

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ОБРАЗОВАНИЯ

"МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ)"

ФИЗТЕХ-ШКОЛА БИОЛОГИЧЕСКОЙ И МЕДИЦИНСКОЙ ФИЗИКИ

КАФЕДРА БИОИНФОРМАТИКИ И СИСТЕМНОЙ БИОЛОГИИ

Выпускная квалификационная работа по направлению

03.03.01

ПРИКЛАДНАЯ МАТЕМАТИКА И ФИЗИКА

на тему:

**Разработка метода для предсказания синергетический комбинаций
малых молекул на основе данных RNA-seq**

Студент _____ Зиганшина Д. О.

Научный руководитель _____ Муртазалиева Х. А.

Зав. кафедрой, регент-профессор, PhD, _____ Бородовский М. Ю.

МОСКВА, 2020

1 Аннотация

Синергизм – тип взаимодействия между двумя или более химическими агентами, который характеризуется тем, что общий эффект лекарств превышает сумму индивидуальных эффектов каждого лекарства. Задача вычислительного предсказания синергетических эффектов остается нерешенной и актуальной во многих областях биомедицины (например, для подбора комбинаций противоопухолевых лекарств или предсказания комбинаций для клеточного перепрограммирования). В работе представлен новый подход для предсказания синергетических комбинаций. Данный метод позволяет подбирать синергетические пары малых молекул для достижения интересующих изменений в фенотипе клетки. В качестве входных данных метод принимает экспрессионные сигнатуры, которые позволяют описать изменение в экспрессии генов между 2 различными клеточными состояниями. С помощью них можно охарактеризовать действие препарата на клеточную линию. В данной работе было опробовано несколько способов вычисления уровня синергии.

Первый способ основан на сравнении экспрессионных сигнатур с учетом биологической значимости генов в контексте интересующих изменений. Значимость гена оценивается линейной комбинацией топологических метрик центральности, рассчитанных для него в генной сети. Была выполнена масштабная оптимизация метода по подбору коэффициентов этого выражения.

Второй способ основан на сравнении наборов обогащенных сигнальных путей, которые также позволяют характеризовать действие препарата на клеточную линию. В данном способе было опробовано 2 варианта подсчета уровня синергии малых молекул. Первый основан на гипотезе, что синергетические малые молекулы, дополняя друг друга, затрагивают необходимые клеточные процессы. Во втором предполагалось, что синергетические пары соединений затрагивают одни и те же процессы.

Была выполнена валидация вышеупомянутых методов в задаче химического перепрограммирования клеток.

Содержание

1	Аннотация	1
2	Обозначения, сокращения, основные определения	4
3	Введение	5
3.1	Основные концепции Хемоинформатики	5
3.1.1	Представление	5
3.1.2	Молекулярные дескрипторы	10
3.1.3	Молекулярные отпечатки	11
3.1.4	Молекулярное подобие	11
3.2	Синергия	12
3.2.1	Область применения синергии	12
3.2.2	Предсказание синергетических пар соединений	13
3.3	Анализ обогащения сигнальных путей	38
4	Материалы и методы	40
4.1	Создание экспрессионной сигнатуры запроса	40
4.2	Вычисление уровня синергии	43
4.2.1	Вычисление уровня синергии на основе сравнения сигнатур	44
4.2.2	Вычисление уровня синергии на основе сравнения обогащенных сигнальных путей	45
4.3	Валидация	48
5	Полученные результаты	51
5.1	Оптимизация метода, основанного на сравнении экспрессионных сигнатур, и его валидация	51
5.2	Валидация метода, основанного на сравнении обогащенных сигнальных путей	57
5.2.1	С использованием коэффициента Танимото	58
5.2.2	С использованием взаимной информации	60
6	Заключение. План дальнейших исследований	62

7	Благодарности	64
8	Список литературы	65

2 Обозначения, сокращения, основные определения

Определение 2.1 (Синергизм).

Синергизм – тип взаимодействия между двумя или более химическими агентами, который характеризуется тем, что общий эффект лекарств превышает сумму индивидуальных эффектов каждого лекарства.

Определение 2.2 (Хемоинформатика).

Хемоинформатика это научная дисциплина на пересечении химии, информатики и математики. Основные задачи в области связаны с построением вычислительных методов для хранения и обработки химической информации и дизайна малых молекул с заданными свойствами.

Определение 2.3 (Дифференциальная экспрессия генов).

Явление дифференциальной экспрессии генов состоит в том, что экспрессия генов в одном клеточном состоянии отличается от их экспрессии в другом клеточном состоянии. Регуляция экспрессии генов может происходить на разных уровнях: репликации, транскрипции, трансляции, а также в процессе созревания иРНК и полипептидных цепей, образующихся в результате трансляции.

Определение 2.4 (Экспрессионная сигнатура).

Набор статистически значимых дифференциально экспрессирующихся генов между 2 клеточными состояниями. Он состоит из 2 списков генов: с повышенной и пониженной экспрессией.

Определение 2.5 (Quantitative Structure-Activity Relationship, (Q)SAR).

Quantitative Structure-Activity Relationship - это одно из интенсивно развивающихся направлений использования математических методов в химии. Под QSAR подразумевается поиск зависимостей между структурами химических соединений и их свойствами. Также часто аббревиатуру QSAR используют для обозначения моделей, в основе которых лежит концепция связи "структура-свойство".

Определение 2.6 (Сигнальный путь).

Сигнальный путь – это последовательность молекул, посредством которых информация от клеточного рецептора передается внутри клетки.

3 Введение

3.1 Основные концепции Хемоинформатики

Хемоинформатика, согласно определению данному Й. Гастайгером [1], - это применение методов информатики для решения химических задач. Одна из основополагающих задач в химии это создание соединений с заданными свойствами. Моделирование (количественных) соотношений "структура-активность"((Q)SAR, Quantitative Structure-Activity Relationships) позволяет выявить взаимосвязь между структурой химических соединений и их активностью, чаще всего биологической. Модели SAR широко используются для виртуального скрининга при разработке лекарств с целью сокращения количества экспериментальных испытаний. Становление химии как науки привело к накоплению огромного количества данных, поэтому возникла необходимость ими оперировать: хранить информацию о миллионах химических соединениях и осуществлять быстрый поиск в этой информации. Более того, количество потенциальных химических соединений почти бесконечно, например, существует более 10^{29} возможных производных н-гексана со 150 заместителями [2]. Для работы с такими объемами информации требовалось создать машиночитаемое представление химических структур. Оно должно быть уникальным и однозначно интерпретируемым.

3.1.1 Представление

Представление структур химических соединений делится на 2 типа:

- внутреннее
- внешнее

Внутреннее представление

Когда говорят о внутреннем представлении структур химических соединений подразумевается машинное представление. Для внутреннего представления обычно используются молекулярные графы. Молекулярный граф — связный граф, находящийся во взаимно-однозначном соответствии со структурной формулой химического соединения таким образом, что вершинам графа соответствуют атомы молекулы, а ребрам графа — химические связи между этими атомами. Такое представление не

хранит информацию о трехмерной структуре молекулы. Обычно граф неориентированный (связи в молекуле не имеют направления) и его вершины помечены (символами атомов) [1]. Если вершины графа непомечены, то он будет отражать только структуру, а не состав молекулы. Две вершины графа могут соединяться несколькими ребрами, так как связь может быть одинарной или кратной.

Граф может быть представлен в виде матрицы разными способами, например матрицей смежности, расстояний, инцидентности.

Матрица смежности для молекулы, состоящей из n атомов, это квадратная матрица размером $n \times n$, содержащая информацию о всех связях в молекуле. Если на пересечении i -ой строки и j -ого столбца стоит 1, то между соответствующими атомами есть связь. Если нет связи между рассматриваемыми атомами, то на соответствующей позиции матрицы стоит 0. То есть матрица смежности является Булевой матрицей. Все диагональные элементы матрицы равны 0 и она является симметричной. Такая матрица является избыточной. Ее можно упростить, устранив дублирование половины матрицы, то есть приведя к верхней треугольной матрице. Также для ясности можно упустить нули и удалить информацию об атомах водорода [1].

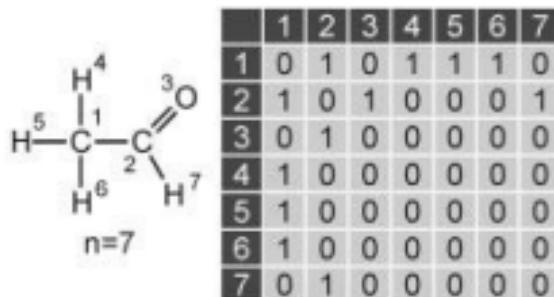


Рис. 3.1: Матрица смежности этанола [1]

Для примера рассмотрим матрицу расстояний, у которой ее элемент представляет кратчайшее расстояние между соответствующими атомами. Расстояние может выражаться в геометрическом расстоянии (ангстремах) или топологическом расстоянии (числе связей) [1].

	1	2	3	4	5	6	7	
1	0	1	0	1	1	1	0	
2	1	0	1	0	0	0	1	
3	0	1	0	0	0	0	0	
4	1	0	0	0	0	0	0	
5	1	0	0	0	0	0	0	
6	1	0	0	0	0	0	0	
7	0	1	0	0	0	0	0	

	1	2	3	4	5	6	7	
1	1		1	1	1			
2	1	1				1		
3		1						
4	1							
5	1							
6	1							
7	1							

	1	2	3	4	5	6	7	
1	1		1	1	1			
2		1				1		
3			1					
4				1				
5					1			
6						1		
7							1	

	1	2	3	4	5	6	7	
1								
2								
3								
4								
5								
6								
7								

Рис. 3.2: a) избыточная матрица смежности этанола; b)матрица после опускания нулей; c) упрощенная до верхней треугольной матрицы; d) после удаления атомов водорода [1]

a)

	C1	C2	O3	H4	H5	H6	H7	
C1	0	1.400	2.190	1.022	1.023	1.022	2.106	
C2	1.400	0	1.123	1.999	1.982	1.999	1.022	
O3	2.190	1.123	0	2.349	2.708	2.995	1.859	
H4	1.022	1.999	2.349	0	1.668	1.661	2.895	
H5	1.023	1.982	2.708	1.668	0	1.668	2.562	
H6	1.022	1.999	2.955	1.661	1.668	0	2.336	
H7	2.106	1.022	1.859	2.895	2.566	2.336	0	

b)

	C1	C2	O3	H4	H5	H6	H7	
C1	0	1	2	1	1	1	2	
C2	1	0	1	2	2	2	1	
O3	2	1	0	3	3	3	2	
H4	1	2	3	0	2	2	3	
H5	1	2	3	2	0	2	3	
H6	1	2	3	2	2	0	3	
H7	2	1	2	3	3	3	0	

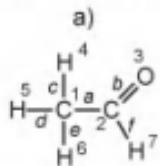
Рис. 3.3: матрица расстояний этанола с a) геометрическим расстоянием; b) топологическим расстоянием [1]

Еще одним представлением является матрица инцидентности. Это матрица размером $n \times m$, где n -число вершин(атомов), m - число ребер (связей). Если на пересечении i -ой строки и j -ого столбца стоит значение 1, то рассматриваемые вершина и ребро инцидентные [1].

Все описанные матрицы не несут информации о типе и порядке связей в молекуле.

Одним из недостатков матрицы смежности является, что число ее элементов равно квадрату числа атомов, а для представления молекулярного графа необходимо, чтобы число элементов в представлении линейно зависело от числа атомов в молекуле. Это достигается с помощью представления таблицей связности, в которой дается список атомов и список связей. Существует много вариантов матриц связности. Например, атомы произвольно нумеруются и в соответствии с индексом заносятся в

	C1	C2	O3	H4	H5	H6	H7
a	1	1	0	0	0	0	0
b	0	1	1	0	0	0	0
c	1	0	0	1	0	0	0
d	1	0	0	0	1	0	0
e	1	0	0	0	0	1	0
f	0	1	0	0	0	0	1



n=7; m=6

	C1	C2	O3	H4	H5	H6	H7
a	1	1					
b		1	1				
c	1			1			
d	1				1		
e	1					1	
f		1					1

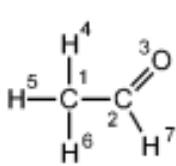
b)

	C1	C2	O3
a	1	1	
b		1	1

c)

Рис. 3.4: а) избыточная матрица инцидентности этанола; б) матрица после опускания нулей; в) после опускания атомов водорода [1]

список атомов. Информация о связях хранится во второй таблице, где для каждой связи записываются индексы атомов, которые она соединяет и ее кратность [1].



Atom list	
1	C
2	C
3	O
4	H
5	H
6	H
7	H

Bond list		
1 st atom	2 nd atom	bond order
1	2	1
2	3	2
2	7	1
1	4	1
1	5	1
1	6	1

Рис. 3.5: таблица связности этанола

Другой формой представления является таблица, в которой первые две колонки содержат информацию о индексах и символах атомов. Далее таблица дополняется колонками, в которых указаны индексы соседних атомов и кратность связи. Такая форма представления избыточна, в ней каждая связь записана дважды. Ее можно упростить, убрав повторение связей и опустив атомы водорода.

atom index	element	1 st index of atom	bond order	2 nd index of atom	bond order	3 rd index of atom	bond order	4 th index of atom	bond order
1	C	2	1	4	1	5	1	6	1
2	C	1	1	3	2	7	1		
3	O	2	2						
4	H	1	1						
5	H	1	1						
6	H	1	1						
7	H	2	1						

Рис. 3.6: таблица связности этанола [1]

Таблицы связности могут быть дополнены другими списками, например содержащими информацию о свободных электронах или заряде атомов, что является еще одним преимуществом по сравнению с матрицами смежности [1].

Методы теории графов находят широкое применение в хемоинформатике.

Внешнее представление

Внешнее представление химических соединений используется в случае долговременного хранения химической информации и обмена ею между приложениями.

Простейшим типом внешнего представления структур химических соединений являются линейные нотации в виде строки символов. Линейные нотации позволяют свободно обмениваться информацией о химических соединениях без необходимости использования специального программного обеспечения. Наиболее популярные типы линейных нотаций:

- SMILES (Simplified Molecular Input Line Entry System);
- SMARTS (расширение SMILES)
- InChI (IUPAC International Chemical Identifier)

В настоящее время наиболее распространённым видом линейных нотаций являются строки SMILES. Преобразование структуры соединения в строку символов SMILES определяется 6 правилами [1] :

- Атомы представлены их атомными символами.
- Атомы водорода опущены

- Соседние атомы расположены рядом друг с другом.
- Двойные и тройные связи характеризуются «=» и «», соответственно.
- Разветвления молекулы представлены скобках.
- Циклы описаны путем выделения цифры для двух "соединяющихся" атомов в цикле.

Синтаксис SMILES позволяет описывать структурные изомеры. SMARTS является расширением SMILES для поисковых запросов в химических базах данных.

Для кодировки химических структур IUPAC предложил универсальную линейную нотацию InChI. То есть это цифровой эквивалент названию соединения по номенклатуре IUPAC. Он содержит следующие уровни информации: связанность, таутомеризм, изотопы, стереохимия, заряд. Этот формат более сложен для интерпретации пользователем, но сохраняет такую возможность.

InChIKey – хешированная версия InChI, созданная для быстрого поиска.

Второй тип внешнего представления структур химических соединений основан на непосредственном кодировании таблицы связности молекулярного графа. Такие распространённые форматы как MOL, SDF и RDF, которые в настоящее время являются стандартными для обмена химической информацией, можно считать способами представления в виде текстового файла матрицы смежности молекулярного графа.

3.1.2 Молекулярные дескрипторы

Традиционный подход к обработке химической информации состоит в отображении химического пространства на дескрипторное пространство, образуемое вычисляемыми для каждого химического объекта векторами молекулярных дескрипторов — числовых характеристик, описывающих химические объекты. Это дает возможность применять методы математической статистики и машинного обучения для работы с химическими объектами. В принципе дескриптором может являться любое число, которое можно рассчитать из структурной формулы химического соединения. Молекулярные дескрипторы можно классифицировать по "размерности": 0D, 1D, 2D, 3D и 4D. 0D дескрипторы описывают совокупную информацию, такую как количество атомов, количество связей, молекулярный вес. 1D дескрипторы описывают число фрагментов в молекуле. В качестве примера можно привести число гидроксильных

групп, нитрогрупп [1]. 2D дескрипторы описывают свойства, которые могут быть вычислены из двумерного представления молекул (например, индексы связности) и 3D-дескрипторы зависят от конформации молекул (например, доступная растворителю площадь поверхности) [3].

