоінформатичних алгоритмів та підходів для отримання таких вирівнювань і кожний зазвичай дає дещо відмінний результат. Отримані вирівнювання матимуть різний ступінь “надійності” відповідно до вимог дослідження. Тому, щоб оцінити вплив алгоритму множинного вирівнювання на оцінку частот нуклеотидних чи амінокислотних заміщень, в роботі використано чотири концептуально різні алгоритми – Clustal Omega (традиційний прогресивний спосіб вирівнювання; Clustal Omega (Sievers et al. 2011), MAFFT (вирівнювання з трансформацією Фур’є; Katoh and Toh 2008), MUSCLE (прогресивне вирівнювання з дальшою оптимізацією; Edgar 2004) та ProbCons (вирівнювання із застосуванням імовірнісних (прихованих) моделей Маркова; Do et al. 2005).

Другим елементом, що робитиме внесок у варіабельність моделей заміщення – це природа генетичних послідовностей. Різні гени можуть мати різну еволюційну історію; якщо гени кодують особливу групу протеїнів з точки зору фізичної хімії (наприклад, трансмембранні білки), то вони теж можуть мати особливий характер амінокислотних заміщень. Ми припускаємо, що один представник з окремої групи білків матиме властивості притаманні цілій групі, та навпаки еволюційна модель одного представника групи описуватиме еволюцію цілої групи. Відповідно, підібрано 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 репрезентуватиме родину транскрипційних факторів, які взаємодіють з ДНК. Продукт гена *sco2706* – глікозилтрансфераза, схожа до низки інших ферментів задіяних в процесі біосинтезу ліпополісахаридів; цей ген містить рідкісний кодон TTA і відтак підлягає контролю з боку тРНК *bldA* (див. Огляд літератури). Ген *sco2706* репрезентуватиме ензими. Нарешті, обрано ген *sco3894*, що кодує трансмембранний білок, імовірну фліппазу (експортер) ліпід-вмісних попередників пептидоглікану бактерій. До кожного з генів, на основі геномів 30 найкраще вивчених бактерій роду *Streptomyces*, створено вибірку ортологічних нуклеотидних та амінокислотних послідовностей, які далі вирівняно за допомогою вказаних вище алгоритмів (Kuzniar 2008).

За допомогою веб сервісу IQ-Tree Web Service (http://iqtree.cibiv.univie.ac.at/) ми здійснили підбір моделі, яка найкраще описує еволюційні процеси в заданих геномних даних, шляхом роботи алгоритмів максимальної вірогідності. У результаті пошуку оптимальної еволюційної моделі для нуклеотидних послідовностей трьох різних вищевказаних груп, виявлено, що кожна з них має власну схему заміщень і, отже, дещо відмінні шляхи молекулярної еволюції. Ортологам

транскрипційного регулятора (*sco1728*) і глікозилтрансферази (*sco2706*) незалежно від типу алгоритму множинного вирівнювання притаманна модель K3Pu, що враховує три параметри в еволюції послідовностей: один параметр для опису швидкості транзицій та два параметри — для швидкості трансверсій. Вирівнювання *sco1728* за алгоритмом MAFFT показало, що оптимальною є модель TVM: змінні швидкості трансверсій та стала щвидкість транзиції. Групі ортологічних генів трансмембранного білка (*sco3894*) відповідає GTR модель. Вона використовує різні частоти нуклеотидів (4 параметра), і різні частоти замін між нуклеотидами (6 параметрів). Всі дані наведені в таблиці 3.1. Слід відзначити, що моделі, ґрунтовані на методі вирівнювання MAFFT у більшості випадків не узгоджуються з тими моделями, які випливають з трьох інших алгоритмів вирівнювання.

Таблиця 3.1

Оптимальні нуклеотидні моделі заміщення для трьох ортологічних груп генів *Streptomyces*

Метод вирівнювання	*sco1728*	*sco2706*	*sco3894*
Clustal Omega	K3Pu	K3Pu	GTR
MAFFT	TVM	K3Pu	GTR
MUSCLE	K3Pu	K3Pu	GTR
ProbCons	K3Pu	K3Pu	GTR

Підбір оптимальних моделей еволюції амінокислотних послідовностей також здійснювався з застосуванням чотирьох різних алгоритмів вирівнювання і груп ортологічних білків, які згадано вище. Згідно з отриманими результатами, підсумованими в таблиці 3.2, кожній групі білків відповідає своя еволюційна модель. Еволюцію ортологів Sco1728

(транскрипційний регулятор родини GntR) найкраще описує матриця Джонса-Тейлора-Торнтона і її розширений варіант за Kosiol і Goldman – JTT і JTTDCMut. Еволюція група ортологів білка глікозилтрансферази Sco2706 найточніше описується матрицею WAG. Оптимальною моделлю для ортологів трансмембранного білка Sco3894 є матриця LG.

Таблиця 3.2

Амінокислотні моделі заміщення для трьох груп білків *Streptomyces*

Метод вирівнювання	sco1728	sco2706	sco3894
Clustal Omega	JTTDCMut	WAG	LG
MAFFT	JTTDCMut	WAG	LG
MUSCLE	JTT	WAG	LG
ProbCons	JTT	WAG	LG

А чи ми справді можемо говорити, що оптимальні моделі притаманні для представників різних груп генетичних послідовностей відповідають оптимальній моделі цілої групи? Для перевірки, ми підібрали співрозмірні вибірки ортологічних послідовностей низки транскрипційних факторів *Streptomyces* різноманітних типів, а саме: *sco1358*, *sco2223*, *sco4159*, *sco5231*, *sco5819* та *sco7512* (Табл. 3.3). Вирівнювали дані лише алгоритмом Clustal Omega, оскільки попередні результати показали відсутність значного впливу алгоритму вирівнювання на підбір моделі.

Оптимальною моделлю для нуклеотидних послідовностей більшості представників транскрипційних факторів є TVM – змінні швидкості трансверсій та стала щвидкість транзиції. Це частково суперечить попереднім результатам (Талб. 3.1), хоча вирівнювання за допомогою MAFFT видавало саме цю модель.

Таблиця 3.3

Оптимальні моделі заміщення для різних ортологічних груп транскрипційних факторів (ТФ) *Streptomyces*

Гени	Тип ТФ	Нуклеотидна модель	Амінокислотна модель
sco1358	LysR	TVM	JTT
sco2223	TetR	TVM	JTT
sco4159	GlnR	TVM	JTTDCMut
sco5231	MutC	TVM	JTT
sco5819	WhiH	K3Pu	HIVb
sco7512	AraC	TVM	JTT

Щодо амінокислотних послідовностей, то оптимальна модель відповідає отриманим результатам (Талб. 3.2). Вийнятком слугує транскрипційний фактор типу WhiH (*sco5819*) оптимальні моделі K3Pu та HIVb для нуклеотидних та амінокислотних послідовностей, відповідно.

Далі перевірено, як вибір моделі заміщення впливає на філогенетичний аналіз реальних даних. Як тестовий приклад вибрано ортологічні гени і відповідні білки плейотропного регулятора типу AraC – AdpA (Takano et al. 2003, Rabyk et al. 2018). Для перевірки нуклеотидних моделей створено кластер з 107 ортологічних послідовностей AdpA, зібраних головно з геномів *Streptomyces*. Ідея дослідження — реконструювати філогенію білків і генів AdpA на основі методу, де всі параметри реконструкції будуть ідентичними, за винятком моделі заміщення (їх буде взято з таблиць 3.1 і 3.2 для генів і білків, відповідно). Тобто, K3Pu або моделі GTR мали б бути оптимальні для GC-багатих генів транскрипційних фаткорів *Streptomyces* (таблиця 3.1). На рисунку 3.1 представлено два альтернативних філогенетичних дерева гена AdpA. Одне дерево побудовано на основі оптимальної моделі GTR (A), інше — на основі неоптимальної моделі HKY (Hasegawa, Kishino and Yano 1985).

Аналогічно, реконструйовано філогенетичні дерева білків AdpA (Rabyk et al. 2018) з набору 200 амінокислотних послідовностей-ортологів цього білка. Як оптимальну модель використано JTT – згідно попередньої оцінки оптимальна для регуляторних білків (табл. 3.2). Як неоптимальну модель використано матрицю WAG, результати наведено на рис. 3.2.

а б

Рис. 3.1 Філогенетичні дерева на основі вирівнювання нуклеотидних послідовностей генів AdpA, побудовані з використанням різних моделей заміщення, GTR (**а**) і HKY (**б**). Дерева побудовано методом максимальної вірогідності на сервері phylogeny.fr (Dereeper et al. 2008), усі параметри за замовчуванням, за винятком вибору моделі заміщень (див. основний текст). Червоним відмічено кладу *Kitasatospora setae*, 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 зміни параметрів реконструкції).

Як випливає із аналізу отриманих реконструкцій, вибір оптимальної матриці має вплив на топологію дерева; помічені відмінності позначено на деревах. Багато змін помітно особливо при реконструкції на основі нуклеотидних послідовностей, при аналізі амінокислотних відмінностей менше. Загалом, наші результати підкреслюють важливість пошук оптимальних моделей заміщень.

а б

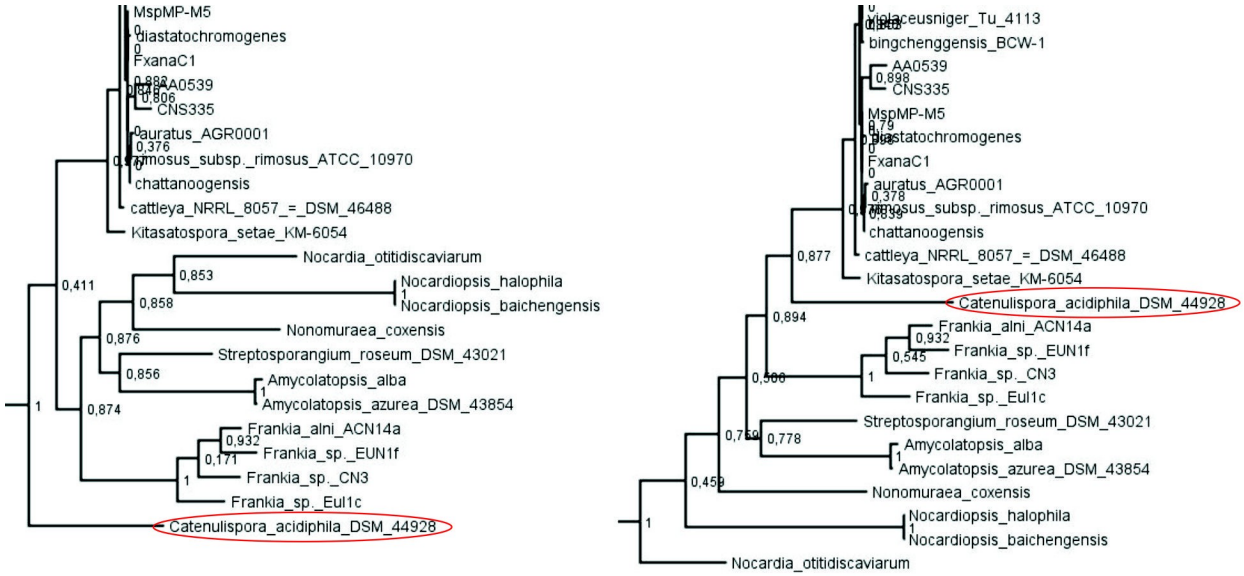


Рис. 3.2 Філогенетичні дерева на основі вирівнювання набору ортологічних білків родини AdpA, побудований з використанням матриць JTT (А) та WAG (Б). Червоним відмічено кладу родини *Catenulispora acidiphila* яка займає відмінну позицію в двох деревах. Інші умовні позначення – див. рис. 3.1.

3.2. Особливості контекстного вживання кодонів у геномах стрептоміцетів

Як розглядалося в розділі 1.4, 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 вживання пар кодонів можна обчислити й оцінити статистично. Спеціалізоване програмне забезпечення Anaconda (Moura 2007) дає змогу проаналізувати секвенований й анотований геном та підрахувати кількість кожної із можливих кодонних пар. Отримані дані підлягають статистичній обробці, яка показує імовірність того, що отримана величина може бути отримана унаслідок випадковості. Пари кодонів які зустрічаються з частотою меншою (за 2SD) ніж статистична випадковість, матимуть «негативний» контекст і вимірятимуться як від'ємне значення різниці кількості дикодону в межах розподілу та кількості цієї пари кодонів. Тоді «позитивний» контекст мають дикодони, частота зустрічності яких вища за випадкову у розподілі. Anaconda будує матрицю розміром 64 на 64 (її фрагмент – на рис. 3.3) де кожна комірка відповідає можливій кодонній парі і отримає забарвлення від

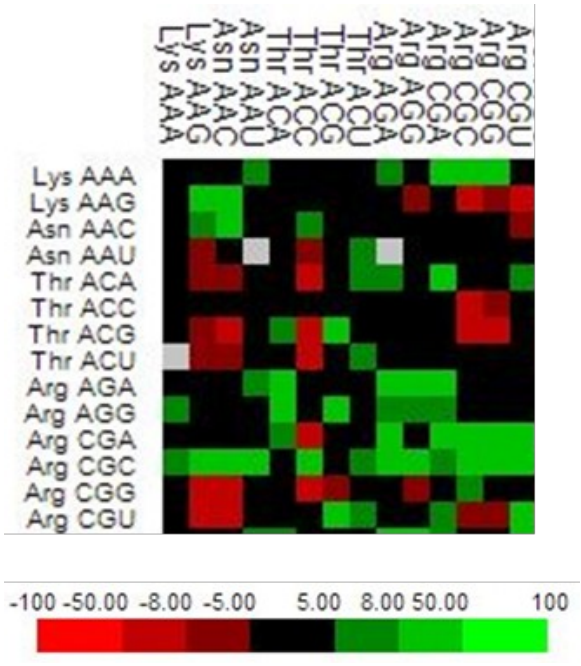


Рис. 3.3 Фрагмент кодонної теплової карти кодонних пар з генома *S. coelicolor*. Шкала відповідає статистичному розподілу від червоного відтінку («негативний» контекст) до зеленого («позитивний» контекст). Сірі комірки вказують на відсутність даних (див. т-ж основний текст).

червоного (негативний контекст) через чорний (відсутність контекстів – випадкова асоціація кодонів) і до зеленого (позитивний контекст).

Отримана “теплова” карта дозволяє візуально оцінити кількість дикодонів і особливості їхнього вживання. Важливо зазначити, застосований нами метод використовує критерій хі-квадрат Пірсона. Тобто, кожна кодонна пара порівнюється як різниця очікуваного та реального. Метод враховує частоти вживання кодонів при розрахунку очікуваних велечин, а отже ми отримуємо нормалізовані дані, які не залежать від GC-складу вхідних даних. Відповідно, у випадку багатих на GC-нуклеотиди геномів *Streptomyces*, зміщення до вживання GC-багатих кодонів не повинен впливати на отримані частоти контекстів дикодонів.

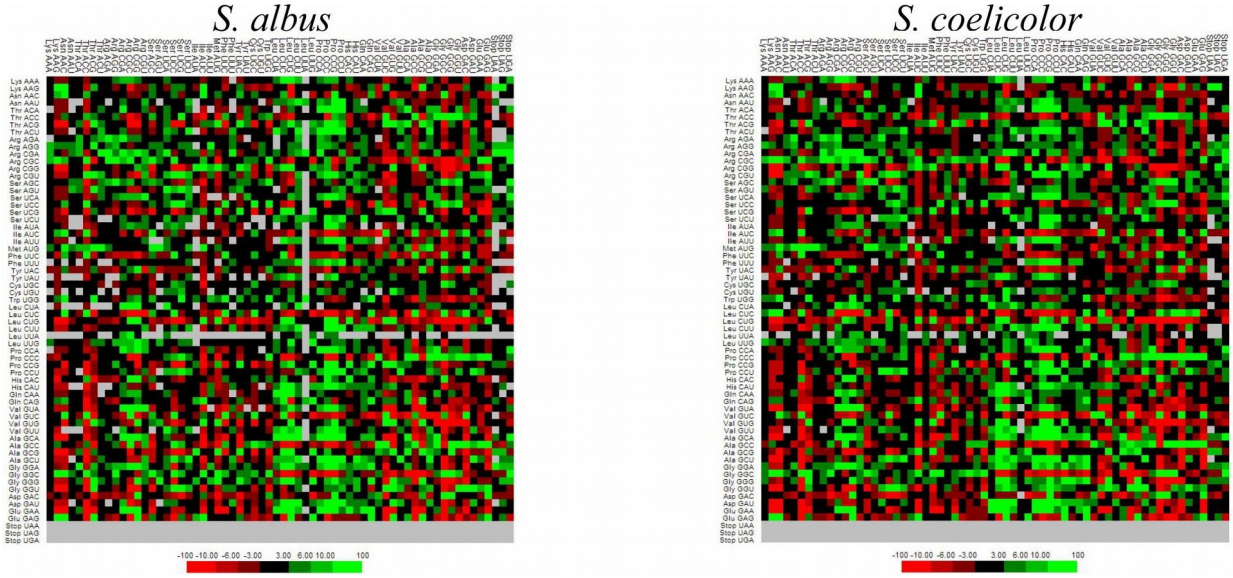


Рис. 3.4 Теплова карта геномів *Streptomyces albus* і *S. coelicolor*. Шкала відповідає такій на рис. 3.3.

Аналіз одного геному за допомогою Anaconda не показує вірогідної картини того, які кодонні пари переважають у геномах певного роду. Тому ми обрали 50 різних видів *Streptomyces* для аналізу й узагальнення контекстного вживання кодонів в геномах цього роду. Результати для видів *S. coelicolor* та

S. albus представлені на рисунку 3.4, результати по ще 6 видам винесено в Додаток 1, а решта — представлені у вигляді узагальненої картини (рис. 3.5).

Для того, щоб говорити про закономірності контекстного вживання кодонів в межах цілого роду стрептоміцетів, необхідно узагальнити отримані дані для кожного з 50 обраних видів. Для цього потрібно «накласти» отримані теплові карти, тобто підрахувати загальну кількість дикодонів для всіх 50 видів та обчислити статистичний розподіл для усередненої теплової карти. Також, необхідно вилучити малозначущі відмінності. Як результат ми отримаємо узагальнену теплову карту для 50 стрептоміцетних геномів (рис. 3.5). На цій карті відображено лише статистично значущі «негативні» й «позитивні» контексти на основі аналізу 50 видів, що, імовірно, відображає особливості вживання дикодонів для всього роду *Streptomyces*.

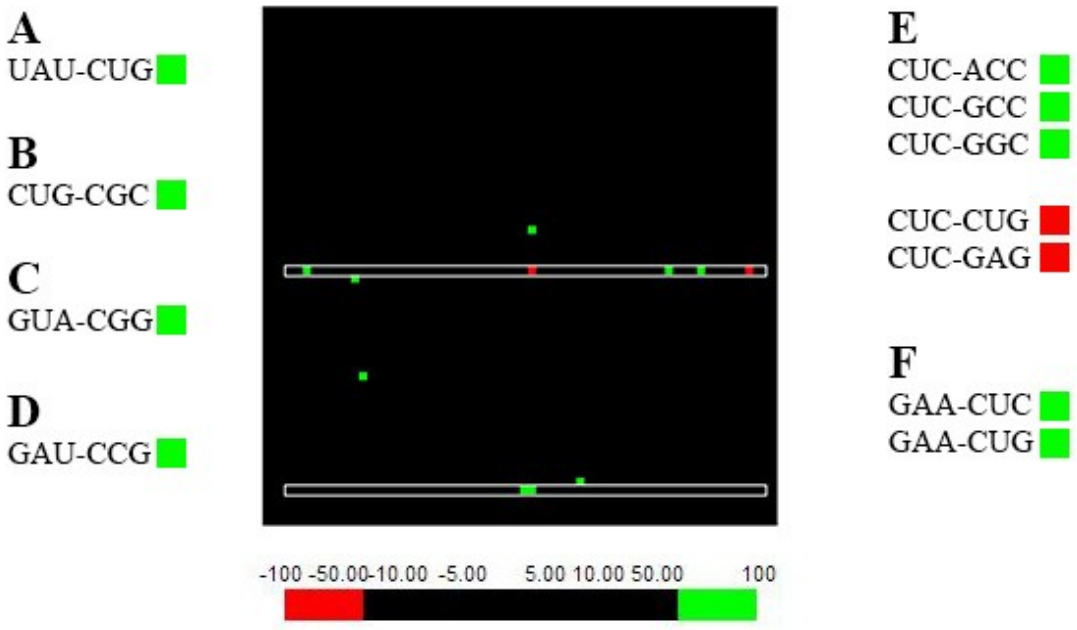


Рис. 3.5 Усереднена теплова карта 50 геномів *Streptomyces*. Для ідентифікації конкретних контекстів дикодонів карта була відфільтрована для відображення кодонів, кількість яких перевищує 50. Червоний – «негативний» контекст та зелений - «позитивний» контекст. Всі інші випадки (відсутність контексту) забарвлено в чорний колір. A, B, C, D – одиночні пікселі (з верху до низу), E і F–виділені рядки карти, відповідно.

