НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»
Факультет біотехнології і біотехніки
Кафедра біоінформатики

«На правах рукопису»
УДК _____

До захисту допущено:
Завідувач кафедри
_____ Світлана ГОРОБЕЦЬ
«___»_____2020 р.

Магістерська дисертація
на здобуття ступеня магістра
за освітньо-науковою програмою «Біотехнології»
зі спеціальності 162 «Біотехнології та біоінженерія»
на тему: “Особливості алгоритму розрахунку кривих для опису та передбачення чутливості культур ракових клітинних ліній до хіміотерапії”

Виконав:
студент VI курсу, групи БМ-81мн
Ємельяновський Микита Ігорович _____

Керівник:
Ас. кафедри біоінформатики, к.ф.-м.н.,
Шевгалішин Роман Львович _____

Консультант з експериментального розділу:
Молодший лідер групи, PhD, провідний науковий дослідник,
Менден Майкл _____

Рецензент:
Зав. лабораторією нанокристалічних структур, к.ф.-м.н., науковий дослідник,
Дереча Дмитро Олександрович _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ – 2020 року

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет біотехнології і біотехніки

Кафедра біоінформатики

Рівень вищої освіти – другий (магістерський)

Спеціальність – 162 «Біотехнології та біоінженерія»

Освітньо-наукова програма «Біотехнології»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Світлана ГОРОБЕЦЬ

«___»_____2020 р.

ЗАВДАННЯ
на магістерську дисертацію студенту

Ємельяновському Микиті Ігоровичу

1. Тема дисертації «Особливості алгоритму розрахунку кривих для опису та передбачення чутливості культур ракових клітинних ліній до хіміотерапії», науковий керівник дисертації Шевгалішин, Роман, Львович, к.ф-м.н., асистент кафедри біоінформатики, затверджені наказом по університету від «___»_____ 20__ р. №_____

2. Термін подання студентом дисертації

3. Об’єкт дослідження

4. Предмет дослідження

5. Перелік завдань, які потрібно розробити

6. Орієнтовний перелік графічного (ілюстративного) матеріалу

7. Орієнтовний перелік публікацій

8. Консультанти розділів дисертації[1]

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
4. Практична частина	Менден М., молодший лідер групи, провідний науковий дослідник Інституту Комп'ютерної Біології, Гельмгольц Центра в Мюнхені		

9. Дата видачі завдання _____

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка

Студент Микита ЄМЕЛЬЯНОВСЬКИЙ

Науковий керівник Роман ШЕВГАЛІШИН

[1] Якщо визначені консультанти. Консультантом не може бути зазначено наукового керівника магістерської дисертації.

РЕФЕРАТ

Магістерська дисертація складається з пояснювальної записки формату А4. Пояснювальна записка містить 106 сторінок, 8 таблиць, 10 рисунків, 40 посилань.

Об'єктом дослідження є вихідні показники CV з ракових клітинних ліній та їх реакції на терапію.

Метою дослідження є оцінка характеру вихідних даних, аналіз роботи існуючого програмного алгоритму та формулювання технічного завдання модифікації вихідного коду програми для покращення її роботи.

Методи, використані в даному дослідженні, включали аналіз вихідних експериментальних даних CV, обчислення біостатистичних показників, *in silico* моделювання клітинних реакцій на хіміотерапевтичні препарати, визначення математичних аспектів алгоритму генерації логістичних кривих тощо.

Магістерську дисертацію було складено на основі літературного огляду науково-технічної літератури та невиданих матеріалів за темою дисертації, обміну ідеями під час проходження переддипломної практики, та проведення власних досліджень *in silico*.

Моделювання, розрахунки та дослідження літератури проводилися на локальному комп’ютері.

Результатом дослідження є опис неочікуваних реакцій ракових клітинних ліній на терапію, що поточна версія алгоритму оболбляє помилково. Було детально вивчено принцип роботи програми розрахунку регресії, вказано на можливі проблеми в її роботі, складено стартап-проект, зроблено висновки.

Прийняття описаних змін до даного програмного пакету може підвищити точність оцінки даних за змінами життєздатностей клітинних ліній. Це може позитивно вплинути на якість розроблюваних видів протиракових терапій, особливо тих, в яких використовується більше одного виду ліків.

ВСТУП

Фармакогеноміка - це сучасна галузь фармакології, яка вивчає вплив геному на терапевтичну відповідь. Вона є перспективним інструментом для розробки спрямованих терапій для лікування патологічних станів, зокрема, онкологічних. Нові види точних терапій, що мають більшу ефективність та менше побічних ефектів, створюються шляхом розробки та впровадження протоколів опису онкологічних фенотипів та генетичних біомаркерів за допомогою молекулярного профілювання. Однак, із збільшенням різновидів хіміотерапевтичних агентів, ускладнюється розуміння взаємодії між різними лікарськими засобами та їх впливу на різні типи раку та організм в цілому. Масштабні скринінги ракових клітинних ліній із розрахунком показників життєздатності клітин та розрахування кривої лінійної регресії є стандартом для оцінки ефективності протиракової терапії. Проте жорсткість алгоритму підгонки кривої призводить до отримання помилкових зведених показників скринінгів, через що клінічно цінні, але нестандартні реакції на ліки, можуть бути не взяті до уваги.

В даній роботі було проаналізовано вихідні дані проекту GDSC та роботу програмного пакету *gdscIC50* після застосування препаратів, та виконано спробу виявлення джерел помилок. Було показано, що до 10 препаратів призводять до збільшення життєздатності у значній кількості ракових клітинних ліній. На прикладі одного з таких препаратів, було визначено біомаркери клітинних відповідей на певні види терапії. Це демонструє, що непередбачувані реакції клітин на ліки, які можуть бути клінічно важливими, зазвичай ігноруються алгоритмами побудови логістичних кривих. Більш глибока молекулярна характеристика цих реакцій може призвести до розробки нових методів терапії, зокрема, використання відповідних сполук у якості хіміо-сенсибілізаторів, що підвищують ефективність інших ДНК-пошкоджуючих агентів.

Об'єктом даного дослідження є вихідні показники CV з ракових CCL та їх реакція на терапію

Мета даного дослідження — підтвердження гомо- чи гетерогенності вихідних даних, знайдення “випадаючих” позицій та формулювання технічного завдання з модифікації вихідного коду програми для уникнення помилок

Методи, використані в даному дослідженні, включали аналіз вихідних експериментальних даних CV, обчислення біостатистичних показників, *in silico* моделювання клітинних реакцій на хіміотерапевтичні препарати, визначення математичних аспектів алгоритму генерації логістичних кривих тощо.

Моделювання, розрахунки та дослідження літератури проводилися на локальному комп’ютері, з використанням спеціалізованого біоінформатичного програмного забезпечення.

Результатом дослідження є опис неочікуваних реакцій CCL на терапію, які не охоплюються поточною версією алгоритму. Також було детально вивчено принцип роботи програми розрахунку кривої та запропоновано зміни до формули алгоритму підгонки кривої, протестовано модель її роботи. Модифікований алгоритм також враховує несподівані та незвичайні реакції на ліки, особливо ті, що призводять до більш швидкого поділу клітин.

Прийняття описаних змін до даного програмного пакету може підвищити точність оцінки даних CV. Це може позитивно вплинути на якість розроблюваних видів протиракових терапій, особливо тих, в яких використовується більше одного виду ліків.

THE INTRODUCTION

Pharmacogenomics is a modern branch of pharmacology that studies the impact of genome on drug response. It is a bleeding-edge tool for the development of targeted therapies, that may be used to treat pathological conditions, in particular, cancer. The new types of precision therapies are created through the development of cancer description phenotype protocols and molecular profiling of genetic biomarkers. They have a potential to be more effective and feature fewer side effects than traditio*NA*l cancer treatments. However, with the increasing number of therapeutic options, it is difficult to evaluate the consequences of drug interactions, their effects on different cancer types and the whole human body. Large-scale screening of cancer cell lines with the calculation of cell viability, followed by curve fitting and IC_{50} calculations - is a standard for evaluation of therapy effectiveness. However, the rigidity of the curve fitting algorithm may negatively impact the generation of summary screening results in clinically valuable but non-standard drug reactions, which may result in erratic drug response data.

In this work, the algorithm for the calculation of the regression curve of cell viability was a*NA*lyzed and an attempt to identify the source of calculation errors was performed. For this purpose, the uncontrolled segmentation of the initial response data of cells to drugs that unexpectedly increased cell viability was involved. Up to 10 drugs were shown to increase viability in a considerable number of cell lines. For one particular exemplar drug, biomarkers of cellular responses to certain types of drug therapy were identified. This finding demonstrates that unpredictable responses to medicines that may be clinically relevant are usually ignored by logistic curve algorithms. A deeper molecular characterization of these responses could lead to the development of new therapies, in particular the use of the according drugs as chemosensitizers that enhance the effectiveness of other D*NA*-damaging agents.

The Master’s dissertation has 106 pages, 27 pictures, 5 tables and 48 references.

The object of this study is a sigmoidal function curve fitting algorithm and the calculation of informative cell viability statistics.

The aim of this study is to identify the source of the mistakes that are sopetimes generated by the curve-fitting algorithm of a software package, to identify and correct code errors that are present in the official version of the said package.

The methods used in this study include the analysis of raw cell viability data, biostatistics calculation, the *in silico* modeling of cellular responses to chemotherapeutic drugs, the determination of mathematical aspects of the curve-fitting algorithm and more.

The master's dissertation was compiled on the basis of a literary review of scientific and technical literature and unpublished materials on the topic of the dissertation, exchange of ideas during internship practice, and by conducting own research *in silico*. Modeling, calculations and research of the literature were performed on a local computer.

The result of the study is a description of the unexpected responses of cancer cell lines to therapy that the current version of the algorithm erroneously falters. The principle of operation of the regression calculation program was studied in detail, possible problems in its work were pointed out, a startup project was drawn up, conclusions were made. Adoption of the described changes to this software package can increase the accuracy of estimating data on changes in cell viability. This can positively affect the quality of the developed types of anticancer therapies, especially those that use more than one type of drug.

Зміст

Пеерлік умовних позначень, символів, скорочень і термінів

Апоптоз - запрограмована загибель клітин

LOF — *англ.,* loss of function, втрата функції

CV – *англ., с*ell viability, життєздатність клітин

CCL – англ, cancer cell lsf
ракові клітинні лінії

IC_{50} — *англ.,* the half maximal inhibitory concentration, концентрація інгібування ракових клітин наполовину

1 Огляд літератури

Існують переконливі докази того, що геноми ракових клітинних ліній (*англ*., cancer cell lines, CCL) можуть суттєво впливати на ефективність протиракової терапії. Наразі існують декілька прикладів геномних змін, що можуть бути використані як молекулярні біомаркери для виявлення пацієнтів, які можуть отримати найбільшу користь від певного лікування. Як приклад, використання лікарських препаратів що вибірково впливають на білки транслокації при хронічній мієлоїдній лейкемії або неналежному функціонуванні таких генів як *BRAF* при злоякісній меланомі, може прискорювати лікування цих захворювань та значно покращити рівень виживаності пацієнтів [1].

В останні роки онкологія досягла значного прогресу в розумінні природи молекулярних змін в CCL. Послідовні зусилля медиків та спеціалістів з біоінформатики вже призвели до появи детальних описів геномних змін, що відбуваються у багатьох підтипах раку. Повний перелік особливостей ракових генів дає глибоке розуміння зародження, еволюції та прогресування раку та слугуватиме поштовхом до розробки нових методів лікування онкологічних захворювань.

Для пришвидшення появи новітніх методів лікування раку, необхідно проводити доклінічні дослідження, які пов'язують геномні особливості раку з функціональними показниками, такими як чутливість до лікарських засобів. CCL, одержані від пухлин клінічних пацієнтів, породжуються багатьма різними типами раку, що мають кожний бути перевірені на резистентність до кількох видів хіміотерапії для оцінки її кумулятивного ефекту на злоякісні утворення [2].

Великий вплив на вибір протиракової терапії має тип перероджуваної тканини та геномні особливості раку. Разом вони є сприятливою системою для експериментальних маніпуляцій і стандартним інструментом дослідження в молекулярній біології та виробництві ліків. Для верифікації даних важливо, аби кілька однакових за дизайном

досліджень використовували одні й ті ж самі дані про лінії CCL, що допоможе пов’язати фармакологічні дані з геномною інформацією та визначити терапевтичні біомаркери. Ці дослідження показали, що фармакогеномічне профілювання ліній CCL може використовуватися як платформа для визначення біомаркерів для розробки нових методів лікування раку [2, 3].

Незважаючи на це, механізм та фармакокінетика терапевтичної відповіді на більшість онкотерапевтичних препаратів не є детально описаною, у зв’язку з чим постає необхідність у високопродуктивних скринінгах поєднань тисяч препаратів та CCL разом. Такі налаштування дозволяють проводити дуже великі обсяги експериментів “всліпу” та без залучення попередніх знань про цитостатичний вплив препаратів [3].

Метою даної магістерської дисертації є створення підґрунтя для модифікації та виправлення потенційних помилок в програмних алгоритмах *gdscIC50* для побудови кривих нелінійної регресії за даними зміни CV у відповідь на застосування до них певних хімічних сполук (*англ*., dose/response curve(-s), DRC(-s)). Вищесказане може сприяти розробці нових та вдосконалення вже існуючих видів комбінаційних онкотерапій.

Завданнями даної магістерської дисертації було провести послідовний аналіз вихідних даних проекту *GDSC*, довести гетерогенність вихідних даних, визначити алгоритм аналізу даних в пакеті *gdscIC50*, та спробувати визначити слабкі місця вихідного коду пакету *gdscIC50*, що призводять до помилок в обробці вихідних даних.

Ціллю даної магістерської дисертації було опанування програмними методами розрахунку і побудови DRCs, які є точним та відносно простим засобом для визначення та оцінки впливу хіміотерапевтичних препаратів на життєздатність CCL.

1.1 Особливості біології ракових захворювань

Реплікація ДНК - це надзвичайно точний процес, при якому помилка копіювання виникає приблизно один раз на 10^7 скопійованих пар нуклеотидів [4]. Крім того, існують численні білки, що виправляють помилки реплікації ДНК, зменшуючи частоту їх виникнення до 1 з 10^{10} пар основ [5]. Проте, постійні процеси поділу, у поєднанні з великою кількістю нуклеотидів, які повинні бути скопійовані для створення нової клітини, призводять до накопичення протягом життя людини великої кількості помилок в геномі, незважаючи на роботу репаративних білків. Процеси канцерогенезу є результатом накопичення цих випадкових соматичних мутацій в ракових генах, що можуть порушувати баланс між поділом клітин і апоптозом - запрограмованою загибеллю клітин. У поєднанні, ці два процеси призводять до неконтрольованого поділу клітин [4-6].

Підвищення вірогідність виникнення раку може виникати за рахунок впливу несприятливих факторів навколишнього середовища, таких як ультрафіолет, тютюновий дим тощо [6]. Ці загрози збільшують коефіцієнт мутації та успадкування конкретних генетичних варіантів зародкових ліній. Наприклад, пацієнти з відсутністю генетичного варіанту гену *BRCA1/2* мають підвищений ризик виникнення раку через порушення репаративного механізму ДНК [7]. Через те, що рак становить близько 15% причин загальної смертності, а процес накопичення мутацій в процесі життєдіяльності ніколи не зупиняється, що кожна людина, рано чи пізно, захворіє на рак. [8]. Отже, ефективна терапія раку має першорядне значення для суспільства та системи охорони здоров'я та є мабуть найбільшим викликом для людства.

1.2.1 Пасажирні та драйверні мутації

Пасажирними мутаціями називають ті соматичні мутації, які виявляються в геномі при дослідженнях ракових захворюваннях, проте не чинять впливу на перебіг захворювання. Через випадковий характер виникнення помилок ДНК при реплікації, пасажирні мутації виникають дуже часто. При дослідженні 12-ти основних типів раку, проведеного групою Кендота, було виявлено, що кожний тип раку мав в середньому близько 200 соматичних пасажирних мутацій та лише 2-6 драйверних мутацій [9]. Пасажирні мутації не впливають на перебіг онкозахворювань, однак вони є однією з причин неоднорідності генетичних профілів різних видів раку, тому їх наявність враховується при розробці хіміотерапій.

Драйверні мутації прискорюють поділ CCL і піддаються позитивній селекції під час виникнення і розвитку раку [10]. Загалом, для виникнення ракового захворювання мають відбутися численні мутації в так званих "*драйверних генах*". Мутації в драйверних генах можуть виявляти активаційний ефект або призводити до втрати функції гену (*англ*., loss of function, LOF). В літературі описані дві великі типи драйверних генів: зміни кількості копій гену(-ів) (*англ*., copy-number alteration, C*NA*) та мутації. В процесах канцерогенезу виникають обидва типи драйверних генів [11]. Відносний внесок C*NA* та мутацій у розвиток раку залежить від типу раку. Наприклад, виникнення раку нирок та колоректального раку обумовлено здебільшого мутаціями, тоді як рак молочної залози та яєчників провокується переважно C*NA*.

C*NA* присутні у більшості ракових захворювань [11]. Вони призводять до активації онкогенів за рахунок збільшення частоти делеції або ампліфікації генів [12]. Наприклад, ген *ERBB2* є дуплікованим у багатьох підтипах раку молочної залози [13]. Іншим типом драйверів ракових захворювань є злиття генів, або злиті гени (*англ*., fusion genes, FGs), які зазвичай є результатом транслокації генів. Відомим прикладом такої транслокації є зворотна транслокація між 9 і 22 хромосомами, яка

призводить до утворення так званої "*філадельфійської хромосоми*", що містить FGs *BCR-ABL1*. Це приклад активізуючого драйверу, оскільки активність *BCR-ABL1* є автономною і значно збільшується за рахунок промотору *BCRs*, порівняно зі звичайним варіантом гену *ABL1*. Така активація призводить до надмірної експресії RAS, активізує сигнальні шляхи, призводить до аномально збільшеного мітозу та неопластичного розширення [14-15]. *BCR-ABL1* також знижує клітинну адгезію до стромальної матриці і зменшує чутливість до апоптотичних подразників [8, 15]. Складні механізми виникнення FGs призводять до складених форм раку, таких як катаегіс, локалізована гіпермутація та хромотрипсис - тисячі кластеризованих хромосомних перебудов, що виникають поряд одна з одною [8-9].

1.3.1 Онкогени та гени-супресори пухлин

Ракові гени, що містять драйверні мутації, поділяють на дві окремі категорії: онкогени та гени-супресори пухлин (*англ.*, tumor suppressor genes, TSP). Онкогени викликають рак, активуючі мутації, які діють на збільшення швидкості клітинного циклу, поділу клітин, уникнення апоптозу та зупинку старіння клітин. Онкогени зазвичай містять так звані «гарячі точки» мутацій - області генів, в яких при ракових захворюваннях часто спостерігають утворення мутацій. Наприклад, близько 80% мутацій *KRAS*, що виявляють при колоректальному раку, знаходяться в кодоні №12 [16].

З іншого боку, TSP є "охоронцями" від раку, що часто піддаються впливу різноманітних LOF, таких як зсув рамки зчитування або нонсенс-мутації [17]. При виникненні LOF-мутацій в TSP, клітина втрачає контроль над клітинним циклом, що зупиняє старіння клітин та апоптоз під час розвитку пухлини.

Історично, першим виявленим TSP був ген *RB1*. LOF-мутації гену *RB1* присутні у широкому діапазоні ракових захворювань [17]. Такі мутації

інактивують білок ретинобластоми (pRb), який зупиняє клітини в першій фазі клітинного циклу G1, під час якої ДНК дублюється [18]. Це відбувається шляхом інгібування активності факторів транскрипції E2F [18].

Однак, функціональність ракових генів залежить від контексту мутації, і один і той самий ген може діяти як TSP або онкоген у різних типах пухлин. Наприклад, *TP53* давно відомий як ген-кодер пухлинного супресору, що має LOF-мутацію у більш ніж 50% усіх видів раку [18]. Протеїн p53, що кодується цим геном, запобігає канцерогенезу, індукуючи апоптоз та активуючи інгібітор p21, для зменшення швидкості поділу клітин [19]. Тим не менш, *TP53* також був визначений як онкоген в контексті певних видів раку, при цьому спостерігаються мутації набуття функцій (*англ*., gain of function, GOF) [19]. Як приклад, можна привести GOF-мутацію R248Q гену *TP53,* що виникає при раку молочної залози [18-19].