3.1.3 Молекулярные отпечатки

Молекулярные «отпечатки» (molecular fingerprints) содержат информацию о присутствии или отсутствии определенных признаков в химическом соединении, например, фрагментов. Могут быть организованы 2 основными способами:

- бинарной строки
- хеш-таблицы

Рассмотрим подробнее молекулярные отпечатки в виде бинарной строки. Каждая подструктура или фрагмент активирует определенное количество позиций (битов) в молекулярном отпечатке. Иногда подструктуре может соответствовать 1 или несколько битов. Алгоритм определяет какие биты были активированы подструктурой. Одна и та же подструктура всегда активизирует одинаковые биты. Если фрагменту соответствует 1 бит, то в случае равенства единице, то фрагмент присутствует в молекуле, иначе его значение равно нулю. Обычно длина бинарной строки 150-2500 битов. Алгоритм работает таким образом, что всегда возможно ассоциировать биты с конкретной подструктурой. Однозначное представление химической структуры строкой позволяет проводить эффективный поиск схожих молекул.

Для хэшированных молекулярных отпечатков нет возможности определить, какие конкретные элементы присутствуют в молекуле.

3.1.4 Молекулярное подобие

В основе принципа молекулярного подобия лежит идея, что структурно схожие молекулы предположительно обладают сходной биологической активностью. Однако это предположение не всегда может быть верным. Например, "activity cliffs в которых незначительная модификация функциональных групп вызывает резкое изменение активности [4]. Поиск структурного сходства молекул основан на доле различных

фрагментов, которые присутствуют одновременно в обеих молекулах. Поиск молекул по такому критерию называется поиском по молекулярному подобию (Similarity Search). В качестве количественной меры молекулярного подобия часто рассматривается величина, возрастающая с уменьшением расстояния между химическими соединениями в дескрипторном пространстве.

Структурное сходство двух молекул чаще всего оценивается путем вычисления коэффициента Танимото (T_c). T_c , также известный как индекс Джаккарда, описывает степень схожести двух множеств. Для парного сравнения используются молекулярные отпечатки молекул. Коэффициент Танимото определяется как :

$$T_c = \frac{bc}{b1 + b2 - bc}, \quad (3.1)$$

где $b1$ - число битов набора первой молекулы, $b2$ - число битов набора второй молекулы, bc - число битов общих для обеих молекул. Значения коэффициента Танимото лежит в пределах от 0 до 1 [3]. Высокие значения T_c указывают на то, что два соединения похожи, но не дает информации о масштабах сходства, например о том, какие конкретные химические группы они разделяют.

3.2 Синергия

Под комбинированной терапией подразумевают одновременное применение нескольких препаратов.

Существует три типа эффектов от комбинации препаратов:

- аддитивный, когда комбинированный эффект эквивалентен сумме независимых эффектов
- синергический, когда комбинированный эффект больше аддитивного
- антагонистический, когда комбинированный эффект меньше аддитивного

Целью комбинированной терапии является достижение синергического или, по крайней мере, аддитивного, но комплементарного эффекта [5, 6]

3.2.1 Область применения синергии

Большинство заболеваний вызвано сложными биологическими процессами. Широко известным примером являются онкологические заболевания. На данный момент

разработана таргетная терапия рака, которая блокирует рост раковых клеток с помощью вмешательства в механизм действия конкретных целевых (таргетных) молекул, необходимых для канцерогенеза, то есть ингибирует критические сигнальные пути рака. Как и ожидается, такая терапия будет более эффективной, чем прежние виды лечения, и менее вредной для нормальных клеток. Однако за резким начальным положительным ответом многих таргетных методов лечения рака часто следует развитие лекарственной устойчивости, приводящей к рецидиву заболевания[7]. Существует множество механизмов, которые могут привести к лекарственной устойчивости, которые включают генетическую и негенетическую гетерогенность, присущую распространенным видам рака, в сочетании со сложными механизмами обратной связи и регуляции, а также динамическими взаимодействиями между опухолевыми клетками и их микроокружением. Любая монотерапия может быть ограничена по своей эффективности, но комбинации лекарств потенциально могут преодолеть эти ограничения [8]. Они состоят из нескольких агентов, каждый из которых обычно используется в клинике как один эффективный препарат. Поскольку агенты в лекарственных комбинациях могут модулировать активность отдельных белков, лекарственные комбинации могут помочь повысить терапевтическую эффективность, преодолев избыточность, лежащую в основе лекарственной устойчивости. Соответственно приводят к более длительным реакциям у пациентов[9].

Кроме того, токсичность и неблагоприятные побочные эффекты, вероятно, снижаются, поскольку дозы комбинаций лекарств обычно ниже, чем дозы отдельных агентов. В настоящее время медикаментозная комбинаторная терапия становится перспективной стратегией лечения многофакторных сложных заболеваний [10].

Также необходимо упомянуть еще одну возможную область применения синергии - химическое перепрограммирование клеток малыми молекулами. Точнее, эта область нуждается в использовании комбинаций малых молекул, так как при переходе из одного клеточного состояния в другое происходят большие изменения в фенотипе и обработки клетки одним соединением недостаточно.

3.2.2 Предсказание синергетических пар соединений

Изначально эффективные комбинации препаратов предлагались на основе клинического опыта и большинство подходов к выявлению синергетических пар соединений

часто носит экспериментальный характер. В исследованиях рака анализ синергизма обычно проводится путем обработки клеточных линий *in vitro* всеми возможными комбинациями соединений. Однако такой подход требует много усилий и затрат. Более того, экспериментальные скрининги накладывают серьезные ограничения на практический размер библиотек и их разнообразие. Вычислительные методы прогнозирования синергии соединений потенциально могут позволить исследователям отбирать наиболее перспективные пары для экспериментального скрининга, и, вследствие сокращения количества изучаемых комбинаций, сократить затраты ресурсов [5].

Большинство существующих методов скрининга нацелены на прогнозирование синергических эффектов двух препаратов, поскольку комбинаторный эффект трех и более препаратов технически сложнее предсказать и практически отсутствуют экспериментальные данные для последующей валидации метода [6]. Также есть наблюдения, что наиболее значительное улучшение достигается при добавлении только одного дополнительного препарата. При дальнейшем добавлении лекарств эффективность постепенно уменьшается и в конечном счете достигает плато [11].

На молекулярном уровне синергетические взаимодействия могут быть реализованы несколькими различными механизмами. Например, соединение может сенсибилизировать клетки к другому соединению, регулируя его поглощение и распределение, моделируя ростовые свойства клетки, ингибируя деградацию соединения, ингибируя пути, индуцирующие резистентность или снижая токсичность другого соединения [5].

Механизмы синергических эффектов не являются универсальными среди различных лекарств или раковых заболеваний [6].

Можно выделить две гипотезы, лежащие в предсказании синергии препаратов [5]:

- гипотеза сходства: соединения, более схожие по вызываемым транскрипционным изменениям, с большей вероятностью будут синергетическими
- гипотеза комплементарности: соединения, которые вызывают наиболее различные, но взаимодополняющие транскрипционные изменения, с большей вероятностью будут синергетическими (так как мы ищем синергетические пары среди препаратов с определенным необходимым эффектом, то можно сказать, что по гипотезе несходства синергетические соединения комплементарно дополняют

друг друга в достижении желаемого эффекта)

На данный момент отсутствуют стандартные подходы к прогнозированию активности пар соединений на основе транскриптомных данных [5].

Поиск синергетических пар можно представить в виде задачи регрессии, когда необходимо предсказать уровень синергии пары препаратов. Эта же задача может быть представлена в виде ранжирования списка пар препаратов по степени синергии [5]. Но также можно рассматривать более простую задачу бинарной классификации, когда необходимо сказать, является пара препаратов синергетической или нет [8].

Данные, используемые для предсказания синергии, можно поделить на несколько типов:

- глубокая молекулярная характеристика клеточных линий, которая включает соматические мутации, изменения числа копий, метилирование ДНК и профили экспрессии генов
- фармакологические признаки препаратов, включая предполагаемые лекарственные мишени и химические свойства препаратов, представленные дескрипторами и молекулярными отпечатками, области медицинских показаний, побочные эффекты и токсикофоры, сигнальные пути
- данные монотерапии, включающие зависимость жизнеспособности клеток от дозы препарата при обработке одиночным препаратом, транскрипционные данные (профиль экспрессии генов клеток после обработки препаратом)

Безусловно, признаки типа клеточной линии существенны, так как они определяют молекулярный контекст. [8]. Признак "сигнальный путь" слабо предсказательный, возможно, потому, что простая ассоциация между лекарствами и путями через целевые белки недостаточно отражает физиологический контекст, в котором работают лекарства [9]. Также плохой плохой предсказательной способностью обладает признак "побочные эффекты потому что он сильно зашумлен, так как существуют некоторые общие побочные эффекты, связанные с большинством лекарств. Эффективность признака побочных эффектов может быть улучшена, если рассмотреть только тяжелые побочные эффекты, связанные с лекарствами [9]. Монотерапия дает существенную информацию о прямом лечебном эффекте на линию раковых клеток, что является основой лекарственного синергизма. Например, если препарат вообще

не действует на определенную клеточную линию, он вряд ли будет синергировать с другими препаратами [6]. Однако такой признак, как зависимость жизнеспособности клеток от дозы препарата, в основном используется для поиска пар противопухолевых препаратов и не подходит для химического перепрограммирования. Также наблюдалась информативность транскрипционных данных, поскольку профили геномной экспрессии очень динамичны и зависят от контекста [12].

Однако различные типы данных дополняют друг друга при прогнозировании комбинаций лекарственных средств [9]. То есть использование ансамбля различных наборов признаков улучшает качество прогноза [8].

Кроме того, по предоставленному набору данных задачу поиска синергетических пар можно реализовывать в двух сценариях [13]:

- для предсказания уровня синергии пар препаратов предоставляются только данные по используемым клеточным линиям, фармакологические признаки препаратов, данные монотерапии
- для предсказания уровня синергии пар препаратов помимо вышеперечисленных данных предоставляется обучающий набор данных, который, например, может состоять из других пар препаратов, для которых будут доступны данные по клеточным линиям, фармакологические признаки препаратов, данные монотерапии и также известно, какие пары препаратов являются синергетическими, какие нет. В данном случае возможно применять методы машинного обучения.

В реальности предоставляемые данные бывают очень разнообразными, поэтому может применяться и смешанный сценарий, то есть использовать обучение с частичным привлечением учителя [12].

Если в работе используется сценарий обучения с учителем, то либо известны синергетические комбинации, либо есть экспериментальные данные жизнеспособности клеток после обработки комбинациями препаратов в зависимости от дозы, по которым рассчитывается синергетический эффект.

Среди подходов к моделированию лекарственной синергии популярны методы машинного обучения. Однако труднодоступность обучающих данных препятствует широкому использованию методов машинного обучения. Используются следующие подходы: регрессия, деревья решений, случайные леса, Гауссовские процессы, SVM, ней-

ронные сети. При сравнении моделей был сделан вывод, что класс алгоритмов показал слабую связь с производительностью метода [8].

Так как предсказание синергии можно отнести к задаче бинарной классификации, то популярны метрики ROC-AUC, accuracy, precision, recall, F1 [9, 12]. Рассмотрим их подробнее. Наша данные поделены на два класса. Их метки принято обозначать как «положительные» и «отрицательные». При классификации мы можем верно определить класс (истинно) или допустить ошибку, то есть отнести к ложному классу, поэтому возможны следующие исходы:

- истинно положительные (TP)
- истинно отрицательные (TN)
- ложно положительные (FP)
- ложно отрицательные (FN)

		Prediction outcome		$TP + FN$
		positive	negative	
Actual value	positive	TP	FN	$TP + FN$
	negative	FP	TN	
		$TP + FP$	$FN + TN$	

Рис. 3.7: Матрица сопряженности возможных результатов бинарной классификации

Одной из наиболее простых метрик является точность (accuracy). Она показывает количество верно классифицированных объектов (истинно положительных и истинно отрицательных) относительно общего количества всех объектов и считается следующим образом:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.2)$$

Метрика accuracy имеет недостаток: она не подходит для несбалансированных классов, где может быть много экземпляров одного класса и мало другого.

Для оценки качества работы алгоритма на каждом из классов по отдельности вводятся метрики precision (точность) и recall (полнота).

Метрика precision показывает сколько из всех объектов, которые классифицируются как положительные, действительно являются положительными, относительно общего количества полученных от модели позитивных меток.

$$precision = \frac{TP}{TP + FP} \quad (3.3)$$

Важность этой метрики определяется тем, насколько высока для рассматриваемой задачи «цена» ложно положительного результата.

Метрика recall показывает, сколько объектов модель смогла правильно классифицировать с позитивной меткой из всего множества позитивных. Она вычисляется по следующей формуле:

$$recall = \frac{TP}{TP + FN} \quad (3.4)$$

Необходимо уделить особое внимание этой оценке, когда в поставленной задаче ошибка не распознать положительный класс высока.

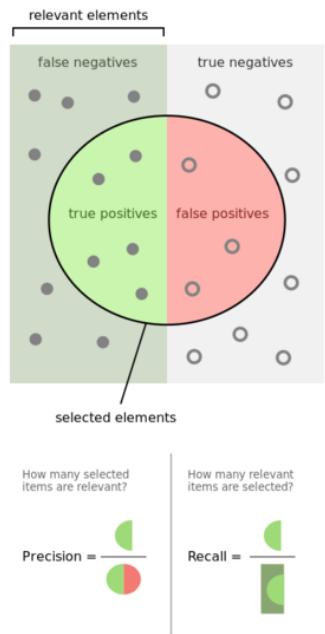


Рис. 3.8: Схематичное представление precision и recall

Precision и recall не зависят, в отличие от accuracy, от соотношения классов и потому применимы в условиях несбалансированных выборок. Если Precision и Recall являются одинаково значимыми, то можно использовать их среднее гармоническое для получения оценки результатов:

$$F1-score = \frac{2 \cdot recall \cdot precision}{recall + precision} \quad (3.5)$$

Для определения бинарной метки (0 или 1) по какому-нибудь вещественному ответу алгоритма (как правило, вероятности принадлежности к классу) необходимо выбрать порог, до которого ставится метка "0" после "1". Порог, равный 0.5 кажется естественным, но он не всегда оказывается оптимальным, например, при отсутствии баланса классов.

Одним из способов оценить модель в целом, не привязываясь к конкретному порогу, является метрика ROC AUC — площадь (Area Under Curve) под кривой ошибок (Receiver Operating Characteristic curve). Данная кривая представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR).

True Positive Rate (Recall) показывает долю верно классифицируемых объектов положительного класса и считается по формуле:

$$TPR = \frac{TP}{TP + FN} \quad (3.6)$$

False Positive Rate показывает, какую долю из объектов отрицательного класса алгоритм предсказал неверно, и определяется как:

$$FPR = \frac{FP}{FP + TN} \quad (3.7)$$

Когда классификатор не делает ошибок ($FPR = 0$, $TPR = 1$) мы получим площадь под кривой, равную единице; в противном случае, когда классификатор случайно выдает вероятности классов, AUC-ROC будет стремиться к 0.5, так как классификатор будет выдавать одинаковое количество TP и FP. Каждая точка на графике соответствует выбору некоторого порога. Площадь под кривой в данном случае показывает качество алгоритма (больше — лучше), кроме этого, важной является крутизна самой кривой — надо максимизировать TPR, минимизируя FPR, а значит, кривая в идеале должна стремиться к точке (0,1).

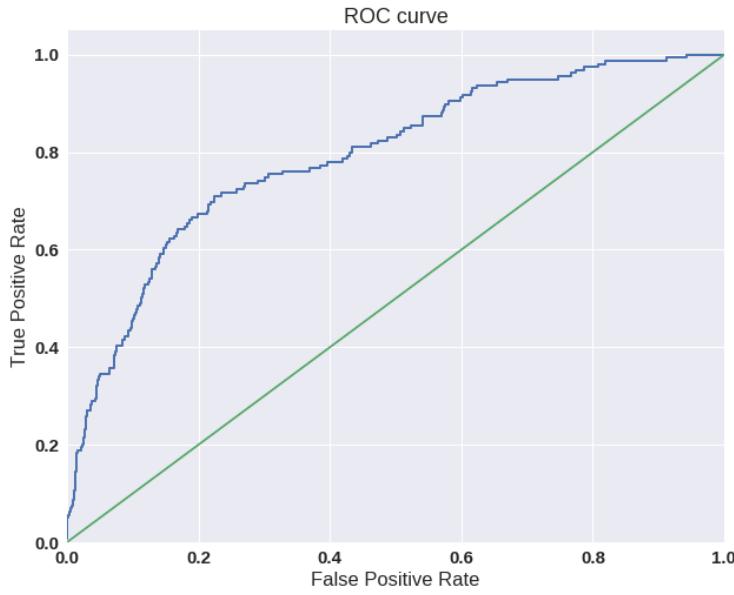


Рис. 3.9: ROC кривая

Для задачи регрессии используется MSE, RMSE, коэффициент корреляции Пирсона [10]. Рассмотрим их подробнее.

Метрика Mean Squared Error (MSE). Измеряет среднюю сумму квадратной разности между фактическим значением и прогнозируемым значением для всех объектов выборки. Возведение во вторую степень необходима, чтобы отрицательные значения не компенсировались положительными. Чем больше разность, тем больше ее вес в этой метрики. Ниже приведена ее формула.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y_i^{\text{pred}})^2 \quad (3.8)$$

Метрика Root Mean Squared Error (RMSE) - это корень от квадрата ошибки. Формула приведена ниже.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_i^{\text{pred}})^2} \quad (3.9)$$

Также в соревновании консорциума DREAM использовалась РС-метрика, которая будет рассмотрена ниже [5], [8], [6], [12].