Виявлено, що в стрептоміцетів 9 «позитивних» та 2 «негативних» контекстів виходять за межі двох стандартних відхилень нормального розподілу. Відповідно, отримані результати можна узагальнити у наступні патерни: «позитивні» — KAU-CYG (A та D рис. 3.5), GWA-CKS (C та F рис. 3.5), CUS-VSC (B та E рис. 3.5), та «негативний» — CUC-SWG (E рис. 3.5). Цікаво, що частина цих патернів 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. Згідно нашого аналізу наукової літератури, пояснення цих патернів наразі відсутні.

3.3. Візуалізація кодонних заміщень у геномах *Streptomyces*

Частина огляду літератури розкриває важливість пошуку математичних закономірностей в доступних геномних даних, а також низку підходів моделювання генетичних послідовностей (розділ 1.2.). Ми ставимо за мету вивчити як еволюціонував геном стрептоміцетів на кодонному рівні організації. Для цього необхідно змоделювати та візуалізувати закономірності кодонних заміщень. Ми вирішили створити власний веб-застосунок обчислення кодонних моделей на основі Xrate (Klosterman et al. 2007). На рисунку 3.6 представлені скріншоти.

Існує багато спеціалізованого програмного забезпечення для здійснення моделювання (Hoban, Bertorelle and Gaggiotti 2012). Ми зупинилися на програмному пакеті DART, зокрема програмі Xrate. Ця програма дає змогу будувати кодонні моделі шляхом тренування вихідної моделі M0 алгоритмом Очікування-Максимізації (EM: Expectation-maximization, Holmes and Rubin 2003). Такий алгоритм має два кроки, які циклічно повторюються поки результат не набуде максимальної статистичної вірогідності. На кроці очікування передбачаються частоти заміщення кодонів виходячи з моделей M0 (перший цикл роботи алгоритму) та наступних

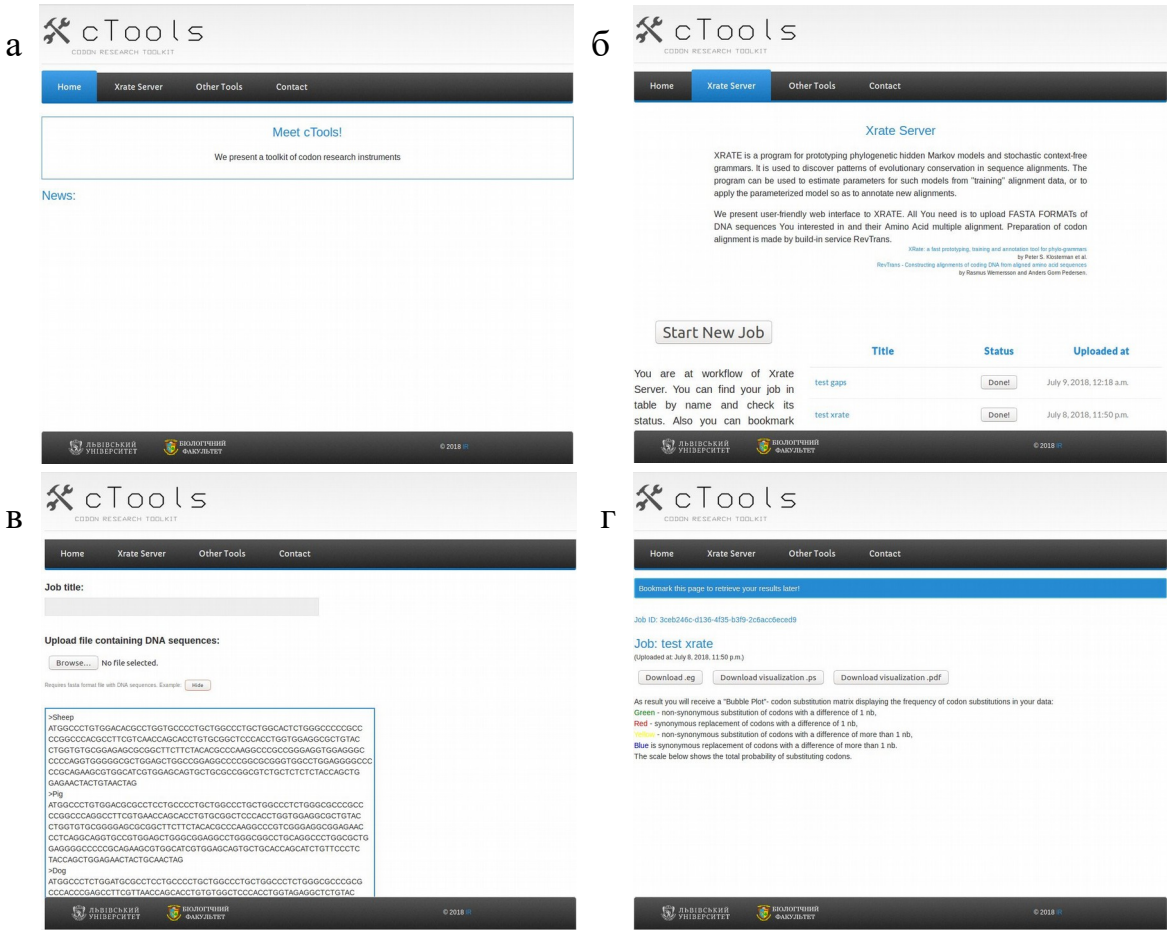


Рис. 3.6 Загальний вигляд веб-застосунку. а — домашня сторінка, з розділом новин; б — сторінка зі списком завантажених робіт; в — сторінка надсилання даних для візуалізації; г — сторінка з результатами (можливість завантажити необхідний формат).

натренованих моделей. Крок максимізації передбачає математичний розрахунок частот заміщення кодонів виходячи з порівняння первісного множинного вирівнювання і отриманої моделі на попередньому кроці.Якщо частоти розраховано статистично правдоподібно – і наступний цикл не покращує їх – то модель можна вважати натренованою. Кодонні моделі отримані в результаті роботи Xrate – це набір частот для всіх можливих переходів від одного кодону до іншого. Такі моделі можна подати як таблиці 61x61 кодон з частотою заміни відповідних кодонів на перетині рядків і стовпчиків. Частоти заміщень можна навести у вигляді кіл (“бульбашок”),

діаметр яких пропорційний частоті заміщення. Такі графіки відомі в англомовній літературі як bubble plots (бульбашкові графіки). Використали мову програмування Python для розробки серверної частини (бек-енд) та Javascript — для веб-кліента (фронт-енд). Веб-застосунок дозволяє завантажити кодонне множинне вирівнювання в форматі Stockholm чи Fasta, та отримати файл (*.eg) з даними натренованої моделі. Далі ці дані можна візуалізувати і отримати зображення “бульбашкового” графіка.

Використовуючи застосунок, ми отримали низку кодонних моделей для різних ортологічних груп генів з різних стрептоміцетів. Процедуру аналізу описано нижче. Спочатку підготували вибірку генів-ортологів, зокрема, ті які використано в розділі 3.1: ген *sco1728* кодує транскрипційний фактор YtrA-типу; *sco2706* – глікозилтрансфераза, схожа до низки інших ферментів задіяних в біосинтезі ліпополісахаридів. Також, використано ген *sven_4640* і його ортологи, що кодують трансмембранний білок-фліппазу II. Кодонні вирівнювання, необхідні для моделювання, отримали за допомогою програми RevTrans (Wernersson and Pedersen 2003), яка транслює нуклеотидні послідовності в білкові, які множинно вирівнюються і транслюються назад в вихідні кодони. Отримані кодонні вирівнювання підлягають аналізу і візуалізації (рис. 3.7 – 3.9).

Слід відзначити кілька особливостей використання bubble-plot-аналізу. В своєму аналізі ми зосередились на ортологах, однак застосунок проаналізує будь-яке кодонне вирівнювання, незалежно від його природи. Відтак, основа умова – отримання множинного вирівнювання, незалежно від філогенетичних зв’язків між послідовностями, що до нього входять. Однак, на нашу думку, масиви ортологічних груп генів несуть найбільше інформації і найлегше пояснити. Імовірно, другою за простотою та інформативністю групою будуть паралогічні родини. Гени, що кодують функціонально конвергентні групи білків – третя перспективна група для аналізу методом бульбашкових графіків. Загалом, будь-яка група послідовностей, яка має

спільний характер кодонних заміщень, може підлягати візуалізації нашим методом.

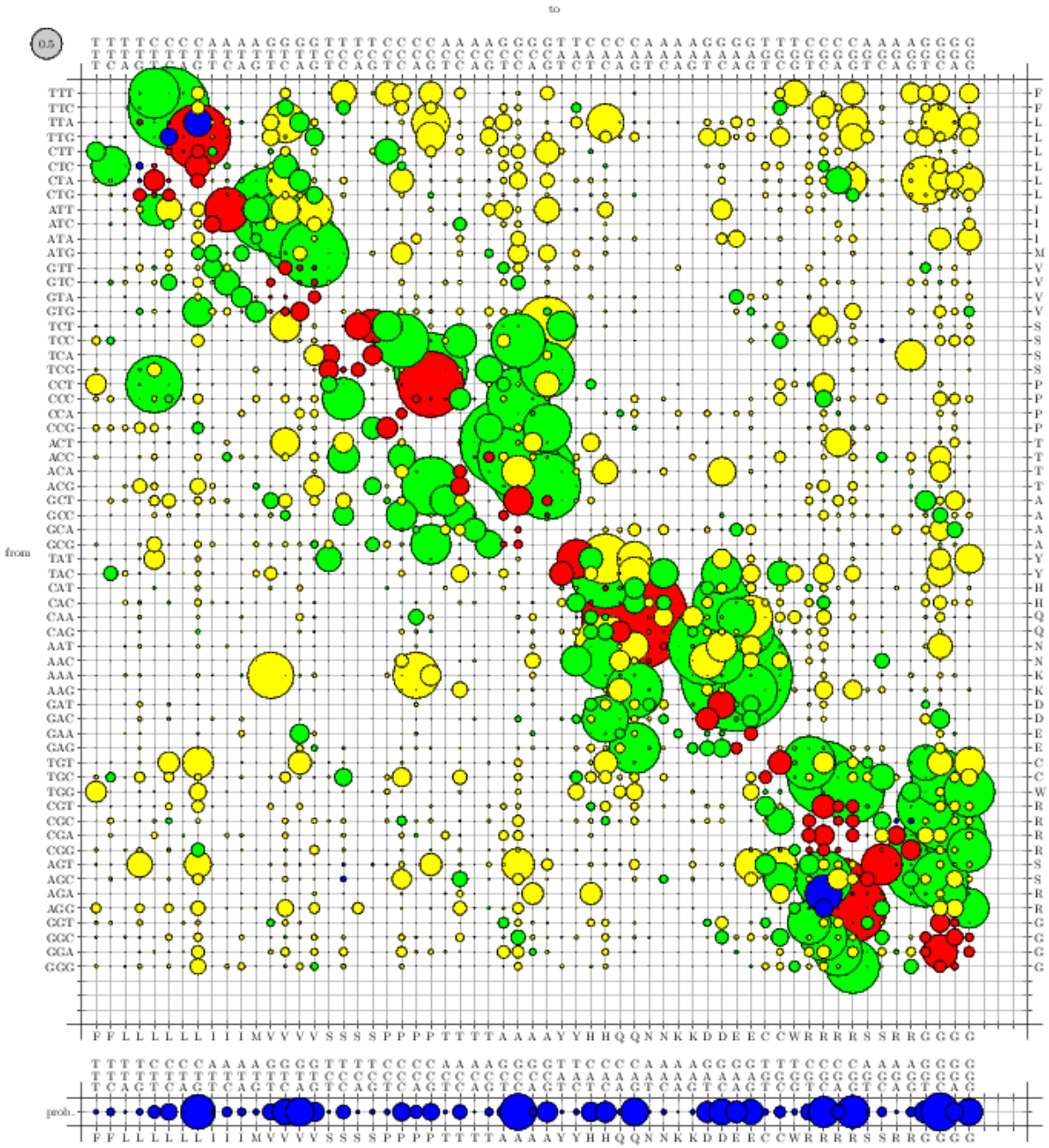


Рис. 3.7 Графік частоти заміщень кодонів в ортологічних генах фліппази II (*sven_4640*). Зелений — несинонімічні заміщення кодонів з різницею в 1 нк. Червоний — синонімічні заміщення кодонів з різницею в 1 нк. Жовтий — несинонімічні заміщення кодонів з різницею більше ніж 1 нк. Синій — синонімічні заміщення кодонів з різницею більше ніж 1 нк. Шкала знизу показує сумарну імовірність заміщення кодонів.

Пілотний аналіз отриманих графіків вказує на найзагальніші особливості еволюції послідовностей. Так, скупчення основної маси заміщень на діагональній лінії графіка вказує на відсутність суттєвих амінокислотних

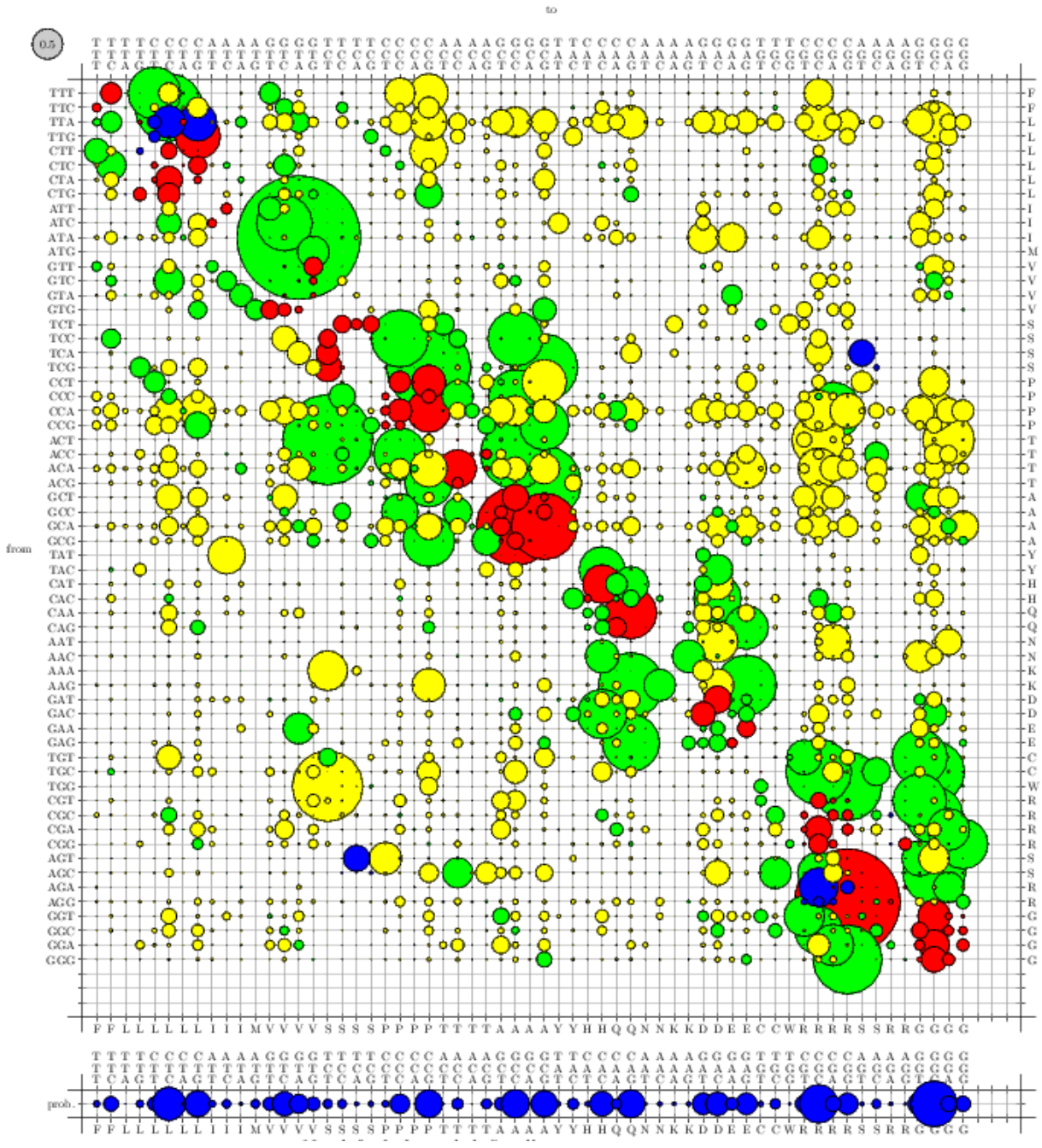


Рис. 3.8 Графік частоти заміщень кодонів в ортологічних генах транскрипційного фактора (*sco1728*). Кольоровий код – див. рис. 3.7. Шкала знизу показує сумарну імовірність заміщення кодонів.

замін. Останні розташовані у кутах графіка. Помітно, що в генах фліппаз домінують саме такі заміщення. Їх також багато у генах транскрипційних факторів родини YtrA (рис. 3.8), і відносно мало у генах глікозилтрансфераз

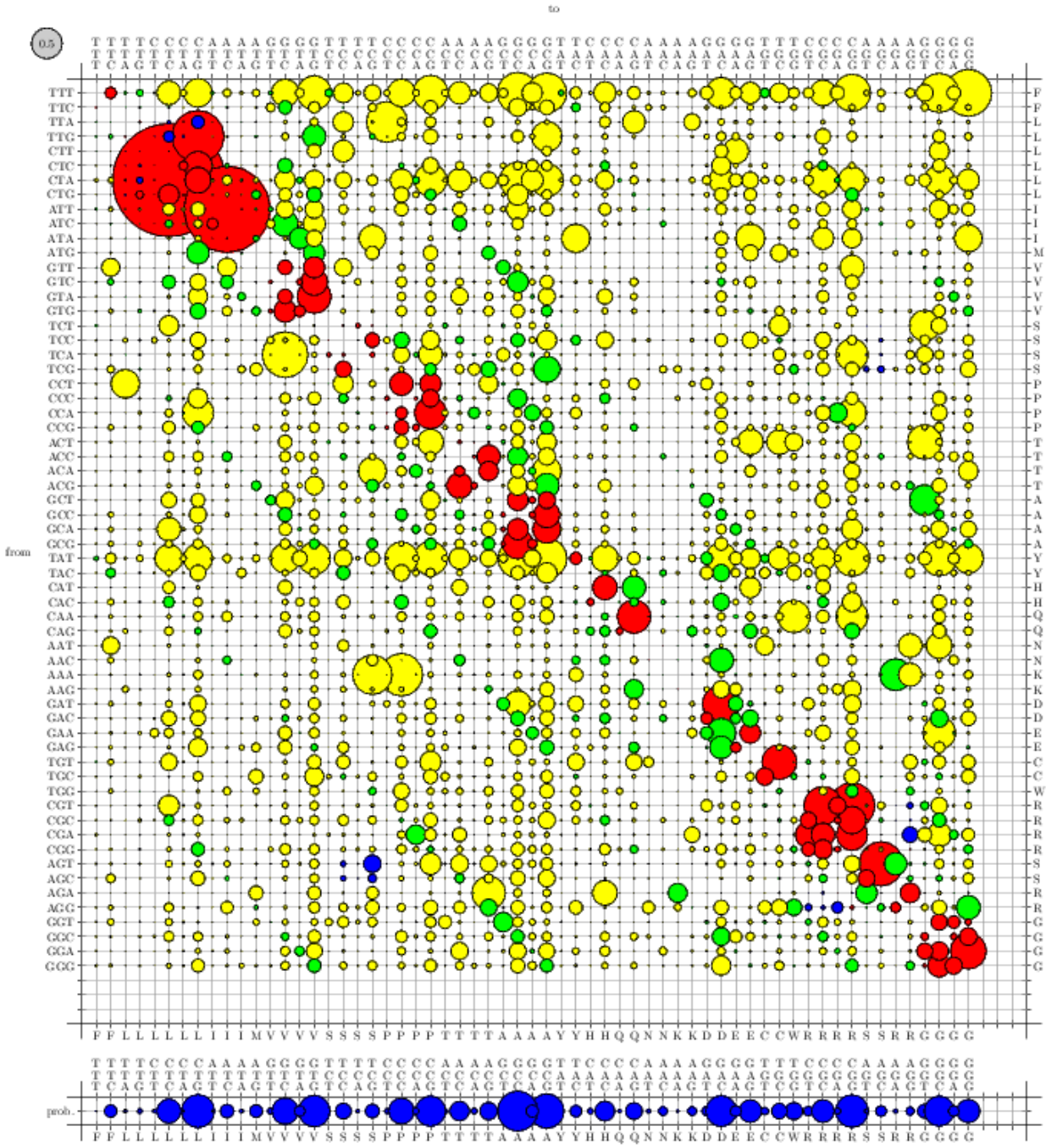


Рис. 3.9 Графік частоти заміщень кодонів в ортологічних генах глікозилтрансферази (*sco2706*). Кольоровий код – див. рис. 3.7. Шкала знизу показує сумарну імовірність заміщення кодонів.