1.4.1 Соматичні драйверні мутації

За природою виникнення, соматичні драйверні мутації поділяються на несинонімічні мутації — заміну одного нуклеотиду та відповідна зміна амінокислотної послідовності білку — та мутації із наявністю невеликих вставок або інделів (*англ*. insertions and deletions, indels), які часто спричиняють мутацію зміщення рамки зчитування. Несинонімічні драйвені мутації можуть мати активуючу функцію. Як приклад, можна навести заміну нуклеотидів у положенні 600 гена *BRAF*, який перетворює валін у глутамінову кислоту і який пов'язаний з виникненням меланоми або колоректальним раком, або здатен викликати LOF-мутацію [19].

Утворення indels викликає утворення LOF-мутацій, такі як делеція 5 пар основ в кодоні 1309 гену *APC* — вона призводить до скорочення амінокислотної послідовності білку, який асоційовано з виникненням сімейного аденоматозного поліпозу (*англ*., familial adenomatous polyposis,

FAP) [20]. FAP є прикладом виникнення мутації у зародкових лініях і можливості успадкування певних драйверних мутацій [21]. Було показано, що деякі indels мають активізуючі властивості, як наприклад делеція в межах рамки зчитування в екзоні 19 *EGFR*, що асоціюється з підвищенням ризику виникнення раку легенів [20].

1.5.1 Характерні ознаки раку

Певні ракові гени, такі як *KRAS*, *RB1, TP53* та відповідні протеїни що ними кодуються, грають значні ролі у багатьох процесах клітинного циклу. Біологічні процеси, що сприяють розвитку пухлин, відомі як характерні ознаки раку, та є описаними в багатьох джерелах (Рис. 1.1) [20].

1.2 Види протиракової терапії

1.2.1 Хіміотерапія

З перших років існування протиракової терапії і до сьогодні, переважна більшість лікарських засобів, що застосовуються для лікування раку, є *хіміотерапевтичними препаратами* - неспецифічними цитотоксичними сполуками, які індукують клітинний стрес і часто вражають ракові клітини сильніше, ніж здорові.

Сутність дії більшості хіміотерапій полягає в тому, що ракові клітини діляться набагато швидше здорових, тому хіміотерапевтичні препарати впливатимуть на них сильніше і в першу чергу. Однак неспецифічність хіміотерапії призводить до виникнення серйозних побічних ефектів, особливо при дії на клітини з високою швидкістю поділу (наприклад, клітини кісткового мозку, волосяні фолікули та клітини травного тракту).

Механізми, за якими працюють хіміотерапевтичні препарати, є досить різноманітними і включають в себе: порушення механізмів мітотичного ділення, як, наприклад, порушення процесу розбору мікротрубочок, пригнічення синтезу ДНК, пошкодження ДНК тощо [21].

1.2.2 ДНК-пошкоджуючі сполуки

Сполуки, що пошкоджують ДНК, є найбільшою категорією хіміотерапевтичних препаратів. Хоча пошкодження і мутації ДНК є основною причиною раку, у вже сформованих CCL механізми репарації є або відсутніми, або погано розвинутими, тому вони не можуть ефективно усувати пошкодження свого геному. Якщо репараційні механізми CCL не спрацьовують, запускаються два інших механізми контролю: повна зупинка клітинного циклу (старіння клітин) або ініціація апоптозу [22]. В обох випадках поділ клітин припиняється. Отже, прийнятною є думка, що пошкодження ДНК зменшує неконтрольований ріст пухлин.

Як зазначено вище, сполуки, що пошкоджують ДНК, є неспецифічним методом хіміотерапії, що завдають шкоди усім живим клітинам. Незважаючи на це, такі сполуки є більш активними проти CCL, ніж проти здорових, оскільки прискорений клітинний цикл і ослаблені механізми відновлення ДНК CCL індукують апоптоз у CCL раніше, ніж у здорових [xxx]. Препарати, що входять до цього класу, також можуть бути складовими цільових терапій, наприклад інгібітори *PARP*, які ковалентно зв'язують *PARP* з ДНК і викликають подвійні розриви в ланцюгу ДНК [22-23].

Отже, сполуки, що пошкоджують ДНК залишаються ключовою терапією для боротьби з раком та є найбільш досліджуваним класом протиракових терапій [23].

1.2.3 Точна онкотерапія

Останніми роками, найбільші зусилля були направлені на розробку більш точних методів лікування раку, для зменшення появи несприятливих наслідків та підвищення загальної ефективності протиракової терапії. У 2015 році Сполучені Штати оголосили про фінансування «Ініціативи з розвитку точної медицини» на суму 215 мільйонів доларів, що має на меті розширення знань про генетичні особливості раку, шляхом розробки нових цільових методів лікування [24].

Невід'ємною складовою точної онкотерапії є молекулярне профілювання та ідентифікація біомаркерів раку для більш точного визначення пацієнтів, які отримують максимум користі від певного лікування — наприклад, через безпосереднє визначення мутацій в геномах CCL. При застосуванні препаратів з урахуванням генетичних біомаркерів пухлин, ефективність лікування значно зростає.

Крім того, впровадження систем доставки ліків із супутньою діагностикою прискорює затвердження препаратів регуляторами. Це

змінило парадигму онкотерапії з "один вид раку → один препарат" на "для кожного випадку і пацієнта - свій препарат" [24].

1.2.4 Цільова терапія

Цільова терапія ракових захворювань є синтезом усіх сучасних знань і вмінь точної онкології, адже вона спрямована на конкретні вузли в зміненій мережі сигналізації раку. Вона була показана як дуже ефективний засіб у лікуванні певних підтипів раку для яких були визначені молекулярні біомаркери, наприклад у лікуванні меланоми *BRAF* V600E, яку зазвичай лікують або інгібіторами *MEK*, або інгібіторами *BRAF*, а крім того, мають більш м'які побічні ефекти, порівняно з широкопрофільними терапіями [25]. На відміну від широкопрофільних терапій, що впливають на всі клітини, що швидко діляться, цільові терапії взаємодіють зі специфічним молекулам, які займають ключове місце в канцерогенезі (Рис 1) [26].

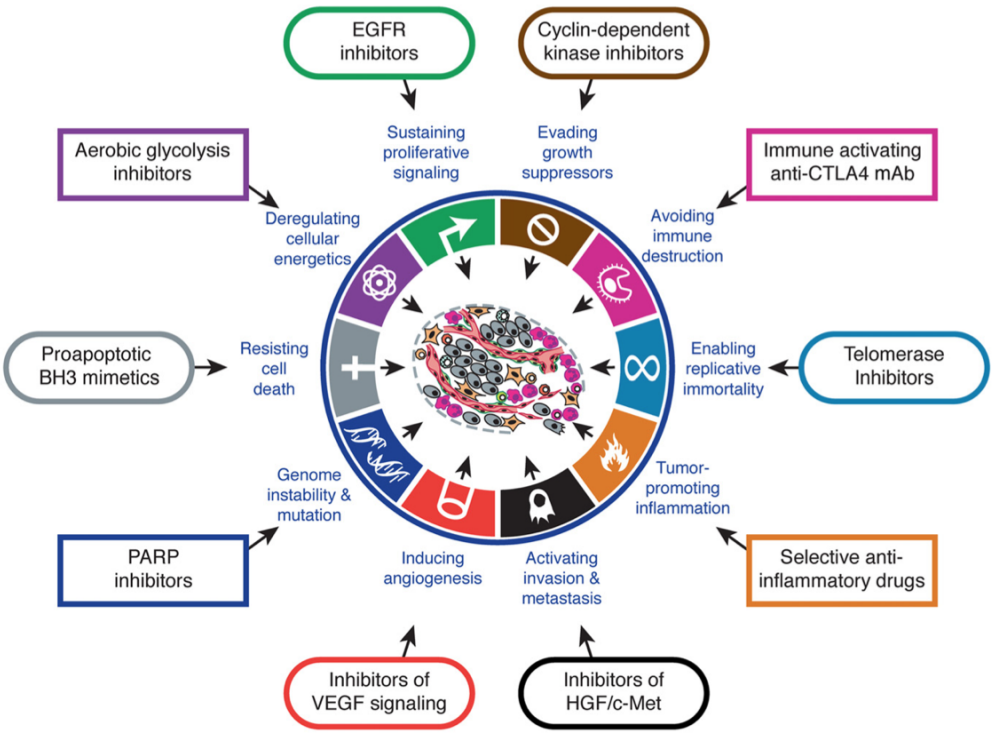


Рисунок 1. Характеристики раку та їх терапевтичні цілі

Додаткова відмінність полягає в тому, що спрямовані терапії частіше за все є цитостатичними (блокують розмноження клітин), тоді як традиційні препарати хіміотерапії є цитотоксичними (вбивають ракові клітини).

Одним із прикладів націленного використання протиракової терапії є використання інгібітору тирозинкінази *іматинібу* для лікування хронічної мієлоїдної лейкемії (*англ*., Chronic Myeloid Leukemia, CML) [27]. Характерною ознакою CML є наявність онкогена *BCR-ABL1*. Цей онкоген був визначений як ідеальна мішень для іматинібу, оскільки він присутній майже у всіх пацієнтів з CML, є унікальним для клітин лейкемії, значно експресується і є критичним для розвитку лейкемії [28]. Іматиніб виявляє свою терапевтичну дію шляхом інгібування транспорту фосфатів, опосередкованих *BCR-ABL1*, до субстратів. Це показало різке зниження CML при перших випробуваннях препарату [29].

Іншим прикладом успішного застосування цільової терапії є лікування нирковоклітинної карциноми (*англ*., Re*NA*l Cell Carcinoma, RCC), яку раніше лікували традиційними хіміотерапіями, що майже не призводили до ремісії. Потенціал цільової терапії для лікування RCC був високо оцінений після дослідження того факту, що більшість хворих на RCC мають мутації в гені *VHL*, які гальмують ріст пухлини через зміну процесу ангіогенезу [30].

Першими цільовими препаратами, які впливають на антиангіогенез та використовуються для лікування РКЦ, були сорафеніб та сунітиніб, чий успіх викликав хвилю терапій, націлених на RCC [30]. Починаючи з 2005 року, у США було затверджено 10 цільових препаратів для терапії RCC, серед яких: інгібітори рецептора фактора росту судинного ендотелію (VEGFR), такі як лентинініб, інгібітори mTOR, еверолімус, та інгібітор КП-1 та PD-1, ніволумаб [29].

Великий успіх у застосуванні цільової терапії сформував сферу сучасної онкології. Однак набута резистентність до препаратів зменшила

очікування вилікувати рак за допомогою єдиної цільової терапії самостійно, навіть якщо пацієнт має для цього усі показання. Це пояснюється тим, що сильний еволюційний тиск, спричинений цілеспрямованими методами терапії, швидко призводить до розвитку резистентності до лікарських препаратів в пухлин. Основна увага дослідників наразі сконцентрована на дослідженні еволюційних механізмів набуття пухлинами резистентності до лікарських препаратів та на розробці нових терапій для передбачення та попередження потенційних резистентностей до лікарських препаратів, ще до того, як вони виникають (Рис. 5) [29].

Рисунок 5: Ознаки раку та їх терапевтичні цілі

1.2.5 Комбінаційна терапія

Комбінаційна терапія (або терапія із застосуванням комбінацій препаратів) є стандартом сучасного лікування раку. Синергетичне або адитивне поєднання декількох схвалених та досліджених препаратів, націлених на окремі пов’язані між собою цільові шляхи, дозволяє швидко розробляти нові якісні види терапії, заощаджуючи при цьому кошти на розробку [31].

Використання комбінацій з декількох препаратів може зменшити вірогідність набуття раковими клітинами стійкості до лікування. Точність та відносно м'які побічні ефекти цільових терапій означають, що вони можуть одночасно атакувати кілька біологічних мішеней в межах одного виду раку, не викликаючи серйозних побічних реакцій та стійкості до певного виду хіміотерапії. Деякі окремі методи терапії раку дозволяють додаткове застосування *хемосенсибілізаторів* - засобів, що посилюють ефект хіміотерапії. Це не тільки збільшує ефективність лікування, але також підсилює основний ефект хіміотерапії, що дозволяє проводити

лікування меншою концентрацією ліків, зменшуючи вираженість побічних ефектів [31-32].

Прикладом комбінованої терапії з використанням хемосенсибілізаторів є використання фітохімічних речовин для лікування пацієнтів, що хворіють на потрійно-негативний рак молочної залози (*англ*., triple-negative breast cancer, TNBC). TNBC зазвичай лікують комбінаціями цитотоксичних препаратів, що часто викликають супутні побічні ефекти, рецидиви захворювання та розвиток резистентності до лікування.

Було показано, що фітохімічні речовини, такі як флавоноїди, здатні сенсибілізувати клітини TNBC до хіміотерапії та зменшувати резистентність до певних видів хіміотерапії [33]. Іншим розповсюдженим типом лікування раку в клінічних умовах є поєднання хіміо- та радіотерапії, що також підвищує ефективність лікування.

Отже, комбінаційна терапія є сучасним та ефективним методом лікування багатьох видів раку, який поєднує більшу ефективність з меншою вираженістю побічних ефектів, порівняно з традиційною терапією.

1.3 Біомаркери раку

Потенціал цільових методів терапії для лікування раку повноцінно розкривається при врахуванні наявності біомаркерів чутливості до лікарських засобів, наприклад онкогену *BCR-ABL1* при раку виду CML та LOF-мутацій у *VHL* TSG у RCC, як було показано раніше. Виявлення біомаркерів є необхідним першим кроком для визначення найбільш сприятливого лікування для певної групи пацієнтів.

Прикладом важливості біомаркерів у терапії раку є відкриття ролі мутованих генів *EGFR* та *KRAS* у аденокарциномі легенів. Пацієнти, які мають мутації *EGFR*, зазвичай добре реагують на терапію препаратами [xxx]. Однак мутації *EGFR* та *KRAS* є несумісними - за наявності однієї мутації, інша не утворюється. Тому пацієнти, які мають мутований ген

KRAS і ген *EGFR* дикого типу, на вищезазначене лікування вже не реагують [34]. Звідси, мутація *KRAS* є біомаркером ефективності певних препаратів для лікування аденокарциноми легенів.

Нещодавний успіх схвалених цільових методів терапії в клініці викликав хвилю нових досліджень з метою виявлення біомаркерів раку та розробки більш ефективних терапій. Крім того, відкриття біомаркерів може значно скоротити термін затвердження нових протиракових засобів.

Отже, цільова терапія може бути ефективно розроблена та призначена правильним пацієнтам лише за умови повного розуміння їх генетичних мутацій шляхом ідентифікації біомаркерів.

1.3.1 Визначення біомаркерів

Біомаркерами називають показники нормальних біологічних або патогенних процесів, що можуть бути об'єктивно виміряні та оцінені. Цим терміном також називають характерні фармакологічні реакції на терапевтичне втручання [35]. Для більш швидкого та точного виявлення біомаркерів слід оцінити ефект цілого ряду препаратів на великій кількості ліній CCL. Такий підхід до скринінгу високопродуктивних препаратів збільшує шанси на знаходження генетичних маркерів чутливості до певного препарату та виявляє генетичну неоднорідності реальних пухлин *in vivo*. Однак, через високу вартість та складність таких досліджень, необхідно досягати певного компромісу між точністю виявлення біомаркерів та доцільністю проведення таких експериментів взагалі.

Для проведення експериментів з виявлення біомаркерів традиційно використовували моделі *in vitro* та *in vivo*. Обидва типи моделей мають властиві їм недоліки і переваги, які будуть розглянуті далі.

1.3.2 Переваги і недоліки моделей *in vitro* та *in vivo*

Загалом, найбільш надійні та клінічно значущі результати для ідентифікації біомаркерів можна отримати просто тестуючі препарати на

людях. Однак, на практиці, такі експерименти вважаються неетичними, особливо без знань про безпечність таких препаратів, адже виявлення біомаркерів найчастіше відбувається на перших стадіях розробки терапій. Тваринні моделі, такі як миші і примати, мають найбільшу схожість з людськими моделями, але їх використання пов'язане з етичними проблемами, та є занадто складним для масштабних експериментів.

Через неможливість застосування повноцінних живих моделей, стандартом для виявлення онкологічних біомаркерів стали масштабні скринінги CCL. За результатами цих скринінгів генерують нові наукові версії та гіпотези, які потім перевіряють на тваринах, підбираючи схеми дозування. Лише після цього, препарати тестують на людях, що обмежує патологічний вплив на здоров’я піддослідних [35].

1.3.3 Проблема гетерогенності реальних пухлин

Більшість існуючих моделей *in vitro*, такі як “безсмертні” лінії CCL, є простими, порівняно з реальними живими моделями. CCL не охоплюють всю складність будови реальних пухлин, проте є більш практичним, етичним та більш економічно-ефективним рішенням.

З точки зору наближення до реальних біологічних моделей, найкращими є первинні пухлинні моделі - зразки, отримані безпосередньо з тканини раку людини. Однак вирощування та підтримування цих зразків для масштабних скринінгів є дуже складними. В даний час розробляються методи аналізу даних отриманих від пацієнтів, для передбачення еволюції пухлини для кожного окремого пацієнта. Проте створення масштабного скринінгу з використанням живих органоїдів все ще є складним завданням, а його переваги перед скринінгами ліній CCL є сумнівними.

Тому, незважаючи на найменшу клінічну значимість, скринінги CCL є найбільш ефективним методом ідентифікації біомаркерів на даний момент: такі клітини мають велику спорідненість з раковими тканинами *in vitro*, а отримані результати мають реальну клінічну значимість [34].

1.3.4 Ракові клітинні лінії

Ракові клітинні лінії (*англ*., cancer cell lines, CCL) - це клітини, отримані з клітин пухлин людини, які є природньо або штучно мутованими та безсмертними для тривалого культивування *in vitro*. Першою раковою клітинною лінією, шо була культивована та використовувалася в дослідженнях, стала HeLa, отримана з клітин хворої на рак шийки матки жінки в 1951 році [35]. Наразі в клінічній практиці використовують понад 1000 ліній CCL у формі одношарових культур *in vitro* [35-36]. Незважаючи на зручність використання, такі культури мають добре задокументовані недоліки, найважливішими з яких є наступні:

1. Геном CCL змінюється з часом, через що вони не зберігають ті генетичні особливості, якими володіли на початку культивування. В таких лініях виявляються значні генетичні та транскрипційні зміни [36], що робить результати, отримані з них, певною мірою нерепрезентативними;

2. В межах CCL, усі клітини є генетично однорідними, через що вони не зберігають генетичну гетерогенність, яку виявляють в реальних пухлинах;

3. CCL культивуються *in vitro* і не містять важливих компонентів пухлинного мікросередовища, які присутні в раковій тканині людини [37].

Однак, простота використання та низька вартість CCL, роблять моделі CCL більш придатними для виявлення біомаркерів, порівняно з іншими методами. Великі панелі CCL мають настільки велику кількість CCL в межах однієї лінії, що з них отримують майже вичерпний спектр клітинної відповіді на лікарські засоби, що опосередковано представляє реальну гетерогенність в межах ракових пухлин.

Отже, використання великих панелей CCL є виправданим та дозволяє зафіксувати ту гетерогенність відповідей на лікування, яка спостерігається в межах клінічної практики [37].

1.4 Масштабні скринінги CCL

1.4.1 Сучасні скринінги ракових CCL

Після розробки і тестування першого скринінгу CCL NCI-60, стало зрозуміло, що для відображення всіх особливостей лікування ракових захворювань, CCL треба дуже багато [37]. Тоді було докладено значних зусиль для збільшення пропускної здатності панелей скринінгів. Прикладом результату таких зусиль є панель Бодмерської лабораторії, яка наразі містить понад 120 ліній клітин колоректального раку [38]. Дослідження показали, що панелі такого розміру можуть точно репрезентувати профілі експресії мРНК та загальні мутації CCL, пов'язані з конкретними підтипами раку, що збільшує шанс виявлення нових біомаркерів [37-38]. Панель ліній клітин колоректального раку також продемонструвала високу відповідність клінічним даним, завдяки наявність в ній диких типів *KRAS*, *NRAS*, *BRAF* та *PIK3CA* [39].