Для проверки предсказанных синергетических пар используются экспериментальные данные и литература [12].

Ниже описаны некоторые методы предсказания синергии, начиная с первого сценария, когда недоступен обучающий набор данных. В качестве примера необходимо

упомянуть соревнования, проводимые консорциумом DREAM Challenges.

Методы предсказания синергии, разработанные на основе наборов данных консорциума DREAM Challenges

В 2012 году был проведен NCI-DREAM Drug Synergy Prediction Challenge: перед участниками была поставлена задача предсказания синергетической и антагонистической активности пар соединений [5]. Перед участниками стояла задача отранжировать 91 пару соединений (все парные комбинации препаратов из 14 соединений) от наиболее синергетических до наиболее антагонистичных при воздействии на клеточную линию OCI-LY3 диффузной крупноклеточной лимфомы (DLBCL). Для этого им были предоставлены следующие данные :

- зависимость жизнеспособности клеток OCI-LY3 от дозы препарата после обработки им (для каждого препарата из набора и включая ДМСО в качестве контрольной среды)
- профили экспрессии генов для 3 биологических копий необработанных клеток и через 6 ч, 12 ч и 24 ч после обработки каждым соединением (для каждого препарата из набора)
- базовый генетический профиль клеточной линии OCI-LY3

Любые дополнительные данные из литературы или экспериментов считались допустимыми, но прямое измерение синергизма соединений было категорически запрещено. В соревновании было предложено 31 метод от участников и метод SynGen от одного из организаторов (этот подход оценивался отдельно от методов, предложенных участниками). Среди методов наблюдалось большое разнообразие, что хорошо показывает отсутствие стандартных подходов к прогнозированию синергетической активности пар соединений на основе транскриптомных данных. Также отсутствие обучающих данных препятствовало использованию методов машинного обучения. Среди 31 команды 10 основывали свои прогнозы на гипотезе о том, что соединения с более высоким сходством транскрипционного профиля с большей вероятностью будут синергетическими (гипотеза сходства), восемь команд предположили обратное (гипотеза несходства). Остальные команды либо использовали комбинацию гипотез

сходства и несходства (комбинированная гипотеза, $n = 4$), либо использовали более сложные гипотезы ($n = 9$).

Для проверки участников организаторами был создан валидационный набор данных, в котором экспериментально оценивали синергизм пар соединений по жизнеспособности клеток OCI-LY3.

Для экспериментальной оценки пользовались моделью excess over Bliss (EOB), которая определяет, является ли комбинированное действие двух соединений значительно большим или меньшим, чем независимое сочетание их индивидуальных эффектов. Также используется модель Bliss additivism, которая считает, что соединения D_x и D_y с экспериментально определенными долями ингибирования (доля клеток, погибших после обработки) f_x и f_y имеют аддитивный эффект, если ожидаемая доля ингибирования f_{xy} , индуцируемая их комбинацией определяется как:

$$f_{xy} = 1 - (1 - f_x)(1 - f_y) = f_x + f_y - f_x \cdot f_y \quad (3.10)$$

Excess over Bliss считается как разница между долей ингибирования комбинации f_z и ожидаемой долей ингибирования f_{xy} при аддитивном эффекте:

$$eob = f_z - f_{xy} \quad (3.11)$$

Если $eob \approx 0$, то пара соединений имеет аддитивный эффект. Затем если $eob > 0$ ($eob < 0$), то пара имеет синергетический (антагонистический) эффект. Оценка активности eob была использована для ранжирования всех пар от наиболее синергетических до наиболее антагонистических.

Для оценки предсказаний команд, использовался модифицированный индекс соответствия, который назвали вероятностным индексом соответствия. Эта метрика количественно определяет соответствие между ранжированием пар соединений в валидационной выборке и ранжированием, предсказанным командой.

Для подробного разбора оценки предсказаний необходимо рассмотреть индекс соответствия.

Проранжируем список из 91 соединения по экспериментально определенному и усредненному по всем репликатам EOB, от наиболее синергетических до наиболее антагонистических пар. Обозначим ранг пары i ($1 \leq i \leq 91$) как o_i . Аналогично для ранжированного списка пар, предсказанного командой, обозначим ранг пары p_i . Заметим, что если $i \neq j$, то $o_i \neq o_j$, $p_i \neq p_j$. Поэтому можем определить s_{ij} :

$$s_{ij} = \begin{cases} 1 & \text{if } (o_i > o_j \ p_i > p_j \text{ или } o_i < o_j \ p_i < p_j) \\ 0 & \text{if } (o_i > o_j \ p_i < p_j \text{ или } o_i < o_j \ p_i > p_j) \end{cases}$$

Индекс соответствия определяется как :

$$c - index = \frac{2}{91 \cdot 90} \sum_{i=1..90, j=i+1..91} s_{ij} \quad (3.12)$$

Для учета шума при экспериментальном ранжировании вводится вероятностный индекс соответствия. Для $i \neq j$ он вычисляется как :

$$sp_{ij} = \begin{cases} \frac{1}{2} + \frac{1}{2} \text{erf}\left(\frac{EOB_i - EOB_j}{\sqrt{sem_{EOB_i}^2 + sem_{EOB_j}^2}}\right) & \text{if } p_i < p_j \\ \frac{1}{2} - \frac{1}{2} \text{erf}\left(\frac{EOB_i - EOB_j}{\sqrt{sem_{EOB_i}^2 + sem_{EOB_j}^2}}\right) & \text{if } p_i > p_j \end{cases}$$

где erf - функция ошибки, EOB_i экспериментальное ЕОВ, усредненное по всем репликатам для i -ой пары, sem_{EOB_i} - среднеквадратическое отклонение ЕОВ для i -ой пары. Предположим, что пара i более синергетическая в среднем, чем j , тогда $EOB_i > EOB_j$ и аргумент функции ошибок положительный, то есть самая функция ошибки будет принимать положительные значения. Если будет пресказано, что пара i более синергетическая чем j ($p_i < p_j$), то $sp_{ij} > \frac{1}{2}$. Однако, если предсказание неверно, то есть $p_i < p_j$, то $sp_{ij} < \frac{1}{2}$. Аналогично в случае $EOB_i < EOB_j$. Если предсказание верно, то sp_{ij} лежит от 0,5 до 1, иначе от 0 до 0,5.

Вероятностный индекс соответствия определяется как :

$$PC - index = \frac{2}{91 \cdot 90} \sum_{i=1..90, j=i+1..91} sp_{ij} \quad (3.13)$$

Максимальный PC-index (PC_{max}) был равен 0,9. Минимальный PC-index (PC_{min}) определялся для предсказания с ранжированием пар противоположным (в обратном порядке) экспериментальному ранжированию пар, он был равен 0,1.

Нормализованный вероятностный индекс соответствия определяется как :

$$PC - index_{norm} = \frac{PC - index - PC_{min}}{PC_{max} - PC_{min}} \quad (3.14)$$

Для проверки результатов использовали вторую метрику - (пересчитанную корреляцию Спирмена). Результаты оценки методов по 2 метрикам были схожи, наблюдались малые отличия для нескольких команд, работавших хуже, чем случайная модель.

Среди 31 методы только три метода были статистически значимы (FDR = 0,05) : DIGRE, IUPUI_CCB and DPST.

DIGRE

Самым лучшим методом был DIGRE (drug-induced genomic residual effect). Он основывается на гипотезе, что при последовательной обработке клеток двумя соединениями транскрипционные изменения, индуцируемые первым препаратом , способствуют эффекту второго. То есть, что синергия обусловлена транскриптомными остаточными эффектами, которые представляют собой транскрипционные изменения, индуцированные первым соединением. Данная гипотеза согласуется с наблюдениями, что последовательность введение лекарств имеет влияние на результат. Несмотря на то, что соединения вводились одновременно в экспериментах, алгоритм моделирует синергию последовательно. В алгоритме можно выделить 3 шага:

- оценка сходства r между 2 соединениями в паре на основе сравнения транскрипционных изменений после обработки одним соединением. Смотрят на перекрытие дифференциально экспрессируемых генов относящихся к восьми сигнальным путям KEGG, связанными с ростом клеток (сфокусированный взгляд) и генов с повышенной экспрессией относящихся к 32 раковым сигнальным путям KEGG (глобальный взгляд)
- аппроксимируют долю выживших клеток после обработки вторым препаратом пары, учитывая влияние транскрипционных изменений, индуцированных первым препаратом. Для этого используют оценку сходства r , рассчитанную на предыдущем шаге.

$$1 - f_{B+A'} = (1 - rf_{2B})(1 - (1 - r)f_B), \quad (3.15)$$

где $f_{B+A'}$ - доля погибших клеток после обработки соединениями В (предполагается, что в клетке уже были индуцированы транскрипционные изменения соединением А), f_B , f_{2B} - доля погибших клеток после обработки соединением В после однократной и двойной дозы соответственно, которые определяются по предоставленной участникам зависимости жизнеспособности клеток от дозы препарата

- определяется доля погибших клеток, после обработки парой препаратов:

$$Z_{B+A'} = 1 - (1 - f_A)(1 - f_{B+A'}), \quad (3.16)$$

где f_A - доля погибших клеток после обработки препаратом А.

Затем оценка синергии определяется как среднее для 2 долей погибших клеток при обработке парой препаратов в разной предполагаемой последовательности ($Z_{B+A'}$, $Z_{A+B'}$)

Для этого метода PC index = 0.61.

Алгоритм SynGen

В основе алгоритма лежит предположение, что активность главных регуляторов (Master Regulators, MR) определенного клеточного фенотипа, которые выводятся Master Regulator Inference algorithm MARINa [14, 15], имеют важное значение для жизнеспособности клеток. MRs определяются как регуляторы, которые необходимы и достаточны для поддержания специфичной для фенотипа сигнатуры экспрессии генов. Целью обработки является:

- подавление активности MRs клеточного состояния
- активация MRs клеточной смерти

Таким образом, сначала SynGen выводит MRs для гибели клеток OCI-LY3 и состояния клеток, а затем идентифицирует соединения, которые наиболее комплементарны в индуцировании первого и отмене последнего. Для определения MRs используются 2 сигнатуры:

- сигнатура "клеточной смерти основанной на профиле экспрессии генов после обработки 14 соединениями, которые обладают заметной токсичностью
- сигнатур "клеточной зависимости связанной с активированным В-клеточным подтипов клеток DLBCL (которые включают OCI-LY3) по сравнению с подтипов В-клеток герминативного центра. Она вычисляется с использованием общедоступных профилей экспрессии генов для клеточных линий подтипа В-клеток герминативного центра (OCI-LY1, OCI-LY7, OCI-LY8, OCI-LY18 и SUDHL5) и для активированной линии подтипа В-клеток OCI-LY3.

Затем SynGen предсказывает синергетические комбинации соединений, выбрав составные пары, которые наиболее комплементарные в реализации или отмене этих MRs-паттернов соответственно. SynGen способен предсказывать только синергетические пары, то есть не предназначен для прогнозирования антагонизма соединений.

В 2015-2016 Dialog for Reverse Engineering Assessments and Methods (DREAM) Challenges в партнерстве с AstraZeneca и Институтом Сэнгера провели AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge [8]. Основными задачами соревнования были:

- предсказать, будет ли известная (ранее протестированная) комбинация лекарств эффективна для конкретного пациента
- предсказать, какие новые (непроверенные) комбинации лекарств будут синергетическими у пациентов
- определение новых биомаркеров, которые могут выявить основные механизмы, лежащие в основе синергии лекарств.

Данные были собраны на основе 11,576 экспериментов (зависимость жизнеспособности клеток от дозы препаратов в паре, представленная в виде матрицы размером 6×6 , в которой первая строка и первый столбец являются данными монотерапии для 2 препаратов) с использованием 85 клеточных линий рака. Оценка синергии пары препаратов рассчитывалась на основе матриц жизнеспособности клеток от доз препаратов в комбинации. Таким образом, набор данных включал высоковоспроизводимые измерения жизнеспособности клеток от доз препаратов в парах и оценки синергизма для 910 попарных комбинаций из 118 препаратов, а также информацию о лекарствах, включая предполагаемые лекарственные мишени и их химические свойства. Также была предоставлена глубокая молекулярная характеристика клеточных линий, включая соматические мутации, изменения числа копий, метилирование ДНК и профили экспрессии генов, измеренные до обработки клеток препаратами. Оценка синергии пары препаратов не была предоставлена по всем клеточным линиям рака.

Соревнование проводилось по 2 направлениям:

- предсказать синергию пары препаратов для конкретной клеточной линии, при этом частично известны уровни синергии для этой комбинации в других кле-

точных линиях (sub-challenge 1, SC1). первое направление было разделено на 2 категории:

- доступны все данные для предсказания синергии (SC1A)
- данные ограничены мутациями и изменение копийности (имитируя текущую возможность клинического анализа), (SC1B).

- предсказать синергию пары препаратов для конкретной клеточной линии, по которым не предоставлены обучающие данные по другим клеточным линиям, то есть используя переносимые признаки, идентифицированные из ранее изученных независимых пар лекарств (sub-challenge 2, SC2).

Участники обоих направлений использовали одни и те же обучающие наборы данных, но методы оценивались на разных тестовых выборках и с использованием разных метрик. Также важно отметить, что для первого направления уровни синергии, предсказанные участниками, должны быть непрерывными, в то время как SC2 требует бинарных предсказаний, указывающих, являются ли два препарата синергетическими или нет.

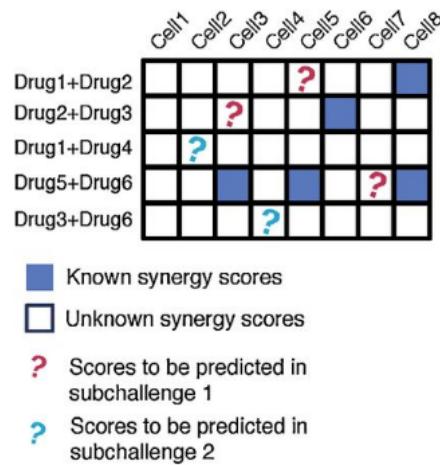


Рис. 3.10: Схематичное представление задач SC1 и SC2

Рассмотрим оценку методов по первому направлению соревнования. Поскольку размер выборки для каждой комбинации препаратов варьируется (например, некоторые комбинации препаратов были протестираны на большем количестве клеточных линий, чем другие), применение корреляции Пирсона непосредственно ко всем показателям синергии потенциально даст больший вес комбинациям лекарств, которые

имеют больше экспериментов. Поэтому для придания равных весов всем комбинациям препаратов в качестве метрики была использована средневзвешенная корреляция Пирсона:

$$p_w = \frac{\sum_{i=1}^N \sqrt{n_i - 1} * p_i}{\sum_{i=1}^N \sqrt{n_i - 1}} \quad (3.17)$$

где n_i - число клеточных линий, обработанных данной комбинацией, $N = 167$ (число комбинаций, используемых для теста), p_i - коэффициент корреляции Пирсона, рассчитанный для предсказанных и наблюдаемых уровней синергии в пределах одной комбинации препаратов по возможным клеточным линиям [6].

Для оценки методов по второму направлению использовали многофакторный дисперсионный анализ (ANOVA).

Команды использовали множество различных подходов к предсказанию синергии лекарств, включая регрессию, деревья решений, случайные леса, Гауссовские процессы, SVM, нейронные сети. Однако производительность методов имела слабую связь с выбранным алгоритмом. Самым лучшим методом был Yuanfang Guan с метриками $p_{wSC1A} = 0,48$, $p_{wSC1B} = 0,45$ и $ANOVA_{SC2} = 74.89$ для обеих категорий первого направления соревнования и второго направления, соответственно. Основываясь на метрике для второго соревнования, метод Y Guan показал значительно лучшие результаты по сравнению с другими командами.

Также для второго направления конкурса был проверен ансамблевый метод, основанный на агрегировании всех представленных моделей. Этим подходом добились скромного улучшения производительности по сравнению с лучшим методом. Это явление называется “мудрость толпы” [8].

Метод Yuanfang Guan

Более подробно разберем самую лучшую модель Yuanfang Guan [6]. Команда разработала три модели:

- модель глобальной синергии (GSM) - использует один обучающий набор и делает прогнозы для всех тестовых образцов сразу
- модель локальной синергии (LSM) - строит обучающий набор для каждого неизвестного показателя синергии в тестовом наборе данных и делает прогнозы отдельно. Обучающий набор данных LSM представляет собой подмножество пар

лекарств в GSM, включая только те пары, когда появляется любой из препаратов в тестируемой паре лекарств.

- модель единого лекарственного средства (SDM) - SDM аналогичен LSM, за исключением того, что обучающий набор данных генерируется для каждого препарата, а не для комбинации препаратов

Каждая из моделей была разработана с использованием алгоритма случайного леса.