з родини Sco2706. Низька “заселеність” діагональної лінії в останньому випадку свідчить про домінування заміщень, які ведуть в білку до нової амінокислоти з фізико-хімічними властивостями, відмінними від вихідної

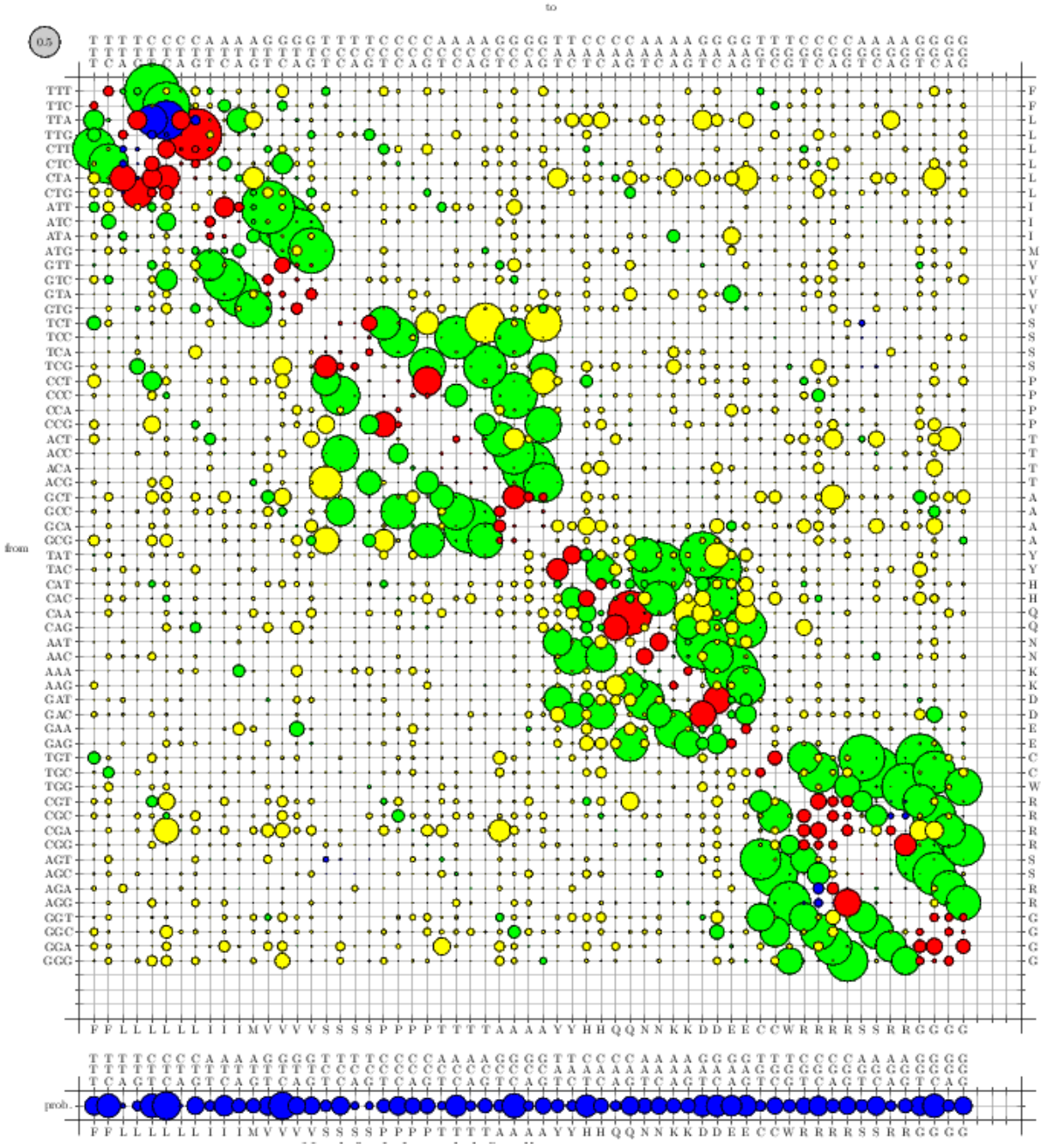


Рис. 3.10 Графік частоти заміщень кодонів в ортологічних генах великої субодиниці рибулозо-5-бісфосфат карбоксилази (*RuBisCo*). Кольоровий код – див. рис. 3.7. Шкала знизу показує сумарну імовірність заміщення кодонів.

амінокислоти. Важливою властивістю графіка є симетричність. А саме, в матриці наведено прямі і зворотні заміщення (напр., CCT→TTA; і TTA→CCT, див. рис. 3.9), і їхні значення різні. Тобто, на відміну від параметризованих чи емпіричних матриць заміщень нуклеотидних чи амінокислотних залишків, які симетричні (див. т-ж огляд літератури), використані у нашій роботі підходи дають змогу отримувати несиметричні значення заміщень – які вказують на переважний напрям заміщень, тобто процес еволюції генетичної послідовності. Побіжний огляд наведених нами результатів (див. рис. 3.7-3.9) вказує на їхню загальну несиметричність. Однак, це не є загальне правило “поведінки” ортологічних масивів даних – деякі ортологічні групи виявляють набагато вищу симетричність. Прикладом такої групи є гени великої субодиниці рибулозо-5-бісфосфат карбоксилази (рис. 3.10). Це вказує на переважання стабілізуючого добору таких послідовностей. Наразі графіки відображають якісну картину кодонних заміщень, і їхня кількісна оцінка вимагає звернення до табличних значень (на основі яких будуються бульбашки). Однак, ми передбачаємо дальший розвиток цього сервісу, який дасть змогу використовувати табличні значення для відображення діаметру бульбашок, а також сумуваня усіх заміщень одного типу (наприклад, для визначення загальної кількості усіх однонуклеотидних синонімічних заміщень). Отже, ми отримали бульбашкові графіки, таблиці 61x61 кодон, де кола відображають частоту заміщення через діаметр. Різні кольори кіл вказують на синонімічність та несинонімічність заміщень. Ми можемо спостерігати першочергово за різними патернами, характерними для білків з різною функцією. Також, кардинально відрізняється частота синонімічних та несинонімічних заміщень в генах різних білків. У фокусі нашого зацікавлення знаходяться гени актинобактерій, однак створений веб-застосунок можна використовувати для будь-яких груп генів.

3.4. Точність трансляції рідкісного лейцинового кодона TTA у стрептоміцетів

В попередніх розділах обговорювалася унікальність геномів стрептоміцетів стосовно екстремально низької частоти вживання кодону TTA. 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 (Hackl and Bechthold 2015). 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, існує суперечність: TTA має бути рідкісний, аби контролювати лише певні гени, і має транслюватися лише однією тРНК (якщо інші тРНК його декодуватимуть – не буде регуляторного ефекту), і тим, що рідкісні кодони зазвичай неточні. Однак, припущення про неточність рідкісних кодонів встановлено більше 20 років тому на обмеженому колі кодонів у модельних евкаріотичних організмів, або бактерій нестрептоміцетного походження. На сьогодні існують реалістичніші моделі трансляції, які дають змогу моделювати точність декодування різних кодонів на основі геномних даних (Shah and Gilchrist 2010). Нами вирішено застосувати ці моделі для передбачення точності трансляції рідкісного у стрептоміцетів лейцинового кодона TTA.

Варто зазначити, що ми будемо оперувати кількістю копій генів тРНК (ККГ) як показником концентрації тРНК в клітині, оскільки в роботі dos Reis (2004) встановлено високу позитивну кореляцію між цими показниками. Отже, концентрація акцепторної тРНК в клітині залежатиме від ККГ цієї тРНК і далі позначатиметься як *tF* (ККГ фокальної тРНК). Тоді *tN* – ККГ

близько-спорідненої тРНК, антикодон якої відрізняється одним нуклеотидом від акцепторної тРНК.

Ми визначили й проаналізували набір генів тРНК шести стрептоміцетних геномів: *S. coelicolor* M145, *S. albus* J1074, *S. ghanaensis* ATCC14672, *S. clavuligerus* ATCC27076, *S. venezuelae* ATCC14115та *S. lividans* TK24. Отримані розрахунки *tF* та *tN* для тРНК, що декодують лейцинові кодони для об'єктів нашого дослідження, наведено в таблиці 3.3., а докладні результати по всіх амінокислотах — у Додатку. В таблиці відсутні дані для лейцинового кодону AAG оскільки в геномах досліджуваних видів не має відповідних генів тРНК, а декодування відбувається за рахунок хиткої (wobble) взаємодії з іншими тРНК (т.зв. псевдо-акцепторні тРНК).

Таблиця 3.3

Концентрації акцепторної (tF) та близькоспоріднених (tN) тРНК в клітинах шести видів *Streptomyces*

Leu	*S. coelicolor*		*S. albus*		*S. ghanaensis*		*S. clavuligerus*		*S. venezuelae*		*S. lividans*	
	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN
UAA	1	3	1	3	1	3	1	3	1	3	1	3
CAA	1	10	1	11	1	10	1	10	1	9	1	10
GAG	2	9	2	9	3	8	2	9	3	8	2	7
UAG	1	2	1	3	1	3	1	3	1	3	1	2
CAG	1	6	1	6	1	6	1	6	1	6	1	6

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 (які мають найвищу імовірність порушити комплементарну взаємодію —

відмінність одним нуклеотидом). Тому ми застосували обчислювальну модель (Shah and Gilchrist 2010), що показує співвідношення споріднених та близько-споріднених тРНК (*tF*/*tN*).

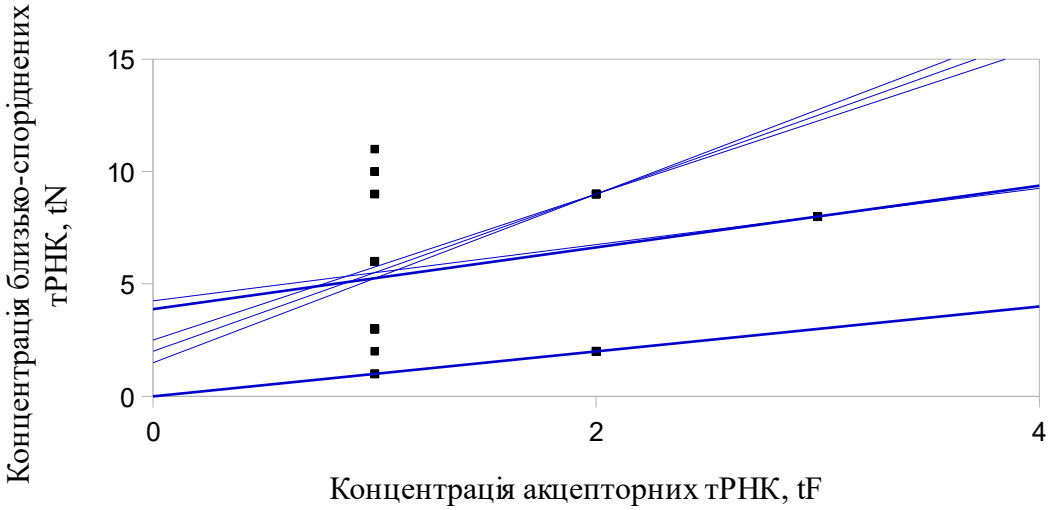


Рис. 3.11 Графік кореляцій між концентраціями фокальної (tF) та близько-споріднених (tN) тРНК для лейцинових кодонів. Синіми лініями позначено лінії регресії між tN й tF для кожного з аналізованих геномів. tF визначалося з ККГ (див. основний текст вище).

На рисунку 3.11 показано кореляційні зв'язки для лейцинових кодонів для досліджуваних стрептоміцетів. Вірогідна позитивна кореляція між концентраціями tF та tN тРНК в клітині для різних Leu кодонів вказує на однакову частоту виникнення помилок незалежно від «популярності» кодону. Схожу картину можна спостерігати для амінокислот з двома синонімічними кодонами (рис. 3.12). Однак, для амінокислот з чотирма синонімічними кодонами ми спостерігали як позитивну (сині лінії),

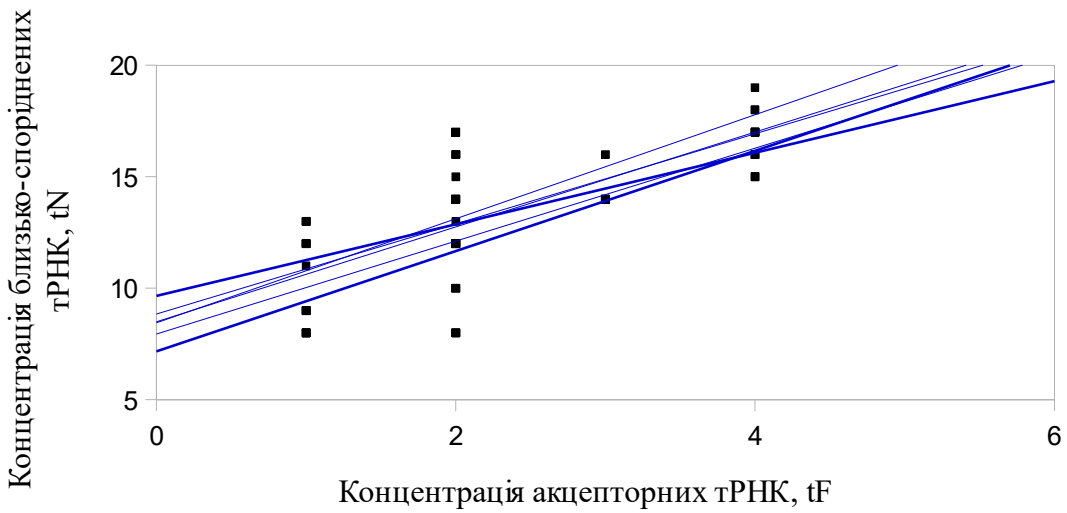


Рис. 3.12 Кореляція між концентраціями tF та tN для амінокислот з двома синонімічними кодонами. Умовні позначення – див. Рис. 3.11.

так і негативну кореляції (червоні лінії, рис. 3.13). А саме, для *S. coelicolor*, *S. lividans* та *S. clavuligerus* характерна негативна кореляція концентрацій

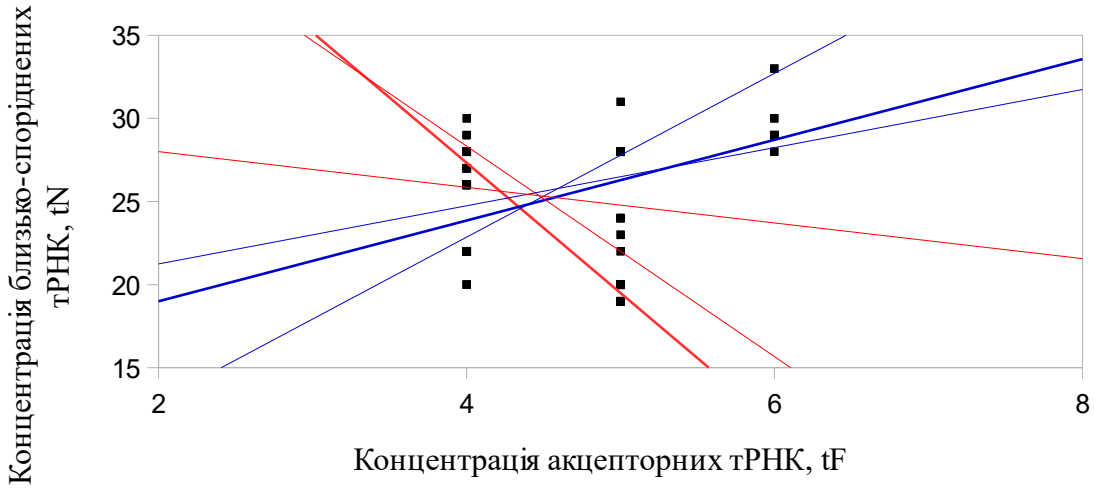


Рис. 3.13 Кореляція між концентраціями tF та tN для амінокислот з чотирма синонімічними кодонами. Сині лінії регресії вказують на позитивну кореляцію, а червоні – на негативну.

акцепторних та близько-споріднених тРНК для групи амінокислот, що мають по чотири акцепторні кодони — червоні лінії на рисунку 3.13. Також, у *S. lividans* немає кореляції для групи амінокислот, що мають по шість акцепторних кодонів (чорна лінія рис. 3.14).

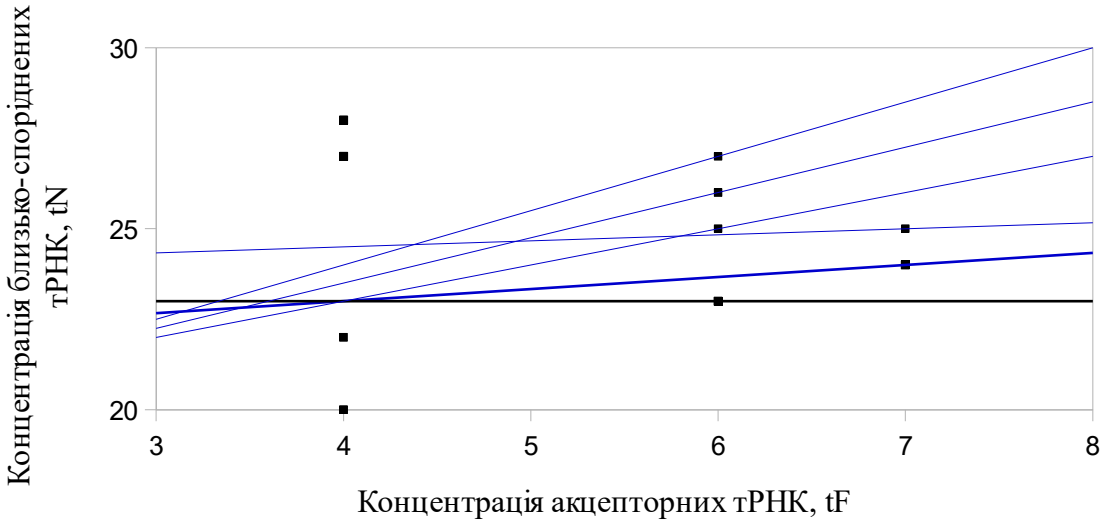


Рис. 3.14 Кореляція між концентраціями tF та tN для амінокислот з шістьма синонімічними кодонами. Чорна лінія регресії вказує на відсутність кореляції у *S. lividans.* Умовні позначення – див. рис. 3.11.

Висновки про точність трансляції можна зробити через порівняння розрахунків швидкостей елонгації кодонів за допомогою акцепторних та псевдоакцепторних тРНК у порівнянні з близько-спорідненими тРНК, які призводять до містрансляції. Псевдоакцепторними тРНК вважатимемо такі тРНК, які не є акцепторними, але можуть правильно декодувати цільовий кодон за рахунок хиткої взаємодії. Оскільки при трансляції акцепторні та псевдоакцепторні тРНК не призводять до виникнення помилок їхній вплив на елонгацію рахуватимемо сумарно, але із врахуванням коефіцієнта імовірності хиткої взаємодії. Відповідно при обрахунку елонгації

послідовності за рахунок близько-споріднених тРНК, ми не враховуватимемо тРНК що відрізнятимуться одним нуклеотидом, але є псевдоакцепторними. Таку вибірку можна розглядати як неакцепторні тРНК. Формули для акцепторних та псевдо-акцепторних тРНК:

$$R_c(i) = a\left(\sum_{j \in SET_c} t_j p_c w_{j,i} + \sum_{j \in SET_p} t_j p_p w_{j,}\right), \quad (3.1)$$

та для близько-споріднених (неакцепторних) тРНК:

$$R_n(i) = a \sum_{j \in SET_n} t_j p_n w_{j,i}, \quad (3.2)$$

де R_c та R_n – це показники елонгації кодону i, a – швидкість нарощування поліпептидного ланцюга (10,992 c^{-1}), t_j – кількість копій генів тРНК з антикодоном j, p_c, p_p та p_n - імовірності приєднання відповідної тРНК (c – спорідненої $6{,}52*10^{-1}$; p – псевдо-спорідненої та n – неспорідненої $6{,}2*10^{-4}$) та $w_{j,i}$ – коефіцієнт-поправка для врахування вобл-взаємодії між кодоном та антикодоном (RR/YY – 0,61; RY/YR – 0,64).