Найбільшого прогресу в аналізі масштабних панелей CCL досягли два проекти: "Геноміка Чутливості Раку до Препаратів" (*англ*., Genomics of Drug Sensitivity in Cancer, GDSC) та "Енциклопедія Ракових CCL" (*англ*., Cancer Cell Line Encyclopedia, CCLE). Перші скринінги проекту GDSC поєднували 130 препаратів та 639 CCL, тоді як наступні містили вже 265 лікарських препаратів та 10000 CCL [40]. Скринінги CCLE поєднували 24 препарати та 479 CCL [41]. Подальшого розвитку скринінги CCLE дістали в проекті "Відкриття та Дослідження Ракових Цілей (*англ*., Cancer Target Discovery and Development, CTD^2), в рамках якого було проаналізовано 545 малих молекул лікарських засобів у поєднанні з 907 лініями CCL [42]. Зі збільшенням розміру сучасних масштабних скринінгів CCL, картина гетерогенності пухлин стає все більш насиченою, що напряму впливає на відкриття і обґрунтованість знань про біомаркери раку [43].

Проект GDSC був заснований для більш глибокого розуміння молекулярних характеристик ракових захворювань та оцінки впливу хімічних сполук на зміни CV CCL. Проведення масштабних скринінгів з

високою пропускною здатністю здійснюється разом проектом GDSC в Інституті Веллком Трест Сангер (WTSI) та Центром молекулярної терапії в загальній лікарні штату Массачусетс (CMT).

База даних чутливості раку до терапії (*GDSC database* або *GDSC archive*) була заснована для зберігання результатів скринінгів проекту GDSC та надання доступу до них усім зацікавленим. Сайт *GDSC* зберігає вихідні дані скринінгів про чутливість CCL до лікарських засобів і пов'язує ці дані з геномною інформацією для виявлення молекулярних біомаркерів реакції на препарати. Дані проекту постійно перевіряються, оновлюються та доповнюються новими результатами скринінгів [43].

Хімічні сполуки для скринінгів GDSC надаються промисловістю, науковими співробітниками, або отримуються від комерційних лабораторій. Для кожної тестованої сполуки обирається такий диапазон концентрацій, за яких вона буде ефективно інгібувати ракові клітини, відповідно до свого профілю дії. Різноманіття цих сполук є надзвичайно великим: сюди входять затверджені препарати, що застосовуються в клініці; препарати, що знаходяться у процесі доклінічної розробки та препарати, що наразі тільки проходять клінічні випробування. Усі ці сполуки охоплюють широкий спектр молекулярних мішеней та процесів, пов'язаних з біологією раку, включаючи вплив на передачу сигналів у клітині, контроль клітинного циклу, відповідь на пошкодження ДНК, нестабільність цитоскелету тощо [44].

Отже, масштабні скринінги відповідей CCL на терапію є ефективним сучасним методом для виявлення та опису біомаркерів.

1.4.3 Особливості скринінгів GDSC та CCLE

Точність висновків, зроблених за даними двох найбільших скринінгів CCL, неодноразово піддавалася сумніву. Відштовхуючись від низького рівня кореляції між даними скринінгів, дослідники виявили велику розбіжність між результатами сканувань, що потенційно робить

обидва скринінги абсолютно ненадійними для відкриття і розробки біомаркерів [45].

У відповідь на ці закиди, дослідники GDSC та CCLE опублікували спільну статтю, де показали високий рівень сумісності між даними обох скринінгів [43]. Було представлено три основні аргументи на користь якості отриманих даних:

1. Чутливість до терапії - це відносне поняття, що залежить від діапазону концентрації застосовуваного препарату. Крім того, класифікація реакції на препарати залежить від способу, за допомогою якого ідентифікують чутливі CCL. Скринінги GDSC та CCLE використовували різні діапазони концентрацій для багатьох лікарських засобів та різні методи визначення чутливості до лікарських засобів. Таким чином, визначення CCL як чутливих в одному дослідженні, і як резистентних - в іншому, не обов'язково вказує на неправильність результатів чи висновків досліджень [43].

2. Незважаючи на відмінності у методології проведення експериментів, усі клінічно значущі біомаркери були однаково визначені та підтверджені в обох скринінгах з високим ступенем узгодженості. Це свідчить про те, що незважаючи на розбіжності у результатах фармакологічних досліджень, обидва скринінги виявили ті самі біомаркери, що підтверджувало якість даних для цілеспрямованої терапевтичної розробки [43].

3. Коефіцієнт кореляції Спірмена (*англ.*, Spearman correlation coefficient, Sp) є некоректною методикою для оцінки узгодженості даних цих двох скринінгів. Sp оцінює монотонне відношення даних на основі їх рангового порядку, а не вихідних необроблених даних. При застосуванні такого коефіцієнту буде враховано довільний порядок тих CCL, що не прореагували на терапію та не є біологічно значущими з точки зору дослідження біомаркерів. Це може викривити оцінку кореляції, особливо враховуючи, що більше 80% пар ліній клітинних клітин взагалі не

відповідають на терапію. Коефіцієнт Пірсона, який оцінює лінійну залежність вихідних даних, може бути більш придатним показником для порівняння результатів двох скринінгів - що підтверджується більш високою кореляцією при повторній оцінці результатів двох скринінгів цим коефіцієнтом [43].

Отже, дані між двома найбільшими скринінгами CCL є узгодженими та такими, що не містять значних відхилень одного від іншого.

1.4.4 Ідентифікація біомаркерів чутливості до лікарських засобів

Використання масштабних скринінгів для визначення чутливості CCL до конкретного препарату є лише першим кроком до визначення біомаркера чутливості. Додатково, проводять глибоку молекулярну характеристику панелі скринінгу CCL для визначення мутацій, зміни кількості копій (C*NA*s), профілів метилювання та рівнів експресії генів, що можуть допомогти пояснити взаємозв'язок між ліками і клітинними лініями. Скринінги GDSC та CCLE надають ці дані для більшості досліджень CCL, аби пришвидшити та полегшити відкриття біомаркерів [45].

1.5. Фармакологічні дані

1.5.1. Показники CV

В проекті GDSC, реакцію CCL на лікарські засоби оцінюють за допомогою аналізу CV (CV), де до клітин в лунках на спеціальній пластині додається певний лікарський препарат у визначеному діапазоні концентрацій та флуоресцентний барвник, чия інтенсивність випромінювання світла прямо пропорційна до кількості АТФ в даній лунці. Після 72 годин очікування, барвник проявляють на спеціальному сканері, вимірюючи інтенсивність флюоресценції в кожній лунці. Значення флуоресценції додатково коригують відповідно до лунок з живильним середовищем, без клітин (РС-0 або РС-1, в залежності від виду живильного середовища, що застосовують) та пустих лунок В (*англ.*, blank — пустий,

порожній), аби скорегувати інтенсивність остаткової флуоресценції. Кількість живих клітин вимірюють відносно контрольної групи, щоб визначити CV, специфічне для кожної комбінації клітини/ліки [46].

Щоб зробити результати скринінгів більш зрозумілими, розраховують формулу кривої нелінійної регресії, що описує значення флуоресценції для кожної комбінації клітина/препарат з мінімальними відхиленнями; з формули такої кривої потім обчислюють зведені показники. Як правило, значення CV буде високим при низьких концентраціях ліків і буде знижуватись за нахилом сигмоподібної кривої при збільшенні доз препаратів. Крім того, будь-який застосовуваний препарат, незалежно від його токсичності, виявить летальну дію на клітини за достатньо великої концентрації, що також враховується в формулі [46].

Формула сигмоїдної кривої, що застосовується в пакеті *gdscIC50* стверджує, що значення CV будь-яких CCL знижується від 1 (або 100% живих клітин) до нуля (або 0% живих клітин). Використання цих суворих припущень допомагає зробити криві нелінійної регресії стійкими до високого рівня шуму у вихідних даних, що позитивно впливає на якість аналізу життєздатності. Після визначення формули нелінійної регресії, з неї обраховуються два важливі підсумкові показники: IC_{50} - концентрація препарату, що знижує CV наполовину (*англ.*, half maximal inhibitory concentration), та *AUC* - площа під кривою DRC (*англ.*, area under the curve) (Рис. 3).

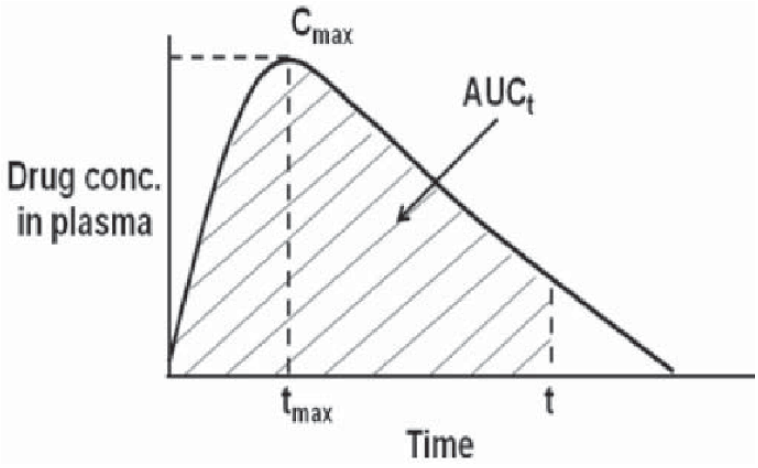


Рисунок 3. AUC — площа під кривою концентрації препарату

Як підсумковий показник для виявлення біомаркерів, найчастіше використовують коефіцієнт IC_{50}, оскільки він включає інформацію про вихідний діапазон концентрації лікарського засобу, яка відсутня у показнику *AUC*. Порівнюючи значення IC_{50} для різних комбінацій DRC, можна визначити чутливість та резистентність біомаркерів раку до певних препаратів**.**

1.5.2 Особливості коефіцієнту *IC50*

Використання будь-якого зведеного коефіцієнту є певним компромісом між наглядністю і точністю репрезентації реальних даних. Значення IC_{50} вважають репрезентативними, коли їх можна обчислити за допомогою інтерполяції - тобто, коли значення IC_{50} лежить в межах просканованого діапазону концентрації ліків. Однак, якщо значення CV не опускається нижче 50% в межах випробуваного діапазону концентрації хімічної сполуки, значення IC_{50} необхідно екстраполювати за межі концентрації препарату або обмежувати, через що показник IC_{50} втрачає свою точність.

В експерименті *GDSC* для розрахунку значень IC_{50} шляхом екстраполяції використовується баєсова сигмоїдна модель. На відміну від *GDSC*, в результатах скринінгів *CCLE* просто наводять максимальну концентрацію ліків, яку перевіряють як IC_{50}, що робить резистентні

клітини такими, що не можна порівнювати з іншими. Підсумковий показник IC_{50} також апроксимує криві реакції на дозу до одиниці, тим самим видаляючи інформацію про нахил кривої, діапазон концентрації та шум у даних.

Коефіцієнт нахилу кривої регресії може бути корисним при оцінці рівня чутливості клітин до препарату. Інформація про діапазон концентрації може бути використана для співставлення одних і тих же комбінацій ліків і клітин, сканованих у різних діапазонах концентрацій, а кількісне визначення шуму в кривій реакції може допомогти визначити достовірність результатів. Крім того, значення IC_{50} плутаються зі змінною швидкістю поділу клітин. Значення IC_{50} відрізняються високою чутливістю до кількості поділів клітин, що виникають під час аналізу, на що можуть впливати природні відмінності клітинного поділу, умов росту та змінної тривалості експерименту.

1.6 Неочікувані фенотипи реакцій на препарати

1.6.1 Маскування неочікуваних реакції на препарати

Показник IC_{50} не може бути використаний для виявлення нішевих реакцій на препарати, а саме:

1. CCL, що збільшують життєздатність. Таке трапляється, якщо препарати прискорюють клітинний цикл CCL. За показником IC_{50} ці відповіді будуть класифіковані як нечутливі, що є помилкою;

2. CCL, з максимальним значенням CV (E_{max}) вище нуля. Це відбувається, якщо CCL розвивають внутрішню стійкість до препарату, динамічно змінюючи свій профіль експресії генів. Показник IC_{50} вважає, що за збільшення концентрації застосовуваної хімічної сполуки, CV буде прямувати до нуля, через що його значення для таких комбінацій CCL та препарату не є репрезентативним;

3. CCL, динаміка значень CV яких включає численні плато. Це відбувається, якщо препарати мають несподівані ефекти. Значення IC_{50} не

беруть до уваги форму DRC, тому такі нестандартні відповіді не враховуються.

1.6.2 Переваги використання вихідних даних для оцінки CV

Шляхом прямого оцінювання даних впливу препаратів на CCL, можна висунути гіпотезу про гомогенність реакцій CCL на препарати, що призведе до більш повного розуміння природи таких нестандартних реакцій та CCL, що їх виявляють. Крім того, при такому оцінюванні зберігається важлива інформація про нахил кривої лінійної регресії, діапазон концентрації препаратів та рівень шуму в реакціях CCL на препарати. Це збільшує об’єм даних, які потрібно проаналізувати приблизно в 7-9 разів (оскільки більшість препаратів присутні в скринінгу в семи, або в дев'яти різних концентраціях), що збільшує час та складність аналізу даних, проте є дуже важливим при перевірці та оцінюванні відповідних наукових гіпотез.

1.6.3 Переваги кластеризації даних для оцінки CV

Неконтрольована сегментація - це ефективний та потужний інструмент для пошуку структури у великих об’ємах даних. Метод неконтрольованої сегментації є надійним інструментом для виявлення гетерогенності вихідних даних ще до проведення будь-якого аналізу. Даний метод часто застосовується у випадках, коли необхідно заздалегідь оцінити, скільки різних типів даних можна виділити з великого масиву інформації [45].

Одним з найбільш розповсюджених методів неконтрольованої сегментації вихідних даних є аналіз основних компонентів даних (*англ*., principal component analysis, PCA) [46]. PCA використовує ортогональне лінійне перетворення для проектування даних на нову систему координат. Таким чином, найбільший за розміром розподіл даних лежить на першій

координати (перший головний компонент), другий за величиною — на другій координати (другий головний компонент) тощо [xxx]. Застосовуючи метод PCA до даних про зміни життєздатностей клітин, отримують групування подібних реакцій на препарати, які потім можна додатково візуалізувати відповідно до інших показників. Інші сучасні методи непідконтрольної сегментації, такі як нейронна мережа, краще підходять для кластеризації складних і більш об'ємних даних, але є набагато більш складними для обчислення [47].

1.6.4 Завдання цього дослідження

Основною метою даної дипломної роботи є виконання аналізу вихідних даних життєздатностей ракових клітин на лікарські засоби, отриманих в рамках проекту GDSC для відокремлення та визначення різних фенотипів реакцій на ліки. Робота заснована на вже існуючій гіпотезі, що вихідні дані скринінгу GDSC можуть містити несподівані реакції на препарати, які є значущими з точки зору фаркмакогеноміки, проте ігноруються або неправильно класифікуються автоматизованим алгоритмом пакету *gdscIC50*.

В ході даної роботи було проведено групування вихідних даних за показником монотонності значень CV, виконано кластеризацію вихідних даних скринінгу GDSC, аналіз шуму даних, обчислено просту лінійну регресію значень CV на діапазон концентрацій хімічних сполук, з обчисленням відповідних коефіцієнтів, проведено гіпергеометричний тест для препаратів та ліній CCL, визначено можливі позиції вихідного коду, що можуть бути відповідальними за неправильну поведінку програми, запропоновано зміни до вихідного коду програми, що можуть виправити її помилки, описано стартап-проект, що може допомогти у комерційній реалізації ідеї даної роботи та зроблено відповідні висновки за результатами роботи.

В подальшій роботі буде використано дані молекулярної характеристики шляхів лікарських препаратів для розуміння залежності змін експериментальних значень CV в CCL від певних біомаркерів раку, виконана спроба пояснити ці особливі реакції та буде виконана спроба об’єднати вже існуючі алгоритми в межах одного програмного коду, з реалізацією додаткових функцій для нестандартних станів та типів даних, для подальшого використання даного продукту в клінічних умовах.

2 Хід роботи

2.1 Отримання клінічних даних для аналізу

Для скринінгу використовували базу даних *GDSC*, що зберігає дані про характеристики більше 1000 CCL, які є представниками спектру всіх найпоширеніших та найгірших видів раку, отриманих від пацієнтів дорослого та дитячого віку. Ці CCL мають епітеліальне, мезенхімальне та гематопоетичне походження, і використовуються в більшості досліджень проекту *GDSC*.

Геномні дані в GDSC синхронізуються з базою даних соматичних мутацій раку (*англ*., catalogue of somatic mutations in cancer, COSMIC) - вільного ресурсу анотації та характеризування соматичних мутацій ракових захворювань [6-8]. Для проведення аналізу показників життєдіяльності, CCL витримують в середовищі з 5% або 10% FBS, і 1% пеніциліном або стрептоміцином; щільність клітин в кожній чашці має бути оптимізована. Оптимальне число комірок для кожної клітинної лінії визначається так, щоб клітини знаходяться в фазі росту для максимального збільшення діапазону вимірювань кінцевих точок. Через 24 години після засіву клітини обробляють хімічними сполуками у відповідних дозах. Після обробки планшети повертають в інкубатор. Через 72 години до усіх комірок додають флуоресцентний барвник та сканують усі планшети, визначаючи інтенсивність флюоресценції. Всі експерименти підлягають жорстким заходам контролю якості, всі скановані комбінації (CCL + хімічна сполука (х/с) в певному діапазоні концентрацій), що не проходять контроль якості, видаляються з подальшого аналізу.

На даний момент, проект GDSC об’єднує два окремих скринінги, GDSC-1 та GDSC-2, що відрізняються наступними показниками (Табл. 1):

Таблиця 1. Показники і характеристики двох скринінгів GDSC.

Характеристика	**GDSC-1**	**GDSC-2**
Рік проведення скринінгів	2009 - 2015	2015 — до сьогоднішнього дня
Кількість лунок в планшеті	96 - 384	1536
Флуоресцентний барвник	Syto60 або Резазурин	Promega CellTiter-Glo
Спосіб внесення препаратів	Апарат для внесення рідин з тонким сталевим наконечником	Акустичний дозатор Echo555 (Labcyte)
Спосіб розведення препаратів перед внесенням до клітин	9-ти кратне розведення сполуки, включаючи 2-кратний етап розведення (діапазон у 256 разів) 5 -ти кратне розведень дози, включаючи 4-кратний етап розведення (256-кратний діапазон)	7-ми кратне розведення сполуки , включаючи етап розведення наважки (1000-кратний діапазон) 7-ми кратне розведення сполуки, включаючи 2-кратне розведення та 4-кратне розведення (1024-кратний діапазон)
Позитивний контроль	Пуста лунка (B)	Пуста лунка (B)
Негативний контроль	Клітини + живильне середовище (NC-0)	Клітини + живильне середовище + ДМСО (NC-1)

Дані розведення, назв лунок позитивного і негативного контролю були використані далі для приведення усіх даних масиву в уніфіковану форму для аналізу.

2.2 Розрахунок кривих доза/відповідь

Усі DRC розраховуються за допомогою пакету функцій *nlme* (*англ.,* nonlinear mixed effects models - нелінійні моделі зі змішаними ефектами), що викликається пакетом *gdscIC50* за потреби. В процесі аналізу результатів скринінгів, для кожної експериментальних комбінацій (1

клітинна лінія + 1 хімічна сполука) розраховується формула логістичної регресії та показники IC_{50} та AUC. Дані життєздатності клітин зі скринінг-планшетів встановлюються для кожної окремої DRC за значеннями інтенсивності флуоресценції через формулу багаторівневої логістичної моделі з фіксованим плато. Як було зазначено раніше, дослідники GDSC постулюють, що форма кривої реакції клітин на препарати є завжди сигмоподібною та S-подібною. Тому, та сама функція застосовується для обрахування усіх скринінгів, зроблених проектом. *nlme* - вільний пакет для обчислення лінійних та нелінійних моделей регресії - використовує два параметри для опису сигмоїдальної кривої та розраховує її формули одночасно для всього масиву або його частини, за вибором дослідника.

Поточний випуск GDSC включає дані про чутливість до лікарських засобів для 503 унікальних хіміотерапевтичних сполук та для 1065 CCL. Панель CCL охоплює 30 різних типів раку, а бібліотека препаратів містить великий арсенал хімічних сполук, орієнтованих на більшість терапевтичних цілей [7].

В середньому, одна хімічна сполука тестується на 525 клітинних лініях, а разом вони формують 570,161 експериментальних комбінацій, що аналізуються в скринінгах GDSC. База даних постійно оновлюється, додавання нових та оновлення існуючих даних відбувається кожні 4 місяці (Рис. 2).