Для второго направления соревнований они использовали ROC-AUC, чтобы оценить эффективность бинарных предсказаний. Обозначили наблюдаемые синергетические оценки > 20 как "1" (наблюдаемая синергия) и оценки < 20 как "0" (наблюдаемая несинергия). Этот порог использовался DREAM. Но, как было упомянуто ранее, для бинарных предсказаний консорциум DREAM использовал многофакторный дисперсионный анализ (ANOVA) для оценки предсказаний.

Метод Ranking-system of Anti-Cancer Synergy

Дополнительно рассмотрим модель, которую назвали Ranking-system of Anti-Cancer Synergy (RACS) [12]. Эту модель обучена с частичным привлечением учителя. То есть используется датасет с малым числом комбинаций, которые известны как синергетические. Для создания обучающего набора данных использовали 26 синергетических пар противоопухолевых препаратов из базы данных Drug Combination Database (DCDB) в качестве меченых. В исследовании было охвачено 33 протестированных препарата с лекарственными мишениями и транскриптомными профилями конкретных клеточных линий после однократной обработки одиночным препаратом. На основе комбинирования 33 препаратов составили 502 пары немаркированных образцов (во всем обучающем датасете 528 пар).

На основе 13 соединений, предоставленных NCI-DREAM, было составлено 78 попарных комбинаций, которые были использованы в качестве тестового набора данных для клеточной линии -клеточной лимфомы человека OCI-LY3.

Также был создан еще один набор данных для валидации метода на основе 142 противоопухолевых препаратов, одобренных FDA или проходящих клинические испытания (из баз данных DrugBank, Therapeutic Target Database и PubMed). После удаления соединений без аннотаций gene ontology (GO) или информации о сигнальных путях базы данных KEGG осталось 118 препаратов. При их комбинирование

было создано 6 877 немеченых пар в качестве тестовых данных для клеточной линии A549 и MCF7.

Изначально было выбрано четырнадцать признаков, охватывающих химическую структуру, фармакологию, функциональные и сетевые свойства лекарственных мишеней, но только семь признаков были идентифицированы как существенно отличающиеся между синергетическими и немеченными парами.

Для предварительного ранжирования был применен полууправляемый метод обучения, включающий многообразный алгоритм ранжирования на основе сходства с меченными парами в пространстве 7 признаков.

Аналогичным образом были протестированы пять параметров, описывающих корреляции между дифференциально экспрессируемыми генами, и два параметра были значительно различны между синергетическими и немеченными образцами. Эти два параметра были использованы в качестве дополнительных транскриптомных фильтров для улучшения предварительного ранжирования. Первый из них DEG_Overlap рассчитывается как:

$$DEG_Overlap = \frac{|A \cap B|}{\sqrt{|A| \times |B|}} \quad (3.18)$$

где А и В представляют собой множества дифференциально экспрессированных генов, возмущенные препаратом x и у соответственно.

Второй параметр назывался Pathway_Coverage и вычислялся по формуле:

$$Pathway_Coverage = \frac{|(A \cup B) \cap N|}{|N|} \quad (3.19)$$

где А и В представляют собой множества дифференциально экспрессированных генов, возмущенные препаратом x и у соответственно, N множество всех генов, связанных с сигнальными путями рака.

Ниже приведена схема метода RACS с предварительным ранжированием и транскриптомными фильтрами.

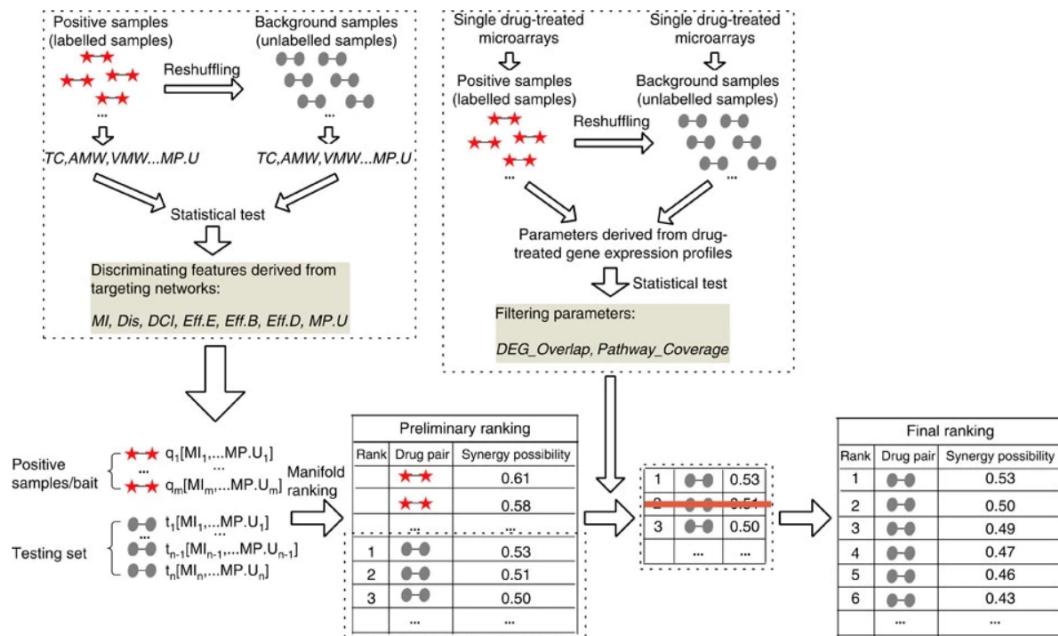


Рис. 3.11: схема RACS

Так как оценка метода была проведена на датасете консорциума DREAM, то метод сравнили с моделями соревнования, о которых было упомянуто выше (среди них выделялись такие методы, как DIGRE, SynGen). В качестве метрик брали ROC-AUC, TPR, PC-index.

У метода RACS PC-индекс равен 0,78, ROC-AUC 0,85, то есть этот метод показал наилучшие результаты. Именно применение транскриптомных фильтров увеличило PC-индекс с 0,69 до 0,78, а значения ROC-AUC с 0,783 до 0,853.

Также метод был проверен на наборе данных для клеточных линий A549 и MCF7, но с применением других метрик.

Стоит отметить, что PC-индекс, предложенный консорциумом DREAM, более строгий по сравнению с ROC-AUC. PC-index устойчив к возмущениям данных. Например, значение ROC-AUC метода DIGER снизилось с 65 до 48% после удаления митомицина С из набора из 14 препаратов. Однако PC-индекс метода DIGER оставался почти неизменным. Также он учитывает шум биологических копий. Однако PC-индекс нельзя было использовать для рака молочной железы или легких, потому что вместо биологических копий использовались множественные комбинации различных концентраций.

Методы с использованием других датасетов

DeepSynergy

Рассмотрим модель DeepSynergy[10], которая является методом глубокого обучения. Модель была разработана для задачи регрессии. DeepSynergy была обучена на наборе данных, который включает 23,062 образца, где каждый образец относится к двум соединениям и клеточной линии. Он охватывает 583 различные комбинации, каждая из которых была протестирована против 39 клеточных линий рака, полученных из 7 различных типов тканей. Пары препаратов были построены из 38 противоопухолевых препаратов (14 экспериментальных и 24 одобренных). Среди 38 соединений данные можно поделить на 2 набора:

- "исчерпывающий состоящий из 22 препаратов, у которых все возможные комбинации были протестированы
- "дополнительный состоящий из 16 препаратов, они были протестированы только в комбинации с препаратами из "исчерпывающего" набора

Для каждого образца была измерена скорость роста клеток при обработке комбинацией препаратов в режиме дозирования 4×4 для 4 биологических копий. Были получены данные монотерапии, то есть зависимость скорости роста клеток при обработке отдельным препаратом для 8 концентраций и 6 биологических копий. На основе этих данных составлялась матрица размером 5×5 , для которой первая строка и первый столбец являются данными монотерапии. Такие матрицы назывались комбинационной поверхностью.

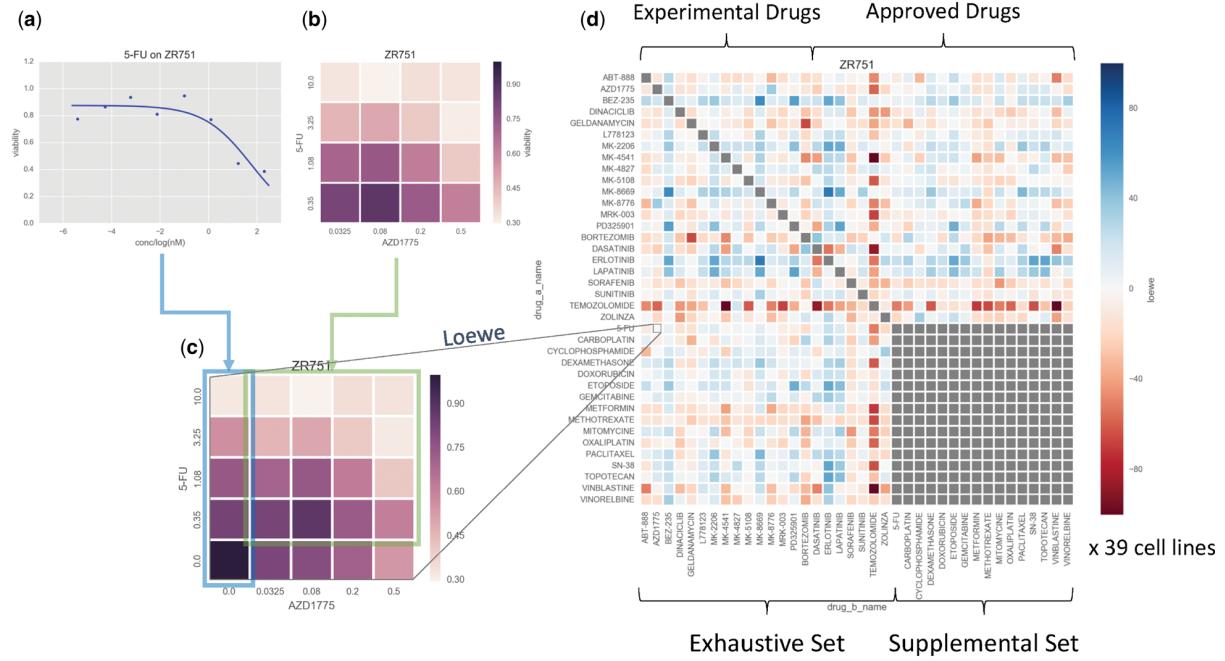


Рис. 3.12: Структура датасета

Уровень синергии комбинации препаратов вычислялся по комбинационным поверхностям с использованием теоретической модели Loewe Additivity [16].

Эффект от препаратов A и B с концентрациями a и b соответственно может быть представлен в виде следующей суммы :

$$E(a,b) = R(a,b) + S(a,b), \quad (3.20)$$

где E(a,b) - наблюдаемый в эксперименте эффект, R(a,b) - аддитивный эффект, определяемый по модели Loewe Additivity, S(a,b) - это величина дополнительного эффекта, он положительный в случае синергетического эффекта и отрицательным в случае антагонистического эффекта пары препаратов. Аддитивный эффект для комбинации препаратов (a, b) вычисляется путем нахождения двух доз a_u и b_u таких, что:

$$E(a_u) = E(b_u) \quad (3.21)$$

Уравнение изоболы (кривой одинакового эффекта):

$$\frac{a}{a_u} + \frac{b}{b_u} = 1 \quad (3.22)$$

Найдя численное решение этих 2 уравнений, считают аддитивный эффект:

$$R_{Loewe}(a,b) = E(a_u) = E(b_u) \quad (3.23)$$

Зная аддитивный и наблюдаемый эффекты, можно посчитать синергетический. Значения синергии варьируются от -326 до 179.

Использовали как химическую информацию о препаратах, так и геномную информацию, отражающую биологию болезни. Авторы вычислили три различных типа химических признаков: extended connectivity fingerprints с радиусом 6, получены с использованием jCompoundMapper, физико-химические свойства с использованием ChemoPy, бинарные признаки основанные на наборе токсикофоров, собранных из литературы. Токсикофоры-это подструктуры, которые токсичны. Пространство химических признаков было уменьшено фильтрацией признаков с нулевой дисперсией. Финальный набор признаков содержал 1309 ECFP_6, 802 физико-химических и 2276 признаков токсикофоров. Клеточные линии были описаны профилем экспрессии генов. Конечный набор был из 3984 геномных признаков.

Модель DeepSynergy - это нейронная сеть с обратной связью, которая принимает входные векторы, представляющие образцы, и выдает одно значение - уровень синергии. Образцы описываются сцепленными векторами, которые включают в себя признаки двух препаратов и одной клеточной линии. То есть нейроны входного слоя получают значения экспрессии генов клеточной линии и химические дескрипторы обоих препаратов в качестве входных данных. Затем информация распространяется по слоям сети DeepSynergy до тех пор, пока выходной блок не выдаст прогнозируемый показатель синергии.

Поскольку сеть не должна различать комбинацию лекарств АВ, представленную в порядке А-В или В-А, авторы удваивают измерения, представляя каждый образец дважды в обучающем наборе. Один раз свойства препарата используются в порядке А-В и один раз в порядке В-А. Для прогнозирования оба способа представления выборки распространяются по сети и усредняются. Наблюдалось, что DeepSynergy учится предсказывать одно и то же значение для комбинации лекарств АВ в порядке А-В и В-А.

Были рассмотрены различные настройки гиперпараметров, а именно разные стратегии нормализации данных, в сочетании с коническими или прямоугольными слоями с различным количеством нейронов. Использовалось два или три скрытых слоя. Кроме того, исследовались различные скорости обучения, а также методы регуляризации. В итоге DeepSynergy имеет коническую архитектуру с двумя скрытыми

слоями, имеющими 8192 нейрона в первом и 4096 во втором скрытом слое.

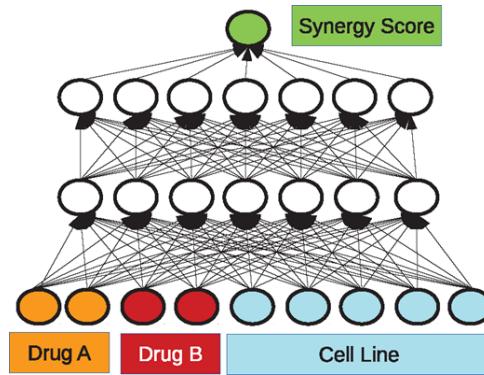


Рис. 3.13: Архитектура DeepSynergy

В качестве основной метрики оценки модели использовалась среднеквадратичная ошибка (MSE), по которой модель оптимизировалась во время обучения. Также использовали корень из среднеквадратичной ошибки (RMSE) и коэффициент корреляции Пирсона. DeepSynergy достигла тестового MSE: 255, RMSE: 15.91, коэффициент корреляции Пирсона: 0.73. Дополнительно DeepSynergy сравнили со следующими моделями: Median Polish, которая использовалась в качестве базовой случайной модели, Gradient Boosting, Random Forests, Support Vector Machines и Elastic Nets. Ниже приведены результаты сравнения на основе вышеперечисленных метрик.

Method	MSE	Confidence Interval	P-value	RMSE	Pearson's <i>r</i>
Deep Neural Networks	255.49	[239.93, 271.06]		15.91 ± 1.56	0.73 ± 0.04
Gradient Boosting Machines	275.39	[258.24, 292.54]	9.6×10^{-17}	16.54 ± 1.37	0.69 ± 0.02
Random Forests	307.56	[286.83, 328.29]	1.2×10^{-73}	17.49 ± 1.63	0.65 ± 0.03
Support Vector Machines	398.39	[371.22, 425.56]	$<10^{-280}$	19.92 ± 1.28	0.50 ± 0.03
Elastic Nets	420.24	[393.11, 447.38]	$<10^{-280}$	20.46 ± 1.29	0.44 ± 0.03
Baseline (Median Polish)	477.77	[448.68, 506.85]	$<10^{-280}$	21.80 ± 1.49	0.43 ± 0.02

Рис. 3.14: Сравнение моделей

DeepSynergy показала MSE, равный 255, в то время как Gradient Boosting, Random Forests достигли 275 и 308 соответственно. Support Vector Machines и Elastic Nets

показали аналогичные результаты с MSE, равным 398 и 420 соответственно. Median Polish, которая использовалась в качестве базовой, достигла наихудшего результата с MSE 478. Относительное улучшение DeepSynergy по сравнению с базовой моделью составляет 53%. DeepSynergy значительно превосходит другие методы машинного обучения.

Так как RMSE, MSE зависят от набора данных, их трудно использовать для сравнения между различными наборами данных. Поэтому для получения сопоставимых показателей прогностических характеристик DeepSynergy использовались метрики, типичные для задач классификации: ROC AUC, PR AUC, точность (ACC), точность (PREC), чувствительность (TPR), специфичность (TNR). У DeepSynergy ROC AUC равен 0,9.

Метод Zhao H.

Для набора данных использовали 184 пары препаратов (на основе 238 препаратов), одобренные FDA orange book. Была собрана молекулярная и фармакологическая информация, связанная с этими препаратами, включая их мишени и соответствующие нисходящие сигнальные пути, области медицинских показаний, терапевтические эффекты, представленные в анатомической терапевтической и химической классификационной системе (ATX), а также побочные эффекты [9].