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

Результати розрахунків для елонгації лейцинових кодонів за допомогою близько-споріднених тРНК наведені у вигляді теплової карти на рисунку 3.15. Карта показує, що рідкісний кодон TTA має низькі показники швидкості декодування неакцепторними тРНК і відповідно має меншу імовірність містранслюватися ніж інші (набагато популярніші) кодони. Отже, кодон може бути рідкісним, йому може відповідати низька ККГ, і все ж він може транслюватися не менш точно, ніж популярні кодони (Rokytskyy et al. 2016).

3.5. Репортерна система для вивчення експресії рідкісного лейцинового кодона TTA у стрептоміцетів

Як неодноразово зазначалось, стрептоміцетам притаманні аномальні ПВК для лейцинового TTA кодону. В попередніх підрозділах розглядалися теоретичні підґрунтя такого явища. Наступним питанням є створення простої системи вивчення вплив ефективності трансляції цього кодона на метаболізм стрептоміцетів.

Описаний Bld-фенотип (від англ. bald – “голомозий”) як результат делеції гена *bldA*, що кодує тРНК для декодування TTA кодону (розділ 1.5.2.). Цей фенотип полягає у відсутності повітряного міцелію та спор на поверхні субстратного міцелію актиноміцетної колонії. Варто зазначити, що за цим фенотипом прихований складний каскад генів, білків та низькомолекулярних сполук, які можуть мати комбінаторний вплив на прояв Bld-фенотипу. Важливо сконструювати систему з найкоротшим шляхом від кодону до фенотипу, що, в ідеалі, постає унаслідок експресії одного гена. Обраний ген має бути таким, шоб можна було легко виявити активність його білкового продукту і очистити його. Це в свою чергу дозволить кількісно судити про вплив рідкісного кодону на трансляцію безпосередньо. Також,

помістивши таку конструкцію в штам з делетованим геном *bldA*, матимемо змогу вивчити вплив відсутності декодуючого елемента при трансляції.

Streptomyces albus J1074 – один з небагатьох стрептоміцетів, що нездатний розшеплювати хромогенний аналог лактози —5-Br-4-Cl-3-β-D-галактопіранозид, або X-Gal (King та Chater 1986). Тому J1074 можна використати як платформу для вивчення ролі TTA. Необхідно знайти ген β-галактозидази актиноміцетного походження, що експресуватиметься в J1074, та створити його TTA-вмісну версію (якщо природний ген не містить цього кодона). Ми обрали ген *S. coelicolor* M145 *sco3479*, що, серед усіх *sco*-генів, найподібніший до гена β-галактозидази *lacZ* з *E. coli*. Ген *sco3479* містить низку Leu кодонів на початку гена, які можна замінити на TTA.

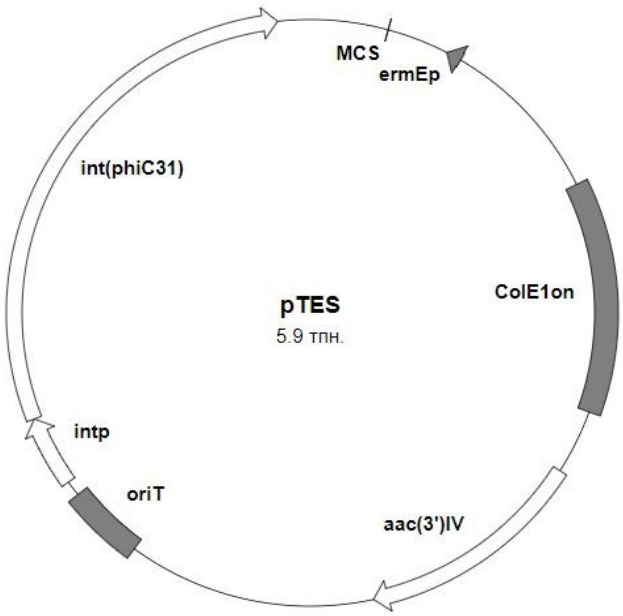


Рис. 3.16 Генетична карта плазміди pTES, на якій позначено її основні функціональні елементи: OriT, ген стійкості до ампіцеліну, ген інтегрази int-phiC31, сильний промотор *ermEp,* та полілінкер.

Ген *sco3479* слугував матрицею для ПЛР). Використовуючи праймери Sco3479upBglII та LacZ_XbaI_TGA2, а також “зворотній” праймер Sco3479rpMfeIBglII (табл. 2.3.) ми отримали три фрагменти *lacZ* із різними сайтами рестрикції на різних кінцях. Це дало змогу сконструювати три різні плазміди на основі вектора pTES (рис. 3.16). Цей вектор містить ген інтегрази int-phiC31, сильний промотор *ermEp* та полілінкер. *Ген sco*3479 із промотором *sco3479p*, утворений за участі праймерів Sco3479upBglII та Sco3479rpMfeIBglII, ми обробили ендонуклеазою рестрикції BglII та лігували із плазмідою pTES, оброблену ендонуклеазами BamHI та BglII. В результаті ми отримали плазміди pOOB109 та pOOB110, що містять ген β-галактозидази у двох можливих орієнтаціях відносно вкороченого промотора гена *ermE*. Перша плазміда, pOOB109 (рис. 3.17), містить фрагмент *sco3479p-sco3479* в одній орієнтації з *ermEp.*

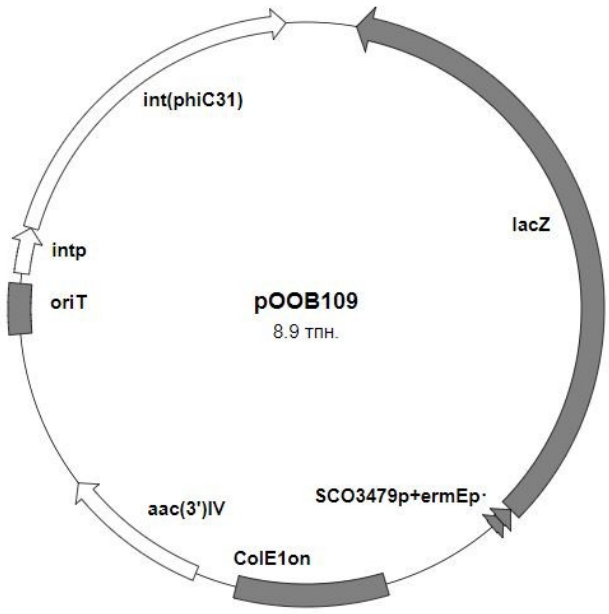


Рис. 3.17 Карта плазміди pOOB109, що сконструйовано на основі pTES. Плазміда pOOB110 (рис. 3.18) сконструйована аналогічно. Відмінність в орієнтації гена *sco3479*, тому він контролюється власним промотором.

Друга плазміда, pOOB110 (рис. 3.18), містить ген фланкований промоторами *ermEp sco3479p*, так що транскрипція гена *sco3479* здійснюється лише з промотора *sco3479p*.

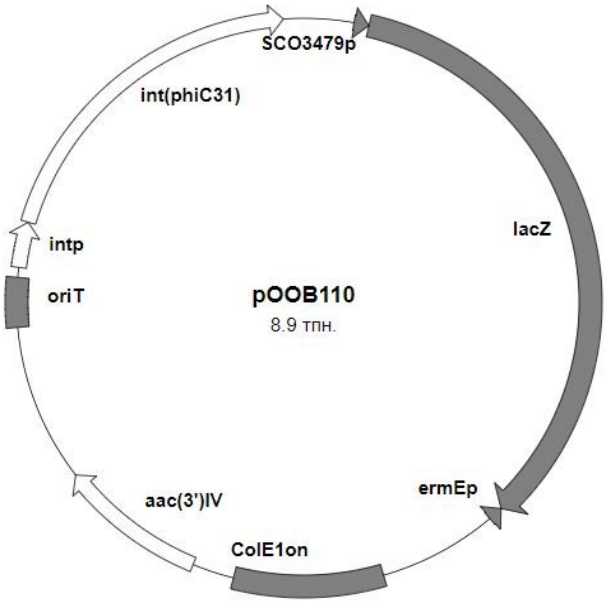


Рис. 3.18 Генетична карта плазміди pOOB110. Умовні позначення – див. рис. 3.17.

Правильність конструкцій перевірено шляхом картування ендонуклеазами рестрикції (рис. 3.19) та секвенування. Згідно описаної методики використовували ендонуклеазу рестрикції EcoRI. Сайти, що розпізнаються цією рестриктазою, містятся на початку гена *lacZ* та полілінкері плазміди pTES. Відповідно, розмір фрагмента залежатиме від орієнтації гена *lacZ* (див. схема на рисунку 3.19). Для плазміди pOOB109, спостерігаємо фрагмент розміром 0.5 т.п.н., а для плазміди pOOB110 — 2.4 т.п.н.

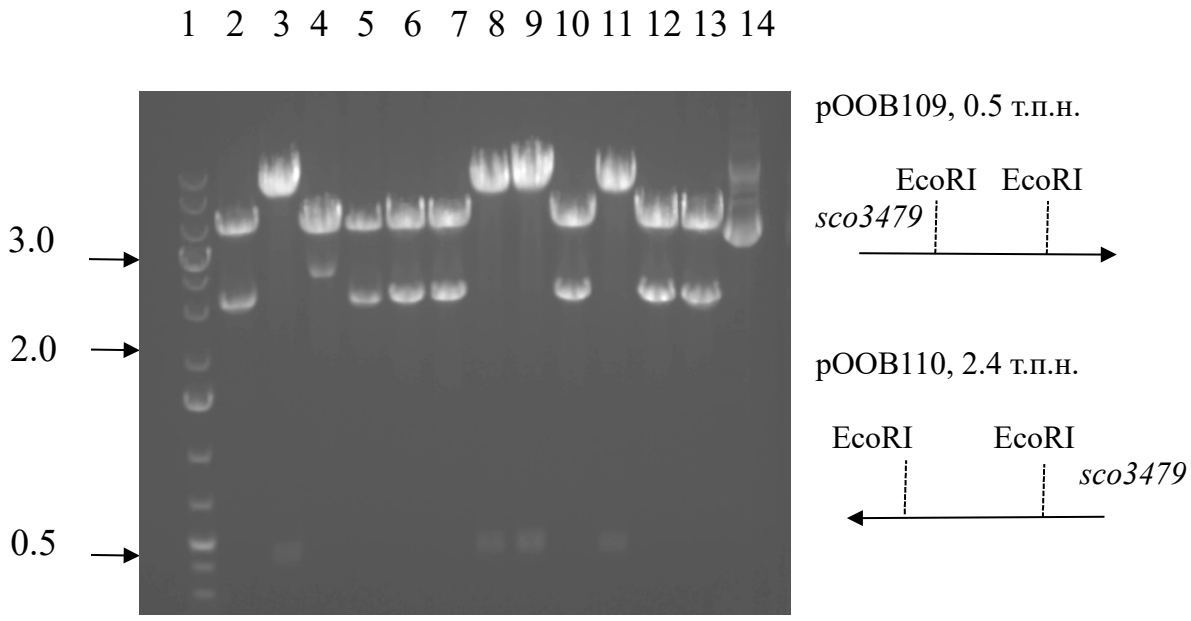


Рис. 3.19 Електрофореграма обробки клонів pTES+*sco3479* ендонуклеазою рестрикції EcoRI. Доріжки: 1 – молекулярний маркер; 14 – нативна плазміда; 2, 5-7, 10, 12, 13 – фрагмент *sco3479* в орієнтації pOOB110 (див. схему праворуч, крайній EcoRI-сайт – з полілінкера вектора); 3,8, 9, 11 – фрагмент *sco3479* в орієнтації pOOB109.

Для негативного контролю ми сконструювали плазміду pOOB114 (рис. 3.20), яка несе не функціональний варіант гена *sco3479*: в 19-ій позиції кодуючої послідовності знаходиться стоп-кодон — TGA. Необхідний ПЛР-продукт ми отримали використовуючи праймери LacZ_XbaI_TGA2, Sco3479rpMfeIBglII та pOOB109 як матрицю. Отриманий фрагмент, оброблений ендонуклеазами XbaI та MfeI, клоновано в сайти XbaI/EcoRI плазміди pTES, і так отримано pOOB114.

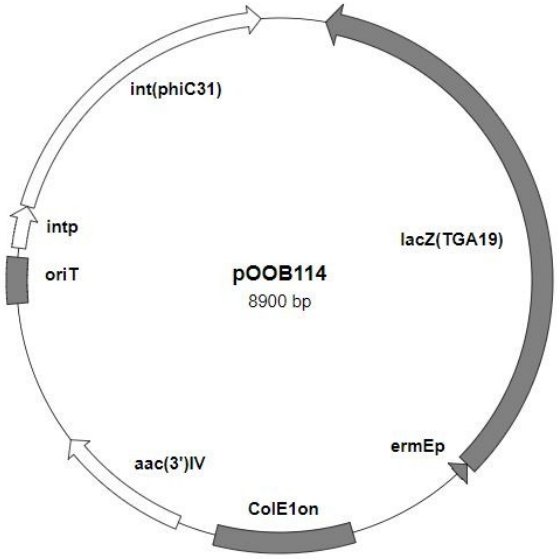


Рис. 3.20 Генетична карта плазміди pOOB114. Умовні позначення – див. рис. 3.17.

Структуру плазміди підтверджено картуванням (рис. 3.21).

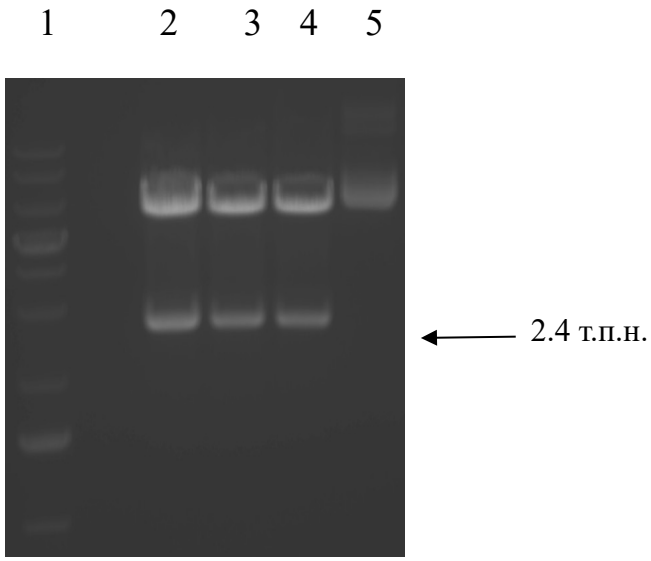


Рис. 3.21. Електрофореграма обробки трьох клонів pTES+TGA-вмісний *sco3479* ендонуклеазами рестрикції XbaI-EcoRI. Доріжки: 1 – молекулярний маркер; 5 – нативна плазміда; 2-4 – рекомбінантна плазміда з геном *sco3479* (позначено стрілкою).

Отже, ми сконструювали низку плазмід (pOOB109, pOOB110 та pOOB114), якими трансформовано кон'югативний штам *E. coli* WM6026. Виявлено, що ген *sco3479* ($lacZ_{sc}$) здатний до невисокого (але помітного) рівня експресії в кишковій паличці, про що свідчить слабке синє забарвлення колоній на індикаторному середовищі (рис. 3.22 А). Далі плазміди перенесено в *S. albus* (рис. 3.22 Б). Видно, що штами без β-галактозидазної активності отримали змогу розщеплювати аналога лактози — X-Gal в середовищі, за рахунок експресії гена $lacZ_{sc}$ із перенесених плазмід.

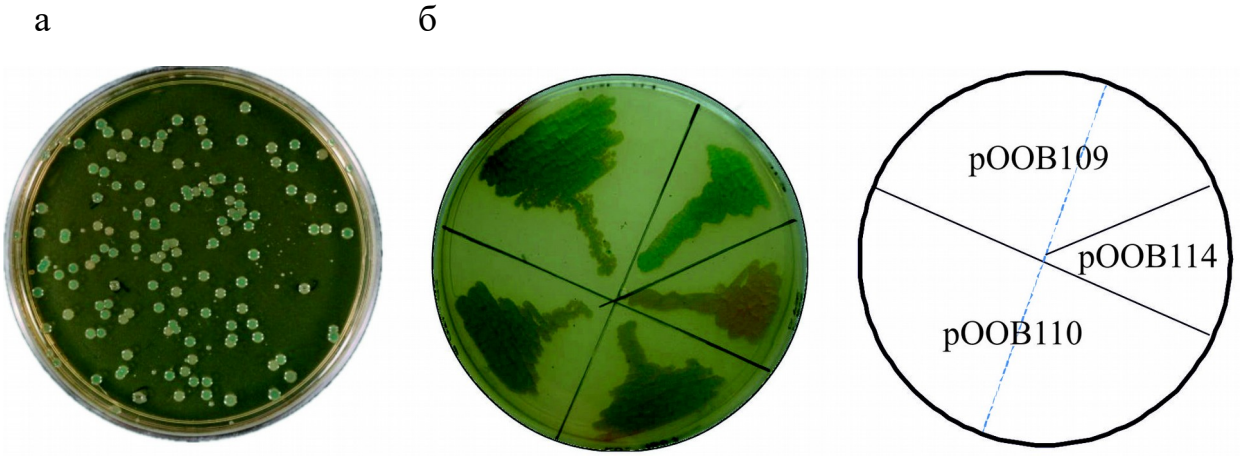


Рис. 3.22 Фотографія газонів колоній *E. coli* (А) та *S. albus* (Б) із внесеними плазмідами pOOB109, pOOB110 та pOOB114 на середовищі TSA з додаванням X-Gal (30 мМ). а — прояв β-галактозидазної активності (сині колонії) *E. coli* pOOB109. б — Прояв β-галактозидазної активності (синє забарвлення) *S. albus* з плазмідами pOOB109, pOOB110 та відсутність активності у контрольному штамі з плазмідою pOOB114.

Наступний крок – потрібно запропоновану вище систему зробити TTA-кодон специфічною, тобто один з лейцинових кодонів в межах кодувальної послідовності гена *sco3479* замінити лейциновим кодоном TTA. Також, варто забезпечити контрольовану експресію репортерного гена.

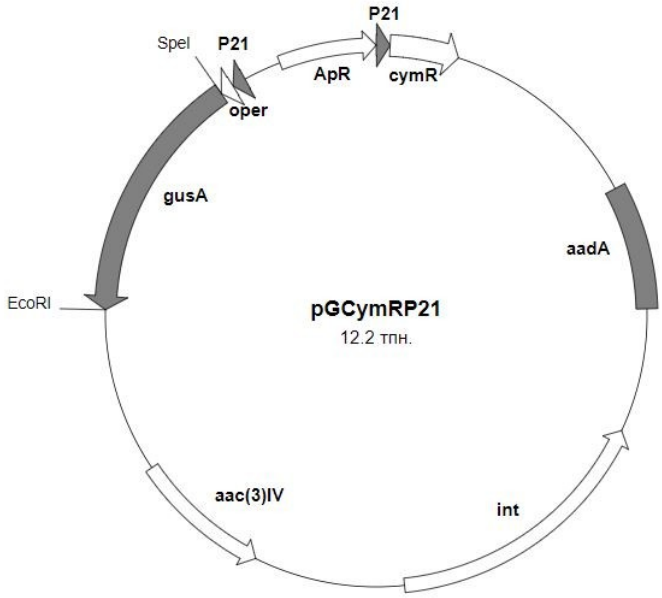


Рис. 3.23 Генетична карта плазміди pGCymRP21. Умовні позначення – див. основний текст.