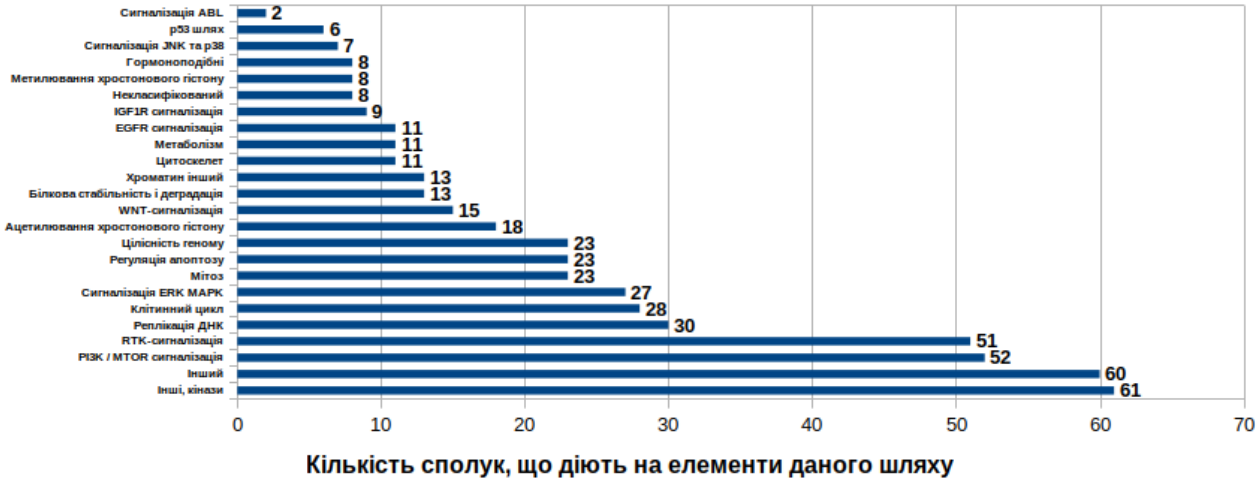


Рисунок 2. Онкологічні профілі, з якими взаємодіють препарати з бібліотеки сполук GDSC

2.3 Дизайн дослідження GDSC

Дані, що були використані у цьому аналізі, були отримані в електронному вигляді безпосередньо з сайту бази даних проекту GDSC в форматі *.csv*. В цьому форматі, дані містяться у вигляді простих текстових файлів, з кінцевим числом рядків (або випадків) та фіксованою кількістю стовпчиків (або критеріїв). Між собою стовпчики розділені комою, крапкою чи спеціальними символами, а перший символ в кожному стовпчику позначає початок рядку. При цьому передбачається, що дані є гетерогенними хоча б за кількістю критеріїв, за якими їх вимірюють, хоча певні значення в стовпчиках можуть бути відсутні, на що вказує спеціальний символ *NA* (*англ*. not available - недоступно) (Рис. 3).

	BARCODE	SCAN_ID	COSMIC_ID	MASTER_CELL_ID	CELL_ID	CELL_LINE_NAME	SEED_DENS	DRUG_ID	TAG	CONC	INTENSITY
1	100541	1765	924238	365	2415	K5	250	NA	UN-USED	NA	68759
2	100541	1765	924238	365	2415	K5	250	NA	B	NA	19639
3	100541	1765	924238	365	2415	K5	250	NA	B	NA	21506
4	100541	1765	924238	365	2415	K5	250	NA	NC-0	NA	170731
5	100541	1765	924238	365	2415	K5	250	1007	L1-D1-S	1.250000e-02	47351
6	100541	1765	924238	365	2415	K5	250	1007	L1-D2-S	6.250000e-03	37567
7	100541	1765	924238	365	2415	K5	250	1007	L1-D3-S	3.125000e-03	59623
8	100541	1765	924238	365	2415	K5	250	1007	L1-D4-S	1.562500e-03	64894
9	100541	1765	924238	365	2415	K5	250	1007	L1-D5-S	7.812500e-04	89824
10	100541	1765	924238	365	2415	K5	250	1007	L1-D6-S	3.906250e-04	125884
11	100541	1765	924238	365	2415	K5	250	1007	L1-D7-S	1.953125e-04	140537
12	100541	1765	924238	365	2415	K5	250	1007	L1-D8-S	9.765625e-05	119293
13	100541	1765	924238	365	2415	K5	250	1007	L1-D9-S	4.882812e-05	151012
14	100541	1765	924238	365	2415	K5	250	NA	NC-0	NA	145067
15	100541	1765	924238	365	2415	K5	250	1024	L2-D1-S	2.000000e+00	27587

Рисунок 3. Зовнішній вигляд масиву даних GDSC-1 з найважливішими для роботи програми показниками, з 24 по 29 рядок

Найважливішими стовпчиками в масиві даних є:

- **BARCODE** – унікальний штрих-код кожного планшету з лунками;
- **SCAN_ID** – унікальний ідентифікатор для сканування пластини зчитувачем при отриманні даних флуоресценції. Пластина може скануватись не один раз, але лише один SCAN_ID пройде внутрішній КК. Тому між опублікованими даними між BARCODE і SCAN_ID існує повна відповідність;
- **COSMIC_ID** – ідентифікатор CCL у зовнішній базі даних COSMIC. Між MASTER_CELL_ID та COSMIC_ID існує відповідність один до одного, тому колонка MASTER_CELL_ID не була використана при формуванні робочої таблиці даних;
- **DRUG_ID** – унікальний ідентифікатор хімічних сполук, які застосовують для обробки клітин. Може приймати значення *NA*, якщо в комірка є контрольною, пустою, або видаляється під час проходження контролю якості;
- **CONC** — концентрація хімічної сполуки, що використовувалася, у мікромолях. Може приймати значення *NA*, якщо в комірка є контрольною, пустою, або видаляється під час проходження контролю якості;

- **INTENSITY** — вимірювання флуоресценції в кінці аналізу. Флуоресценція є результатом усього скринінгу, з якого обчислюється показник CV;
- **TAG** — ідентифікатор речовин, які додаються до комірки. На одну комірку може бути більше одного тегу на одну комірку, наприклад, при наявності двох комірок з однаковими комбінаціями клітинна лінія/хімічна сполука, проте з різними поживними середовищами та/ чи додатками до них.

Решта стовпчиків слугує для запису службової інформації та може бути уникнена в даному експерименті, але позбуватися їх не можна через потенційні втрати важливої інформації [9].

До та після кожної послідовності титрувань знаходяться комірки негативного (B) та позитивного контролю (PC-0 або PC-1 в масивах GDSC-1 та GDSC-2, відповідно). Вони необхідні для того, аби зробити корекцію рівня випромігювання з експериментальної лунки на фонове випромінювання та більш точно розрахувати значення життєздатності клітин (*англ.*, cell viability, CV) За ними виконується розрахунок міри виживаності клітин за формулою 1:

$$CV_{drug} = \frac{Idrug - Iblank}{Icontrol - Iblank}$$

Формула 1: Визначення міри виживаності клітин

де CV_{drug} – загальне виживання клітин в даній комірці, I_{drug} – міра флуоресценції для оброблюваної комірки, $I_{control}$ – середня інтенсивність в лунках на даній пластині, яку обробляли ДМСО; I_{blank} – міра флуоресценції з пустих комірок (поправка на фонову флуоресценцію), що розташовані до та після титрованої послідовності [10].

CV є показником, що обчислюється першим та присутній в усіх без виключення аналізах даних життєздатності. За ним будують графіки

впливу препаратів на CCL, розраховуючи залежність CV від концентрації сполуки в звичайній чи в логарифмічній формі.

Отже, в даному експерименті необхідно було представити характеристики кривих DCR у відповідності до концентрацій препаратів.

3 Результати

Отримані проміжні результати показують, що існуючі методи аналізу можуть бути непридатними для правильної оцінки значущих показників впливу терапії на CCL для певних комбінацій CCL та сполук. Далі було ініційовано перевірку всіх даних з масиву, аби виключити можливість виявлення інших типів помилок даних.

3.1 Нормалізація даних

3.1.1 Оцінка монотонності вихідних даних

Перш за все, була проведено оцінка монотонності вихідних даних.

За передбаченнями консорціуму GDSC, оцінка даних пакетом *gdscIC50* відбуватиметься з найменшою кількістю помилок за умови використання лише монотонних даних, адже відхилення реальних значень від прямої лінійної регресії буде мінімальним та майже не буде містити слідів шуму. Монотонними вважалися ті результати, в яких кожне значення градієнту формули лінійної регресії для певної комбінації було або нижче за попереднє, або вище, або постійно дорівнювало 0, за формулою:

$$((x_i - x_{i+1},\ y_i - y_{i+1}) \geq 0) \quad \vee \quad ((x_i - x_{i+1},\ y_i - y_{i+1}) \leq 0$$

Формула 1. Математичний критерій монотонності значень вибірки

,що означає наступне: при збільшенні або зменшенні незалежної змінної, залежна змінна має змінюватися строго пропорційно до зміни залежної змінної.

Ця формула, записана мовою програмування R, виглядає наступним чином:

$$Monotonic = all(dif(m) > 0) \mid all(dif(m) < 0) \mid all(dif(m) = 0),$$

Формула 2. Програмний запис критерію монотонності значень вибірки

,що означає наступне: Вектор даних є монотонним, якщо кожне наступне значення в цьому векторі є ***або*** більшим за всі попередні, ***або*** менше за всі попередні, ***або*** дорівнює всім попереднім. Це припущення додатково автоматично видаляє з масиву даних ті значення, середньоквадратична похибка яких є надто великою, адже до монотонних даних просто застосовувати лінійну регресію.

Для монотонних значень життєздатності, в окремий масив даних була додана колонка, що приймає лише три значення (Табл. 1):

Таблиця 1. Показники монотонності та їх значення в аналізі даних проекту *GDSC*

Значення	**-1**	**0**	**+1**
Які значення CV представляє	Монотонні, що *спадають*	Монотонні, що *не змінюються*	Монотонні, що *зростають*

В даному дослідженні нас особливо цікавлять CCL, чиї значення CV є монотонно зростаючими Тому з даного масиву даних було отримано лише ті групи даних, що мали значення “+1” в колонці монотонності. Ці дані були візуалізовані із застосуванням графічного пакету ggplot2 для R.

Дані життєздатності клітин, що пройшли даний другий етап фільтрації, виглядають майже як лінійна залежність. Для візуалізації профільтрованих даних були побудовані графіки залежностей CV від $\log_{10}$(CONC) для 100 випадково відібраних монотонно зростаючих комбінацій. На графік було також поміщено значення $IC_{50,}$ які були обраховані консорціумом GDSC, для візуальної оцінки їх коректності.

Візуалізація усіх даних виконувалася за допомогою спеціально розробленої комп'ютерної програми для роботи з пакетом для мови R *ggplot2* (Рис. 7).

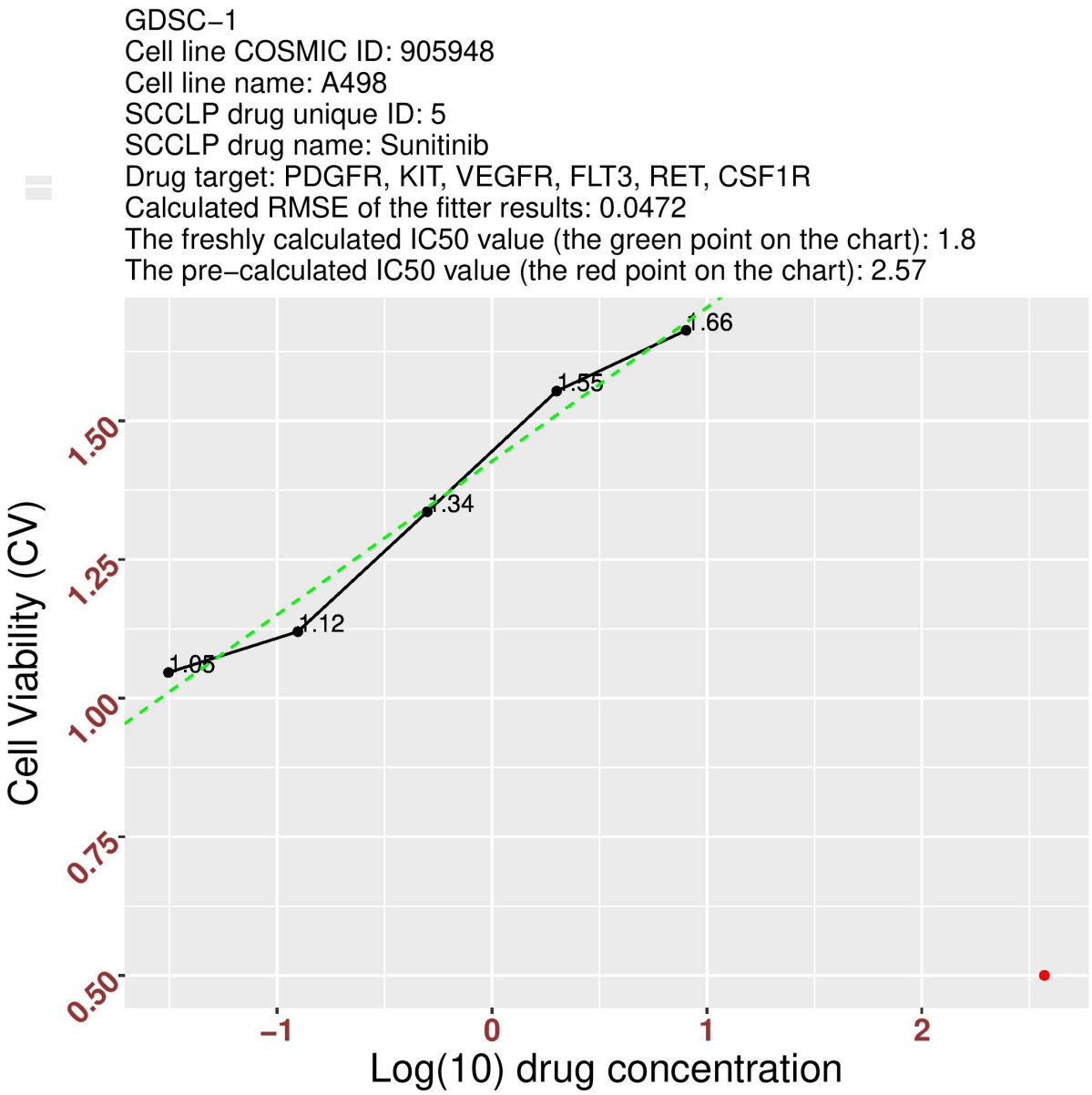


Рисунок 7. Монотонний графік залежності життєздатності клітин від концентрації застосовуваного препарату.

На вісі абсцис показано значення концентрації х/с, що додається до CCL, на вісі ординат — значення CV, розраховані для окремої пари CCL та х/с. Зеленим пунктиром показано графік лінійної регресії, що була побудована для значень CV даної комбінації.

Показана комбінація пройшла жорсткий фільтр оцінки монотонності даних. Зелена переривчаста лінія - пряма простої лінійної регресії до цих даних. Червоною точкою показане значення IC50, розраховане консорціумом GDSC, що вочевидь, не є правильним для даної послідовності даних. Загалом, з усіх векторів значень CV що монотонно зростали, значення IC_{50} не було обраховано коректно в жодному разі. Це змусило нас провести більш ґрунтовний аналіз даних на наявність головних компонент.

3.1.2. Базовий принциповий компонентний аналіз

PCA здебільшого провели на підмножині комбінацій які монотонно зростали або зменшувалися, тобто лише для чистих даних, як це було задумано розробниками *GDSC*. Така підмножина становить близько 19% від загального числа даних. Отримані результати показують значну відмінність у характері CV що зростають, порівняно із CV що спадають, що є загальним домінувальним трендом (Рис. 8).

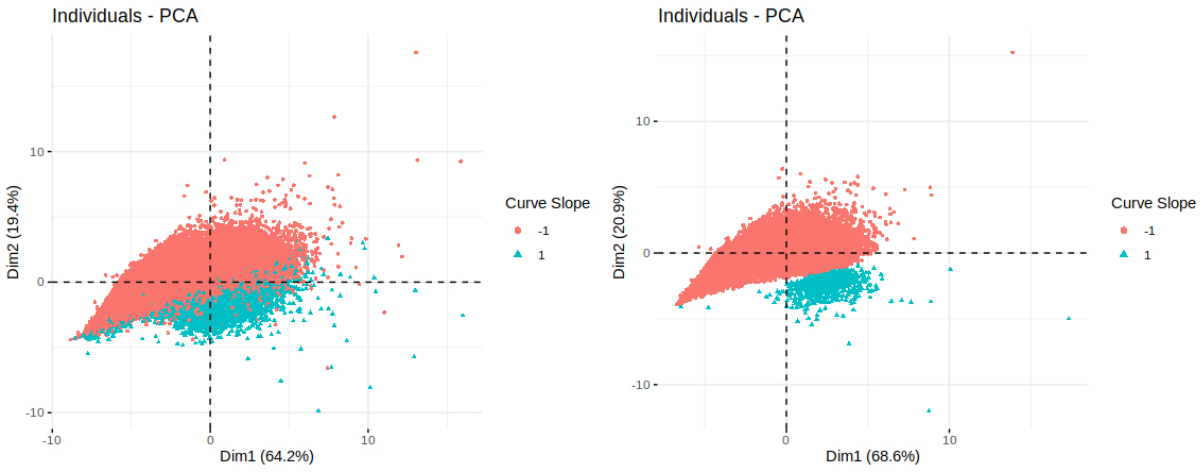


Рисунок 8. Результати PCA для всіх значень вибірки (графік зліва) та для даних, що монотонно змінюються (графік зправа). Застосовано кольоровий фільтр за знаком показника монотонності

З графіку вище видно, що характер усіх комбінацій зі зростаючою життєздатністю (блакитний колір), є дуже відмінними за природою від характеру комбінацій зі спадаючою життєздатністю (рожевий колір). Навіть до застосування фільтру за монотонністю, блакитні дані кластеризуються дуже компактно та окремо від великого кластеру звичайних реакцій. Додатково було проведено швидка перевірка даної гіпотези — з тих комбінацій, що під час PCA кластеризувалися в кластер зростаючих значень, були випадково обрані 5 комбінацій, значення CV з яких нанесли на графік та виконали побудову простої лінійної регресії до кожної комбінації. За результатами даного порівняння, було підтверджено наявність у вихідних даних GDSC CCL, що збільшують значення CV при застосуванні до них певних хімічних сполук. Подібні хімічні сполуки, що збільшують життєздатність окремих CCL, були умовно називаються CCLP (*англ*., cancer cell lines proliferators, підсилювачі ракових клітинних ліній).

3.1.3 Фільтрація за градієнтом “шуму” вихідних даних

У вихідних даних експерименту містяться 224,202 комбінації ліків-клітин: 189,046 з них вимірювали в 9 концентраціях лікарських засобів (точки титрування, серія розведення в 2 рази), 34,305 вимірювали в 5 точках титрування (серія 4-кратного розведення), а 851 з них вимірювали в обох концентраціях. При аналізі подібних даних за допомогою методу PCA виникнуть помилки, адже дані знаходяться в різних концентраційних групах, що спотворить результати аналізу. Тому перед початком аналізу даних було проведено зменшення кількості титраційних точок для всієї групи комбінацій з 9 титраційними точками до 4-кратного ряду розведення 5 точок титрування шляхом видалення парних значень та відповідного зміщення залишкових значень.

Беручи до уваги, що кожен експеримент знаходиться в межах одного планшета та титрується послідовно, у вихідних даних флуоресценції можуть бути наявні так звані *шуми* (*англ*., noise) — окремі значення, що не

вписуються у загальну картину розподілу даних, і які здатні змінити значення коефіцієнтів регресії та призвести до неправильної інтерпретації результатів експерименту [11]. Для зменшення впливу шуму на дані, їх необхідно “профільтрувати” — видалити з них ті CV, що вибиваються із діапазону більшості . Для цього було використано фільтрацію даних за “відстанями Кука” для оцінки істотності окремих значень флуоресценції для побудови регресійної кривої (Фор. 3-4).

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2}$$

Формула 3. Формула “відстані Кука” для окремих значень з масиву даних.

$$D_i > 4 \times \frac{\sum_{j=1}^{n} D_j}{n}$$

Формула 4. Формула межі значень “відстаней Кука”, за якою приймається рішення про подальший аналіз або відкидання певного значення.