Для аннотаций лекарственных мишеней использовали данные о взаимодействии соединений и белков из базы данных STITCH, DrugBank и базы данных терапевтических мишеней TTD. Далее исследовали пути, на которые возможно воздействует препарат через мишень, информация о сигнальных путях была получена из базы данных KEGG. Каждый препарат был связан с путями, в которых состоят его мишени. Пара препаратов может быть представлена в виде вектора, состоящего из пар признаков. Например, в случае признака "мишень" препарат 1 связывает два белка p1, p2, препарат 2 связывает три белка p3, p4, p5, комбинация препарата 1 и препарата 2 может быть представлена в виде следующих пар признаков: (p1, p3), (p1, p4), (p1, p5), (p2, p3), (p2, p4), (p2, p5), аналогично для других признаков.

Затем был проведен поиск тех пар признаков, которые наиболее часто встречаются в синергетических парах. Для пары препаратов (d_i, d_j) признака F (например, мишень), f_i связан с d_i и f_j с d_j , где f_i и $f_j \subset F$. Пара препаратов (d_i, d_j) может быть

представлена парой признаков (f_i, f_j) . Для каждой пары признаков рассчитывается оценка обогащения s_{ij} следующим образом:

$$s_{ij} = \frac{N_{ij}}{N'_{ij}} \quad (3.24)$$

где N_{ij} - сколько раз пара признаков встречается в синергетических комбинациях лекарств, N'_{ij} - сколько раз пара признаков встречается в фоновом наборе всех возможных попарных комбинаций между лекарственными средствами, участвующих в синергетических комбинациях лекарств. Таким образом, все пары признаков могут быть ранжированы по s_{ij} , то есть по обогащению в синергетических парах.

Затем для выбранного признака F обучали модель предсказывать оценку обогащения для пары признаков комбинации препаратов. Для этого использовали кросс-валидация. В ходе кросс-валидации все комбинации препаратов были случайным образом разделены на пять групп одинакового размера без перекрытия, четыре из которых использовались в качестве обучающего набора и использовались для расчета оценки обогащения для каждой пары признаков, в то время как оставшаяся группа использовалась в качестве валидационного набора, и процедура повторялась в течение пяти раз. Оценка F1 была принята в качестве показателя производительности модели. Порог, выше которого был достигнут самый высокий балл F1 при перекрестной валидации, использовался для предсказания эффективности. То есть пара препаратов считалась синергетической, если оценка обогащения ее пары признаков была выше порогового значения. Таким образом рассмотренная задача относилась к классификации и для сравнение важности признаков использовали метрику ROC-AUC. По ней было отобрано наиболее три важных признака: область терапии, мишень и область показаний. Затем, интегрируя эти признаки, аналогичным образом предсказываем синергетические пары [9].

Среди рассмотренных методов наибольшего результата достигли подходы, для которых был доступен обучающий набор данных. Методы, которые могли использовать только молекулярные данные клеточных линий, фармакологические признаки, данные монотерапии и гипотезы механизма синергии, показывали результаты немного выше случайных моделей [5], что говорит о плохом понимании механизмов, лежащих в основе синергии.

Рассмотрим еще один метод предсказания синергетических пар малых молекул, используемый в подходе *L1000CDS*² [17]. В основе работы этого инструмента лежит концепция Connectivity Map [18]. Рассмотрим 2 клеточных состояния, например здоровое и больное, либо 2 типа клеток. На основе данных RNA-seq для этих 2 состояний можно найти сигнатуру, состоящую из генов с повышенной и пониженной экспрессией. *L1000CDS*² производит поиск малых молекул, которые имитируют или обращают сигнатуру, подданную в качестве входных данных. Этот поиск осуществляется на основе сравнения поданной сигнатуры с сигнатурами клеточных линий, химически возмущенных, из базы данных LINCS-1000. Сигнатуры рассчитываются из данных экспрессии с помощью их собственного метода Characteristic Direction (CD). Для приоритизации малых молекул *L1000CDS*² вычисляет косинусное расстояние между вектором входной сигнатуры и сигнатурами в LINCS-L1000.

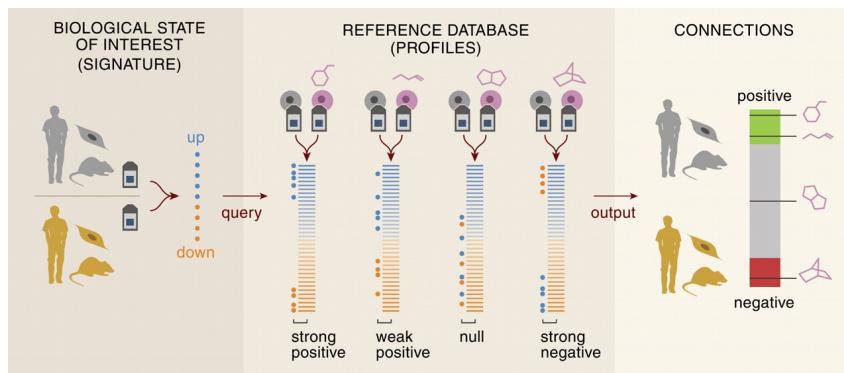


Рис. 3.15: Концепция Connectivity Map[18]

При поиске синергетических комбинаций *L1000CDS*² сравнивает каждую возможную пару из 50 лучших подобранных сигнатур и вычисляет уровень синергии каждой пары, как ортогональность между двумя сигнатурами. Идея состоит в том, что если два возмущения ортогональны, то они воздействуют через два независимых пути и вероятно несут комплементарный синергетический эффект.

3.3 Анализ обогащения сигнальных путей

Повышению доступности технологии секвенирования привело к накоплению большого количества данных. Однако их интерпретация в контексте биологических процессов довольно трудная задача.

Геномное секвенирование показало, что большая часть генов, определяющих основ-

ные биологические функции, является общей для всех эукариот. Знания о биологической роли таких общих белков в одном организме часто могут быть перенесены на другие организмы [19]. Для организации накопленных знаний о функциях групп генов используются библиотеки наборов генов. Каждая такая библиотека состоит из набора списков связанных генов, где каждый набор генов связан с определенным функциональным термином, таким как название сигнального пути или фактора транскрипции [20].

При анализе дифференциальной экспрессии определяются статистически значимые гены с повышенной или пониженной экспрессией (также можно говорить об этих генах в терминах сильной корреляции или антикорреляции с одним из фенотипов). Однако в большинстве случаев представляют интерес затронутые ими биологические процессы. Используя библиотеки наборов генов, можно определить, из каких наборов гены наиболее представлены среди дифференциально экспрессированных генов. Функциональные термины, соответствующие этим наборам генов, говорят о процессах, чья активность претерпела изменения.

Самый первый метод, который разработал такой подход, назывался анализом обогащения набора генов (GSEA). Принимая список генов, ранжированных по корреляции их профилей экспрессии с одним из двух фенотипов, GSEA стремится оценить значимость избыточного представления независимо определенного набора генов, S , в сильно коррелированных или антикоррелированных генах списка. Для оценки степени обогащения, метод GSEA вычисляет величину значения обогащения, проходя по списку, увеличивая совокупную сумму, когда ген находится в S , и уменьшая ее, если ген не принадлежит списку S . Размер приращения зависит от корреляции генов и фенотипов. Значение обогащения (ES) определяется как максимальное отклонение от нуля совокупной суммы [21].

Было разработано много других инструментов анализа обогащения набора генов с сохранением этой первоначальной концепции. Также стоит упомянуть о самих ресурсах, содержащих ранее упомянутые наборы генов. К самым известным относится Gene Ontology, KEGG.

4 Материалы и методы

Данный метод разработан для задачи поиска малых молекул, способных вызвать нужные изменения в фенотипе клетки. Поскольку для достижения больших изменений не достаточно одного соединения, в этой задаче актуальна проблема поиска синергетических комбинаций химических соединений. Наш подход позволяет оценить уровень синергии каждой пары соединений и, соответственно, приоритезировать список пар малых молекул по уровню синергии. В его основе лежит работа с экспрессионными сигнатурами и генными сетями.

Экспрессионные сигнатуры позволяют охарактеризовать разницу между двумя состояниями клеток. Мы используем базу данных L1000FWD [22], которая содержит 42809 сигнатур, возмущенных малыми молекулами на разных клеточных линиях. Для каждой малой молекулы из рассматриваемого набора мы находим индуцированные ею сигнатуры в базе L1000FWD. Таким образом, для конкретного набора малых молекул создается набор сигнатур из L1000FWD. Чтобы найти пары малых молекул, которые вызовут желаемые изменения в экспрессии генов, мы будем сравнивать соответствующие пары сигнатур с желаемой сигнатурой. На основе этого сравнения будет рассчитан уровень синергии малых молекул.

Метод можно использовать в 2 режимах:

- "прямой" - позволяет искать синергетические комбинации малых молекул для индуцирования нужных изменений в фенотипе клетки
- "обратный" - позволяет искать синергетические комбинации малых молекул для обращения изменений в фенотипе клетки

Рассмотрим метод подробнее. В нем можно выделить 2 основных этапа:

- создание экспрессионной сигнатуры запроса
- вычисление уровня синергии пары малых молекул

4.1 Создание экспрессионной сигнатуры запроса

Для описания желаемых изменений в фенотипе клетки используется экспрессионная сигнатура. На основе данных RNA-seq для начального и конечного состояния клетки

проводится анализ дифференциальной экспрессии с использованием пакета edgeR, затем отбираются гены по значениям logFC и p-уровню значимости. Полученная генная сигнатура далее будет использована для сравнения. При сравнении сигнатур мы хотим учитывать биологическую значимость генов в данном контексте, то есть нас интересуют сигнатуры, которые схожи по наиболее значимым генам. Таким образом, одна из основных целей этой работы - создать метод, который оценивает биологическую значимость гена в данном контексте, и при пересечении сигнатур гены будут иметь вес в соответствии с их биологической значимостью.

Чтобы определить важность генов в сигнатуре запроса, мы строим 2 генные сети, одна для генов с повышенной экспрессией, вторая для генов с пониженной экспрессией. При построении сети информация о взаимодействиях берется из базы данных STRING [23], которая включает в себя как известные, так и предсказанные взаимодействия. Эти взаимодействия включают в себя прямые и косвенные связи; они создаются на основе вычислительного прогнозирования, переноса знаний между организмами, взаимодействий, агрегированных из других (первичных) баз данных. Пять основных источников базы данных STRING: предсказания в контексте генома, высокопроизводительные лабораторные эксперименты, коэкспрессия, автоматизированный интеллектуальный анализ текстов, прежние сведения в базах данных. К сожалению, API запрос в базу данных STRING позволяет узнать о взаимодействиях генов между собой для наборов, содержащих не более 2000 генов, поэтому сети были ограничены 2000 узлами. Для большинства рассмотренных сигнатур количество генов с повышенной или пониженной экспрессией в сигнатуре больше 2000. Поэтому производился отбор 2000 генов с большими значениями модуля logFC, про которых есть сведения в базе STRING.

Поскольку была показана связь между биологической значимостью белка и топологическими метриками, такими как betweenness и closeness [24], [25], то в данном подходе уровень значимости рассчитывается по метрикам центральности, среди которых есть betweenness (Btw) и closeness (Cln).

Betweenness (Btw) для вершины v определяется следующим образом:

$$Btw(v) = \sum_{\substack{s \neq v \neq t \in V \\ s \neq t}} \frac{\sigma_{st}(v)}{\sigma_{st}}, \quad (4.1)$$

где σ_{st} - количество кратчайших путей от вершины s до вершины t , $\sigma_{st}(v)$ - количество кратчайших путей от вершины s до вершины t через вершину v .

Closeness (Cln) для вершины i определяется следующим образом:

$$Cln_i = \frac{1}{\sum_j d_{ij}}, \quad (4.2)$$

где d_{ij} - расстояние от вершины i до j . Если нет пути между этими вершинами, то расстояние считается равным 0.

Также при расчете уровень значимости гена были включены другие топологические метрики: pagerank (Prk), eigenvector centrality (Egv), Katz centrality (Ktz), eigentrust (Egt).

Pagerank (Prk) для вершины v определяется следующим образом:

$$Prk(v) = \frac{1-d}{N} + d \sum_{u \in \Gamma^-(v)} Prk(u) d^+(u), \quad (4.3)$$

где $\Gamma^-(v)$ - набор соседних вершин, которые ссылаются на вершину v , $d^+(u)$ - число выходящих ребер из вершины u , d - коэффициент затухания.

Eigenvector centrality (Egv) определяется как собственный вектор \mathbf{X} матрицы смежности для наибольшего собственного значения, то есть является решением:

$$\mathbf{Ax} = \lambda \mathbf{x}, \quad (4.4)$$

где \mathbf{A} - матрица смежности, λ - наибольшее собственное значение.

Katz centrality (Ktz) определяется как решение неоднородной линейной системы:

$$\mathbf{x} = \alpha \mathbf{Ax} + \mathbf{1}, \quad (4.5)$$

где \mathbf{A} - матрица смежности, α - коэффициент ослабления.

Eigentrust (Egt) определяется следующим образом:

$$Egt = \lim_{n \rightarrow \infty} (\mathbf{C}^T)^n \mathbf{c}, \quad (4.6)$$

где $c_i = \frac{1}{|V|}$, элементы матрицы \mathbf{C} представляют собой нормализованные значения доверия:

$$c_{ij} = \frac{\max(s_{ij}, 0)}{\sum_j \max(s_{ij}, 0)} \quad (4.7)$$

В качестве s_{ij} использовались значения достоверности предсказания взаимодействия генов из базы данных STRING.

Все выше указанные метрики реализованы в пакете Graph-tool. Все топологические метрики и значения logFC, взятые по модулю, были нормализованы так, чтобы их значения были распределены от 0 до 1.

Уровень биологической значимости гена на основе топологических метрик центральности и значения модуля logFC вычислялся следующим образом:

$$\text{inf_score} = (a_1 \cdot |\log FC| + 1) \cdot (a_2 \cdot \text{Prk} + 1) \cdot (a_3 \cdot \text{Btw} + 1) \\ \cdot (a_4 \cdot \text{Egv} + 1) \cdot (a_5 \cdot \text{Cln} + 1) \cdot (a_6 \cdot \text{Ktz} + 1) \cdot (a_7 \cdot \text{Egt} + 1),$$

где $a_1, a_2 \dots a_7$ - числовые коэффициенты. Они были подобраны на основании масштабной валидации на базе данных CFM [26], которая будет подробно освещена дальше.

Если для гена не были рассчитаны топологические метрики, то уровень значимости рассчитывается только по значению logFC:

$$\text{inf_score} = (a_1 \cdot \log FC + 1) \quad (4.8)$$

Ниже представлена схема основных этапов создания сигнатуры запроса.

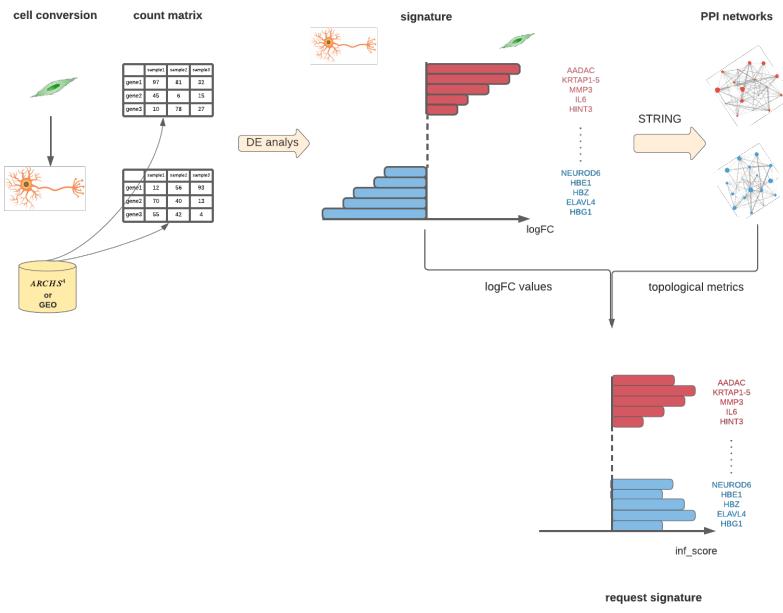


Рис. 4.1: создание экспрессионной сигнатуры запроса

4.2 Вычисление уровня синергии

Для каждой пары сигнатур из вычисляется уровень синергии. Он считается 2 способами:

- Вычисление уровня синергии на основе сравнения сигнатур
- Вычисление уровня синергии на основе сравнения обогащенных сигнальных путей

4.2.1 Вычисление уровня синергии на основе сравнения сигнатур

Каждая сигнатура из базы данных L1000FWD представлена списком генов с повышенной экспрессией и списком генов с пониженной экспрессией. Нашей целью являются пары малых молекул, в которых обе малые молекулы, действуя в синергии, вызовут желаемые изменения в экспрессии генов. Поэтому сначала мы хотим предсказать сигнатуру, которую бы индуцировала пара молекул. Эта предполагаемая экспрессионная сигнатура строится по соответствующей паре сигнатур из L1000FWD, в которых каждая возмущена одной малой молекулой из пары. У этих 2 сигнатур мы объединяем списки генов с повышенной экспрессией и списки генов с пониженной экспрессией. Если при таком объединении ген окажется и в числе с повышенной экспрессией для одной сигнатуры и в числе с пониженной экспрессией для другой сигнатуры, то он удаляется из сигнатуры, полученной объединением. Такую сигнатуру мы будем называть "сигнатурой пары".