Таким чином, за основу ми обрали плазміду pGCymRP21 (рис. 3.23), яка містить добре описаний механізм куматної репресії за допомогою гена cymR та операторної ділянки (Horbal et al. 2014). Кумат, доданий в середовище, зв’язуватиметься з репресором CymR, що вестиме до вивільнення оператора. Зняття репресії з оператора запустить експресію цільового гена, у випадку плазміди pGCymRP21 це — *gusA*, β-глюкоронідаза. Отже, необхідно замінити ген *gusA* на варіант гена *sco3479* із TTA кодоном на гексагістидиновою послідовністю на С-кінці.

Необхідний фрагмент ампліфікували із плазміди pOOB109, використавши праймери LacZ_XbaI_TTA1 та Sco3479_6Hisrp. Отриманий фрагмент довжиною 3017 п.н. є ген *sco3479* з TTA кодоном замість CTC в 8-ій позиції та шістьма гістидиновими кодонами CAC після стоп кодону (рис 3.24). Також, ген має 2 необхідні сайти рестрикції ендонуклеазами XbaI та MfeI і має змогу лігуватися по сайтах рестрикції SpeI та EcoRI в pGCymRP21.

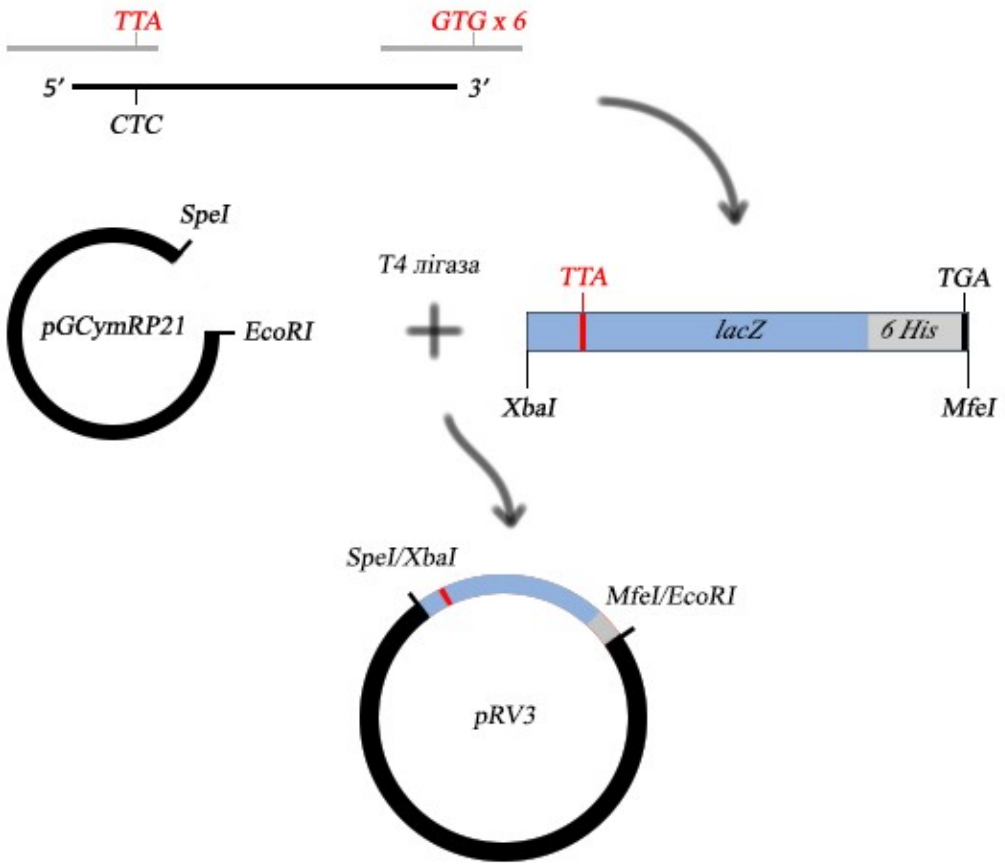


Рис. 3.24 Конструювання фрагменту *sco3479* із TTA кодоном та його клонування у вектор *pGCymRP21*.

В результаті ми отримали плазміду pRV3 (рис 3.24) з геном *sco3479*, що містить лейциновий TTA кодон у 81 позиції (замість CTC) та 6-His таг. Ген знаходиться під кумат-індуцибельним промотор *cmtR-P21*, а відповідно синтез β-галактозидази відбуватиметься за наявності кумату у середовищі. Плазміда містить маркерні гени резистентності до ампіциліну, апраміцину та спектиноміцину, що було використано в ролі маркерів під часконструювання та селекції конструктів. Докладну генетичну карту наведено на рис. 3.25.

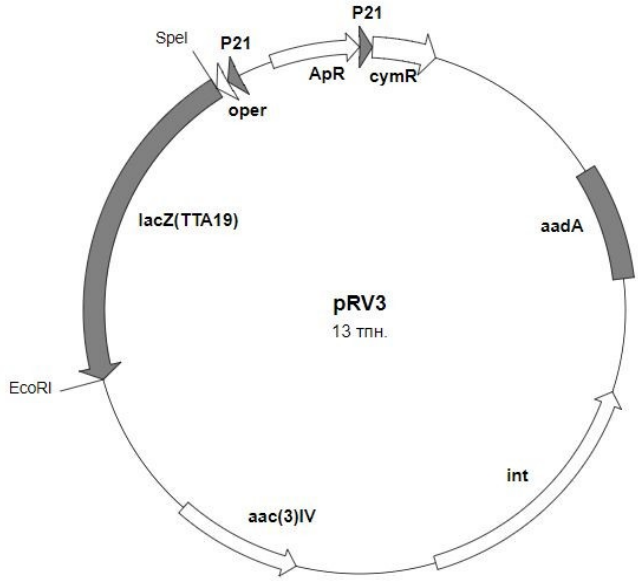


Рис. 3.25 Генетична карта плазміди pRV3. Умовні позначення – див. Вище

Аналогічно створено плазміду pRV4 (рис 3.26). Тут як донорний фрагмент для лігування отримано з використанням праймерів Sco3479upXbaI та Sco3479_6Hisrp. Відповідно ми отримали плазміду із геном *sco3479* дикого типу, що має 6-His таг.

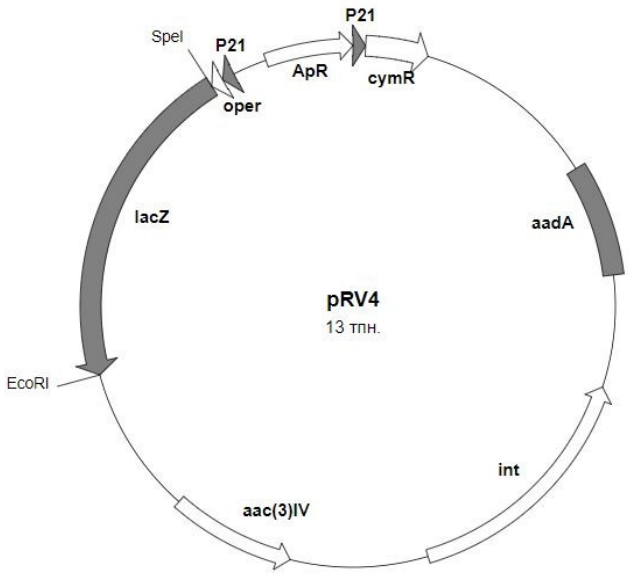


Рис. 3.26 Генетична карта плазміди pRV4. Умовні позначення – див. Вище

Далі pRV3 та pRV4 перенесено за допомогою електропорації в кон'югативний штам *E. coli* WM6026, а звідти – в *S. albus* J1074 (SAM2). Ми довели коректність нуклеотидної послідовності генів *sco3479* у плазмідах pRV3 й pRV4 за допомогою секвенування.

Картували плазміди ендонуклеазами рестрикції XbaI-EcoRI, в результаті дії яких, утворюється два фрагменти плазміди: один довжиною 2.6 т.п.н. (послідовність гена *sco3479*), а інший – близько 10.5 т.п.н. (рис. 3.27, доріжки 2, 3 та 5).

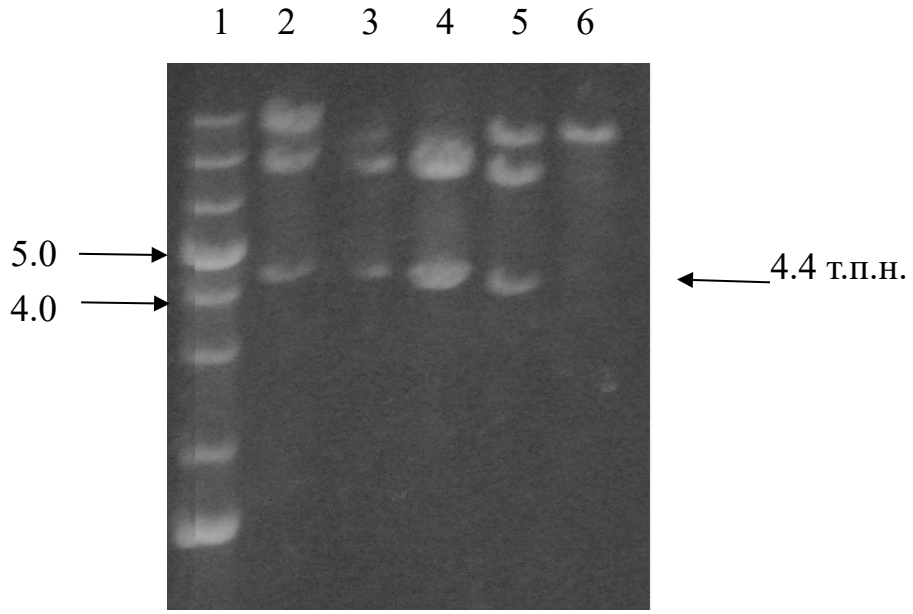


Рис. 3.27 Електрофореграма обробки п’яти клонів pGCymRP21+TTA-вмісний *sco3479* ендонуклеазами рестрикції XbaI-EcoRI. Доріжки: 1 – молекулярний маркер; 6 – нативна плазміда; 2-5 – рекомбінантна плазміда з геном *sco3479.*

Сконструйовані штами стрептоміцета вирощували на двох типах індикаторного середовища – триптон-соєвому агарі з X-Gal, де за наявності та відсутності індуктора CymR-cmt-P21 репортерної системи. Нами виявлено, що рекомбінантні (pRV3, pRV4-вмісні) штами розщеплюють X-Gal в середовищі з утворенням синьої сполуки 5,5'-дибром-4,4'-дихлоріндиго

(рис. 3.28), лише за наявності індуктора. Отже, репортерна система суворо контролюється на рівні ініціації транскрипції; вона має очікуваний відгук на появу індуктора, у вигляді експресії гена *sco3479* та розщеплення індикаторного субстрату.

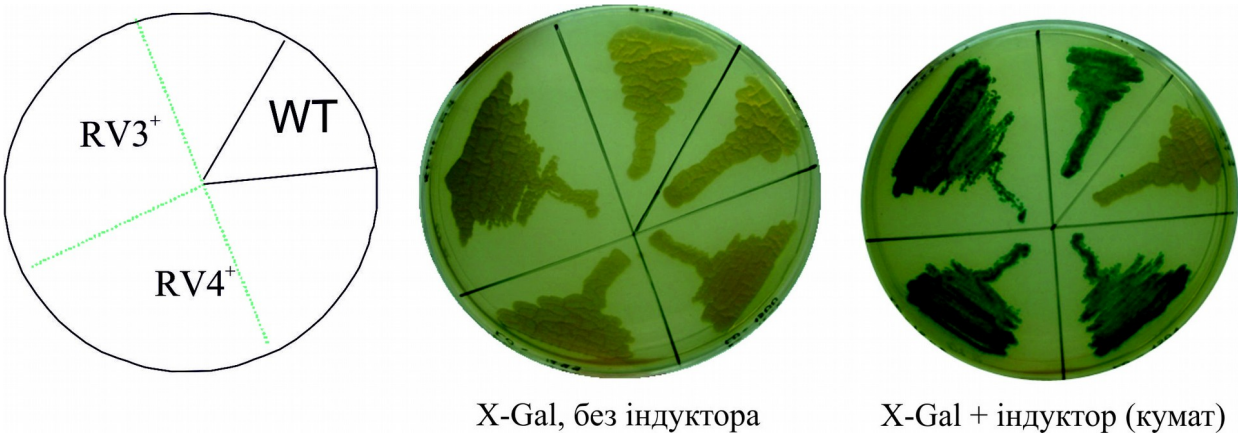


Рис. 3.28 Ріст *S. albus* J1074 із внесеними плазмідами pRV4 (RV^+), pRV3 ($RV3^+$) та дикого типу (WT), на X-Gal-вмісному середовищі з куматом або без нього.

Очікується, що репортерна система даватиме якісний фенотип – забарвлення міцелію як непрямий доказ трансляції, її відсутності або містрансляції гена β-галактозидази. Наявність TTA кодон-залежної складової репортерної системи дає змогу перевірити її функціонування в штамі *S. albus* ОК3 з делетованим геном $тРНК^{Leu}_{TTA}$ (Koshla et al. 2017). На відміну від рисунку 3.13, тут утворення забарвленого продукту можливе лише за наявності плазміди pRV4, що містить дикий варіант *sco3479* (CTC кодон у 8й позиції). Плазміда pRV3 не веде до синього забарвлення міцелію унаслідок відсутності в штамі ОК3 лейцил-тРНК з необхідним антикодоном для кодона TTA, а будь який прояв синього забарвлення свідчитиме про містрансляцію кодону TTA. Справді, нами показано (рис. 3.29), що в штамі ОК3 експресується лише варіант *sco3479* дикого типу (кодон CTC). Експресія

фенотипу у випадку pRV4-вмісного штаму спостерігається лише за умов індукції транскрипції *sco3479*. Це слугує непрямим доказом того, що нездатність штаму ОК3-pRV3^{+} експресувати Lac^{+}-фенотип пов'язано із блокуванням трансляції мРНК *sco3479*.

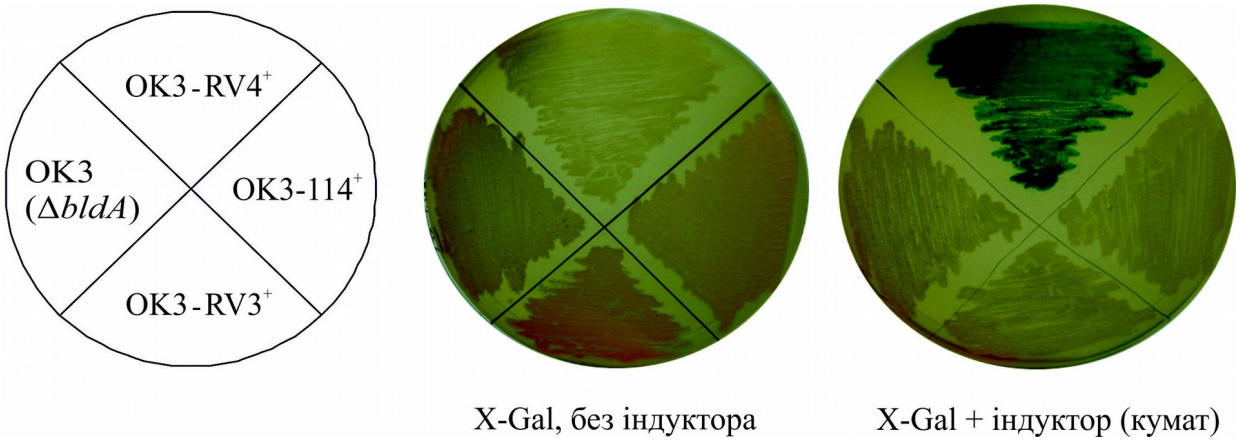


Рис. 3.29. Газони штамів *S. albus* , що містять плазміди pRV4 (ОК3-RV4^{+}), pRV3 (ОК3-RV3^{+}), pOOB114 (ОК3-114^{+}) та дикого типу (ОК3), на TSA з X-Gal (30 мМ), без (ліворуч) та із додаванням (праворуч) кумату.

Сьогодні в літературі описано низку випадків фенотипової експресії у *bldA*-мінус мутантах ознак, що контролюються TTA-вмісними генами (Trepanier et al. 2002). В усіх вищезгаданих випадках така експресія пояснюється містрансляцією UUA-вмісних мРНК. Однак, доказів містрансляції саме UUA-вмісних транскриптіів наведено не було. Сконструйована нами репортерна система відкриває покращені можливості для дослідження містрансляції, які ми вирішили продемонструвати. А саме, якщо містрансляція дійсно відбувається, то її можна буде виявити як появу у популяції штаму *S. albus* ОК3 з плазмідою pRV3 колоній синього кольору. За умови дотримання усіх контрольних заходів, поява таких колоній слугуватиме першим непрямим (генетичним) доказом того, що UUA-вмісний транскрипт *sco3479* зазнає містрансляції. У результаті тривалого (14 діб)

інкубування pRV3-вмісного штаму ОК3 ми справді спостерігали появу блідо-синіх колоній на газоні вищезгаданого штаму на середовищі з X-Gal та індуктором транскрипції sco3479, куматом (рис. 3.30). За відсутності індуктора колоній синього кольору не спостерігали. Отже, синьо забарвлені колонії зявляються тільки за умови транскрипції гена *sco3479*. Низка контрольних експериментів виключили можливість того, що кумат стимулює деградацію хромогенного субстрату та інші альтернативні пояснення отриманих результатів.

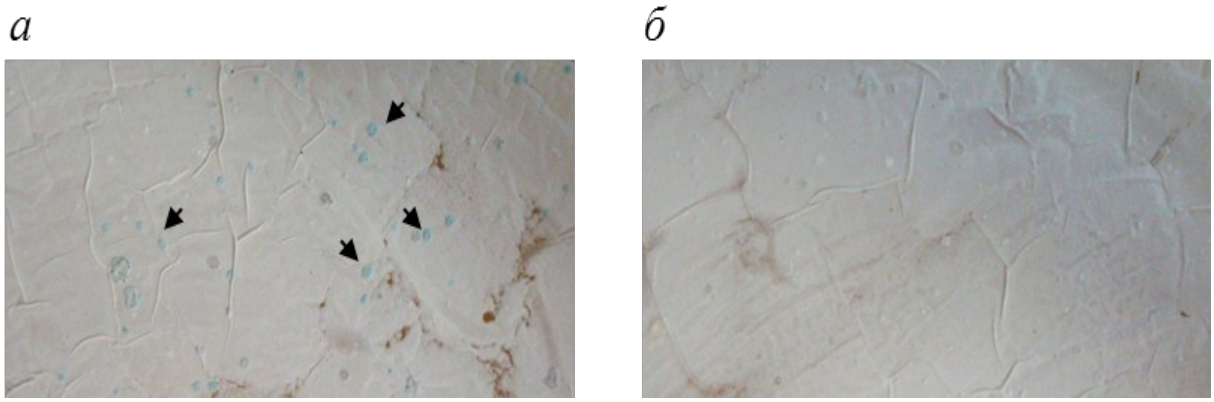


Рис. 3.30 Фотографія газонів штамів *S. albus* ΔbldA (ОК3) RV3^{+} на середовищі TSA з апраміцином, X-Gal (100 мМ) й куматом (***а***), або без нього (***б***). Стрілками позначено кілька репрезентативних колоній S. albus, що здатні розщеплювати хромогенний субстрат.

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 дальше її використання дасть змогу відповісти на низку питань про особливості трансляції кодона TTA у геномах стрептоміцетів. Як приклад, нами отримано перші докази, що непрямо вказують на містрансляцію

TTA-вмісних генів у *S. albus*. Створена система дасть змогу виявляти нові генетичні та фізіологічні фактори, що стимулюють чи гальмують містрансляцію кодона TTA у стрептоміцетів.

РОЗДІЛ 4. ОБГОВОРЕННЯ РЕЗУЛЬТАТІВ

У цій роботі викладено перші спроби систематичного дослідження особливостей кодонного складу геномів *Streptomyces* на основі методів обчислювальної та експериментальної біології. Оскільки це питання наразі не досліджувалось, нами виконано комплекс досліджень, починаючи від оцінки особливостей заміщень на рівні нуклеотидів та амінокислотних залишків, і далі на рівні кодонної структури. Отримані результати привели до низки висновків щодо структури й функції кодонного складу стрептоміцетів, а також окреслили низку нових питань у галузі нашого дослідження. Обговоренню останніх присвячено цей розділ.