Для кожного значення CV в масиві даних було розраховано окреме значення “відстаней Кука”. Після цього, з масиву даних видалили усі рядки, чиї показники “відстані Кука” були більше за 0.8 (за Форм. 3: $D_i > 4 \times$ кількість вимірювань (5 для всіх даних), отже ⅘ = 0.8).

Після проведення фільтрації даних за “відстанню Кука”, був обчислений новий відносний коефіцієнт шуму даних, що не є настільки суворим як фільтр даних за критерієм монотонності та залишає більше значень для аналізу, при східній якості фільтрування сторонніх значень. В

цьому показнику спочатку визначається абсолютна (за модулем) різниця значень життєздатності клітин у векторі даних (титрування одним препаратом), від якого піднімається число диапазону значень усіх CV для однієї комбінації (Формула 5):

$$\eta_{drug} = \left(\sum_{i=1}^{n-1} |CV_{tp_i} - CV_{tp_{i+1}}| \right) - \text{range}(CV_{tp_1}, CV_{tp_2}, ..., CV_{tp_n})$$

Формула 5: Відносний коефіцієнт шуму η

,де CV_{tpi} - CV в *i*-й точці титрування певною х/с, CV_{tpi+1} - CV в точці титрування *i+1,* *n* — загальна кількість точок титрування, *range* — діапазону значень усіх CV для однієї комбінації.

Показник $\eta = 0$, якщо реакція на препарат є монотонною; він зростає тим швидше, чим менш монотонними стають дані. Після обчислення значень *η,* було виконано встановлення межі, відповідно до якої виконувалась фільтрація. Була взята вибірка з 10 комбінацій, з рівнями шуму від 0 до 4.5, з кроком рівня шуму 0.5. Дані комбінації були проаналізовані вручну для того, аби емпірично встановити верхню межу рівня шуму та відповідно нижню межу надійності даних. За результатами цього аналізу, межа значень коефіцієнту шуму була встановлена на рівні $\eta = 0.5$.

Даний фільтр виключає близько 15% усіх значень масиву даних через високий рівень шуму та близько 8.5% значень у вибірках з низьким рівнем шуму в даному експерименті, при цьому дані результати майже повністю співпадають з кількістю профільтрованих вручну значень (14.7% та 9.1% значень у вибірках з високим та низьким рівнем шуму, відповідно) [14].

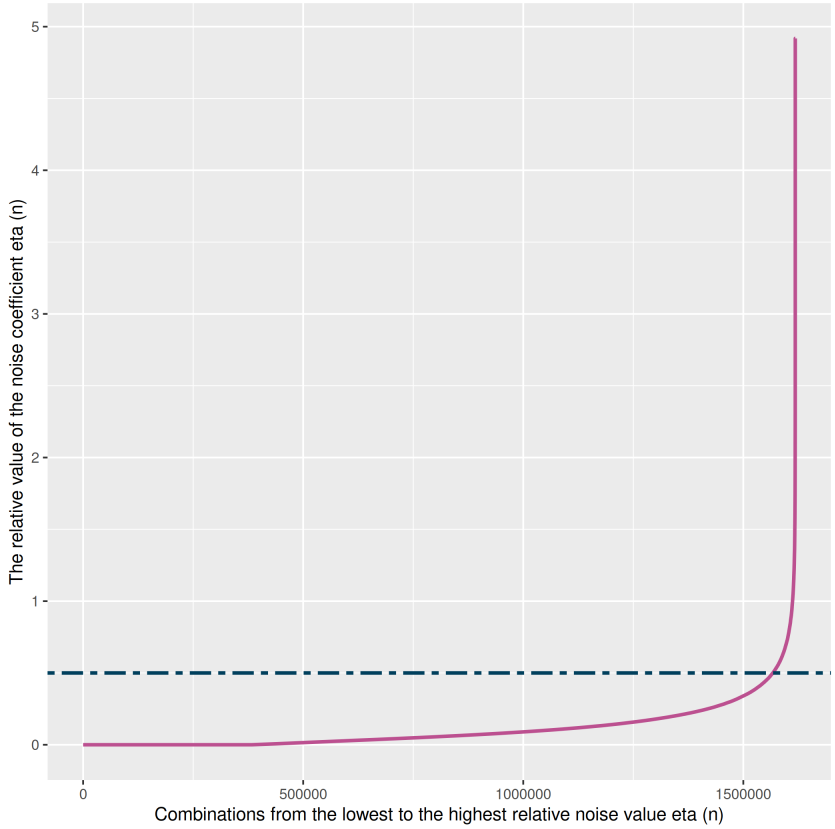


Рисунок 2.2. Графік розподілу значень коефіцієнту шуму η за значеннями в масиві вихідних даних GDSC. Пунктирною лінією позначено поріг значень η = 0.5, який є межею для визначення аномальних клітинних відповідей

3.1.4 Фільтрація за градієнтом лінійної регресії

Дані, що залишилися в масиві після 2-х фільтрацій, було згруповано за ідентифікаторами штрих-коду планшету, клітинної лінії та лікарського засобу. Для виявлення декількох взаємодіючих геномних особливостей, що впливають на кожну реакцію на препарат, до результатів регресійного аналізу було додано дані про транскрипційні особливості клітин або тканини, що мають відношення до даного типу раку після чого був

проведений гіпергеометричний тест на визначення збагачення результатів скринінгу.

Для визначення геномних маркерів реакцій на препарати використовують два аналітичні підходи. Підраховані коефіцієнти чутливості до лікарських засобів - IC_{50} та *scale* (нахил кривої доза/відповідь) - співвідносять зі змінами, що відбуваються в геномі CCL: з точковими мутаціями, індукціями та інделами. Для цього необхідно визначити індивідуальні геномні особливості, пов'язані з чутливістю до лікарських засобів, масштаб виявленого ефекту та статистичну значущість кожної асоціації генів та лікарських засобів [7].

Після розрахунку формули простої лінійної регресії з методами лінійного моделювання, було визначено розподіл коефіцієнтів нахилу регресійної прямої для всього масиву даних та побудований відповідний графік. Було вирішено встановити мінімальне значення, за яким проводитиметься фільтрація, на рівні $m = 0.3$, усі комбінації з $m < 0.3$ були відфільтровані з масиву (Рис. 2.1).

Після видалення з даних шуму, лінійна регресія, подібна до тої, що використовувалась у формулі дистанції Кука, була обчислена знову за наступною формулою:

$$\boldsymbol{y = mx + n + e,}$$

де $\boldsymbol{y}$ - залежний фактор лінійної регресії, $\boldsymbol{m}$ - коефіцієнт лінійної регресії, $\boldsymbol{x}$ - незалежний фактор лінійної регресії, $\boldsymbol{n}$ - значення $\boldsymbol{y}$, за якого $\boldsymbol{x} = 0$ (також показує, за якого значення $\boldsymbol{x}$, пряма регресії перетинатиме вісь ординат), $\boldsymbol{e}$ - значення довільної помилки моделі

Знак коефіцієнту лінійної регресії m має великий біологічний сенс. Він показує, гинуть клітини під дією певної хімічної сполуки (значення $m < 0$), чи є вони резистентними до певних хімічних сполук (значення $m = 0$ або m

~ 0) , та чи призводить дана хімічна сполука до прискорення поділу CCL ($m > 0$).

Додатково, чисельний показник m вказує на швидкість відповідно інгібування (при $m < 0$) або збільшення CV ($m > 0$) CCL. Наприклад, при m = -1.3 клітини діляться більш повільно та за менших концентрацій хімічного агенту, ніж при m = -0.9.

Для визначення хімічних сполук, що збільшують CV, було розраховано лінійну регресію за усіма значеннями CV в масиві даних та побудовано графік розподілу значень m. Далі було здійснено емпіричний підбір такої нижньої границі значень m, вище якої знаходяться такі комбінації CCL та х/с, що ракові клітини в них демонструють значну швидкість поділу. За результатами аналізу було вирішено встановити нижню границю значень градієнту лінійної регресії m на рівні m = 0.3 (Рис. 4).

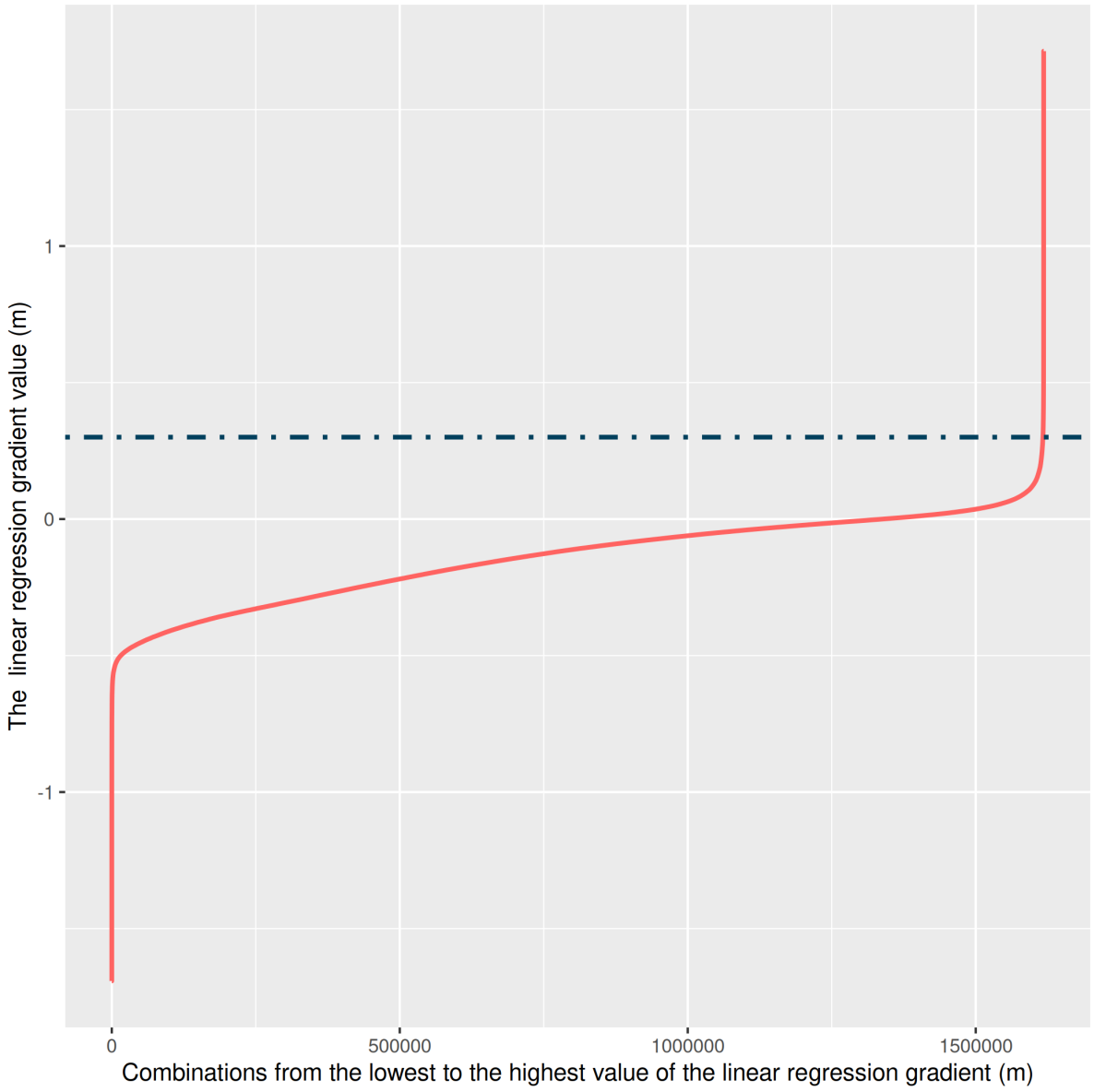


Рисунок 4. Графік розподілу значень градієнту лінійної регресії m в масиві вихідних даних.

Після цього було візуалізовано розподіл значень за масивом вихідних даних та наведено емпірично встановлені показники для відносного коефіцієнту шуму даних η та градієнту лінійної регресії m. Було реалізовано невелику програму, що самостійно обчислює обидва значення та будує кольоровий графік в RStudio. В подальшій роботі буде виконано спробу повністю автоматизувати процес фільтрації даних за коефіцієнтами шуму та регресії (Рис. 6).

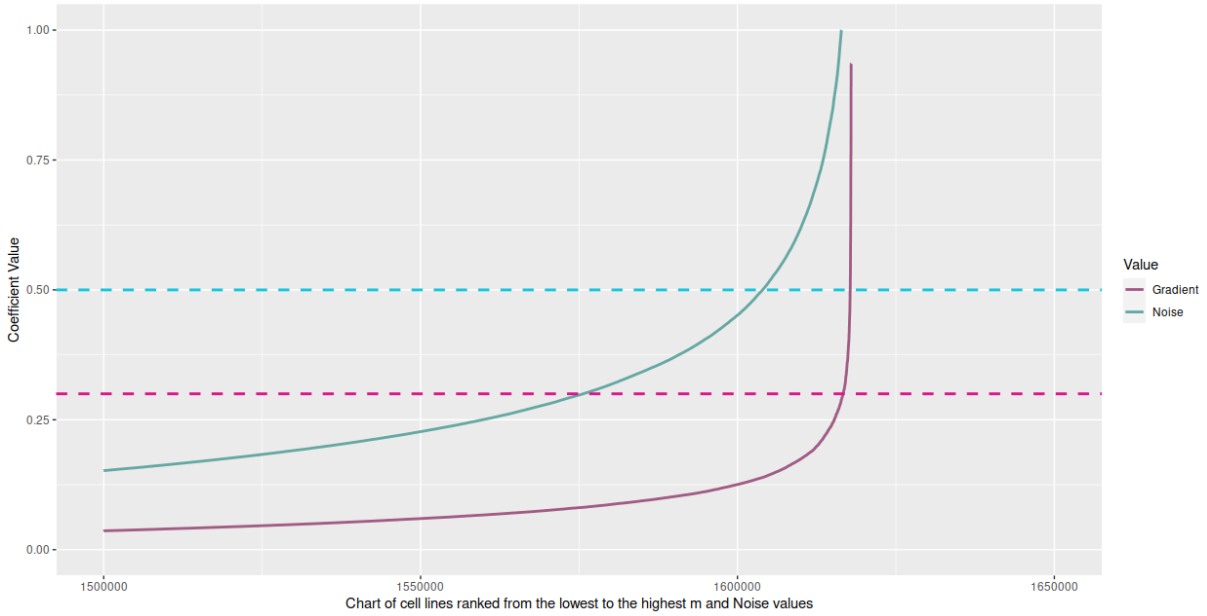


Рисунок 6. Спільний графік розподілу значень шуму даних та градієнту концентрації по всьому масиву даних життєздатності клітин.

На рис. 6, фіолетовий графік - графік розподілу коефіцієнтів градієнту концентрації m ; синій графік - значення коефіцієнту шуму η. Рожева пунктирна лінія - нижня межа придатності значень коефіцієнтів градієнту концентрації m, синя пунктирна лінія - верхня межа значень коефіцієнту шуму даних η. Значення, що знаходяться між цими двома пунктирними лініями, і є даними клітин, що збільшують життєздатність при додаванні хімічних сполук.

Було проведено фільтрація даних масиву за граничними значеннями η та m. Усі рядки із значеннями $m < 0.3$ та $\eta > 0.5$ були видалені з масиву даних, залишаючи близько 15% від початкового числа рядків, залишаючи в масиві даних лише комбінації з нестандартною відповіддю на хіміотерапію.

3.1.5 Молекулярна характеристика ракових клітинних ліній

Пан-ракова (PANCAN, від *англ*., pan-cancer — “про всі види раку”) матриця бінарних подій (BEM, від *англ*., binary event matrix) — це текстова база даних консорціуму *GDSC*, що містить інформацію про 961 CCL та більше 300 ракових генів, що в них знаходяться. Кожен ген в цій мариці позначений або як змінений (1), або як ген дикого типу (0). Зміни в BEM можуть бути наслідками виникнення соматичних ракових мутацій, злиття генів або зміни кількості копій генів. PANCAN BEM є результатом грунтівного визначення та молекулярної характеристики генетичних особливостей CCL та пасажирських мутацій.

В даній роботі, інформацію з такої матриці останньої версії було додано до основного масиву даних, у відповідності до значень штрих-коду (**BARCODE**), номеру CCL (**COSMIC_ID**) та номеру х/с (**DRUG_ID**). Злиття даних допомогло в останньому етапі даної роботи, що полягав у побудові масиву даних про хімічні сполуки та відповідні їм CCL, що найбільш специфічно та активно збільшують життєздатність CCL.

3.1.5 Тест на гіпергеометричний розподіл

Тест на гіпергеометричний розподіл було проведено після додавання до числових даних інформації про молекулярну характеристику та мутації.

Гіпергеометричний тест підходить для всіх експериментів, для яких необхідно розв’язати наступне завдання:

“Існують певні об’єкти, що можуть мати або характеристику ***G,*** або характеристику ***B***. Всього існує ***Q*** таких об’єктів. Випадково вибираємо з цього загального числа об’єктів ***Q*** малу вибірку об’єктів ***q***. Яка вірогідність того, що певна кількість об’єктів в малій вибірці ***q*** матиме характеристику ***G****?”*.

В даній роботі це питання було розширено та сформульовано так:

“Існує певна кількість CCL ***Q***. До них належить також клітинна лінія ***G***. Також існує певна хімічна сполука ***X***, що діє на усі CCL з числа ***Q***, в тому числі і на ***G***.
Тепер ми випадково відбираємо з числа CCL ***Q*** меншу кількість CCL ***q***.
Яка вірогідність того, що в малій вибірці ***q*** наявні лише клітини лінії ***G***?”

Або, іншими словами:
“Наскільки хімічна сполука ***X*** є специфічною до клітинної лінії ***G***?”

Вибірку даних скринінгів, з життєздатністю з градієнтом лінійної регресії $m > 0.3$ (1023 комбінацій CCL та х/с) були піддано гіпергеометричному тесту, для визначення нерівномірності розподілу вибірок ліків, CCL та градієнту їх регресійних прямих.

3.1.6 Алгоритм роботи програми gdscIC50

Для отримання графічної інформації про CV, побудови графіку, розрахунку значень IC_{50} та *AUC*, використовується програмний пакет *gdscIC50*, що розробляється та підтримується безпосередньо консорціумом *GDSC*. Частковий алгоритм роботи даного пакету був отриманий методом зворотної розробки та перевірений на практичних прикладах [14].

Після проведення усіх дослідів, стало очевидно, що *gdscIC50* обраховує значення IC_{50} для значень, що збільшують свою життєздатність, неправильно. Для аналізу вихідного коду, *gdscIC50* була завантажена з репозиторію *GDSC* на *GitHub* за допомогою інструментарію *devtools* для мови програмування *R* та середовища розробки *RStudio*.

Програма *gdscIC50* представляє собою три інструкції мовою програмування R, які послідовно обробляють вихідні дані у вигляді масиву. Дані передаються від одного скрипту до іншого з використанням зовнішніх функцій пакету *nlme* для розрахунку регресії, там де це потрібно. Кожна інструкція містить функції, що мають бути викликані безпосередньо користувачем програми для отримання результатів, отже за ступенем автоматизації дана програма є *напівавтоматичною*.

Детальний опис цих інструкцій та найважливіших частин вихідного коду, може бути використаний для пошуку та виправлення помилок та пропозиції нових функцій:

1. nlme_fit_prep виконується першою та відповідає за перетворення даних у належний вигляд для розрахування кривої регресії. Ця підпрограма видаляє рядки, де відсутні дані про ліки, про відповідь на лікування або визначальні теги.