Далее мы сравниваем сигнатуру пары с сигнатурой запроса. Как сигнатуру пары, так и сигнатуру запроса мы представляем в виде 2 генных векторов: для генов с повышенной или пониженной экспрессией.

В случае "прямого" режима мы создаем генное пространство для генов с повышенной экспрессией из сигнатуры пары и сигнатуры запроса. В нем определяем вектор генов с повышенной экспрессией сигнатуры пары и вектор генов с повышенной экспрессией сигнатуры запроса. Считаем взвешенное косинусное расстояние между этими векторами. Затем аналогично рассчитываем взвешенное косинусное расстояние между векторами для генов с пониженной экспрессией сигнатуры пары и сигнатуры запроса. Уровень синергии определяется как: 1 - среднее значение этих рассчитанных косинусных расстояний. Вычитание из 1 необходимо для того, чтобы чем выше синергетический эффект, тем больше был уровень синергии.

В случае "обратного" режима мы создаем генное пространство для генов с повышенной экспрессией из сигнатуры пары и генов с пониженной экспрессией сигнатуры запроса. В нем определяем вектор генов с повышенной экспрессией сигнатуры па-

ры и вектор генов с пониженной экспрессией сигнатуры запроса. Считаем взвешенное косинусное расстояние между этими векторами. Затем аналогично рассчитываем взвешенное косинусное расстояние между векторами для генов генов с пониженной экспрессией сигнатуры пары и генов с повышенной экспрессией сигнатуры запроса. Уровень синергии определяется как: 1 - среднее значение этих рассчитанных косинусных расстояний.

В обоих случаях мы считаем взвешенное косинусное расстояние. Вес для гена определяется следующим образом. Если сигнатура запроса содержала этот ген, то для него был рассчитано значение `inf_score`, показывающее биологическую значимость гена. В качестве веса берется значение `inf_score`. Если сигнатура запроса не содержала этот ген, то в качестве веса берется 1.

На рисунке 4.2 представлена схема вычисления уровня синергии на основе сравнения сигнатур в "обратном" режиме, поскольку он более сложный для понимания.

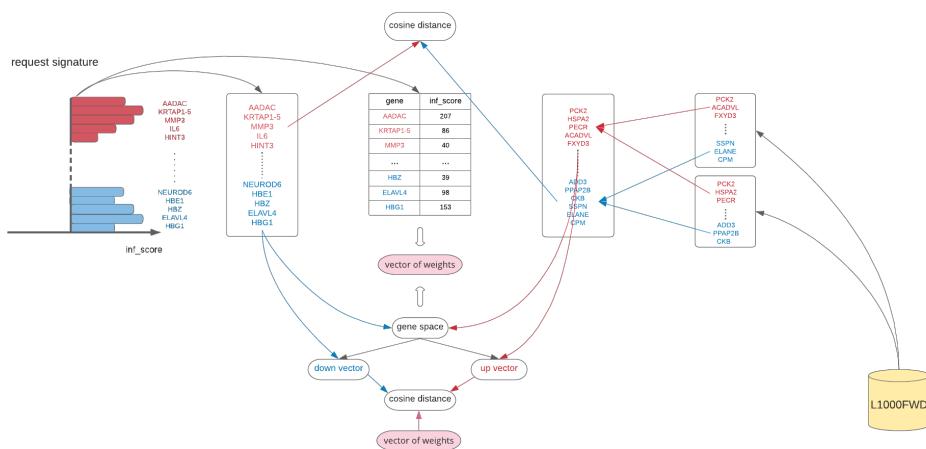


Рис. 4.2: схема вычисления уровня синергии на основе сравнения сигнатур в "обратном" режиме

Для обоих режимов чем выше уровень синергии, тем больше синергетический эффект.

4.2.2 Вычисление уровня синергии на основе сравнения обогащенных сигнальных путей

В этом подходе вместо каждой сигнатуры использовались сигнальные пути, полученные в результате анализа обогащения генов этой сигнатуры с повышенной / понижен-

ной экспрессией. Таким образом каждая сигнатура была представлена 2 списками сигнальных путей: активированные и подавленные. Анализ обогащения проводился с помощью инструмента Enrich [27]. В данном методе были опробованы 2 способа вычисления уровня синергии на основе сравнения сигнальных путей:

- оценка схожести с помощью вычисления коэффициента Танимото
- оценка схожести с помощью вычисления взаимной информации

Вычисление уровня синергии с помощью коэффициента Танимото

Аналогично предыдущему методу вычисления уровня синергии, для каждой пары сигнатур из базы данных объединялись списки активированных/подавленных сигнальных путей. Таким образом каждая пара сигнатур характеризовалась 2 списками сигнальных путей: активированными и подавленными.

На следующем этапе оценивалось сходство списков сигнальных путей сигнатуры запроса и сигнатуры пары вычислением коэффициентом Танимото. Коэффициент Танимото измеряет степень схожести двух множеств и вычисляется по следующей формуле:

$$Tc = \frac{bc}{b1 + b2 - bc}, \quad (4.9)$$

где $b1$ - число элементов первого множества, $b2$ - число элементов второго множества, bc - количество элементов в пересечении множеств.

В "прямом" режиме вычисляется коэффициент Танимото для множества активированных сигнальных путей сигнатуры запроса и множества активированных сигнальных путей сигнатуры пары. Также аналогично вычисляется коэффициент Танимото для множества подавленных сигнальных путей сигнатуры запроса и множества подавленных сигнальных путей сигнатуры пары. В качестве оценки уровня синергии берется среднее значение вычисленных коэффициентов Танимото.

В "обратном" режиме вычисляется коэффициент Танимото для множества активированных сигнальных путей сигнатуры запроса и множества подавленных сигнальных путей сигнатуры пары. Также аналогично вычисляется коэффициент Танимото для множества подавленных сигнальных путей сигнатуры запроса и множества активированных сигнальных путей сигнатуры пары. В качестве оценки уровня синергии берется среднее значение вычисленных коэффициентов Танимото.

На рисунке 4.3 представлена схема вычисления уровня синергии в "обратном" режиме.

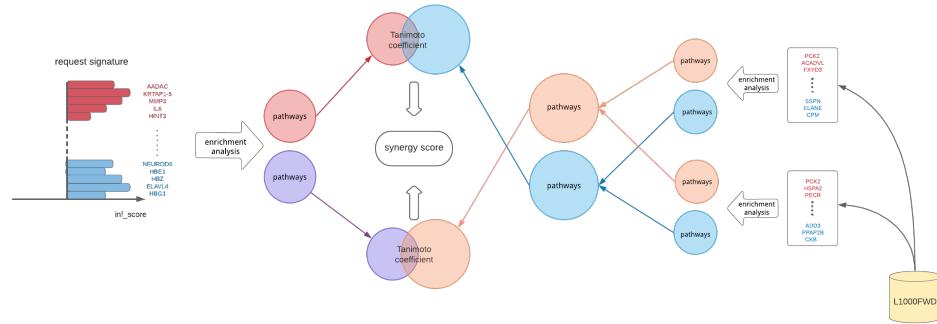


Рис. 4.3: Схема вычисления уровня синергии при использовании коэффициента Танимото в "обратном" режиме

Вычисление уровня синергии с помощью взаимной информации

Взаимная информация также позволяет измерить степень схожести 2 множеств.

В случае "прямого" режима сначала для каждой сигнатуры из пары находится пересечение ей соответствующего набора активированных сигнальных путей с набором активированных сигнальных путей сигнатуры запроса. Далее для этих 2 полученных наборов сигнальных путей считается взаимная информация. По аналогии находится пересечение подавленных сигнальных путей сигнатуры из пары и подавленных сигнальных путей сигнатуры запроса. Для этих 2 наборов сигнальных путей также считается взаимная информация. Уровень синергии считается как среднее значение рассчитанных величин взаимной информации.

В случае "обратного" режима сначала для каждой сигнатуры из пары находится пересечение ей соответствующего набора активированных сигнальных путей с набором подавленных сигнальных путей сигнатуры запроса. Далее для этих 2 полученных наборов сигнальных путей считается взаимная информация. По аналогии находится пересечение подавленных сигнальных путей сигнатуры из пары и активированных сигнальных путей сигнатуры запроса. Для этих 2 наборов сигнальных путей также считается взаимная информация. Уровень синергии считается как среднее значение рассчитанных величин взаимной информации.

На рисунке 4.4 представлена схема вычисления уровня синергии в "обратном" режиме.

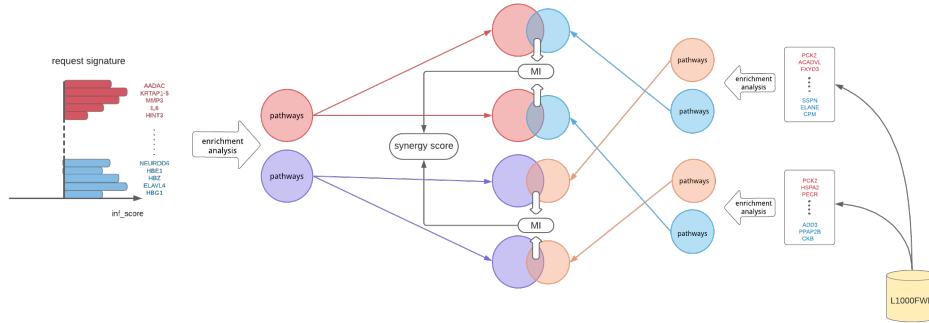


Рис. 4.4: Схема вычисления уровня синергии при использовании взаимной информации в "обратном" режиме

4.3 Валидация

Данный метод разрабатывался в применении к задаче химического перепрограммирования. Для валидации использовалась база данных CFM, которая содержит протоколы химического перепрограммирования клеток. В каждом протоколе приведена синергетическая комбинация малых молекул с указанием исходной и конечной клеточной линии. Также для веществ были дополнительно указаны идентификационные номера (compound identification number, cid) в базе данных PubChem.

Для всех малых молекул из базы CFM по их cid были найдены SMILES в базе данных PubChem. Среди соединений базы L1000FWD есть вещества, которые по химическому строению схожи с малыми молекулами из CFM и, предположительно, обладают схожей активностью. Поэтому при пересечении соединений базы CFM и L1000FWD для последующей валидации было необходимо учесть такие соединения. Для этого для каждого вещества из базы CFM были найдены схожие по строению вещества из базы L1000FWD. Структурное сходство двух молекул оценивалось путем вычисления коэффициента Танимoto. В качестве молекулярных отпечатков использовались Morgan Fingerprints. Все пары соединений с коэффициентом Танимoto больше 0.9 считались схожими. При анализе химического сходства использовалась библиотека RDKit [28].

Для набора малых молекул в каждом протоколе были найдены сигнатуры из L1000FWD, возмущенные этими соединениями. Любую пару малых молекул из одного протокола мы считаем синергетической. Все пары, в которых малые молекулы принадлежат разным протоколам, мы считаем несинергетическими.

Так как среди рассматриваемых соединений должны быть представлены как синергетические, так и несинергетические пары, мы работали с клеточными переходами, для которых в базе CFM было больше 2 протоколов, и в каждом протоколе было больше 1 молекулы. При работе с конкретным переходом сначала рассчитывалась сигнатура запроса на основе анализа дифференциальной экспрессии между начальным и конечным типом клетки. Использовались данные RNA-seq из базы ARCHS4 [29] для начального и конечного типа клетки. Затем собирали набор сигнатур по протоколам в CFM для этого перехода. Для каждой пары сигнатур из этого набора был рассчитан уровень синергии. Рассчитанные значения уровня синергии делились на 2 выборки: для синергетических и несинергетических пар. Далее мы сравнивали распределения этих выборок, используя тест Колмогорова-Смирнова, реализованный в библиотеке SciPy. Он используется для проверки гипотезы о принадлежности двух независимых выборок одному закону распределения. Если статистика Колмогорова-Смирнова значима ($p < 0,05$), то эта гипотеза должна быть отвергнута. Нашей задачей было добиться, чтобы метод хорошо разделял синергетические пары от несинергетических по уровню синергии. Соответственно, для нас важными показателями были p -value и разница между средним значением уровня синергии для несинергетических пар и средним значением уровней синергии для синергетических пар. На рисунке 4.5 ниже приведена схема валидации.

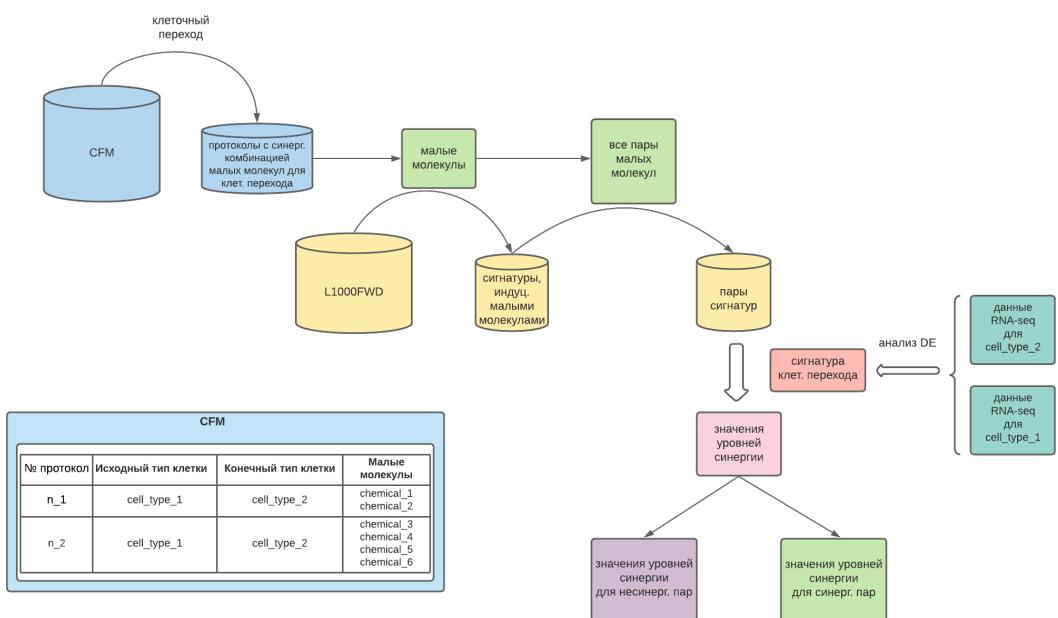


Рис. 4.5: Схема валидации

Нашей задачей было подобрать коэффициенты в выражении inf_score , при которых метод работал на всех клеточных переходах. При валидации на каждом клеточном переходе на большинстве наборов коэффициентов $p\text{-value}$ было достаточно низким, поэтому нас интересовала максимальная разница средних значений уровня синергии для несинергетических пар и синергетических пар. Соответственно разница средних значений уровня синергии для несинергетических пар и синергетических пар рассматривалась как функция коэффициентов. Нашей задачей было найти коэффициенты при максимуме этой функции. Для этого использовали байесовский оптимизатор, поскольку он подходит для задач, где целевая функция неизвестна. Таким образом, для каждого перехода были найдены оптимальные коэффициенты. Далее коэффициенты в выражении inf_score были определены как средние значения найденных оптимальных коэффициентов по всем возможным клеточным переходам в базе CFM.

5 Полученные результаты

При разработке данного метода его валидация проводилась на задаче химического перепрограммирования. Валидация метода была выполнена на 6 клеточных переходах из базы данных CFM. Ниже приведена таблица 5.1 с указанием для них исходной и конечной клеточной линии, а также количества протоколов в базе CFM.

Таблица 5.1: Рассмотренные клеточные переходы базы данных CFM

№ клет. перехода	Исходная клеточная линия	Конечная клеточная линия	количество протоколов
1	фибробласты	индуцированные кардиомиоциты	19
2	фибробласты	индуцированные нейроны	8
3	фибробласты	индуцированные нейральные стволовые клетки	6
4	фибробласты	индуцированные β -клетки поджелудочной железы	6
5	мезенхимальные стволовые клетки	индуцированные нейроны	5
6	фибробласты	индуцированные плюрипотентные стволовые клетки	4

5.1 Оптимизация метода, основанного на сравнении экспрессионных сигнатур, и его валидация

Одним из важных этапов разработки метода было определение значений коэффициентов в выражении `inf_score`. При помощи байесовской оптимизации были определены коэффициенты в выражении `inf_score` для каждого клеточного перехода. Эти коэффициенты соответствовали максимуму разницы средних значений уровня синергии для синергетических пар и несинергетических пар. В качестве оптимальных значений коэффициентов метрик взяли средние значения определенных коэффициентов метрик при байесовской оптимизации по всем клеточным переходам. Ниже в

таблице приведены эти значения коэффициентов.