Гомологічні генетичні послідовності в межах одного роду в тій чи іншій мірі мають спільні тенденції еволюції, які можна описати певним набором правил, або моделлю заміщень. Зокрема, нуклеотидні послідовності мають певні кількісні закономірності у переході одних нуклеотидів до інших у конкретній позиції. Основний мотив, що спонукає до обрахунку оптимальної моделі еволюції для певної групи послідовностей – це бажання отримати точніші результати досліджень, що опираються на використання еволюційних моделей. У нашому випадку ми обрахували оптимальні моделі для роду *Streptomyces*, які можна використати, наприклад, в подальших філогенетичних дослідженнях.

Оскільки, еволюційні аналізи здійснюються на множинно-вирівняних ортологічних послідовностях, необхідно встановити наявність чи відсутність впливу самого алгоритму множинного вирівнювання на кінцевий результат. Для цього, ми обрали низку різних алгоритмів, які використовуть кардинально відмінні підходи до множинного вирівнювання.

Як видно з результатів підбору нуклеотидних та амінокислотних моделей (таблиці 3.1 та 3.2) вибір алгоритму не має значного впливу на кінцевий результат. Відмінності виявлено при використанні алгоритму

MAFFT при вирівнюванні ортологів гена *sco1728*, що репрезентує родину транскрипційних факторів (табл. 3.1). Таку відмінність можна пояснити появу додаткових трансверсійних пар у кінцевому вирівнюванні при трансформації Фур'є. Це дозволило програмі для підбору моделей IQ-Tree уточнити 3-параметричну модель K3Pu до моделі TVM, що передбачає перевагу трансверсій. Щодо моделей амінокислотних заміщень для ортологів гена *sco1728*, то тут вирівнювання за Clustal Omega та MAFFT алгоритмами дали в результаті JTTDCMut, а за MUSCLE та ProbCons — JTT (табл. 3.1). Ця відмінність несуттєва, оскільки еволюційні моделі JTT та JTTDCMut – лише класична та розширена варіації однієї моделі (JTT).

Ми використали три різні вибірки ортологічних послідовностей різних білків, які відрізняються за своїми функціями в мікроорганізмі. Ми керувалися ідеєю, що один представник з групи покаже модель оптимальну для цілої групи. Для цього ми обрали представника для кожної з обраних груп. Ген *sco1728* кодує транскрипційний фактор YtrA-типу з родини GntR. Це білок, який взаємодіє з ДНК. Продукт гена *sco2706* – глікозилтрансфераза, і так репрезентує ензими. Ген *sco3894* кодує трансмембранний білок, імовірну фліппазу (експортер) ліпід-вмісних попередників клітинної стінки бактерій. Ідеєю було перевірити чи впливатимуть функції та організація досліджуваних білків на еволюційну модель. Згідно отриманих результатів (таблиці 3.1 та 3.2) заміщення у різних групах білків (і їхніх генів) найкраще описуються різними еволюційними моделями. Для групи транскрипційного фактора Sco1728 оптимальною є модель JTT, для ензиму Sco2706 – WAG та модель LG – для трансмембранного білка Sco3894 (фліппаза). Стосовно послідовностей генів цих білків: *sco1728* та *sco2706* еволюціонують згідно моделі K3Pu, а *sco3894* — GTR. Варто зазначити, що у всіх подальших дослідженнях множинне вирівнювання ми здійснювали за допомогою прогресивних алгоритмів Clustal Omega та MAFFT.

Оскільки ми додаково дослідили вплив вибору оптимальної моделі на філогенетичний аналіз транскрипційного фактора, ми визначили оптимальні моделі для ряду транскрипційних факторів різного типу. Результат (табл. 3.3) вказує, що оптимальною моделлю для інших транскрипційних факторів є TVM.

Виходячи з вищевказаного, різні білки, особливо відмінні за своєю функцією, зазнають відмінного добору – різного за інтенсивністю чи типом. Не виключена й дія нейтральних процесів, як от генетичний дрейф. Таким чином, можна говорити про неможливість виведення однієї загальної оптимальної моделі еволюції генетичних послідовностей для певного роду. Відповідно, кожна вибірка ортологічних послідовностей матиме власну оптимальну модель, яку слід ідентифікувати.

Наступним питанням для дослідження був вплив знання про оптимальну модель еволюції на результат філогенетичного аналізу. Ми порівняли філогенетичні дерева побудовані з використанням оптимальної та неоптимальної моделей. Для цього ми обрали вибірку ортологів транскрипційного фактора AdpA. На рисунку 3.1 представлені фрагменти філогенетичних дерев побудовані на основі множинного вирівнювання нуклеотидних послідовностей. Ми використали оптимальну модель GTR для трансмембанних факторів, згадану вище, та неоптимальну модель Hasegawa-Kishino-Yang (HKY). Для амінокислотних послідовностей це були моделі JTT (оптимальна) та WAG (неоптимальна) (рис. 3.2). Як результат ми можемо констатувати зміну положення видів на дереві, а відповідно про важливість використання саме оптимальних моделей еволюції для отримання правдоподібних філогенетичних дерев.

Далі ми розпочали вивчення особливостей вживання кодонів в геномах стрептоміцетів. Перше питання, яке нами було висвітлено – це закономірності розміщення кодонів один відносно одного – тобто контекстна залежність вживання кодонів у генах. У цьому керунку нами проаналізовано

зустрічність усіх можливих пар кодонів у геномі одного виду. Для такого дослідження ми використали спеціалізоване програмне забезпечення Anaconda. Воно дозволяє отримати графічну теплову карту нормалізованих показників розподілу кількості кодонних пар в анотованому геномі (рис. 3.3). Дикодони, які зустрічаються з частотою меншою ніж статистична випадковість, матимуть «негативний» контекст, і навпаки частота вища за очікувану нестиме «позитивний» контекст.

Таким підходом, ми проаналізували 50 різних видів стрептоміцетів (див. Додаток, рис. 3.4) та узагальнили отримані карти, як одну усереднену матрицю. Виходячи із отриманої теплової карти (рис. 3.5) всі точки, які показують “негативний” чи “позитивний” контекст, імовірно, відображають особливості вживання дикодонів для всього роду *Streptomyces*. Найбільшу наукову цікавість становлять “екстремальні” контексти, які виходять далеко за межі випадкових варіацій нормального розподілу. Відфільтрувавши низькі показники, ми отримали 11 пар кодонів які можуть нас зацікавити. Це 9 точок із “позитивним” контекстом: UAU-CUG, CUG-CGC, GUA-CGG, GAU-CCG, CUC-ACC, CUC-GCC, CUC-GGC, GAA-CUC та GAA-CUG, а також 2 точки із “негативним” контекстом: CUC-CUG, CUC-GAG.

Тлумачення кодонних контекстів залишається незадовільно розв’язаним питанням. У випадку геномів *Streptomyces* можна відзначити такі загальні риси. По-перше, з 11 виявлених контекстів 9 базуються на лейцинових кодонах. Лейцин – найуживаніша амінокислота в протеомах бактерій (Itzkovitz and Alon 2007), водночас одна із найдегенеративніших (6 синонімів) на кодонному рівні. Імовірно, кодонні контексти для високопопулярних кодонів слугують для оптимізації структури мРНК. По-друге, ми бачимо, що N_2 ($N_1N_2N_3$-$Z_1Z_2Z_3$) у всіх випадках аденін чи урацил. Це може бути певним правилом врегулювання вживання кодонів з A/T нуклеотидами у GC-багатому геномі, хоча, як згадувалося в літературі, контекстне вживання не залежить від кодонного складу досліджуваного

об'єкта. По-третє, виявлені контексти можна узагальнити у вигляді правил-патернів: «позитивні» — KAU-CYG (A та D рис. 3.5), GWA-CKS (C та F рис. 3.5), CUS-VSC (B та E рис. 3.5), та «негативний» — CUC-SWG (E рис. 3.5), де згідно номенклатури ми позначили будь-який нуклеотид, крім урацилу за допомогою “V”, а A/U, G/C, U/G та A/C — як “W”, “S”, “K” та “Y” відповідно.

Отримані патерни показали, що в дикодонах не можна виділити однозначний вплив нуклеотида в певнії позиції. Тим не менше, 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. Варто зазначити, що необхідно припустити і протилежний сценарій згідно якого, ПВК серед синонімічних кодонів спричиняє високі показники не випадковості розподілу як кодонів, так і нуклеотидів. У такому випадку, патерни контекстного вживання кодонів, можна розглядати як відображення ПВК. Насамкінець, наш аналіз засвідчив, що найрідкісніший для стрептоміцетів кодон UUA не виявляє ані позитивної, ані негативної асоціації з наступним кодоном при аналізі цілих геномів стрептоміцетів. Втім, цей результат все ж не можна вважати остаточним доказом того, що такі асоціації відсутні взагалі. По-перше, вибірка TTA-вмісних генів у геномах стрептоміцетів вкрай мала – наприклад, у геномі *Streptomyces coelicolor* серед 2.5 млн кодонів лише 145 – TTA. Складно вивчати і статистично оцінювати поведінку такого малого масиву даних у межах набагато більшої вибірки. Далі, вибірка TTA-вмісних генів – функціонально дуже гетерогенна у геномах стрептоміцетів, для більшості з них точну функцію не встановлено (Li et al. 2007; Chandra and Chater 2008). Імовірно, що багато TTA^+ генів у межого певного генома не експресуються, а тому мають особливий еволюційний темп і режим порівняно з генами первинного метаболізму. Ми припускаємо що вивчення проблеми контекстного

вживання рідкісних кодонів потребуватиме створення окремих масивів генів, які буде згруповано за функціональним принципом – наприклад, за експериментально встановленим рівнем експресії, чи типом кодованого білка тощо.

Згідно нашого аналізу, на сьогодні не існує у відкритому доступі знарядь візуалізації кодонних заміщень. Для дослідження закономірностей вживання кодонів у геномах *Streptomyces* ми створили застосунок (наразі офлайн-версію) для побудови графічних моделей кодонних заміщень. Створений застосунок дає змогу нам судити про закономірності заміщення одних кодонів іншими в межах вибірки ортологічних генів роду *Streptomyces*. Результатом роботи створеного сервіса є матриці 61x61 представлені на рисунках у розділі 3.3. Важливим є розуміння інформації, яку містить така матриця, а саме що можна сказати про еволюційні процеси у досліджуваній вибірці.

Матриці представлені у вигляді таблиць, де по осях розміщено кодони, які відсортовано за фізико-хімічними властивостями (кодони для споріднених амінокислот – поруч). На перетині рядків та стовпчиків знаходяться кола з діаметром пропорційним імовірності заміщення кодону, на діагоналі кола відсутні (імовірність консервації – незаміщення – кодона). Кола забарвлені відповідно до типу заміщення. Зелений — несинонімічні заміщення кодонів з різницею в один нуклеотид. Червоний — синонімічні заміщення кодонів з різницею в 1 нк. Жовтий — несинонімічні заміщення кодонів з різницею більше ніж 1 нк. Синій — синонімічні заміщення кодонів з різницею більше ніж 1 нк.

Першочергово, результат можна інтерпретувати за патерном. Ми розглянемо детальніше фрагменти представлених результатів (розділ 3.3.) та спробуємо витлумачити еволюційні процеси, що імовірно спричинили ці патерни.

Групу трансмембраних білків ми репрезентували геном *sven_4640*, що кодує білок-фліппазу II. Виходячи із функцій цієї групи білків, вони мають бути відносно консервативними, тобто мати консервативні ділянки, що відповідають за взаємодію із субстратом та місця кріплення до клітинної мембрани. Крім цього, трансмембранні білки мають ділянки, в яких зміна амінокислотних залишків не матиме значного впливу на функціонування і такі зміни зберігатимуться. Це можна спостерігати на рис. 4.1, як скупчення заміщень біля діагоналі, які згруповано в межах кодонів, що кодують подібні

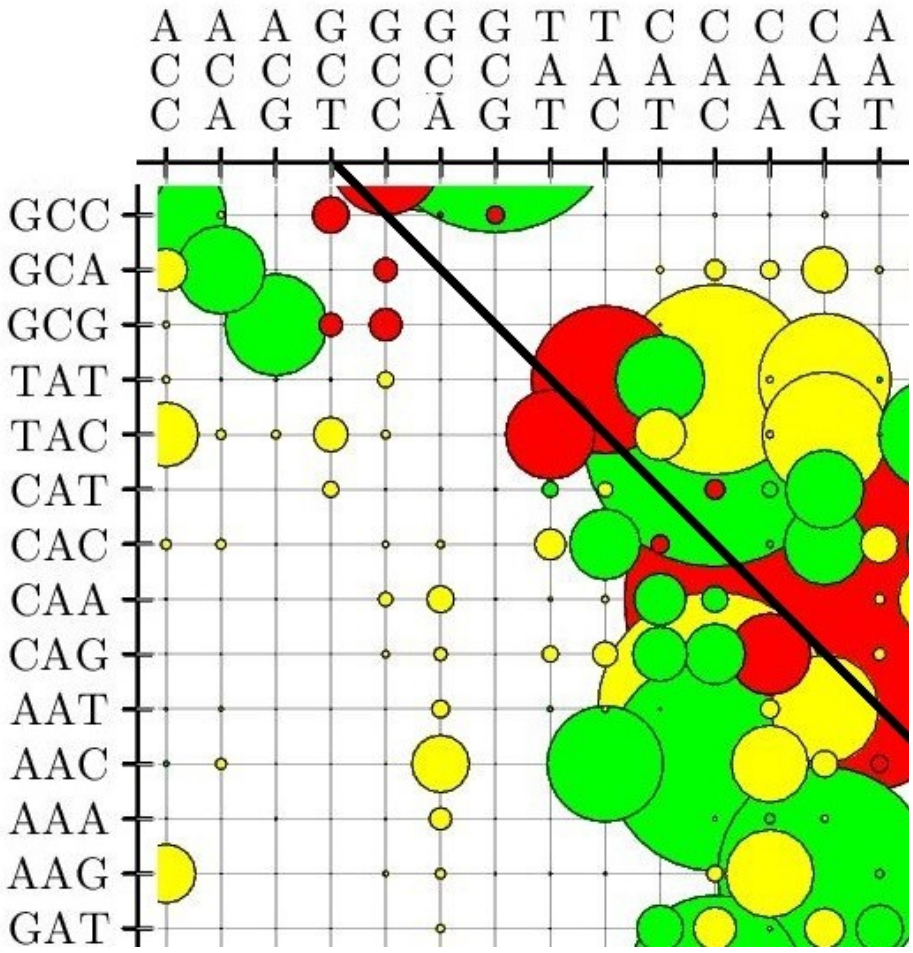


Рис. 4.1 Фрагмент графіка частоти заміщень кодонів в ортологічних генах фліппази II (*sven_4640*). Кольоровий код – див. Рис. 3.7. Чорна лінія — діагональ матриці.

за фізико-хімічними властивостями амінокислоти. Значна частка всіх кодонних заміщень – синонімічні та несинонімічні з різницею в один нуклеотид. Інші типи заміщення кодонів розсіяні по всій матриці, незначні за частотою і формують так званий “фоновий шум” – масив рідкісних подій, які не мають виразних тенденцій.

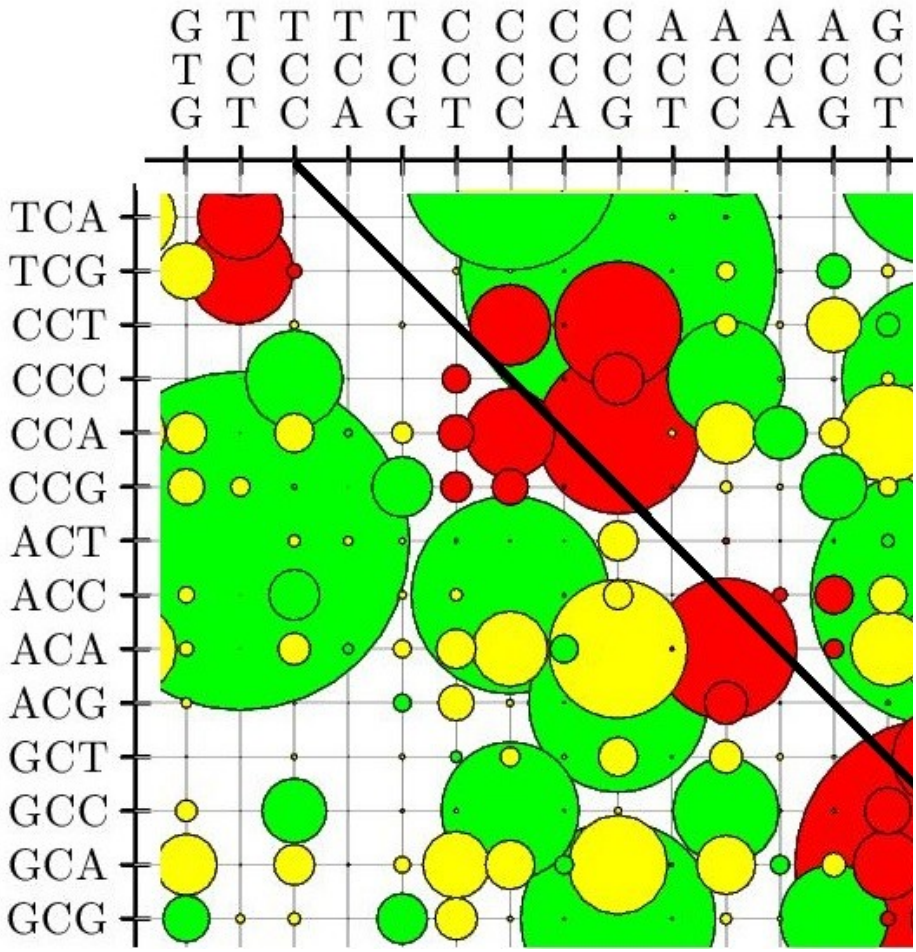


Рис. 4.2 Фрагмент графіка частоти заміщень кодонів в ортологічних генах транскрипційного фактора (*sco1728*). Кольоровий код – див. рис. 3.7. Чорна лінія — діагональ матриці.

Схожу картину можемо спостерігати і для ортологів гіпотетичного транскрипційного фактора *Sco1728* (рис 4.2), що репрезентуватиме групу транскрипційних факторів. Основна відмінність – вища, порівняно з іншими,

частота несинонімічних заміщень кодонів із різницею більше одного нуклеотиду. Це можна пояснити тиском рушійного добору на транскрипційні фактори цієї родини порівняно з трансмембранними білками, що відкриває шлях до накопичення суттєвіших заміщень в їхній послідовності. Можна припустити, що група далі змінюється а її представники мають відмінні функції.

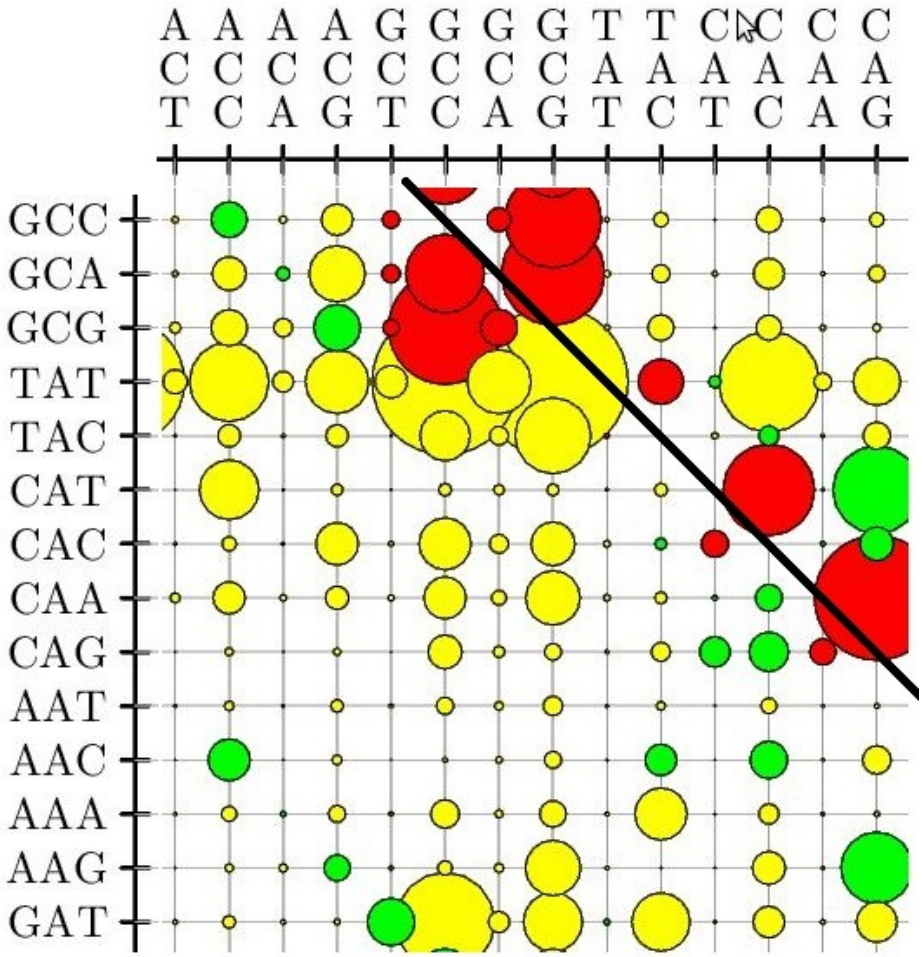


Рис. 4.3 Фрагмент графіка частоти заміщень кодонів в ортологічних генах глікозилтрансферази (*sco2706*). Кольоровий код – див. рис. 3.7. Чорна лінія — діагональ матриці.