Далі виконується нормалізація вихідних даних шляхом пошуку співпадінь і закономірностей у вмісті комірок із застосуванням пошуку за регулярними виразами. Функція *CondenseScreenData* відповідає за

вилучення побічних даних з масиву та переведення його в формат, доступний для роботи *nlme_fit.* Після цього вона групує дані за DRUGSET_ID та SCAN_ID [14, 15].

```
condenseScreenData <- function(screen_data, neg_control, pos_control){
 average_controls <- averageControlData(screen_data,
                                        neg_control = neg_control,
                                        pos_control = pos_control)
 drugset_layouts <- screen_data %>%
  select_(~DRUGSET_ID, ~POSITION, ~TAG, ~DRUG_ID, ~CONC) %>%
  distinct()
 drugset_layouts <- drugset_layouts %>%
  group_by_(~DRUGSET_ID) %>%
  do(condensed_layouts = condenseDruggedLayout(.)) %>%
  tidyr::unnest(condensed_layouts)
 screen_data <- screen_data %>% select_(~RESEARCH_PROJECT,
                  ~BARCODE,
                  ~SCAN_ID,
                  ~DATE_CREATED,
                  ~SCAN_DATE,
                  ~CELL_ID,
                  ~COSMIC_ID,
                  ~MASTER_CELL_ID,
                  ~CELL_LINE_NAME,
                  ~SEEDING_DENSITY,
                  ~DRUGSET_ID,
                  ~ASSAY,
                  ~DURATION,
                  ~POSITION,
                  ~INTENSITY) %>%
  distinct()
 condensed_screen_data <-
  inner_join(screen_data, drugset_layouts,
          by = c("DRUGSET_ID", "POSITION")) %>%
  inner_join(average_controls, by = c("SCAN_ID"))
 return(condensed_screen_data)
}
```

2. nlme_fit розраховує модель концентрації і відповіді на дані GDSC, будує залежність між дозою препарату і відповіддю на нього, розраховує криву відповіді, розраховує значення RMSE, мінімізує помилки невідповідності даних до кривої, тим самим стабілізуючи її. Даний скрипт

розрізняє велику вибірку даних і малу, для останньої він застосовує інші налаштування. По закінченню роботи цей скрипт передає готові дані на наступний скрипт *nlme_fit_stats*.

Найважливішою частиною даного скрипту є, безумовно, *fitModelNlmeData* — велика за кількістю рядків функція, що включає в себе функції *logist3*, *logist4* та функції обчислення інтегралу. Сама функція *fitModelNlmeData*, здебільшого, просто робить виклик інших під-функцій *logist*, тому записується у простий вигляд:

```
     fitModelNlmeData <- function
(nlme_data, isLargeData = TRUE) {
       gDat <-
groupNlmeData(nlme_data)
       nlme_model <- fitModel(gDat,
bLargeScale = isLargeData)
       return(nlme_model)
     }
```

Проте найважливіша частина даної функції - *logist3* є двопараметричною функцією, що визначає відповідь клітинної лінії на препарат і має наступний вигляд:

```
logist3 <- stats::selfStart( ~ 1/(1 + exp((xmid - x)/scal)),
                             initial = function(mCall, LHS,
data){
xy <- stats::sortedXyData(mCall[["x"]], LHS, data)

if(nrow(xy) < 3) {

 stop("Too few distinct input values to fit a logistic")

 }

xmid <- stats::NLSstClosestX(xy, 0.5 )
scal <- stats::NLSstClosestX(xy, 0.75 ) - xmid
value <- c(xmid, scal)
names(value) <- mCall[c("xmid", "scal")]
value

},

parameters = c("xmid", "scal"))
```

Далі ця функція переходить у *logist4* та малу функцію обчислення інтегралу. Фактично, один виклик даної функції являє собою всю програму nlme_fit, що завершується в один цикл і передає дані далі, до nlme_stats.

3. nlme_fit_stats отримує готові дані від nlme_fit, які переводяться в візуальний та придатний для аналітики формат. Скрипт розраховує коефіцієнти моделі nlme_fit, значення IC_{50}, *AUC*, *AUC*Trap та доступну статистику залежності клітинної відповіді від концентрацій препаратів. Після обрахування статистики, запускається програма *ggplot2*, яка будує графіки [16].

Графік клітинної відповіді є основним результатом роботи програми, що також включає значення IC_{50}, *AUC* та інші. Для його побудови запускається функція *plotResponse*, що “стягує” обраховані значення з масиву даних та додає стилі *ggplot2* (пробіли залишені для простоти інтерпретації даних).

Далі ініціюються обрахування *AUC* та RMSE, із додатковими перевірками на правильність вхідних значень:

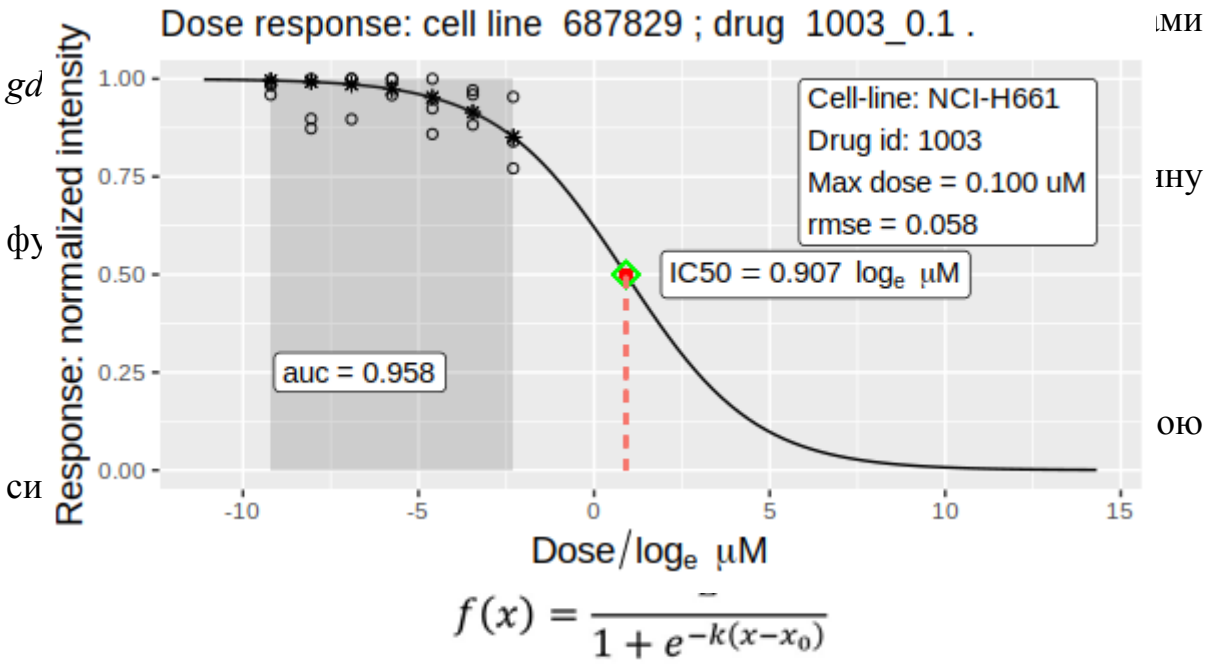
***AUC* <- unique(plot_data$*AUC*)**

stopifnot(length(*AUC*) == 1)

rmse <- unique(plot_data$RMSE)

stopifnot(length(rmse) == 1)

Результатом роботи цієї функції є графік наступного вигляду (рис. 4):



Формула 6: Логістична функція

, де x_0 – значення на осі абсцис всередині графіку, L – максимальне значення (асимптота) функції, k – швидкість росту.

Типовий графік, що генерує дана функція, представлений на Рис. 5:

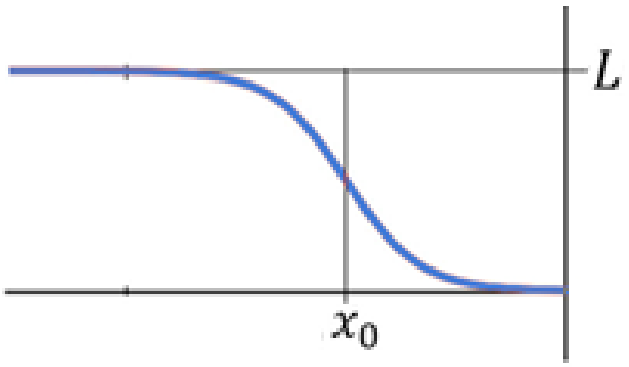


Рисунок 5. Типовий графік логістичної функції, що генерує програма *gdscIC50*

Формула логістичної кривої, яку застосовує для побудови графіків програма *gdscIC50*, є досить жорсткою, та поводить себе некоректно при вхідних значеннях виміряної люмінесценції, що збільшуються (що вказує на збільшення CV при збільшенні концентрації лікарського засобу). В контексті коду *gdscIC50* такі дані інтерпретуються неправильно та призводять до хибно-позитивних результатів — коли програма має “передбачити” падіння виживаності клітин, екстраполюючи існуючі дані, незважаючи на фактичну відсутність таких даних [17]. За попередніми припущеннями, це може бути пов’язано із жорстким дотриманням логістичної кривої програмою. Компонент скрипту *nlme_fit logist3* може бути відповідальним за подібну поведінку програми, адже саме він містить формулу логістичної кривої. Даний аспект роботи програми має бути досліджений у майбутньому - за необхідності, програма може бути дороблена.

Пропозиції по зміні вихідного коду програми

Обчислення значень логістичної кривої не має сенсу для тих CCL, що збільшують свою життєздатність. Методика обрахування значення IC_{50} передбачає, що значення CV в певний момент часу сягне нуля, чого не можна зрозуміти з наявних даних, особливо тоді, коли ці дані є монотонно-зростаючими.

Дана крива є стійкою до шуму, в контексті точності вимірювальних даних, через “жорсткість” описання її формули у вихідному коді програми — крива може змінювати лише значення куту нахилу (*scal*), та положення точки перегину (*xmid*), проте не загальну формулу даної кривої, та не значення її асимптоти (Рис. 777).

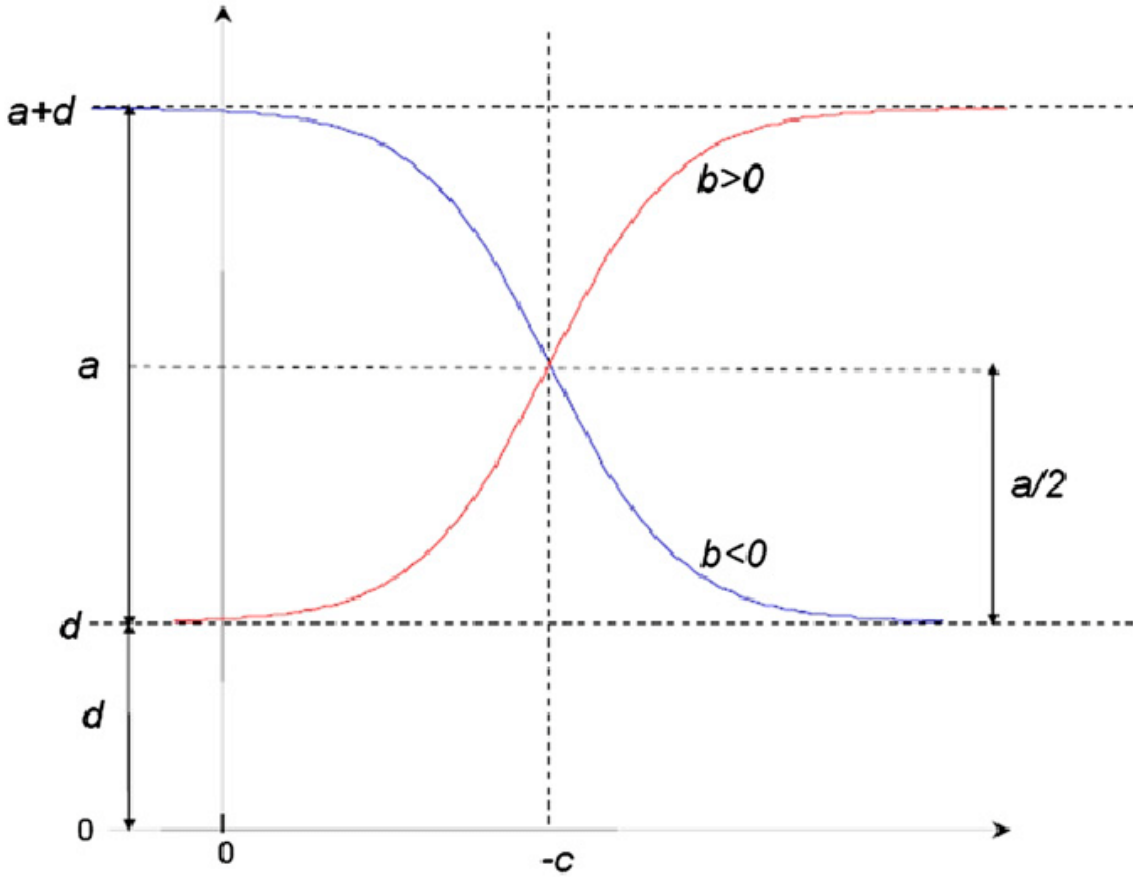


Рис. 777. Проста логістична крива та її загальні змінювані параметри.

$$f(x) = \frac{a}{1 + e^{-b(x+c)}} + d$$

Крива, подібна до тієї, що зображена на малюнку, може змінювати наступні параметри:

- ***a*** — асимптота логістичної кривої, або крива до якої вона нескінченно наближається;
- ***c***— точка перегину функції, після проходження якої вона сповільнює прискорення свого росту і починає загортатися до значення асимптоти;
- ***scal*** — кутовий коефіцієнт дотичної, або значення "крутизни" схилу функції. Чим меншим є коефіцієнт scal, тим більшим є значення 1/scal, та тим більш компактною буде функція в своєму основному об'єму. Scal не показаний на даному графіку.

В пакеті *gdscIC50* рівняння логістичної кривої записується як:

Що є відображенням формули 5, записаної у простий лінійний вигляд мовою програмування R. Запишемо цю формулу ще раз, позначаючи в ній ті коефіцієнти, які відсутні в формулі 5:

(~ 1/(1 + exp(-(x - xmid)/scal)),

В даній функції довільне значення ***a*** було замінено на **1**, а показник ступеня експоненти є від'ємним за замовчуванням. Такі зміни до оригінальної формули роблять функцію більш жорсткою, подібною до Рис. 777.2

На нашу думку, програма gdscIC50 має бути доопрацьована у наступних аспектах:

1. В коді програми має бути реалізований обов’язковий попередній розрахунок коефіцієнтів лінійної регресії із перевіркою їх значень.
2. При виявленні коефіцієнтів регресії, що вказують на зростаючі значення CV, програма має залишити помітку про дану клітинну лінію та хімічну сполуку, аби потім перерахувати для них певні коефіцієнти ефективності терапії, концентрації тощо.
3. Після обчислення показників IC50 для традиційних реакцій комбінацій, необхідно окремо розрахувати подібні коефіцієнти для значень, що збільшуються. Таким коефіцієнтом може бути градієнт лінійної регресії, за умови його корекції відповідно до нахилу прямої та інших показників, або інша значуща метрика.

Отже, очевидно, що програма gdscIC50 допускає помилки при обчисленні окремих значень IC50 для CV, що зростають. Консорціум GDSC має бути сповіщений про небезпеку застосування старої версії програми без значних змін вихідного коду, адже за помилковими значеннями IC50 проводять дослідження, тестування ліків та експериментальне лікування онкозахворювань, багато неприбуткових та комерційних лабораторій та клінік світу.

Разом з цим, очевидно, що увага, яка буде приділена виправлення даної помилки та повного тестування роботи програми *gdscIC50* не буде достатньою, враховуючи її великий вплив на правильність отримуваних клінічно-діагностичних даних по всьому світу. Певною мірою, ця проблема є характерною для всього наукового безкоштовного програмного забезпечення, адже подібні проекти програють у якості та оперативності підтримки комерційному програмному забезпеченню. Разом з тим, важливо забезпечити рівний доступ до такого ПЗ усім, хто його потребує,

адже інакше продукт просто не набере необхідної популярності та впливу на дослідження по всьому світу.

В наступному розділі було зроблено спробу знайти шляхи вирішення обидвох цих проблем, зробити аналіз потенційних споживачів, очікуваного прибутку та шляхів залучення інвестицій для розробки комерційного наукового ПЗ.

3 СТАРТАП ПРОЕКТ

3.1 Резюме: конкретизація бізнес-ідеї, мети стартапу, об'єкту дослідження, місця розробки у інноваційному ланцюжку цінностей

Загальна характеристика розробки:

Тема: Особливості алгоритму розрахунку кривих для опису та передбачення чутливості культур CCL до хіміотерапії;

Мета проекту: Розробити новітнє програмне забезпечення для аналізу чутливості CCL до терапії;

Суб'єкт замовлення: Інститут Комп'ютерної Біології, Мюнхенського Центру Спілки ім. Гельмгольца;

Об'єкт дослідження: дані масштабного скринінгу чутливості CCL до терапії від консорціуму GDSC;

Місце розробки у інноваційному ланцюжку цінності: ідея знаходиться на етапі розробки, проводяться додаткові перевірки нового алгоритму та оцінка складності його практичної реалізації.

Місце товару у міжнародній класифікації товарів: Клас 42: Наукові або технологічні послуги та дослідження і розробки, що їх стосуються;

Цінність: Автоматичного підбір протиракової терапії та визначення біомаркерів;

Гранична корисність товару: здешевлення виготовлення сировини для очищення стічних вод та збільшення її ефективності.

Таблиця 3.1 – Резюме стартап-проекту

Показник	Характеристика
1. Сутність ідеї	Розробка новітнього програмного алгоритму для аналізу геномних маркерів ракових пухлин та автоматичного підбору протиракової терапії, а також пакету програмного забезпечення для нього.
2. Наявність аналогів або прототипів ідеї	Повних аналогів на ринку збуту не існує, часткові некомерційні аналоги і робочі

	прототипи є в наявності
3. Основна потреба, яку задовольнить реалізований стартап	Потреба в аналізі гетерогенних впливів CCL на хіміотерапію
4. Ступінь розробленості технології реалізації	Початкова, не реалізована у вигляді кінцевого продукту
5. Класифікація продукту стартапу за міжнародною класифікацією послуг	МКТП 11-2020 (NCL11-2020): Клас 42: Наукові або технологічні послуги та дослідження і розробки, що їх стосуються; базовий номер послуги: 420257 (дослідження в медицині, medical research)
6. КВЕД, до якого може належати дане виробництво	КВЕД-2010, Клас 72.11
7. Очікувана потужність стартапу	мале підприємство
8. За масштабом виробництва	масове (електронна дистрибуція)
9. За рівнем спеціалізації	вузькопрофільне
10. За ресурсами, що споживатимуться	Інформаційномістке, працемістке, капіталомістке
11. За чисельністю персоналу	середнє
12. Органи управління при реалізації стартапу	Багаторівневі (національні, міжнародні та транснаціональні)
13. Бажане географічне розташування ● потужностей стартапу; ● офісу стартапу; ● збутової мережі; ● постачальників комплектуючих.	● Потужності стартапу: серверне обладнання знаходиться у Німеччині, послуги хостингу надає філіал компанії Amazon — AWS. ● Офіс стартапу: м. Київ, офісна будівля з офісом типу “open space” на 50 чол. ● Постачальники комплектуючих - сервери закупляються напряму в CloudFlare та їх технологічних партнерів; програмне забезпечення та електронні матеріали завантажуються через мережу інтернет та/або напряму з інтернет-сайтів продавців спеціалізованого ПЗ
14. Місце ідеї у ланцюжку цінностей інноваційного	Дослідження та розробка (Research and Development, R&D)

процесу	
15. Гранична корисність ідеї стартапу	Підвищення точності філогенетичної характеристики ракових захворювань
16. Бізнес-модель стартапу	B2B
17. Конкуренти вітчизняні (ціна, на якому етапі реалізації знаходяться, основні конкурентні переваги, фактори успіху)	Відсутні
18. Конкуренти іноземні (ціна, на якому етапі реалізації знаходяться, основні конкурентні переваги, фактори успіху)	Повних аналогів проекту не існує, часткові аналоги розробляються двома науково-дослідними університетами США: Університет БеССLі (США, Каліфорнія), Стенфордський Університет (США, Каліфорнія); дані продукти є безкоштовними; факторам успіху є заявлена кількість функцій та точність роботи ПЗ
19. Ключові фактори успіху стартапу	Стабільність у потребах ринку досліджень раку та/або пошуку протиракових терапій
20. Споживачі (основні на етапі впровадження, групи, орієнтовна чисельність)	- Дослідники ракових захворювань в некомерційних дослідницьких установах США та Західної Європи; - Працівники R&D відділень біофармацевтичних компаній; - Працівники комерційних лабораторій з дослідження онкозахворювань.
21. Планова кількість продукту розробки для першого етапу реалізації	140-150 ліцензій на підтримку програмного забезпечення на рік; верхня межа кількості одночасно активних ліцензій на підтримку ПЗ — 350 ліцензій до кінця першого року активності стартапу; верхня межа кількості одночасно активних копій ПЗ — необмежено.
22. Мінімальна кількість виробництва за методом точки беззбитковості	125 ліцензій/рік (перший рік), 200 ліцензій/рік (усі наступні роки).
23. Споживачі на етапі розвитку	Національні дослідницькі медичні установи; комерційні лабораторії; поодинокі дослідники.
24. Споживачі на етапі зрілості	Національні дослідницькі медичні установи; комерційні лабораторії; поодинокі дослідники.
25. Конкурентна ціна на	2000-2500 дол. США/рік за ліцензію на

продукт стартапу	підтримку ПЗ, в залежності від версії програми
26. Плановий рівень рентабельності при реалізації продукту	110-130 копій/рік
27. Капіталовкладення в проект	21000-23000 дол. США/рік (25000 долл. США/ перший рік)
28. Період повернення капіталовкладень у проект	менше 1 року
29. Джерела фінансування	внутрішні, зовнішні, іноземні

Продовження табл. 1

Показник	**Характеристика**
30. Основні компоненти продукції стартапу (їх доля у готовому товарі, ступінь готовності компонентів у наявному виробництві)	ПЗ “agwo5” та “agwo5 pro” (40%) мережа технічної підтримки клієнтів (60%)
31. Потенційні постачальники складових компонентів розробки (виділити вітчизняних і закордонних, плановий обсяг замовлень, наявна потужність постачальника)	Постачальники технічних засобів (сервери, маршрутизатори, пакети програмного забезпечення тощо) відсутні; усі технічні питання забезпечення запису, обробки і зберігання значущої інформації вирішуються напряму корпорацією Amazon, як власником сервісу Amazon Web Services.
32. Планове місце реалізації результату розробки (місце, планова доля реалізації продукту через це місце)	Прямі інтернет продажі
33. Наявність посередників при реалізації (так, ні, орієнтовні посередники, форми оплати їх діяльності)	Посередникі не потребуються та відсутні
34. Методи просування результатів розробки на ринок	контекстна реклама в соціальних мережах для дослідників; реферальна мережа (знижки та/або прямі виплати для клієнтів, що

	приведуть за собою ще одного клієнта)

Терміни:

1.Продукт: програмне забезпечення.