Таблица 5.2: Рассмотренные клеточные переходы базы данных CFM

№ клет. перехода	coeff Btw	coeff Cln	coeff Egt	coeff Egv	coeff Ktz	coeff logFC	coeff Prk
1	9.8498	4.6486	9.4933	9.5533	9.7345	1.0621	9.4915
2	4.1706	4.2732	5.3341	3.5209	6.2944	2.7864	9.9804
3	9.8498	4.6486	9.4933	9.5533	9.7345	1.0621	9.4915
4	9.5985	9.6596	9.2349	1.2538	9.53	1.4785	7.2369
5	9.0555	8.3232	9.1446	9.6125	9.8218	9.867	7.4246
6	8.7357	1.6625	9.9294	9.8643	9.3573	1.2376	1.2547
opt. coeff	8.5433	5.5359	8.7716	7.2264	9.0787	2.9156	7.4799

В целом наблюдается некоторое постоянство коэффициентов для метрик. Например, для метрик betweenness(Btw), eigentrust (Egt), Katz centrality (Ktz), |logFC| наблюдается маленький разброс коэффициентов. По полученным коэффициентам можно сказать, что большой вес имеют метрики betweenness(Btw), eigentrust (Egt), eigenvector centrality (Egv), Katz centrality (Ktz), pagerank (Prk). Метрика |logFC| обладает наименьшим весом.

Также в таблице 5.3 для этих валидаций приведены величины, которые показывают успешность метода : разница средних значений уровня синергии для синергетических пар и несинергетических пар, р-значение, определяемое тестом Колмогорова-Смирнова. Разница средних значений уровня синергии для синергетических пар и несинергетических пар рассчиталась как:

$$d = \overline{\text{synergy_score}_{\text{syn}}} - \overline{\text{synergy_score}_{\text{not_syn}}} \quad (5.1)$$

Ожидается, что разница средних значений уровня синергии для синергетических пар и несинергетических пар для всех клеточных переходов будет положительна, поскольку, чем выше уровень синергии, тем выше синергетический эффект.

Поскольку список пар ранжируется по вычисленному уровню синергии, то следующие метрики: число синергетических пар среди первых 50 пар, число синергетических пар среди первых пар, составляющих 5% от всех пар, доля синергетических пар среди первых пар, составляющих 5% от всех пар, - были показательны.

Таблица 5.3: Результаты валидации на коэффициентах, определенных при байесовской оптимизации

№ кл. перех.	разн. ср. знач.	p_value	число синерг. пар в топ 50	число пар	число синерг. пар в топ 5%	число пар в топ 5%	доля синерг. пар в топ 5%
1	0.0218	0	15	1236378	33287	61819	0.538
2	0.0086	8.095e-12	37	71631	2581	3582	0.721
3	0.0469	0	26	475800	17560	23790	0.738
4	0.0084	4.409e-10	22	74691	1552	3735	0.416
5	0.0124	3.455e-42	40	49770	1525	2488	0.613
6	-0.0199	1.990e-60	37	128778	3590	6439	0.558

Стоит сразу отметить, что p-значение, определяемое тестом Колмогорова-Смирнова, достаточное низкое, что позволяет отвергнуть гипотезу, что выборки значений уровня синергии для синергетических и несинергетических пар принадлежат одному распределению. Таким образом метод разделяет синергетические и несинергетические пары. Для 5 клеточных переходов разница средних значений уровней синергии синергетических пар и несинергетических пар положительна. Она отрицательна только для перехода из фибробластов в индуцированные плюрипотентные стволовые клетки.

На рисунке 5.1 показаны распределения значений уровня синергии для синергетических пар и несинергетических пар.

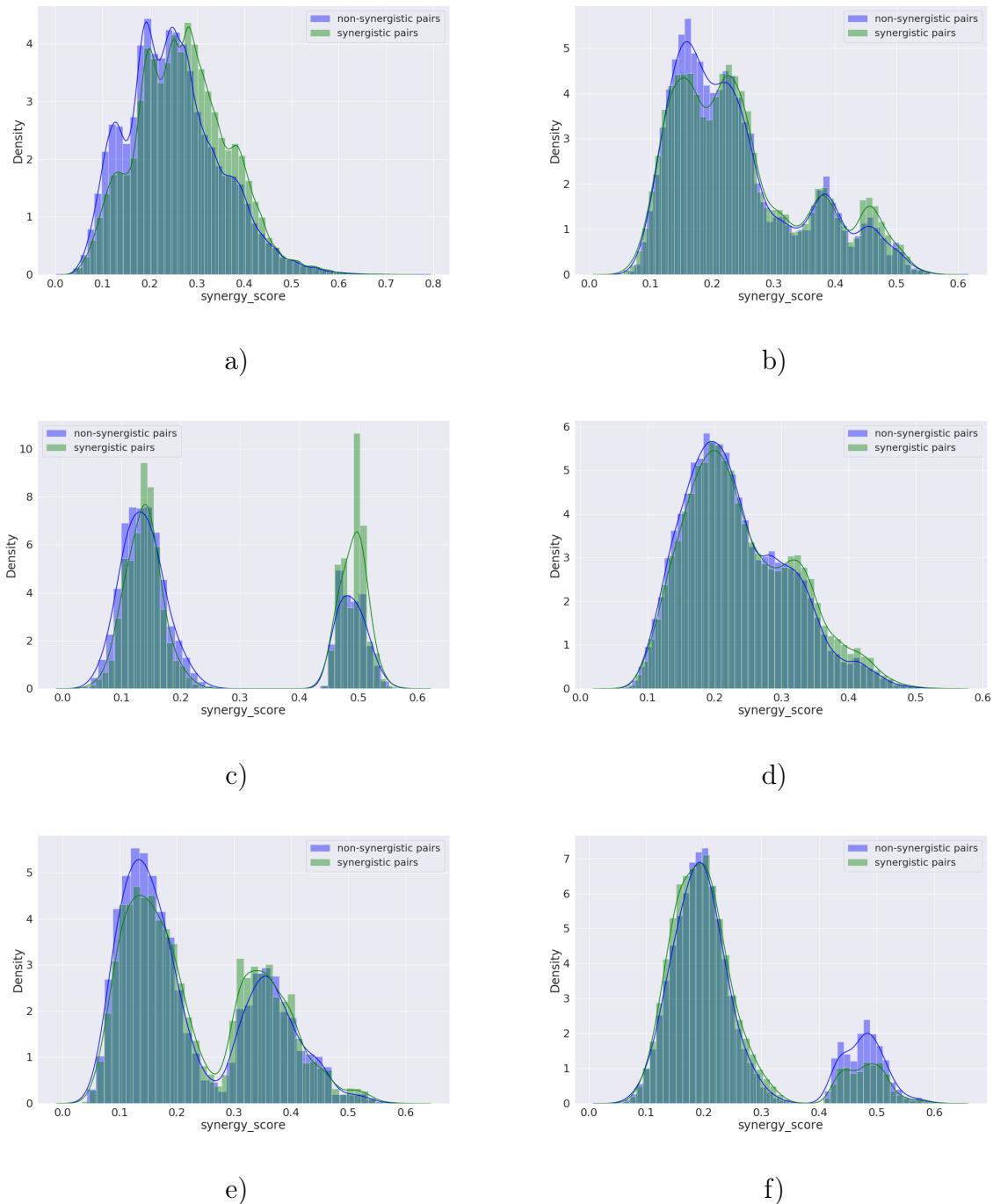


Рис. 5.1: Распределение значений уровня синергии синергетических и несинергетических пар : a) фибробласти -> индуцированные кардиомиоциты, b) фибробласти -> индуцированные нейроны c) фибробласти -> индуцированные нейральные стволовые клетки, d) фибробласти -> индуцированные β -клетки поджелудочной железы, e) мезенхимальные стволовые клетки -> индуцированные нейроны, f) фибробласти -> индуцированные плuriпотентные стволовые клетки

Чтобы проверить оптимальные коэффициенты в выражении inf_score , была прове-

дена валидация с этим набором коэффициентов на 6 клеточных переходах. В таблице приведены результаты валидации.

Таблица 5.4: Результаты валидации на оптимальных коэффициентах

№ кл. перех.	разн. ср. знач.	p_value	число синерг. пар в топ 50	число пар	число синерг. пар в топ 5%	число пар в топ 5%	доля синерг. пар в топ 5%
1	0.0203	0	11	1236378	33108	61819	0.536
2	0.0115	4.667e-10	37	71631	2566	3582	0.716
3	0.0422	0	23	475800	16889	23790	0.710
4	0.0068	8.492e-08	18	74691	1471	3735	0.394
5	0.0068	4.840e-38	42	49770	1558	2488	0.626
6	-0.0221	1.595e-60	34	128778	3559	6439	0.553

При валидации на оптимальных коэффициентах для всех переходов p-значения также достаточно низкие. Характер смещения выборки значений уровня синергии для синергетических пар относительно выборки значений уровня синергии несинергетических пар остался прежним.

На рисунке 5.2 показаны распределения значений уровня синергии для синергетических пар и несинергетических пар.

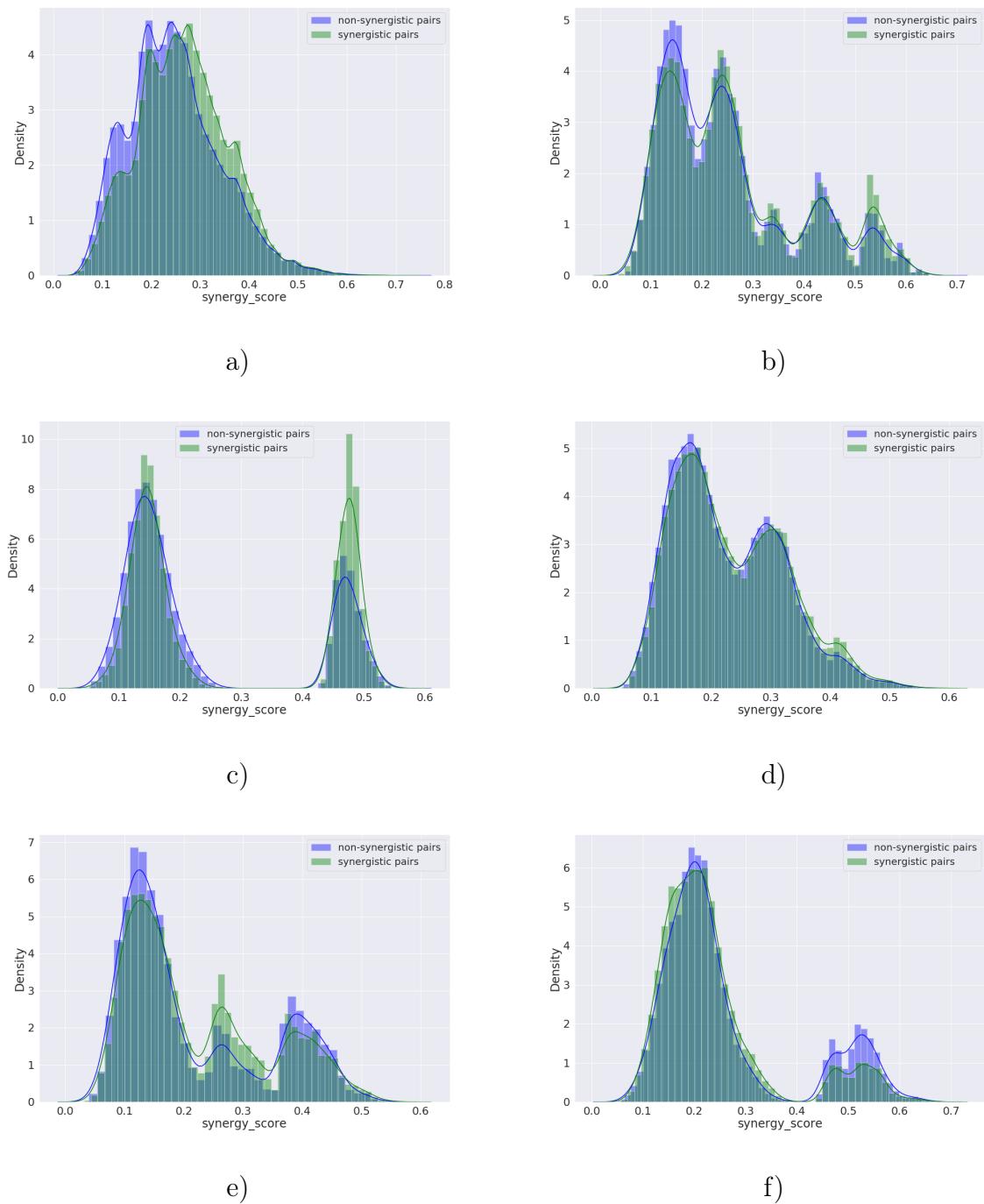


Рис. 5.2: Распределение значений уровня синергии синергетических и несинергетических пар : a) фибробласти -> индуцированные кардиомиоциты, b) фибробласти -> индуцированные нейроны c) фибробласти -> индуцированные нейральные стволовые клетки, d) фибробласти -> индуцированные β -клетки поджелудочной железы, e) мезенхимальные стволовые клетки -> индуцированные нейроны, f) фибробласти -> индуцированные плuriпотентные стволовые клетки

Сравним результаты валидации на оптимальных коэффициентах и коэффициентах,

определенных при байесовской оптимизации. В таблице 5.5 приведены относительные изменения следующих величин : разница средних значений уровня синергии для синергетических пар и несинергетических пар, число синергетических пар в топ 50 пар, число синергетических пар в топ 5% пар.

Таблица 5.5: Сравнение результатов валидации

№ клет. перехода	отн. изм. разн. ср. знач.	отн. изм. числа синерг. пар в топ 50 пар	отн. изм. числа синерг. пар в топ 5% пар
1	-0.069	-0.267	-0.005
2	0.337	0	-0.006
3	-0.100	-0.115	-0.038
4	-0.190	-0.182	-0.052
5	-0.452	0.05	0.021
6	0.111	-0.081	-0.009

Как и ожидалось, для большинства переходов результаты ухудшились, поскольку наборы коэффициентов, полученные при байесовской оптимизации, были подобраны под каждый переход. Если смотреть на относительное изменение разницы средних значений, то сильнее всего эта величина снизилась для перехода из мезенхимальных стволовых клеток в индуцированные нейроны (на 45%). Для остальных она упала меньше, чем на 20%, или повысилась. Для числа синергетических пар в топ 50 пар снижение составляло не более 30 %. Снижение числа синергетических пар в топ 5% пар не превышало 6%. Эти результаты показывают, что оптимальные коэффициенты подходят всем рассмотренным клеточным переходам.

5.2 Валидация метода, основанного на сравнении обогащенных сигнальных путей

Также была проведена валидация с целью проверки эффективности метода при вычислении уровня синергии на основе сравнения обогащенных сигнальных путей. Было проверено 2 способа расчета уровня синергии.

5.2.1 С использованием коэффициента Танимото

Данный подход проверяет гипотезу, что синергетические малые молекулы, дополняя друг друга, затрагивают сигнальные пути, полученные в результате анализа обогащения генов с повышенной экспрессией и генов с пониженной экспрессией сигнатуры запроса. То есть, синергетические малые молекулы совместно охватывают все нужные клеточные процессы. Была выполнена валидация метода с разными метриками оценки его эффективности. В таблице 5.6 показаны ее результаты. Р-значения до-

Таблица 5.6: Результаты валидации

№ кл. перех.	разн. ср. знач.	p_value	число синерг. пар в топ 50	число пар	число синерг. пар в топ 5%	число пар в топ 5%	доля синерг. пар в топ 5%
1	-0.0011	7.522e-193	23	1236378	25300	61819	0.409
2	-0.0011	2.684e-05	22	71631	2356	3582	0.658
3	0.0052	6.020e-244	35	475800	18451	23790	0.778
4	0.0018	1.096e-10	38	74691	1523	3735	0.408
5	0.0001	1.950e-02	31	49770	1457	2488	0.586
6	-0.0022	2.949e-26	27	128778	3967	6439	0.616

статочно низкие для всех клеточных переходов, как и в предыдущем методе. Однако разница средних значений синергии синергетических и несинергетических пар отрицательна для 3 клеточных переходов. По этому показателю можно сказать, что данный подход хуже метода, основанного на экспрессионных сигнатаурах и генных сетях.

На рисунке 5.3 показаны распределения значений уровня синергии для синергетических пар и несинергетических пар.