Ензими представлено ортологічною групою глікозилтрансферази *Sco2706,* характеризуються в основному синонімічними заміщеннями (рис 4.3). Такий патерн можна пояснити дією сил стабілізуючого добору – відкидається більшість змін амінокислотної послідовності. Несинонімічні заміщення розподілені між майже всіма кодонами, що вказує на відсутність чітко напрямленого добору. Водночас, жовті кола більшого діаметру знаходяться в рядах з GC-багатими кодонами, що корелює із GC-багатим геномом стрептоміцетів загалом і показує тенденцію до поступового збагачення кодонома функціонально близькими GC-вмісними кодонами.

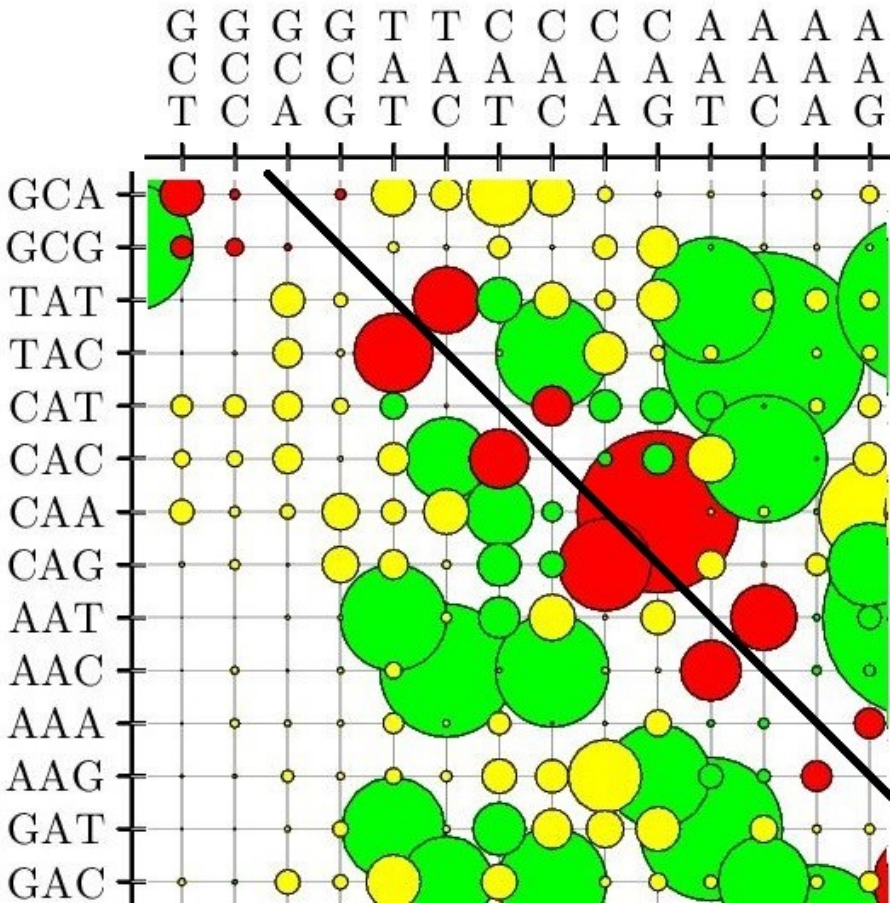


Рис. 4.4 Фрагмент графіка частоти заміщень кодонів в ортологічних генах великої субодиниці рибулозо-5-бісфосфат карбоксилази (*RuBisCo*). Кольоровий код – див. рис. 3.7. Чорна лінія — діагональ матриці.

Усі розглянуті вище масиви послідовностей репрезентують білки, функція яких описана, як імовірна і встановлена на подібності будови білків до інших чия функція вивчена. Також, відсутня інформація про належність цих білків до первинного чи вторинного метаболізму. Це накладає відбиток на характер їхньої еволюції. А саме, якщо це неважливі гени/білки, то в них можна очікувати відмінні типи і частоти заміщень. Тому як контроль, або як приклад консервативної групи ензимів, ми обрали отрологічну вибірку добре вивченого гена великої субодиниці рибулозо-5-бісфосфат карбоксилази (*RuBisCo*) із фотосинтезувальних мікроорганізмів. Фрагмент патерну показано на рисунку 4.4. Ми можемо спостерігати значну симетричність патерна, де прямі та зворотні заміщення відбуваються з однаковою чи схожою імовірністю. Така симетричність говорить про консервативність послідовності: хоча в ній спостерігається накопичення несинонімічних заміщень, загалом для цієї родини білків швидкості прямих і зворотних заміщень практично урівноважують одна одну. Кінцевий результат – родина білків залишається практично незмінною – її представники виконують ідентичну функцію. Отже, порівняння прямих й зворотніх заміщень є ще одним способом тлумачити кодонні бульбашкові графіки.

Щодо недоліків застосованого методу, то тут слід відзначити втрату позиційної інформації про заміщення кодону у послідовності. Тобто ми спостерігаємо певне заміщення, описуємо його частоту, але де саме вона відбувається (початок чи кінець відкритої рамки) – сказати цей метод не може. Однак, він вказує напрям заміщення, оскільки обрахунки опираються на філогенетичні дані.

Питання щодо високого GC-складу геномів *Streptomyces* є відкритим. Якщо опиратися на факти, то ми маємо загальну тенденцію поступового збагачення геномів AT-парами серед бактерійних геномів (Hershberg and Petrov 2010; Jee et al. 2016). Також філогенетична гілка стрептоміцетів — наймолодша на дереві класу *Actinobacteria* (Chandra and Chater 2014). Спроба

пояснити, чому стрептоміцети мають високий вміст GC в противагу загальній тенденції, призвела до появи низки теорій.

Згідно теорії запропонованої Гершбергом та Петровом, як можливе пояснення збагачення геномів бактерій A/T парами, причиною змін на нуклеотидному рівні виступають спонтанні мутаці. А оскільки стрептоміцети появилися відносно недавно, високий GC-склад все ще знизиться вході подальшої еволюції. Виходячи з цього, мутації мають виникати рівномірно вздовж послідовності геному, не зважаючи чи вона кодувальна чи ні. В геномах стрептоміцетів мала кількість некодувальних ділянок. Ми не можемо спостувати чи підтвердити таку теорію, як наприклад, статистично достовірно порівняти імовірність заміщень в кодувальних та некодувальних ділянках.

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, а точніше певний рушійний добір на кодонному рівні.

Виникає питання про переваги у використанні GC-багатих кодонів, які би несли селективне значення. У стрептоміцетів, велика кількість довгих генних послідовностей, 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 лишніх рибосомальних сайтів зв'язування. Так, імовірність появи A/T-багатих сайтів зв'язування з рибосомою в непризначеному місці менша в генах з переважанням GC-

багатих кодонів. Це в свою чергу несе перевагу в еволюційному розумінні — помилкові транскрипти лишня витрата ресурсів. Відповідно, можемо поставити під сумнів постулат, що зміщення в ПВК – наслідок високого GC-складу стрептоміцетів. Навпаки, імовірніше, 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 мутаційним процесам протидіє тиск природного добору, який зумовлює збагачення кодувальних послідовностей GC-багатими кодонами. Глибше розкриття цього аспекта еволюції *Streptomyces* потребує кількісного аналізу процесів нуклеотидних заміщень у кодувальній та некодувальній частинах геномів, і виходить за рамки цієї роботи.

Незважаючи на високий GC-склад геномів стрептоміцетів, кодон TTA вживається рідше ніж передбачено статистичним розподілом. Як описано в попередніх розділах (див. Огляд літ-ри.) таке явище зумовлює існування певного регуляційного механізму в стрептоміцетів. Існування такого механізму в свою чергу передбачає, що кодон TTA матиме високу точність трансляції. Також, важливим є факт відсутності трансляції цього кодону за відсутності акцепторної тРНК. Це означає, що інші тРНК (псевдоакцепторні) не здатні транслювати TTA. Також, в середовищі клітини низька концентація акцепторної тРНК “непопулярного” кодону через відсутність необхідності частої трансляції. Логічно припустити, що створюється певна конкуренція за декодування рідкісного кодону. Адже, серед всіх тРНК необхідно знайти акцепторну з низькою концентрацією. Це збільшує шанси, що рідкісний кодон декодується іншою тРНК, що призведе до містрансляції.

Підсумувавши визщесказане, ми можемо вивести певну модель: містрансляція кодону має обернену залежність від концентрації акцепторної тРНК. Тобто, “популярні” кодони транслюються точніше ніж “непопулярні”. Така закономірність викликає неузгодження необхідності TTA бути точним

рідкісним кодоном, щоб виконувати функцію регулятора, та мати вищі шанси до містрансляції (як кожен непопулярний кодон).

Отже, перед нами постало завдання перевірити, чи може рідкісний кодон бути водночас точним, використавши нову модель для обрахунку точності трансляції. Модель Шаха та Гілкріста (Shah і Gilchrist 2010) передбачає врахування впливу всіх тРНК в клітині які можуть декодувати цільовий кодон. Таким чином, ми враховували залежність містрасляції не тільки від концентрації акцепторної тРНК, а й від сумарної концентрації тРНК, які здатні містранслювати кодон. Оскільки, концентрація тРНК в середовищі клітини корелює із кількістю копій генів тРНК в геномі, ми змогли використати біоінформатичний підхід до вивчення точності трансляції, оперуючи лише інформацією про геном. А саме, підрахувати ККГ акцепторної тРНК - *tF* та ККГ близько-споріднених тРНК *tN*. Відношення цих величин показало присутню кореляцію у більшості випадків (рис. 3.11 — 3.14). Ці дані дозволяють зробити припущення, що чим більше копій гена акцепторної тРНК, а відповідно вища її концентрація, тим більша кількість генів близько-споріднених тРНК.

Як результат, ми маємо нове формулювання моделі містрансляції у стрептоміцетів: точність декодування кодону залежить від співвідношення фокальної тРНК *tF* до близько-споріднених *tN*. Така модель узгоджується з тим, що рідкісний кодон TTA є точним, оскільки йому притаманна низька концентрація конкурентних тРНК. Для перевірки цього припущення ми розрахували швидкість елонгації шести лейцинових кодонів для шести стрептоміцетних видів за допомогою усіх можливих типів (і близько-споріднених) тРНК, що веде до містрансляції кодонів (рис. 3.15). Порівнявши отримані показники, ми можемо зробити висновок, що рідкісний кодон TTA містранслюється з меншою імовірністю ніж інші лейцинові кодони. Такий результат узгоджується із регуляторною функцією цього кодона.

Попередні результати показали теоретичні засади важливості вивчення кодоного складу геномів стрептоміцетів. Далі ми запропонували систему для вивчення експресії кодонів в експериментальних умовах. Система базується на стрептоміцетному гені *sco3479*, продукт якого виявляє β-галактозидазну активність. Репортерна складова системи – це здатність розщеплювати X-Gal з утворенням кольорової сполуки. Важливо знайти штам стрептоміцетів, який природно нездатний розщеплювати хромогенні аналоги лактози. Загалом розроблення репортерів на основі β-галактозидази для стрептоміцетів цікавило науковців давно, але на перешкоді стояли відсутність генів-гомлогів *lacZ*, а також висока галактозидазна активність усіх модельних штамів актиноміцетів. Нам вдалося подолати першу перепону на основі аналізу генома *S. coelicolor*, а також ідентифікувати штам *S. albus* J1074 як X-Gal-мінус вид. Ген β-галактозидази внесений у плазміду pGCymRP21 з геном репресора CymR та відповідно кумат-індуцибельним промотором (див. рис. 3.21). Основна ідея дослідження кодонів за допомогою такої системи – це вносити кодони, які нас цікавлять, замість їхніх синонімів у послідовність репортерного гена β-галактозидази *sco3479*. Ми продемонстрували функціональність такої системи на прикладі порівняння експресії двох версій цього гена – дикого типу та його алеля, що містить синонімічне заміщення CTC8→TTA. Виявлено, що CTC-версія гена експресується у мутанті *S. albus* з делетованим геном *bldA*, що кодує $tRNA^{Leu}_{UAA}$. Водночас, TTA-вмісна версія репортерного гена не експресується в цьому мутанті, імовірніше всього, унаслідок абортивної трансляції кодона UUA. Ми також отримали перші докази містрансляції TTA-вмісної алелі гена *sco3479*.

Хоча на рівні фенотипу створена репортерна поводиться очікувано, нами наразі не отримано остаточного доказу того, що експресію TTA-вмісної алелі гена порушено саме на рівні трансляції. Відповідні експерименти (пряме порівняння рівнів транскрипції та трансляції CTC^+ й TTA^+ репортерних генів та білків) є предметом дальшої роботи в НДЛ-42. Цінність

наразі виконаної роботи – передусім у створенні першої експериментальної системи вивчення функції кодона TTA, що підлягає повному транскрипційному контролю, має простий, однозначний гено- та фенотип. Усі наявні дані, які викладено у цій роботі, дають підстави нам вважати, що таку систему створено, і це відкриває нові можливості для вивчення кодон-залежних регуляторних механізмів у 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. Використання різних моделей для філогенетичного аналізу вестиме до філогенетичного дерева з невеликими змінами в топології окремих гілок.
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, увигляді “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.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Adachi, J. and Hasegawa, M. 1996. “Model of amino acid substitution in proteins encoded by mitochondrial DNA.” *J. Mol. Evol.* 42(4):459–468.
2. Baltz, R.H. 2010. “Streptomyces and Saccharopolyspora hosts for heterologous expression of secondary metabolite gene clusters.” *J Ind Microbiol Biotechnol* 37(8):759–772. doi:10.1007/s10295-010-0730-9.
3. Barka, E.A., Vatsa, P., Sanchez, L., Gaveau-Vaillant, N., Jacquard, C., Klenk, H.P., Clement, C., Ouhdouch, Y. and van Wezel, G.P. 2015. “Taxonomy, physiology, and natural products of Actinobacteria.” *Microbiol Mol Biol* 80(1):1–43. https://doi.org/10.1128/MMBR.00019-15.
4. Bernard, E.J., Azad, Y., Vandamme, A.M., Weait, M. and Geretti, A.M. 2007. “HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission.” *HIV Med.* 8(6):382-7.
5. Bibb, M. 1996. “Colworth Prize Lecture. The regulation of antibiotic production in Streptomyces coelicolor A3(2)” *Microbiology* 142:1335 – 1344.
6. Bibb, M.J., Findlay, P.R. and Johnson, M.W. 1984. “The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences.” *Gene*. 30(1-3):157-66.
7. Bilyk, B. and Luzhetskyy, A. 2014. “Unusual site-specifc DNA integra-tion into the highly active pseudo-attB of the *Streptomyces albus* J1074 genome.” *Appl Microbiol Biotechnol* 98(11):5095–5104. doi:10.1007/s00253-014-5605-y.

8. Bulmer, M. 1988. “Codon usage and intragenic position.” J *Theor Biol.* 133(1):67-71.
9. Chang, P.C., Kim, E.S. and Cohen, S.N. 1996. “Streptomyces linear plasmids that contain a phage-like, centrally located, replication origin.” *Mol Microbiol.* 22(5):789-800.
10. Chargaff, E. 1951. “Structure and function of nucleic acids as cell constituents.” *Fed Proc.* 10(3):654-9.
11. Chater, K. and Chandra, G. 2008. “The use of the rare UUA codon to define “expression space” for genes involved in secondary metabolism, development and environmental adaptation in *Streptomyces*.” *J Microbiol* 46:1–11
12. Chater, K.F. 1993. “Genetics of differentiation in *Streptomyces*.” *Annu. Rev. Microbiol.* 47:685 – 713.
13. Chater, K.F. 2006. “*Streptomyces* inside-out: a new perspective on the bacteria that provide us with antibiotics.” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 361:761 – 768.
14. Chater, K.F. and Wilde, L.C. 1980. “*Streptomyces albus* G mutants defective in the SalGI restriction-modifcation system.” *J Gen Microbiol.* 116(2):323–334.
15. Chen, C.W., Huang, C.H., Lee, H.H., Tsai, H.H. and Kirby, R. 2002. “Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes.” *Trends Genet* 18:522-529.
16. Chevance, F.F.V., Hughes, K.T. 2017. “Case for the genetic code as a triplet of triplets.” *Proc Natl Acad Sci U S A* 114:4745–4750.
17. Curran, M.E., Splawski, I., Timothy, K.W., Vincent, G.M., Green, E.D. and Keating M.T. 1995. “A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome.” *Cell* 80(5):795-803.

18. Dayhoff, M., Schwartz, R. and Orcutt, B. 1978 “A model for evolutionary change in proteins.” A*tlas of Protein Sequence and Structure* 15:345–352.
19. Dereeper, A., Guignon, V., Blanc, G., Audic, S. et al. 2008. “Phylogeny.fr: robust phylogenetic analysis for the non-specialist.” *Nucleic Acid Res* 36:465–469. https://doi.org/10.1093/nar/gkn180.
20. Do, C., Mahabhashyam, M., Brudno, M. and Batzoglou, S. 2005. “PROBCONS: Probabilistic Consistency-based Multiple Sequence Alignment.” *Genome Res.* 15:330–340.
21. Edgar, R. 2004. “MUSCLE: multiple sequence alignment with high accuracy and high throughput” *Nucleic Acids Res.* 32(5):1792–1797.
22. Eyre-Walker, A. and Bulmer, M. 1993 “Reduced synonymous substitution rate at the start of enterobacterial genes.” *Nucleic Acids Res.* 21(19):4599-4603.
23. Felsenstein, J. 1981 “Evolutionary trees from DNA sequences: a maximum likelihood approach.” *J. Mol. Evol.* 17(6):268–276.
24. Fernandez-Moreno, M.A., Caballero, J.L., Hopwood, D.A. and Malpartida , F. 1991. “The act cluster contains regulatory and antibiotic export genes, direct targets for translational control by the bldA tRNA gene of *Streptomyces.*” *Cell* 66:769 – 780.
25. Goodman, D.B., Church, G.M., Kosuri S. 2013. “Causes and effects of N-terminal codon bias in bacterial genes.” *Science*. 342(6157):475-9. doi: 10.1126/science.1241934.
26. Goldman, N. and Yang, Z. 1994. “A codon-based model of nucleotide substitution for protein-coding DNA sequences” *Mol. Biol. Evol.* 11:725–736.
27. Gonnet, G., Cohen, M. and Benner, S. 1993. “Exhaustive matching of the entire protein sequence database” *Science* 256(5062):1443–1445.

28. Gouy, M. 1987. “Codon contexts in enterobacterial and coliphage genes.” *Mol Biol Evol.* 4(4):426-44.
29. Grantham, R. 1974. “Amino acid difference formula to help explain protein evolution.” *Science* 185(4154):862-864.
30. 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.
31. Guthrie, E.P. and Chater, K.F. 1990. “The level of a transcript required for production of a *Streptomyces coelicolor* antibiotic is conditionally dependent on a tRNA gene.” *J Bacteriol* 172:6189–6193. doi:10.1128/jb.172.11.6189-6193.1990.
32. Gutman, G.A. and Hatfield, G.W. 1992. “Nonrandom utilization of codon pairs in *Escherichia coli.*” *Proc Natl Proc Natl Acad Sci USA* 89(22):10915-10919.
33. Hackl, S. and Bechthold, A. 2015. “The gene bldA, a regulator of morphological differentiation and antibiotic production in *Streptomyces*.” *Arch Pharm* 348:455–462. doi:10.1002/ardp.201500073.
34. Halpern, A. and Bruno, W. 1998. “Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies.” *Mol. Biol. Evol.* 15. (7):9107.
35. Hartl, D.L., Moriyama, E.N. and Sawyer, S.A. 1994. “Selection intensity for codon bias.” *Genetics* 138(1):227-234.
36. Hasegawa, M., Kishino, H. and Yano, T. 1985. “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.” *J. Mol. Evol.* 22(2):160–174.