2.Технологія: Впровадження новітнього алгоритму розрахунку регресії даних для створення швидкої, ефективної та надійної роботи програми для аналізу життєздатності CCL;

3.Кваліфікація персоналу: Випускник технічних спеціальностей або науровий співробітник; повна вища освіта; володіння спеціалізованими знаннями з функціонального програмування; зі спеціалізованими знаннями з молекулярної біології та генетики.

4.Споживач: Науково-дослідні лабораторії та приватні лабораторії з розробки протиракової терапії.

5.Ринок збуту: Західна Європа, США.

6.Конкурентні переваги: швидкість, надійність роботи, порівняно висока функціональність, наявність цілодобової професійної технічної підтримки.

3.2. Аналіз зовнішнього та внутрішнього середовища стартапу

Таблиця 3.2 – Аналіз загроз і можливостей зовнішнього середовища

Фактори	**Загрози**	**Можливості**
Економіка		
Реформа податкового законодавства	Можливі несприятливі умови для ведення бізнесу на/з території України, необхідність перереєстрації бізнесу в іншій країні для збереження темпів росту та/або величини чистого прибутку	Потенційне спрощення та роз’яснення податкового права дозволить оптимізувати податкові витрати

Висока волатильність ринку біоінформатичного програмного забезпечення	Можливість швидко “спіймати хвилю” та збільшити прибуток стартапу в кілька разів за короткий час	Нестабільність фінансових умов, можливе банкруцтво стартапу через нестачу амортизаційних фондів
Політика		
Державне регулювання/вплив на стартап	1. Підтримка наукової ініціативи на законодавчому рівні	Посилення державного втручання в економіку, обмеження на отримання прибутку тощо
Додатковий контроль за дотриманням стартапом усіх державних норм і правил	Спрощення взаємовідносин з державою, зменшення часу на реєстрацію, заключення договорів тощо	Законодавче обмеження самостійності управління стартапом
Програми доступних інвестицій і кредитів для технологічних компаній	Спрощення умов для ведення бізнесу	Поява додаткових конкурентів на ринку збуту
Науково-технічний прогрес		
Виникнення і розвиток науки і технології	Збагачення продукту стартапу новими функціями та збільшення його загальної ефективності, збільшення швидкості роботи програмного забезпечення, посилення позицій стартапу на	Посилення конкуренції на ринку збуту через адаптування тих самих технологій конкурентами; “відтік ідей” в інші країни, стартапи, ініціативи; недоброчесна конкуренція
Обмін ідеями і досвідом із закордонними колегами та співробітниками		

Швидка модернізація програми за рахунок більш простого доступу до актуальної науково-технічної інформації	ринку	

Таблиця 3.3 – Аналіз факторів зовнішнього оперативного середовища

Фактор	**Переваги**	**Недоліки**
Конкуренти	Безкоштовна версія програми	Відсутність певного функціоналу Повільна розробка та впровадження нових функцій
Постачальники	Майже не потребуються, за виключенням початку роботи стартапу Великий вибір на ринку послуг постачальників	Висока ціна на послуги Залежність від сторонньої інфраструктури
Посередники	Відсутні в даному стартапі (перевага)	
Споживачі	Потенційна наявність великого попиту Готовність сплачувати майже будь-які суми за якісну розробку	Вузька категорія та спеціалізація споживачів Відсутність ефективних способів реклами товару

За результатами аналізу факторів зовнішнього і зовнішнього оперативного середовищ формуємо перелік зацікавлених сторін (табл. 3.4) для визначення потенційних загроз у балах у процесі впровадження.

Таблиця 3.4 – Аналіз зацікавлених сторін

Зацікавлена сторона	Вплив її на реалізацію проекту	Що саме цікавить в проекті?	Загальний коефіцієнт впливу на проект
Суб'єкти зовнішнього оперативного середовища			
Розробник:	Забезпечує впровадження результатів стартап проекту	Зацікавлений у впровадженні змін у стартап проекті	70
Постачальник:	Забезпечує необхідним серверним обладнанням та постачає необхідний сервіс	Зацікавлений в пришвидшені розвитку стартапу та більшій кількості проданих копій програмного забезпечення	15
Споживачі:	Забезпечують використання придбаної продукції	Зацікавлені в кращій якості та функціоналу продукту	15
Посередники: *Не залучаються тому не впливають на розвиток стартап проекту*			
Політичні структури:	Беруть участь у формуванні бюджету НД інститутів	Зацікавлені у розвитку науки в країні в цілому, готові виділяти під неї гранти	10
Суб'єкти економічного середовища	Банки надають кредитні кошти Інвестори та акціонери розпоряджаються власністю	Зацікавлені у збільшенні кількості та якості досліджень	3
Власник географічних об'єктів	Не впливають на проект		
Суб'єкти демографії	Технічно, можуть мати вплив на споживачів, проте	Зацікавлені у високій якості продукції, можуть впливати на вид	3

	цей вплив навряд чи значний на прикладі даного продукту	продукції, що обирається покупцями	
Суб'єкти культурного середовища	Мають вплив на ставлення споживачів до продукції, проте вплив не є значним в даному прикладі	Зацікавлені у високій якості продукції	3
Суб'єкти НТП	Забезпечують розвиток НДР	Зацікавлені у впровадженні нових технологій НДР, сприяють цьому	10

Переваги та недоліки внутрішнього середовища наведено в табл. 3.5.

Таблиця 3.5 – Переваги і недоліки внутрішнього середовища

Фактор	**Переваги**	**Недоліки**
Відсутність власної матеріальної бази	+ незалежність від розташування + делегування відповідальності за матеріальну базу	- залежність від постачальника матеріальної бази - порівняно висока вартість оренди матеріальної бази
Персонал	+ Відносно невелика кількість персоналу	- Високі вимоги до кваліфікації співробітників
Висока складність кінцевого продукту з точки зору його розробки	+ Складність копіювання продукту конкурентами	- Велика вірогідність виникнення проблем чи помилок

3.3 Визначення ключових факторів успіху проекту

На даний момент в світі не існує прикладів програмного забезпечення, що здатно автоматично, швидко та якісно проводити аналіз змін CV при дії на них хімічних сполук. Дослідження в даному напрямку не є частими, а через те, що такі проекти завичай ведуться лише кількома дослідниками, готові продукти не мають зрозумілого та простого інтерфейсу, який допоміг

би збільшити коло користувачів та позитивно вплинути на результати їх досліджень. При цьому, опитування показали, що більшість дослідників готові платити гроші заради економії часу на написання власного спеціалізованого ПЗ.

До сьогоднішнього дня, в світі не існувало повністю автоматизованого, простого, точного, приємного у користуванні та, одночасно з цим, безкоштовного програмного забезпечення, що могло б автоматизувати процес отримання значущої інформації про особливості мутаційного профілю, можливі побічні реакції та біомаркери CCL.

Запропонований нами продукт, за умови його реалізації, дозволить повернути всі втрачені кошти менше ніж за рік, та, при цьому, не бути прив'язаним до певної географічної локації, а керувати написанням, розповсюдженням та підтримкою програмного забезпечення виключно через інтернет.

Тому можна вважати, що при появі даного продукту на науковому ринку, він займає лідируючу позицію за попитом.

Оцінка конкурентоспроможності продукції проводиться методом Шонфільда. Оцінка показника якості продукції відбувається за 5-ти бальною шкалою. Коефіцієнт значущості показника для замовника лежав у межах 0…1 (табл 3.6 рис 3).

Таблиця 3.6 – Оцінка характеристик продукції

Характеристика ключових факторів	**Коефіцієнт вагомості характеристики**	**Оц інка характеристик**		
		Наша продукції	**Конкурент А**	**Конкурент Б**
Функціонал продукту	***0,2***	***10***	***4***	***6***
Ціна	***0,1***	***2***	***5***	***4***

Простота експлуатації	***0,1***	***9***	***7***	***1***
Ефективність роботи продукту	***0.6***	***10***	***8***	***8***
Загальний незважений бал:	***1***	***31/40***	***24/40***	***19/40***

З урахуванням коефіцієнту вагомості характеристики визначається бальна оцінка кожної характеристики для нашої продукції і для конкурентів:

Характеристика	**Бальна оцінка характеристик**		
	Наша продукції	**Конкурент А**	**Конкурент Б**
Функціонал продукту	*10 * 0.2=**2***	*4 * 0.2=**0.8***	*6 * 0.2=**1.2***
Ціна	*2 * 0.1=**0.2***	*5 * 0.1=**0.5***	*4 * 0.1=**0.4***
Простота експлуатації	*9 * 0.1=**0.9***	*7 * 0.1=**0.7***	*1 * 0.1=**0.1***
Ефективність роботи продукту	*10 * 0.6=**6***	*8 * 0.6=**4.8***	*8 * 0.6=**4.8***

На підставі отриманих бальних оцінок будується графік порівняння конкурентних переваг нашого підприємства з конкурентами.

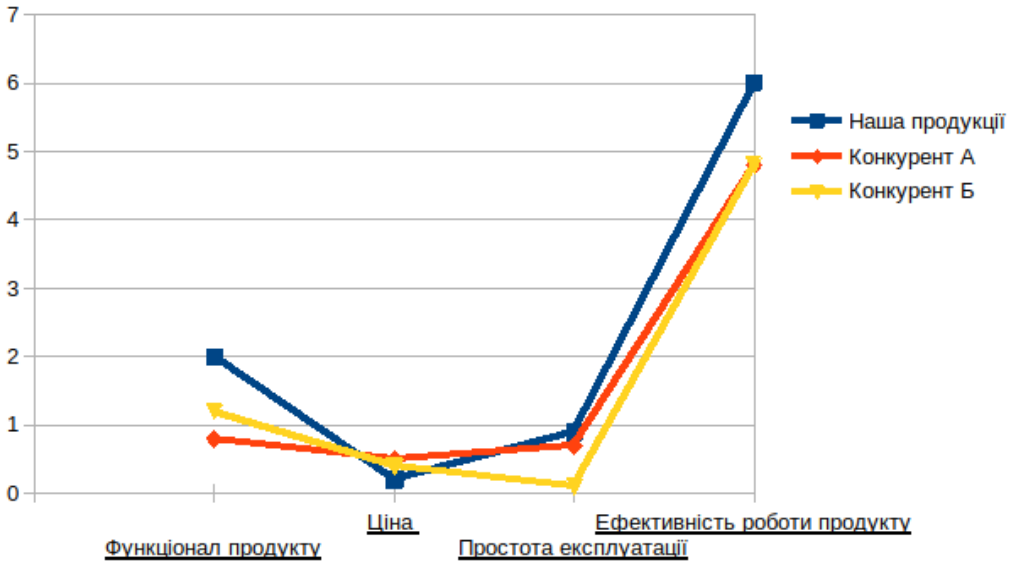


Рисунок 3 – Графічне порівняння конкурентних переваг підприємства з конкурентами

Отже, розроблений продукт є конкурентоспроможним. Ключовим фактором проекту є ефективність технології та можливість експортувати дослідження за кордони України.

На основі аналізу ключових факторів успіху стартап-проекту формують можливі варіанти розвитку інноваційної ідеї та визначають перспективний напрям її розвитку (табл. 7).

Таблиця 3.7 – Варіанти розвитку ідеї стартапу

Варіант	**Стислий опис можливого розвитку**
1. Випуск спеціальних версій ПЗ	Версія ПЗ загального призначення надається безкоштовно та без підписки; при купівлі підписки надається персональна виділена лінія технічної підтримки, що поступово дописує необхідні саме цій компанії функції ПЗ
2. Уніфікація ПЗ	Одна універсальна версія, доступна за підпискою; технічна підтримка надається безкоштовно

3. Зміна типу оплати за продукт	ПЗ доступне без підписки і оплати, технічна підтримка надається безкоштовно усім користувачам. Якщо продукт використовується в комерційних цілях (приватні лабораторії, лікарні, госпіталі тощо), за користування продуктом стягується певний відсоток від загального прибутку установи-користувача

3.4 Визначення потенційних споживачів

Метою оцінки потенційних споживачів є визначення перших клієнтів, які придбають дану стартап-розробку. В табл. 3.8 наведено основні критерії для вибору потенційних споживачів.

Таблиця 3.8 – Класифікація потенційних споживачів

Критерій	**Значення**
1. Юридична особа	
1. Форма власності	Державне, приватне
2. КВЕД	Розділ 72 Наукові дослідження та розробки; Група 72.1 Дослідження й експериментальні розробки у сфері природничих і технічних наук
3. За потужністю (малі, середні, великі)	Середні, великі, малі
4. За масштабом виробництва (одиничні, серійні, масові)	Одиничні, масові
5. За рівнем спеціалізації (вузькопрофільні, багатопрофільні, комбіновані)	Вузькопрофільні, комбіновані
6. За ресурсами, що споживаються (працемісткі, матеріаломісткі, капіталомісткі, інформація)	Залежні від інформації, капіталомісткі
7. За чисельністю персоналу (малі, середні, великі)	Малі, середні
8. За сферою діяльності (виробничі, комерційні, фінансові, посередницькі, страхові…)	Дослідні, посередницькі, виробничі, комерційні
9. За приналежністю капіталу і контролю (національні, іноземні,	Іноземні, спільні багатонаціональні

спільні багатонаціональні,…)	
10. За географічним розташуванням	Закордонні
11. За віддаленістю органів управління (національні, міжнародні, офшорні, транснаціональні,…)	Міжнародні, транснаціональні
12. За характером господарської діяльності (промислові, сільськогосподарські, транспортні, будівельні, фінансово-кредитні, страхові, туристичні, консалтингові,…),	Науково-дослідницькі, інженерні
13. За рівнем технологічної цілісності (провідні, дочірні, філії,…)	Провідні
14. За долею іноземного капіталу (з іноземними інвестиціями (більше 10%), іноземне підприємство (100%)).	З іноземними інвестиціями (більше 90%)
15. За формуванням статутного капіталу (унітарні, корпоративні)	Унітарні, корпоративні
16. За організацією виробничих процесів (періодичні, безперервні)	Періодичні (дослідження і розробка), безперервне (виробництво готової біотехнологічної продукції)
17. За роботою протягом року (сезонні, позасезонні)	Позасезонні
18. За географічним розташуванням на території України	По всій території України (проте, здебільшого будуть розташовуватися за межами України)
19. За наявністю вільних ОбЗ (коштів)	З наявними ОбЗ
20. За динамікою розвитку регіону розташування юридичної особи: − Регіон − Чисельність населення − Динаміка росту регіону − Структура регіону − Правові обмеження торгівлі	Регіон: Київ; Чисельність населення: 2967000 чол.; Динаміка росту регіону:

Оскільки в нас модель продажів B2B, ми працюємо виключно з юридичними особами. Тому при дослідженні потреб наших споживачів ми застосовуємо метод спостереження.

Таблиця 3.9 – Основні групи потенційних споживачів та їх потреби

Категорія (група) клієнтів	**Потреби, які він задовольняє за допомогою Вашого продукту**
1. Приватні дослідницькі лабораторії	Швидке отримання аналізу експериментальних даних і отримання надійних автоматично обрахованих результатів та висновків
2. Державні або приватні некомерційні дослідницькі центри	Швидке отримання аналізу експериментальних даних і отримання надійних автоматично обрахованих результатів та висновків
3. Малі лабораторії з дослідження ракових захворювань	Надійний інструмент для валідації експериментальних даних і отримання точних результатів і висновків
Відкоригована ідея стартап проекту	
Створення безкоштовного програмного забезпечення з відкритим вихідним кодом та введення підписки на технічне обслуговування коду та розв’язання проблем на стороні клієнтів	

Таблиця 3.10 – Паспорт потенційного клієнта

Характеристика	**Значення**
Організаційно-правова форма	Будь яка
Класифікація за потужністю	Середнє, мале
Класифікація за чисельністю персоналу	Мале, середнє
Класифікація за обсягом виробництва	Серійне
Класифікація за сезонністю виробництва	Позасезонне
Розташування	Місто
Вид продукту, який потрібен даному споживачеві	Програмне забезпечення, медичні препарати, медико-діагностичні тести
Призначення придбаної розробки -за призначенням -інше	За призначенням (аналітичні розрахунки)
Кваліфікація персоналу підприємства -робочі	Вчені-дослідники, експерти, лабораторні працівники,

-службовці -керівники	керівники наукових груп, кваліфіковані лаборанти
Потенційний обсяг споживання розробки	Обсяг визначити неможливо, проте за частотою: - користування товаром : 15-20 разів на місяць; - за зверненнями в технічну підтримку: 5-10 разів на місяць.
Хто приймає рішення про придбання розробки (узагальнена характеристика працівника)	1. Директор відділу досліджень та розробки 2. Головний науковий співробітник підприємства

3.5 Ціна інноваційної пропозиції на ринку

Ціноутворення – це процес обґрунтування, затвердження та перегляду цін і тарифів, визначення їх рівня, співвідношення та структури.

Методи ціноутворення, що ґрунтуються на врахуванні витрат називаються витратними. В методі повних витрат, ціна розраховується, виходячи із суми постійних і змінних витрат на одиницю продукції й запланованого прибутку з урахуванням нижнього порогу ціни.

$$\boldsymbol{Ц = С + П,}$$

де ***Ц*** – ціна одиниці товару, грн;

С – собівартість одиниці товару, грн;

П – величина прибутку, яку бажає отримати підприємство від реалізації одиниці товару, грн.

3.5.1 Основні фонди підприємства

Згідно Податкового кодексу термін експлуатації наступних основних фондів та амортизаційні відрахування наведено в таблиці 3.11

Будівлі було надано засновником проекту безкоштовно.