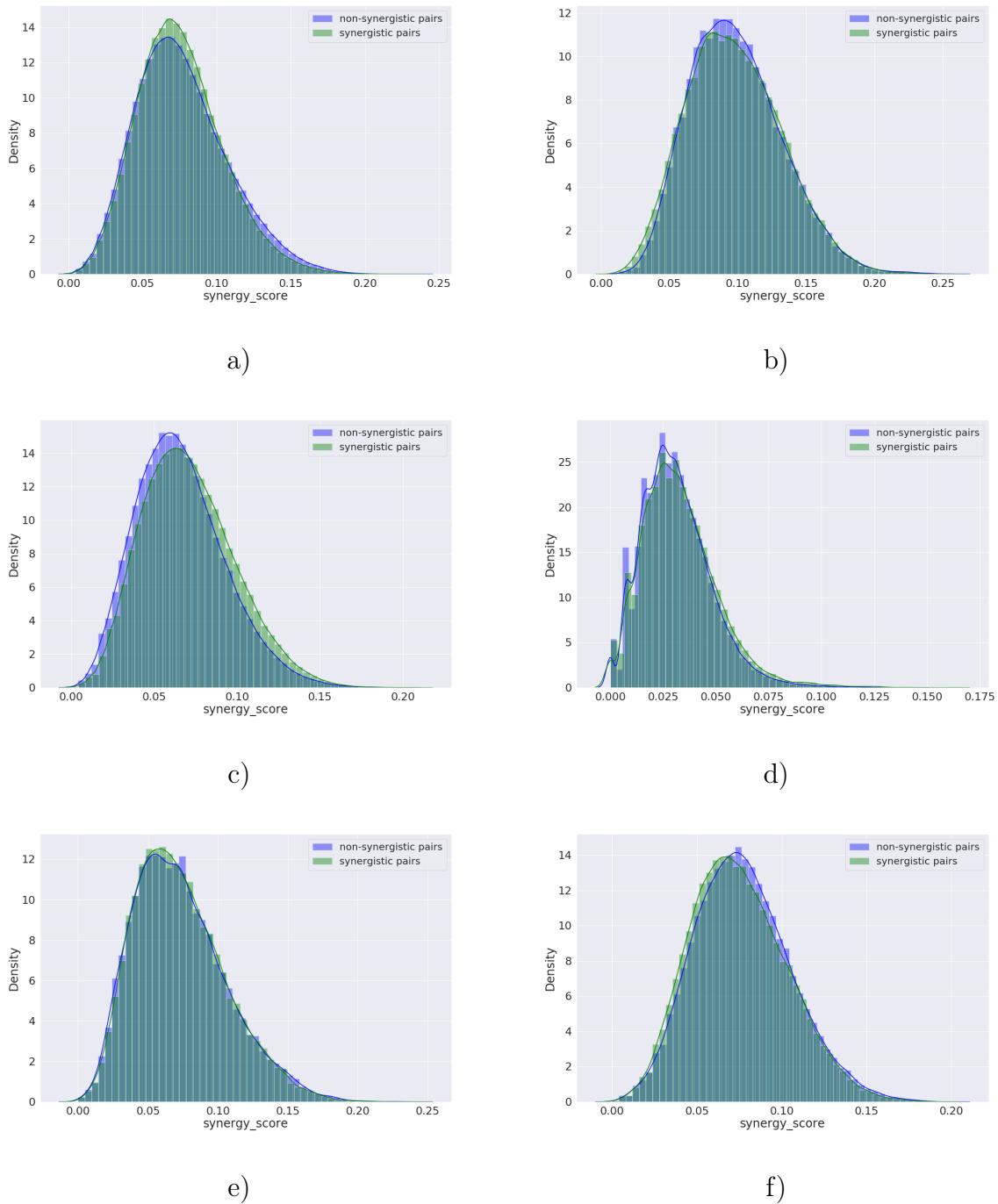


Рис. 5.3: Распределение значений уровня синергии синергетических и несинергетических пар : a) фибробласти -> индуцированные кардиомиоциты, b) фибробласти -> индуцированные нейроны c) фибробласти -> индуцированные нейральные стволовые клетки, d) фибробласти -> индуцированные β -клетки поджелудочной железы, e) мезенхимальные стволовые клетки -> индуцированные нейроны, f) фибробласти -> индуцированные плuriпотентные стволовые клетки

5.2.2 С использованием взаимной информации

Данный подход основан на гипотезе, что синергетические малые молекулы активируют или подавляют одни и те же сигнальные пути (гипотеза сходства). Была проведена валидация метода и ее результаты приведены в таблице 5.7.

Таблица 5.7: Результаты валидации

№ кл. перех.	разн. ср. знач.	p_value	число синерг. пар в топ 50	число пар	число синерг. пар в топ 5%	число пар в топ 5%	доля синерг. пар в топ 5%
1	0.0573	0	29	1236378	33222	61819	0.537
2	0.0097	3.840e-21	41	71631	2565	3582	0.716
3	-0.2091	0	24	475800	15174	23790	0.638
4	0.0156	4.804e-06	22	74691	1200	3735	0.321
5	0.0023	1.7023e-11	40	49770	1708	2488	0.686
6	-0.0257	1.202e-18	36	128778	4067	6439	0.632

Для этого метода есть 2 отрицательных значения для разницы средних значений уровня синергии синергетических и несинергетических выборок. То есть данных подход показал лучше результаты, по сравнению с тем, который основан на гипотезе комплементарности.

На рисунке 5.4 показаны распределения значений уровня синергии для синергетических пар и несинергетических пар.

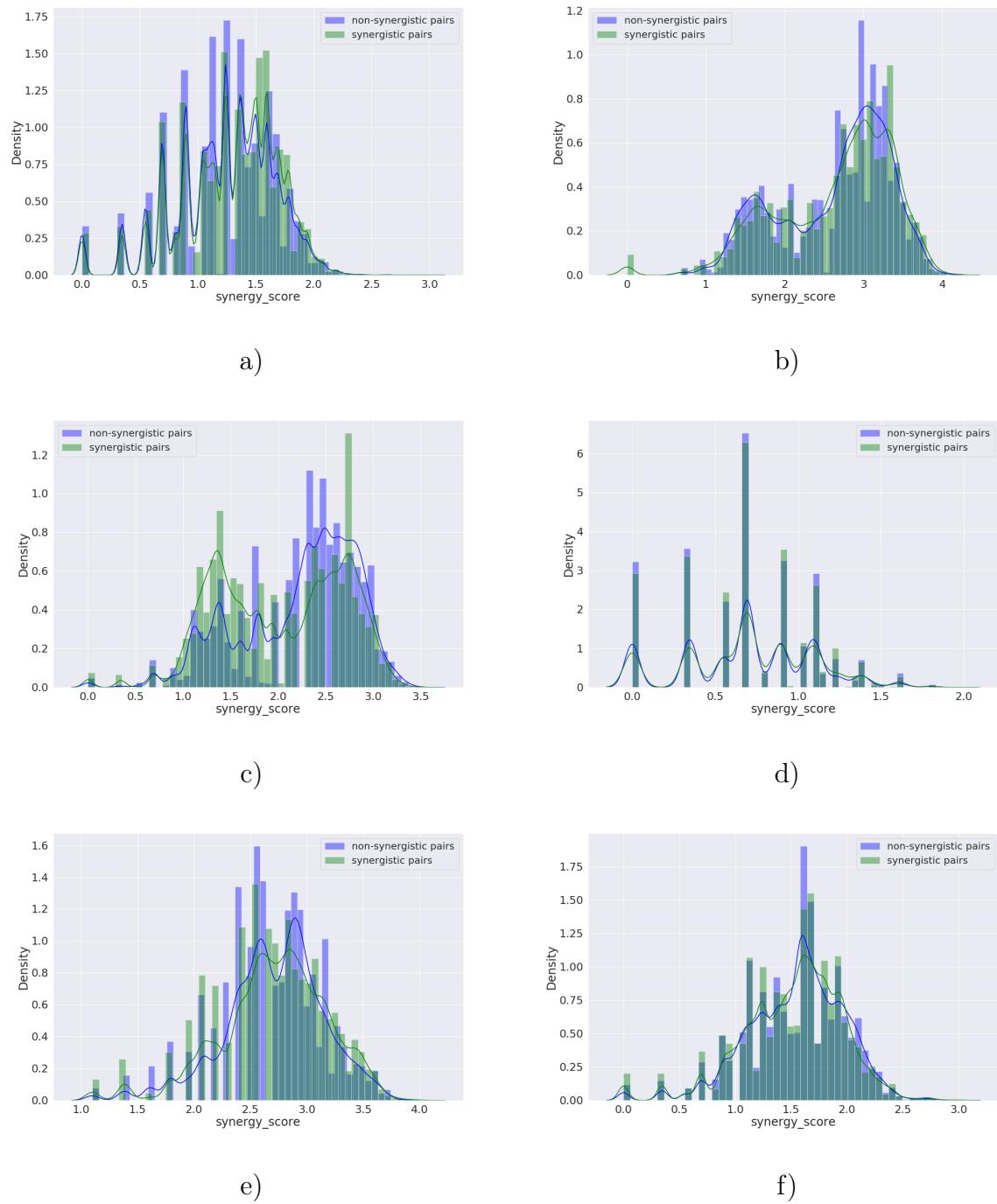


Рис. 5.4: Распределение значений уровня синергии синергетических и несинергетических пар : a) фибробласти -> индуцированные кардиомиоциты, b) фибробласти -> индуцированные нейроны c) фибробласти -> индуцированные нейральные стволовые клетки, d) фибробласти -> индуцированные β -клетки поджелудочной железы, e) мезенхимальные стволовые клетки -> индуцированные нейроны, f) фибробласти -> индуцированные плuriпотентные стволовые клетки

6 Заключение. План дальнейших исследований

В данной работе был реализован метод предсказания синергетических пар на основе данных RNA-seq. Было опробовано несколько способов вычисления уровня синергии пары малых молекул. Их валидация была проведена на задаче химического перепрограммирования.

Первый способ основан на сравнении экспрессионных сигнатур с учетом топологии генных сетей. Одним из важных этапов его разработки был подбор оптимальных коэффициентов в метрике значимости гена. С этой целью для каждого рассматриваемого клеточного перехода были найдены коэффициенты с помощью байесовского оптимизатора. В качестве оптимальных значений использовались средние значения коэффициентов, определенных по всем переходам. Для проверки оптимальных коэффициентов было выполнено сравнение результатов валидации метода с использованием набора коэффициентов, полученных при байесовской оптимизации для каждого перехода, и оптимальных коэффициентов. При сравнении разницы средних значений уровня синергетических и несинергетических пар относительное снижение составляло не более 45%. Для числа синергетических пар в топ 5% понижение не превышало 5%. На основе этого можно сказать, что оптимальные коэффициенты подходят под все рассмотренные клеточные переходы.

Второй способ основан на сравнении наборов обогащенных сигнальных путей. Для него было проверено 2 варианта вычисления уровня синергии, первый из которых основан на гипотезе комплементарности, второй на гипотезе сходства. Для этих методов была выполнена валидация, первый подход уступил по её результатам второму.

Если ориентироваться на разницу средних значений уровня синергии синергетических и несинергетических пар, то метод, основанный на сравнении экспрессионных сигнатур с учетом топологии генных сетей, показал наилучшие результаты. Если сравнивать долю синергетических пар в топ 5%, то методы, основанные на сравнении экспрессионных сигнатур и сравнении обогащенных сигнальных путей, дали похожие результаты.

В дальнейшем для улучшения метода планируется учитывать типы клеточных линий, на которых были индуцированы сигнатуры. Это необходимо из-за того, что клеточный ответ на возмущение малой молекулой сильно варьируется на разных клеточных линиях. Также планируется провести валидацию метода на других дан-

НЫХ.

7 Благодарности

В первую очередь я бы хотела выразить искреннюю признательность и благодарность моему научному руководителю, Муртазалиевой Халимат Асадулаевне, за помочь на всех этапах выполнения дипломной работы. Во-вторых, хочу поблагодарить заведующую лабораторией биоинформатики клеточных технологий, Медведеву Юлию Анатольевну, и заместителя заведующего лабораторией, Ступникова Алексея Ильича, за полезные замечания и советы при написании дипломной работы.

8 Список литературы

- [1] Johann Gasteiger и Thomas Engel. *Chemoinformatics: a textbook*. John Wiley & Sons, 2006.
- [2] Christopher Lipinski и Andrew Hopkins. «Navigating chemical space for biology and medicine». в: *Nature* 432.7019 (2004), с. 855—861.
- [3] Ling Xue и Jurgen Bajorath. «Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening». в: *Combinatorial chemistry & high throughput screening* 3.5 (2000), с. 363—372.
- [4] Lo и др. «Machine learning in chemoinformatics and drug discovery». в: *Drug discovery today* 23.8 (2018), с. 1538—1546.
- [5] Mukesh Bansal и др. «A community computational challenge to predict the activity of pairs of compounds». в: *Nature biotechnology* 32.12 (2014), с. 1213—1222.
- [6] Hongyang Li и др. «Network propagation predicts drug synergy in cancers». в: *Cancer research* 78.18 (2018), с. 5446—5457.
- [7] Lei Huang и др. «Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction». в: *Bioinformatics* 35.19 (2019), с. 3709—3717.
- [8] Martin Kuiper и др. «Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen». в: (2019).
- [9] Xing-Ming Zhao и др. «Prediction of drug combinations by integrating molecular and pharmacological data». в: *PLoS Comput Biol* 7.12 (2011), e1002323.
- [10] Kristina Preuer и др. «DeepSynergy: predicting anti-cancer drug synergy with Deep Learning». в: *Bioinformatics* 34.9 (2018), с. 1538—1546.
- [11] Francesco Napolitano и др. «Automatic identification of small molecules that promote cell conversion and reprogramming». в: (2020).
- [12] Yi Sun и др. «Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer». в: *Nature communications* 6.1 (2015), с. 1—10.

- [13] Pingjian Ding и др. «Incorporating Multisource Knowledge To Predict Drug Synergy Based on Graph Co-regularization». в: *Journal of Chemical Information and Modeling* 60.1 (2019), с. 37—46.
- [14] Celine Lefebvre и др. «A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers». в: *Molecular systems biology* 6.1 (2010), с. 377.
- [15] Maria Stella Carro и др. «The transcriptional network for mesenchymal transformation of brain tumours». в: *Nature* 463.7279 (2010), с. 318—325.
- [16] Giovanni Y Di Veroli и др. «Combenefit: an interactive platform for the analysis and visualization of drug combinations». в: *Bioinformatics* 32.18 (2016), с. 2866—2868.
- [17] Qiaonan Duan и др. «L1000CDS 2: LINCS L1000 characteristic direction signatures search engine». в: *NPJ systems biology and applications* 2.1 (2016), с. 1—12.
- [18] Justin Lamb и др. «The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease». в: *science* 313.5795 (2006), с. 1929—1935.
- [19] Michael Ashburner и др. «Gene ontology: tool for the unification of biology». в: *Nature genetics* 25.1 (2000), с. 25—29.
- [20] Edward Y Chen и др. «Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool». в: *BMC bioinformatics* 14.1 (2013), с. 1—14.
- [21] Aravind Subramanian и др. «GSEA-P: a desktop application for Gene Set Enrichment Analysis». в: *Bioinformatics* 23.23 (июль 2007), с. 3251—3253. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm369. eprint: <https://academic.oup.com/bioinformatics/article-pdf/23/23/3251/16860704/btm369.pdf>. URL: <https://doi.org/10.1093/bioinformatics/btm369>.
- [22] Zichen Wang и др. «L1000FWD: fireworks visualization of drug-induced transcriptomic signatures». в: *Bioinformatics* 34.12 (февр. 2018), с. 2150—2152. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty060. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/12/2150/25047844/bty060_l1000fwd-supp-info-01072017.pdf. URL: <https://doi.org/10.1093/bioinformatics/bty060>.

- [23] Damian Szkłarczyk и др. «STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets». в: *Nucleic Acids Research* 47.D1 (нояб. 2018), с. D607–D613. ISSN: 0305-1048. DOI: 10.1093/nar/gky1131. eprint: <https://academic.oup.com/nar/article-pdf/47/D1/D607/27437323/gky1131.pdf>. URL: <https://doi.org/10.1093/nar/gky1131>.
- [24] Matthew W Hahn и Andrew D Kern. «Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks». в: *Molecular biology and evolution* 22.4 (2005), с. 803–806.
- [25] Maliackal Poulo Joy и др. «High-betweenness proteins in the yeast protein interaction network». в: *Journal of Biomedicine and Biotechnology* 2005.2 (2005), с. 96.
- [26] Sizykh A и др. «CFM: a database of experimentally validated protocols for chemical compound-based direct reprogramming and transdifferentiation». в: *F1000Research* (2021).
- [27] Maxim V. Kuleshov и др. «Enrichr: a comprehensive gene set enrichment analysis web server 2016 update». в: *Nucleic Acids Research* 44.W1 (май 2016), W90–W97. ISSN: 0305-1048. DOI: 10.1093/nar/gkw377. eprint: <https://academic.oup.com/nar/article-pdf/44/W1/W90/18788036/gkw377.pdf>. URL: <https://doi.org/10.1093/nar/gkw377>.
- [28] «RDKit: Open-Source Cheminformatics Software.» в: (2020). URL: <https://www.rdkit.org>.
- [29] Alexander Lachmann и др. «Massive mining of publicly available RNA-seq data from human and mouse». в: *Nature communications* 9.1 (2018), с. 1–10.