37. Henikoff, S. and Henikoff, J.G. 1992 “Amino acid substitution matrices from protein blocks.” *Proc Natl Acad Sci USA* 89(22):10915-10919.
38. Herrmann, S., Siegl, T., Luzhetska, M., Petzke, L., Jilg, C., Welle, E., Erb, A., Leadlay, P.F., Bechthold, A. and Luzhetskyy, A. 2012. “Site-specific recombination strategies for engineering actinomycete genomes.” *Appl Environ Microbiol* 78(6):1804–1812. doi:10.1128/AEM.06054-11.
39. Hershberg, R. and Petrov, D.A. 2010. “Evidence that mutation is universally biased towards AT in bacteria.” *PLoS Genet.* 6(9):e1001115. doi:10.1371/journal.pgen.1001115.
40. Hoban, S., Bertorelle, G. and Gaggiotti, O.E. 2012. “Computer simulations: tools for population and evolutionary genetics.” *Nat Rev Genet.* 13(2):110-122. doi:10.1038/nrg3130.
41. Holmes, I. and Rubin, G.M. 2002. “An expectation maximization algorithm for training hidden substitution models.” *J Mol Biol.* 317(5):753-764.
42. Hopwood, D. 1967. “Genetic analysis and genome structure in *Streptomyces coelicolor*.” *Bacteriol. Rev.* 31:373 – 403.
43. Horbal, L., Fedorenko, V. and Luzhetskyy, A. 2014. “Novel and tightly regulated resorcinol and cumate-inducible expression systems for *Streptomyces* and other actinobacteria.” *Appl Microbiol Biotechnol* 98(20):8641–8655. doi:10.1007/s00253-014-5918-x.
44. Itzkovitz, S., Alon, U. 2007. “The genetic code is nearly optimal for allowing additional information within protein-coding sequences.” Genome Res. 17(4):405–412.
45. Jee, J., Rasouly, A., Shamovsky, I., Akivis, Y., Steinman, S.R., Mishra, B. and Nudler, E. 2016. “Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing.” *Nature* 534(7609):693-696.
46. Jones, D., Taylor, W. and Thornton, J. 1992. “The rapid generation of mutation data matrices from protein sequences.” *Comput. Applic. Biosci.* 8:275–282.

47. Jukes, T. and Cantor, C. 1969. “Evolution of protein molecules. In Munro, H.N. Mammalian protein metabolism.” *New York: Academic Press* :21–123.
48. Katoh, K. and Toh, H. 2008. “Recent developments in the MAFFT multiple sequence alignment program.” *Brief. Bioinf.* 9(4):286–298.
49. Kimura, M. 1980. “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.” *J. Mol. Evol.* 16(2):111–120.
50. Kimura, M. 1981. “Estimation of evolutionary distances between homologous nucleotide sequences.” *Proc Natl Acad Sci USA* 78(1):454–458.
51. King, A.A. and Chater, K.F. 1986. “The expression of the *Escherichia coli lacZ* gene in *Streptomyces*.” *J Gen Microbiol* 132(6):1739–1752. doi:10.1099/00221287-132-6-1739.
52. Klosterman, P.S., Uzilov, A.S., Bendaña, Y.R., Bradley, R.K., Chao, S., Kosiol, C., Goldman, N. and Holmes, I. 2008. “XRate: a fast prototyping, training and annotation tool for phylo-grammars.” BMC Bioinf **7**:428. doi:10.1186/1471-2105-7-428.
53. Knight, S., Kim, R., Pain. D. and Dancis, A. 1999. “The Yeast Connection to Friedreich Ataxia” *Insights from Model Systems* 64(2):365–371. doi:http://dx.doi.org/10.1086/302270.
54. 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 tRNALeuUAA gene bldA.” *Arch Microbiol* 199(8):1175-1183. doi: 10.1007/s00203-017-1389-7.
55. Kosiol, C. and Goldman, N. 2005 “Different versions of the Dayhoff rate matrix.” *Mol. Biol. Evol.* 22:193–199.
56. Kuzniar, A., van Ham, R.C., Pongor, S. and Leunissen, J.A. 2008. “The quest for orthologs: finding the corresponding gene across genomes.” *Trends Genet* 24:539–551. doi:10.1016/j.tig.2008.08.009.

57. Kwak, J., McCue, L.A. and Kendrick, K.E. 1996. “Identification of bldA mutants of *Streptomyces griseus*.” *Gene* 171:75 – 78.
58. Lawlor, E., Baylis, H. and Chater, K. 1987. “Pleiotropic morphological and antibiotic deficiencies result from mutations in a gene encoding a tRNA-like product in *Streptomyces coelicolor* A3(2).” *Genes Dev* 1:1305–1310. doi:10.1101/gad.1.10.1305.
59. Le, Q. and Gascuel, O. 2008. “An Improved General Amino Acid Replacement Matrix.” *Mol. Biol. Evol.* 25(7):1307–1320.
60. Leskiw, B.K., Lawlor, E.J., Fernandez-Abalos, J.M. and Chater, K.F. 1991. “TTA codons in some genes prevent their expression in a class of developmental, antibiotic-negative, Streptomyces mutants.” *Proc Natl Acad Sci USA* 88:2461–2465.
61. Leskiw, B.K., Mah, R., Lawlor, E.J. and Chater, K.F. 1993. “Accumulation of bldA-specifed tRNA is temporally regulated in *Streptomyces coelicolor* A3(2).” *J Bacteriol* 175:1995–2005. doi:10.1128/jb.175.7.1995-2005.1993.
62. Li, W., Wu, J., Tao, W. et al. 2007. “A genetic and bioinformatic analysis of *Streptomyces coelicolor* genes containing TTA codons, possible targets for regulation by a developmentally significant tRNA.” *FEMS Microbiol. Lett.* 266:20–28.
63. Merrick, M. 1976. “A morphological and genetic mapping study of bald colony mutants of *Streptomyces coelicolor*.” *J. Gen. Microbiol.* 96:299–315.
64. Moura, G., Pinheiro, M., Arrais, J. et al. 2007. “Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure.” *PLoS One* 2:e847.
65. Muse, S.V., and Gaut, B.S. 1994. “A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.” *Mol Biol Evol.* 11(5):715-724.

66. Musialowski, M.S., Flett, F., Scott, G.B., Hobbs, G., Smith, C.P. and Oliver, S.G. 1994. “Functional evidence that the principal DNA replication origin of the *Streptomyces coelicolor* chromosome is close to the dnaA-gyrB region.” *J Bacteriol.* 176(16):5123-5.

67. Napolitano, M.G., Landon, M., Gregg, C.J., Lajoie, M.J., Govindarajan, L., Mosberg, J.A., Kuznetsov, G., Goodman, D.B., Vargas-Rodriguez, O., Isaacs, F.J., Söll, D., Church G.M. 2016. “Emergent rules for codon choice elucidated by editing rare arginine codons in *Escherichia coli*.” *Proc. Natl. Acad. Sci. USA*. 113(38):E5588-97. doi: 10.1073/pnas.1605856113.

68. Pettersson, B.M. and Kirsebom, L.A. 2011. “tRNA accumulation and suppression of the bldA phenotype during development in *Streptomyces coelicolor*.” *Mol Microbiol* 79:1602–1614. doi:10.1111/j.1365-2958.2011.07543.x.

69. Piepersberg, W. 1993. “*Streptomyces* and *Corynebacteria*.” *Biotechnology* : 434-468.

70. Plotkin, J.B. and Kudla, G. 2011, “Synonymous but not the same: the causes and consequences of codon bias.” *Nat. Rev. Genet.* 12:32–42.

71. 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

72. Redenbach, M., Kieser, H.M., Denapaite, D., Eichner, A., Cullum, J., Kinashi, H. and Hopwood, D.A. 1996. “A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome.” *Molecular microbiology* 21:77-96.

73. 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.” *Springer-plus* 5:982. doi:10.1186/s40064-016-2683-6.

74. 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.

75. Rokytzkyy, 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.

76. Schneider, A., Cannarozzi, G.M. and Gonnet, G.H. 2005. “Empirical codon substitution matrix.” BMC Bioinformatics 6:134 https://doi.org/10.1186/1471-2105-6-134.

77. Segata, N., Bernigen, D., Morgan, X.C. and Huttenhower, C. 2013. “PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes.” *Nat. Commun.* 4:2304.

78. Shah, P. and Gilchrist, M.A. 2010. “Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias.” *PLoS Genet* 6:e1001128. doi:10.1371/journal.pgen.1001128.

79. Shah, P. and Gilchrist, M.A. 2011. “Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift.” *Proc. Natl. Acad. Sci. USA* 108(25):10231–10236.

80. Shpaer, E.G. 1986. “Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation.” *J Mol Biol.* 188(4):555-64.

81. Shuman, H.A. and Silhavy, T.J. 2003. “The art and design of genetic screens: *Escherichia coli*.” *Nat Rev Genet.* 4(6):419-431.

82. Sievers, F., Wilm, A., Dineen, D. et al. 2011. “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.” *Mol. Syst. Biol.* 7:539.

83. Takano, E., Tao, M., Long, F., Bibb, M.J., Wang, L., Li, W., Buttner, M.J., Bibb, M.J., Deng, Z.X. and Chater, K.F. 2003. “A rare leucine codon in adpA is implicated in the morphological defect of bldA mutants of *Streptomyces coelicolor*.” *Mol Microbiol* 50:475–486.

84. Tamura, K. 1992. “Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases.”*Mol. Biol. Evol.* 9(4):678–687.

85. Tamura, K. and Nei, M. 1993. “Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.” *Mol. Biol. Evol.* 10(3):512–526.

86. Trepanier, N. K., Jensen, S.E., Alexander, D.C. and Leskiw, B.K. 2002. “The positive activator of cephamycin C and clavulanic acid production in *Streptomyces clavuligerus* is mistranslated in a bldA mutant.” *Microbiology* 148(3):643 – 656.

87. Trifinopoulos, J., Nguyen, L-T., Haeseler, A. and Minh, B. 2016. “W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis.” *Nucleic Acids Research* 44(1):232–235. doi: http://doi.org/10.1093/nar/gkw256.

88. Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., Pilpel Y. 2010. “An evolutionarily conserved mechanism for controlling the efficiency of protein translation.” *Cell*141(2):344-54.

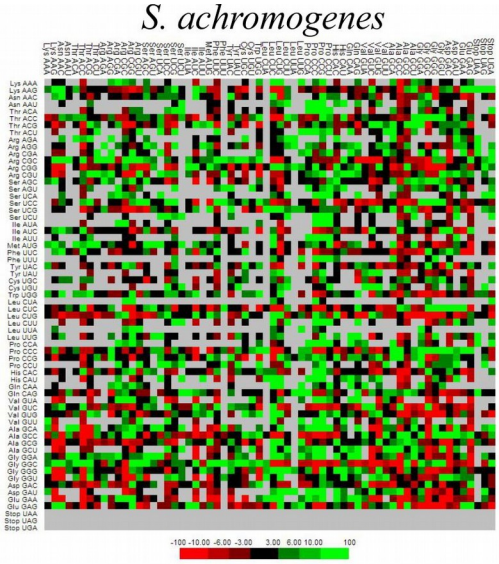
89. Chandra, G. and Chater, K.F. 2014. “Developmental biology of *Streptomyces* from the perspective of 100 actinobacterial genome sequences.” *FEMS Microbiol Rev.* 38(3):345-379. doi: 10.1111/1574-6976.12047.

90. Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G.F., Chater, K.F. and van Sinderen, D. 2007 “Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum.” *Microbiol Mol Biol Rev.* 71(3):495-548.
91. Wernersson, R. and Pedersen, A.G. 2003. “RevTrans: multiple alignment of coding DNA from aligned amino acid sequences.” *Nucleic Acids Res.* 31:3537–3539.
92. Whelan, S. and Goldman, N. 2001. “A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.” *Mol. Biol. Evol.* 18:691–699.
93. White, J. and Bibb, M. 1997. “bldA dependence of undecylprodigiosin production in *Streptomyces coelicolor* A3(2) involves a pathway-specific regulatory cascade.” *J. Bacteriol.* 179:627 – 633.
94. Wolanski, M., Donczew, R.,† Zawilak-Pawlik, A. and Zakrzewska-Czerwinska, J. 2015. “oriC-encoded instructions for the initiation of bacterial chromosome replication.” *Front Microbiol* 5:735. doi:10.3389/fm.
95. Yang, Z. 1997. “PAML: a program package for phylogenetic analysis by maximum likelihood.” *Comput Appl Biosci* 13(5):555-556.
96. Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. 2000. “Codon-substitution models for heterogeneous selection pressure at amino acid sites.” *Genetics* 55(1): 431–449.
97. Yang, Z., Nielsen, R. 1998. “Synonymous and nonsynonymous rate variation in nuclear genes of mammals.” *J. Mol. Evol.* 46:409–418.
98. Yang, Z., Nielsen, R. 2002. “Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.” *Mol. Biol. Evol.* 19:908– 917.
99. Yang, Z., Nielsen, R. and Hasegawa, M. 1998. “Models of amino acid substitution and applications to mitochondrial protein evolution.” *Mol. Biol. Evol.* 15:1600– 1611.

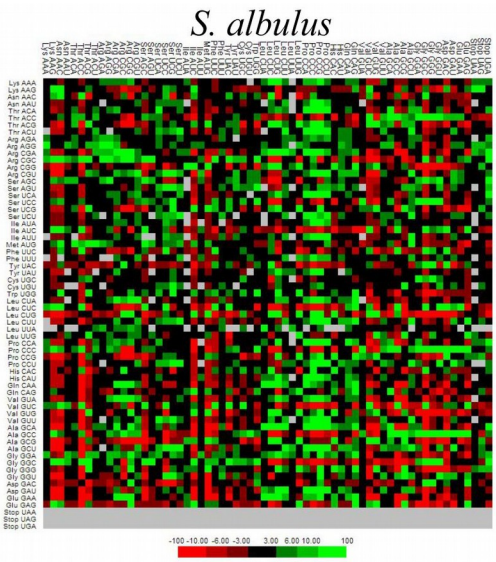
100. Yarus, M. and Folley, L.S. 1985. “Sense codons are found in specific contexts.” J *Mol Biol.* 182(4):529-40.
101. Zaburannyi, N., Rabyk, M., Ostash, B., Fedorenko, V., Luzhetskyy, A. 2014. “Insights into naturally minimised *Streptomyces albus* J1074 genome.” *BMC Genom.* 15:97. doi:10.1186/1471-2164-15-97.
102. Сиволоб, А.В. *Генетика : підручник*. 2008. Київ: Видавничо-поліграфічний центр "Київський університет".

Додаток А

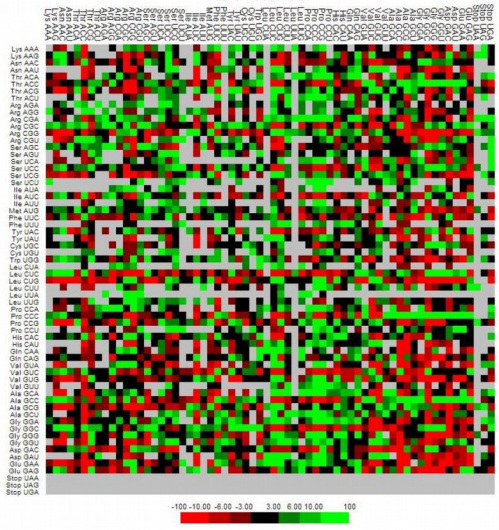
S. achromogenes



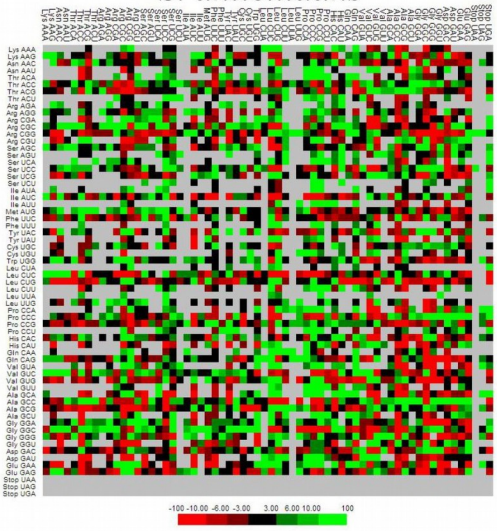
S. albulus



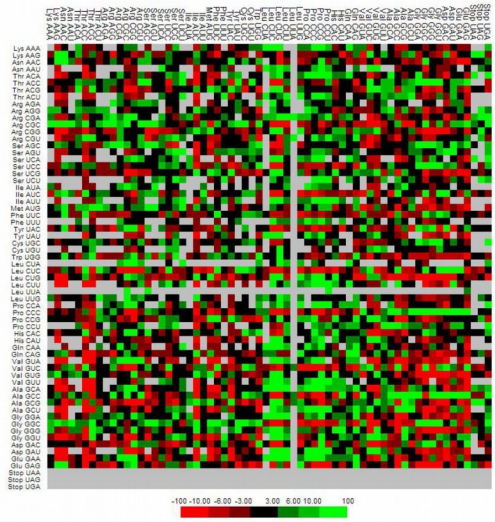
S. alni



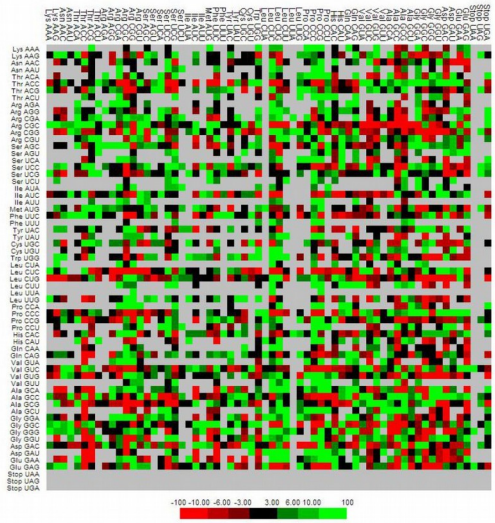
S. antioxidans



S. atratus



S. atriuber



Додаток Б

D2	*S. coelicolor*		*S. albus*		*S. ghanaensis*				*S. venezuelae*		*S. lividans*	
	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN
F	1	11	1	13	1	12	1	13	1	13	1	9
Y	1	8	1	9	1	9	1	9	1	9	1	8
H	1	12	1	12	1	12	1	12	1	13	1	12
Q	2	10	3	14	3	16	2	15	3	14	2	10
N	2	12	2	13	2	12	2	12	2	12	2	12
K	4	17	4	19	4	18	4	18	4	17	4	17
D	2	16	2	16	2	16	2	14	2	17	2	14
E	4	15	4	16	4	16	4	17	4	16	4	15
C	1	8	2	8	2	8	1	8	2	8	1	8

D4	*S. coelicolor*		*S. albus*		*S. ghanaensis*				*S. venezuelae*		*S. lividans*	
	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN
A	4	29	4	28	5	31	4	30	5	28	4	27
V	6	28	6	30	6	33	4	29	6	29	4	28
T	4	27	4	28	4	26	4	26	4	26	4	27
P	5	19	4	20	4	20	5	20	4	22	5	19
G	5	22	5	23	5	24	5	24	5	24	5	20

D6	*S. coelicolor*		*S. albus*		*S. ghanaensis*				*S. venezuelae*		*S. lividans*	
	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN
S	4	27	4	28	4	27	4	27	4	28	4	27
L	6	25	6	27	7	25	6	26	7	24	6	23
R	4	19	4	20	4	22	4	20	4	18	4	19

D3	*S. coelicolor*		*S. albus*		*S. ghanaensis*				*S. venezuelae*		*S. lividans*	
	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN
I	1	16	2	17	1	15	1	17	1	16	1	14

D1	*S. coelicolor*		*S. albus*		*S. ghanaensis*				*S. venezuelae*		*S. lividans*	
	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN	tF	tN
M	5	0	6	0	4	0	5	0	4	0	5	0
W	1	0	1	0	1	0	1	0	1	0	1	0

tF ККГ фокальної тРНК

tN ККГ близько-спорідненої тРНК

Di групи синонімічних кодонів

Додаток В

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.