Таблиця 3.11 – Вартість основних фондів

№	Найменування	Кількість, шт.	Вартість, грн (загальна)	Норма амортизації, %	Амортизаційні відрахування, грн
1	Офісна будівля (приміщення, підсобне приміщення)	1	0	5	0
2	Серверне обладнання (купівля в AWS та підписка на обслуговування на 1 рік)	1	25000	2	500
3	Комп'ютерне обладнання для розробників Acer Zett (ноутбук, спеціальна клавіатура, мишка, додаткове програмне забезпечення)	5	62500 (12500*5)	4	5000
4	Набір комп'ютерного обладнання для робітника технічної підтримки Aoyool Voice (ноутбук, спеціальна клавіатура,	10	175000 (17500*10)	4	11000

	мишка, гарнітура, автоперемикач, програмне забезпечення)				
	Загальна вартість		**262500**		**16500**

3.5.2. Оборотні фонди підприємства

В даного підприємства відсутні оборотні фонди, окрім спожитої електроенергії, адже мова йде розробку програмного забезпечення, що не є матеріальним активом (табл 3.12)

Таблиця 3.12 – Кількість необхідних матеріалів для проведення дослідження

Назва реактиву	**Кількість, необхідна для дослідження, л**	**Ціна за 1 л, грн**	**Ціна, витрачена для дослідження, грн**
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-

Витрати на електроенергію та потужність обладнання наведена в таблиці 3.13

Таблиця 3.13 – Спожита електроенергія

Електрообладнання	**Потужність, кВт·год**	**Час використання електрообладнання, год**	**Час використання обладнання за 2 місяці Год**	**Використана потужність, кВт·год**
Веб-сервер AWS	0.850	24	1440	1224
Ноутбуки та периферія	0.15	8	352	52.8
Освітлення	1.4	10	440	616
Додаткові побічні витрати	0.8	5	300	240
Разом	3.2	47	2532	**2132.8**

Тариф на електричну енергію для юридичних осіб становить 2,5385 грн за кВт·год електроенергії.

$$E=2132.8\times2,5385=\mathbf{5414}\text{ грн}$$

5. Витрати на ФОП:

$$ФОП = ЗП + Нарахування.$$

ЗП працівників складає 12000 грн./місяць для розробників програмного забезпечення та 11000 грн./місяць для співробітників технічної підтримки. Період дослідження було 2 місяці, тобто:

$ЗП = (12000\times5+11000\times10)\times2 = (60\ 000 +110\ 000)\times2 = 340\ 000$ грн/2 місяці

$$ФОП = 340\ 000 \cdot 1,22= \mathbf{414800}\text{ грн},$$

де 1,22 – це нарахування на заробітну плату в розмірі **22** %.

Вартість оборотних засобів підприємства наведена в таблиці 3.14

Таблиця 3.14 – Оборотні засоби підприємства

№	**Оборотні засоби**	**Ціна, грн/рік**
1.	Витрати на матеріали	0
2.	Упаковка	0
3.	Витрати на електроенергію	2132.8
4.	Витрати на водопостачання	0
5.	ФОП	414800
Загальна вартість		**416 932.8**

Калькуляція на проведення НДР наведена у таблиці 3.15.

Таблиця 3.15 – Калькуляція на проведення НДР.

№	**Статті калькуляції**	**Сума, грн**
1	Заробітна плата	340 000
2	Нарахування на заробітну плату	74 800
3	Матеріали	0
4	Витрати на електроенергію	2 132.8
5	Витрати на водопостачання	0
6	Амортизація	16 500
Всього		**433 432.8**

3.5.3 Розрахунок собівартості НДР

$$Срік = ОбЗ + А = 416\ 932.8 + 16\ 500 = 433\ 432.8\ грн/рік.$$

Де *С* – собівартість НДР,

ОбЗ – оборотні засоби

А – амортизаційні витрати

Прибуток – це частина виручки від реалізації продукції, яка залишилась на підприємстві після компенсації витрат на виробництво і

реалізацію та інших обов’язкових платежів. Ціна на річну підтримку програмного забезпечення складає 79718.76 грн. (~$3000 USD).

Так як плановий випуск продукції 150 ліцензій на рік тому

Спит = 433 432.8 /150 = 2889.552 грн/шт.

Річний прибуток підприємства:

П= К – С

П = (150·79718.76) – (2889.552·150) = 11 957 814 — 433 432.8 = 11 524 381.2 грн/рік

Очікуваний прибуток з одиниці продукції: 76 829.21 грн за реалізацію 1 копії продукту.

Отже, за витратним методом прогнозована ціна продукту становитиме:

Ц = С + П = 19.26 + 76 829.21 = 76 848.47 грн/копія.

Капіталовкладення за рік:

$$К = ОФ + ОбЗ = 262\ 500 + 416\ 932.8 = 679\ 432.8 \text{ грн.}$$

Рентабельність:

Р= (П/К) × 100;

Р = (11 524 381.2 / 679 432.8) × 100 = ~1696.18 %

Термін повернення капіталовкладень:

$$T_{пов.к.} = \frac{K}{П} = \frac{679432.8}{11524381.2} = 0.06\ року = 21\ день.$$

Фондовіддача виробничих фондів:

ФВ = (Ц×В)/ОФ;

ФВ = (76848.47 × 150) / 416 932.8 = 11527270.5 / 416932.8 = 27.65 грн./грн.

Продуктивність праці:

ПП= В/(Чсп× Т);

ПП= 150/(15 × 8)= 1.25 грн./ос.

Коефіцієнт економічної ефективності:

Е= П/К;

E= 11 524 381.2 / 679 432.8 =~ 16.96

*Слід зазначити, що значення терміну повернення капіталовкладень (0.06 року ~ 21 день) не слід сприймати буквально, через велику ціну одиниці продукції (в даному випадку, річної ліцензії на технічну підтримку, яка сплачується одразу в повному об'ємі). Тому значення "0.06 років" слід сприймати як "капіталовкладення в проект будуть повернуті менше, ніж за рік, за умови продажу запланованої кількості ліцензій на технічну підтримку".

Таблиця 3.16 – Техніко-економічні показники проекту

Показники	**Одиниця виміру**	**Значення**
1. Річний обсяг реалізації ідеї, технології, методики	од.	**150**
2. Середньорічна чисельність персоналу за списком	Осіб	**15**
3. у тому числі - розробників - інженерів технічної підтримки	Осіб	 **5** **10**
4.Середньорічний виробіток робітника	Т/особу	**10**
5. Капіталовкладення у проект: - всього - на одиницю продукції	 Грн. Грн/од.	 **679 432.8** **4 529.6**
6. Повна собівартість - всього - на одиницю продукції	 Грн. Грн/од.	 **433 432.8** **2 889.6**
7. Відносний прибуток		

- всього - на одиницю продукції	Грн. Грн./од.	**11524381.2** **76829.208**
8. Рентабельність	%	**1696.18**
9. Період повернення капіталовкладень	Років	**0.06 року = ~ 21 день.***
10. Фондовіддача виробничих фондів	Грн./грн.	**27.65**
11. Фондоємність	Грн./грн.	**0.036**
12. Продуктивність праці	Грн./особу	**1.25**
13. Коефіцієнт економічної ефективності		**16.96**

3.6 Концепція бізнес-моделі проекту та карта бізнес-процесів реалізації проекту

Таблиця 3.17 –Карта бізнес-процесів виконання стартап-проекту

Стадія реалізації стартап проекту	**Бізнес-процеси**	**Характеристики**		
		Задіяні ресурси	**Орієнтовна тривалість процесу**	**Верхня межа фінансових витрат**
Розробка ідеї стартапу (15 тисяч гривень)	Розробка ідеї; Аналіз можливостей ринку; Формування команди; Перевірка потреб споживача; Розробка схем експерименту	Інформаційні, людські, ЗМІ, пошук в інтернеті та фінансові операції	48 год; 10 год; 8 год; 2 год; 48 год;	2 тис грн; 1 тис грн; 4 тис грн; 1 тис грн; 7 тис грн;
Реалізація ідеї	Оформлення патенту;	Людські, фінансові.	40 год; 24 год;	10 тис грн; 2 тис грн;

(25 тис грн)	Заключення договору про намір з банком; Заключення договору про намір з виробником; Заключення договору про намір з точкою збуту.		36 год; 36 год;	10 тис грн; 3 тис грн;
Впровадження у виробництво (500000 грн.)	Запуск договорів; Виготовлення	Фінансові, людські.	40 год 100 год	- 500000 грн
Масова реалізація		-	-	-
Закриття або продаж проекту (якщо передбачено)		-	-	-

Визначено фактори і елементи бізнес-процесів методом системного аналізу (табл. 3.18).

Таблиця 3.18 – Системний аналіз бізнес-процесів стартапу

Функції	**Елементи**								
	Автор	**Команд**	**Банк**	**Юрист**	**Бухгал**	**Марке**	**Розроб**	**Реклам**	**Спожи**

		а р о з р о б н и кі в			**т е р**	**т о л о г**	**н и к**	**н е а г е н с т в о**	**в а ч**
Розробка ідеї	+								
Аналіз ринку	+					+			
Формування команди	+						+		
Перевірка потреб споживача	+	+				+			
Розробка схеми експерименту	+	+							
Оформлення патенту	+	+							
Заключення договору про намір з банком	+		+	+					
Заключення договору про намір з розробником	+						+		
Заключення договору про намір з точкою реклами	+			+				+	
Запуск договорів	+	+							
Виготовлення		+					+		
Споживче тестування									+

3.7 Ризики стартап – проекту та методи управління ними

У розділі визначені найбільш ймовірні ризики, які можуть виникнути при реалізації даного проекту

Таблиця 3.19 – Ризики інноваційної розробки

Назва процесу стадії реалізації стартап проекту	**Бізнес-процеси**	**Зовнішні ризики**	**Внутрішні ризики**
Розробка загальної ідеї	Розробка ідеї стартапу;	Зміна потреб ринку, втрата	Брак ідей, фінансування,

стартапу та основного продукту стартапу		актуальності стартапу	мотивації на пошук ідеї
	Аналіз ринку та умов для входження на ринок;	Втрата актуальності в цієї області науки, або розробка схожого за ідеєю/функціями продукту конкурентами	Неможливість задоволення потреб потенційних споживачів через брак необхідних ідей/технічного обладнання/експертизи
	Перевірка потреб споживача;	Спрощення вже наявних технологій, через що не має сенсу у розробці даного стартапу	Відсутність компетенції та кваліфікації у сфері, що пріоритетна для споживача
	Набір персоналу, формування команди	Відсутність кваліфікованого персоналу на ринку, перетягування кваліфікованих кадрів конкурентами	Брак знань та вмінь у вже зібраної команди; компетентність менеджерів та/або експертів
	Розробка теорій алгоритмів та побудова тестової демонстраційної моделі	Відсутність інтересу у інвесторів, відсутність інвестицій	Демонстраційна модель не працює або спростовує сформульовано теорію
Реалізація ідеї	Оформлення патенту на науково-технічну розробку	Наявність вже зареєстрованого патенту за такою самою розробкою	Спростування уявлень, що вказані в патенті, тестами та перевірками
	Заключення договору про реалізацію ідеї з промисловістю	Інфляція та банкрутство банку Занепад даної промислової сфери, сектору	Фінансова неспроможність через брак менеджменту

		економіки, або банкруцтво конкретного виробника	
	Заключення договору про намір з банком	Ліквідація банку як фінансової установи, блокування джерела фінансування	Проблеми застосування отриманого фінансування за призначенням: проблеми найму персоналу, закупівлі техніки та залучення експертизи
Активна розробка та підтримка програмного продукту	Впровадження анонсованих та запланованих до реалізації функцій програмного забезпечення	Блокування розповсюдження ПЗ регулюючим органом чи адміністрацією за порушення первых законів	Відставання від запланованого графіку реалізації функцій
	Підтримка клієнтів в режимі 24/7 онлайн	Відсутність попиту на підтримку наживо, падіння популярності ПЗ та стартапу в цілому	Некомпетентність служби технічної підтримки у вирішенні проблем на стороні клієнтів
Масова реалізація	Розповсюдження програмного забезпечення до зацікавлених клієнтів	Локальні проблеми з інтернет зв’язком в регіоні розташування клієнтів – неможливість своєчасного оновлення та	Відсутність суворого контролю якості продукції, що випускається на ринок, відтік клієнтської бази

		підтримки програмного забезпечення з боку компанії	
Закриття або продаж проекту	Реалізація продукту іншому фізичному чи юридичному лицю	Низька оцінка вартості проекту з боку покупця, втрата самостійності при купівлі	Низька оцінка вартості проекту з боку стартапу, що продає, втрата прав на продукт

Ступінь впливу на дохід підприємства та ймовірність настання ризиків наведено в таблиці 3.20.

Таблиця 3.20 – Ризики інноваційної розробки та ймовірність їх настання

Види ризиків	**Назва ризику**	**Ймовірність настання**	**Вплив на очікуваний результат**
Зовнішні ризики			
Демографічний	Міграція кваліфікованого персоналу за кордон	Висока ймовірність	Низький рівень
Науково-технічний	Зміни трендів на ринку біоінформатичного ПЗ	Середня ймовірність	Високий рівень
	Наявність в Україні вже готового патенту	Низька ймовірність	Високий рівень
Ринковий	Втрата інтересу до товару або актуальності його призначення	Висока ймовірність	Високий рівень

	Поява більш дешевих та більш технологічних продуктів	Середня ймовірність	Середній рівень
	Зняття продукту з реалізації чи/та закриття проекту	Середня ймовірність	Високий рівень
	Банкрутство інвестиційних фондів, що підтримуються стартап	Низька ймовірність	Низький рівень
Правовий	Проблеми комерціалізації продукту через обмеження певних ліцензій на ПЗ	Низька ймовірність	Високий рівень
	Проблеми з реєстрацією та/чи захистом авторських прав на розробку	Середня ймовірність	Високий рівень
Внутрішні ризики			
Комерційний	Неможливість використання магнетиту для вирощування грибів	Низька ймовірність	Високий рівень
Фінансовий	Недостатнє фінансування з боку інвестиційних фондів чи нестача запланованих ресурсів	Низька ймовірність	Середній рівень
	Неспроможність оплати по кредитах та	Низька ймовірність	Високий рівень

	заборгованостях		
Організаційний	Слаба злагодженість у роботі відділу розробки та відділу технічного обслуговування	Середня ймовірність	Високий рівень
	Проблеми узгодження швидкості розробки ПЗ та адаптації протоколів для однакового детального рівня його підтримки	Середня ймовірність	Високий рівень
Технічний	Невідповідність реальної реалізованої технології заявленій	Низька ймовірність	Високий рівень
	Необхідність доопрацювання ПЗ, що знаходиться у використанні	Середня ймовірність	Низький рівень
Інформаційний	Слабка реклама та публічне сповіщення про продукт	Середня ймовірність	Середній рівень

Таблиця 3.21 – План заходів з управління ризиками

Назва ризику	**Назва методу управління ризиком**	**Відповідальні виконавці**	**Період виконання / застосування методу**	**Очікувані результати від впроваджен**

				ня методів управління
Організаційний	Ухилення від ризику	Автор	Затримка постачання матеріально-технологічних ресурсів	Знаходження нових постачальників
Технічний	Попередження ризику	Команда розробників	Необхідність доопрацювання технології	Покращення технології
Демографічний	Попередження ризику	Автор	Масовий потік кадрів за кордон	Забезпечення кращих умов праці
Комерційний	Попередження	Комада розробників	Відмова від реалізації продукції	Створення більш вигідних умов для компанії-покупців
Інформаційний	Попередження	Комада розробників	Мала інформації про даний засіб	Розповсюдження засобами ЗМІ інформації

Висновок стартап-проекту:

Представлена в стартап-проекті бізнес-ідея є унікальною за рахунок поєднання безкоштовного основного продукту та якісної експертної підтримки з високим очікуваним попитом, та нестандартним способом монетизації проекту (плата стягується лише за користування підпискою на дистанційну технічну підтримку і обслуговування, при цьому основний продукт доступний без оплати).

За результатом проведеного аналізу визначено, що даний стартап є конкурентоспроможним на світовому ринку біоаналітичного та біоінформатичного програмного забезпечення, та забезпечує ефективне, швидке та просте у використанні отримання значущої інформації про чутливість ракових CCL до препаратів.

Висновки

Протягом дипломної роботи було досліджено будову і особливості збереження інформації в базі даних Genomics of Drug Sensitivity in Cancer, визначено процес роботи програми *gdscIC50* та її вихідний код. В процесі виконання даної роботи було засвоєно основи роботи з мовою програмування R та середовищем розробки RStudio. Було проведено реверс-інжиніринг програми *gdscIC50*, визначено особливості її роботи.

Мова програмування R, якою було написано дану програму, має потужний та виразний синтаксис, що дозволяє створювати компактні та легкі в модифікації програми, такі як *gdscIC50*. Сама програма є такою, що написана з дотриманням правил чіткості і виразності коду, диференційована за призначенням на три окремі скрипти та працює без навантажень на систему.

Разом з цим, очевидно, що певні місця програми потребують переробки, адже застосування даної програми з певними типами даних призводить до помилкових результатів. Спроба знайти код програми, що відповідає за її неправильну роботу із клітинними лініями, що збільшують свою виживаність зі збільшенням концентрації препаратів, та замінити його таким, що працює вірно, має бути виконана далі.

Вихідний код даної програми можна комерціалізувати, як це описано в розділі стартап-проекту. Дозволяючи користування програмою без додаткової плати, підприємство-виробник може брати плату за дистанційну підтримку клієнтів та вирішення проблем технічного характеру. Такий підхід не буде стримувати розвиток науки і техніки, проте дозволить українським технологічним компаніям конкурувати із західними аналогами.

Список використаних джерел

1. Laszlo T. Algorithms for robust nonlinear regression with heteroscedastic errors / T. Laszlo, E. Laszlo. // Inter*NA*tio*NA*l Jour*NA*l of Bio-Medical Computing. – 1996. – №42. – С. 181–190.
2. Ermon S. Machine Learning 2: Nonlinear Regression / Stefano Ermon. – Stanford University: Stanford University Press, 2014. – 51 с.
3. Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms / [*m*. J. Sorich, J. O. Miners, R. A. McKinnon та ін.]. // J. Chem. Inf. Comput. Sci.. – 2003. – №43. – С. 2019–2024.
4. Finding Near-Optimal Bayesian Experimental Designs via Genetic Algorithms / *m*.Hamada, H. F. Marz, C. S. Reese, A. G. Wilson. // The American Statistician. – 2001. – №55. – С. 175–181.
5. Lange O. F. Generalized Correlation for Biomolecular Dy*NA*mics / O. F. Lange,

H. Grubmüller. // Proteins: Structure, Function, and Bioinformatics. – 2006.

– No62. – С. 1053–1061.
6. An integrated map of genetic variation from 1,092 human genomes / [G. R. Abecasis, A. Auton, L. D. Brooks та ін.]. // *NA*ture. – 2012. – №491. – С. 56–65.
7. Molecular Biology of the Cell / [B. Alberts, A. Johnson, J. Lewis та ін.]. – Garland: Garland Science, 2002. – (4th edition).
8. 3. Direct and immune mediated antibody targeting of ERBB receptors in a colorectal cancer cell-line panel / [S. Q. Ashraf, A. *m*. Nicholls, J. L. Wilding та ін.]. // Proceedings of the *NA*tio*NA*l Academy of Sciences. – 2012. – №109. – С. 21046– 21051.
9. The multitude and diversity of environmental carcinogens / [D. Belpomme, P. Irigaray, L. Hardell та ін.]. // Environmental Research. – 2007. – №105. – С. 414–429.
10. A Landscape of Pharmacogenomic Interactions in Cancer / [F. Iorio, T. A. Knijnenburg, D. J. Vis та ін.]. // Cell. – 2016. – №166. – С. 740–754.
11. Cermi*NA*ra N. The Complete Visual Guide to Sublime Text 3: Getting Started and Keyboard Shortcuts [Електронний ресурс] / Nicholas Cermi*NA*ra. – 2014. – Режим доступу до ресурсу: https://scotch.io/bar-talk/the-complete-visual-guide-to-sublime-text-3-getting-started-and-keyboard-shortcuts.
12. van Buuren S. Flexible Imputation of Missing Data / Stefan van Buuren. – Amsterdam: Chapman and Hall/CRC, 2018. – 416 с. – (Chapman & Hall/CRC Interdisципli*NA*ry Statistics)
13. Wickham H. R for Data Science / Hadley Wickham. – Boston: O'Reilly Media, 2016. – 520 с.

14. Wickham H. R Packages / Hadley Wickham. – Wellington: O'Reilly Media, 2015. – 202 с.

15. Wickham H. ggplot2 v 3.3.0: the R package [Електронний ресурс] / Hadley Wickham // TidyVerse. – 2019. – Режим доступу до ресурсу: https://ggplot2.tidyverse.org/reference/.

16. Kieran H. Data Visualization: A Practical Introduction / Healy Kieran. – Princeton: Princeton University Press, 2018. – 296 с.

17. Peng R. D. R Programming for Data Science [Електронний ресурс] / Roger D. Peng // LeanPub. – 2019. – Режим доступу до ресурсу: https://leanpub.com/rprogramming.