1 Аннотация

Одним из первых систематических подходов для определения новых применений существующих лекарств - подход Connectivity Map. Он сравнивает сигнатуру запроса, которая представляет собой разницу между двумя интересующими состояниями, с сигнатурами, вызванными различными возмущениями (малыми молекулами). Однако этот подход не учитывает биологическое значение генов в сигнатуре. Мы разработали подход Connectivity Map, основанный на топологических метриках белок-белковых взаимодействий и данных регуляторных сетей. Он учитывает биологическую роль генов и может быть использован для подбора химических веществ по желаемому механизму действия. Наш инструмент находит малые молекулы, которые могут спровоцировать рассматриваемые клеточные изменения. В качестве входных данных используются экспрессионные сигнатуры состояний, которые мы хотим обратить или имитировать. Наша основная гипотеза: если сигнатура из базы данных сильно пересекается с сигнатурой запроса, то это вещество вероятно вызовет аналогичный клеточный ответ. При ранжировании малых молекул на основе косинусного расстояния предполагается, что молекулы с наименьшим расстоянием имеют высокую вероятность вызвать желаемый переход. В рамках нашего подхода была проведена масштабная оптимизация для подбора метрики оценки влиятельности генов. Количество малых молекул с экспериментально подтвержденным механизмом действия, вызывающих клеточные превращения, в верхней части ранжированного списка использовалось в качестве метрики качества.

Содержание

2 Обозначения, сокращения, основные определения

WLN - Wiswesser

SLN - Sybyl

HBLC - high betweenness and low connectivity

CMap - Connectivity Map

HDAC - Деацетилазы гистонов

HSP90 - белок теплового шока 90

MCF7 - клеточная линия рака груди

CD - Characterictical Direction

CDK1-2 - циклин-зависимые киназы 1-2

GSK3β - Киназа гликоген синтазы 3 β

STAT - signal transducer and activator of transcription protein

GSEA - gene set enrichment analysis

ES - enrichment score

pr - pagerank centrality

bw - betweenness

ev - eigenvector centrality

cl - closeness

kz - Katz centrality

hs - Hits centrality

et - eigentrust

FC - кратное изменение

inf_score - коэффициент влиятельности

fb - фибробласты

heart - индуцированные кардиомиоциты

neuron - индуцированные нейроны

neural - индуцированные нейральные стволовые клетки

beta - индуцированные панкреатические бета клетки

ips - индуцированные плюрипотентные стволовые клетки

mes - мезенхимальные стволовые клетки

TGF-β - трансформирующий фактор роста бета

NF$\varkappa$B - ядерный фактор «каппа-би»

MAPK - митоген-активируемая протеинкиназа

PI3K - Фосфоинозитид-3-киназы

Akt - протеинкиназа B

mTORC - Мишень рапамицина млекопитающих

T24 - клеточная линия мочевого пузыря

MDA 468 - клеточная линия карциномы груди

MDA 435 - клеточная линия карциномы груди

SAHA - субероиланилидгидроксамовая кислота

ER - рецептор эстрогена

E2 - 17β-эстрадиол

LNCaP - клеточная линия рака простаты

PPAR - Рецепторы, активируемые пероксисомными пролифераторами

AD - Болезнь Альцгеймера

3 Введение

3.1 Представление химических веществ

Проблема представления химических молекул в виде, пригодном для анализа и обработки на компьютере, является острой. В современной хемоинформатике используется множество различных представлений химических молекул таких как: номенклатурное представление, линейные нотации, различные матричные представления. Так как каждый из форматов имеет свои сильные и слабые стороны, не существует единого стандарта хранения химических соединений. Более того, для различных задач используются различные форматы. В данной главе будут рассмотрены основные представления химических веществ и форматы их хранения.

3.1.1 Линейная нотация

Наиболее популярными среди линейных нотаций молекул являются следующие: Wiswesser (WLN), ROSDAL, SMILES и Sybyl (SLN) [1]. Большинство из вышеперечисленных нотаций в настоящее время используются редко, поэтому имеет смысл подробно остановиться только на нотации SMILES.

- SMILES (Simplified Molecular Input Line Entry System)
  Данная нотация получила широкое распространение и применяется повсеместно для хранения химических молекул. SMILES основана на шести простых правилах представления молекул:
  - Атомы представлены своими химическими символами
  - Атомы водорода автоматически заполняют свободные валентности и опускаются
  - Соседние атомы расположены подряд
  - Двойная и тройная связи представляются символами ”=”и "#" соответственно

– Боковые цепи молекулы заключаются в скобки

– Атомы циклов, находящиеся на концах разорванной при построении линейной нотации молекулы, обозначаются одним и тем же номером

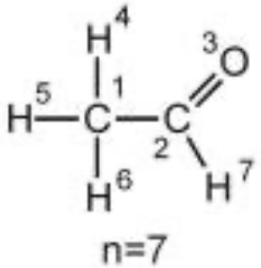
Данный вариант представления химических структур имеет ряд достоинств: наиболее простое линейное представление, возможность быстрого обмена данными, поддержка структуры Маркуша, стереохимии и некоторых других опций. Однако существует несколько серьезных недостатков данного формата: неоднозначность декодирования, а также некоторые проблемы с представлением ароматических соединений.

В настоящее время существует множество разновидностей форматов SMILES. Наиболее интересным для текущей задачи является формат USMILES (Unique SMILES), позволяющий однозначно восстанавливать химическую структуру молекул после конвертации из формата USMILES.

3.1.2 Молекулярный граф

Часто молекулы представляются в виде молекулярных графов. Под термином молекулярный граф мы понимаем следующее: молекулярный граф — связный неориентированный граф, находящийся во взаимно-однозначном соответствии со структурной формулой химического соединения таким образом, что вершинам графа соответствуют атомы молекулы, а рёбрам графа — химические связи между этими атомами. Как правило в органической химии при представлении молекул опускаются атомы водородов, а также символы атомов углеродов.

В компьютере молекулярные графы часто представляются в виде матриц. На рисунке ниже показано, как молекулярный граф представляется в виде матрицы. Каждому атому присваивается уникальный номер. Если два атома, например атом 1 и атом 5 на рисунке, связаны, то в матрице в полях [1,5] и [5,1] будет стоять единица, в противном случае, если атомы не связаны, стоит ноль.



	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

Рисунок 3.1 – Представление молекулярного графа в виде матрицы

3.1.3 Матричное представление молекул

В настоящее время широко используется матричное представление молекул. Существует множество различных матриц, таких как: матрицы смежности, матрицы расстояний, матрицы связей и т.д. Несмотря на разнообразие матриц, существует два основных принципа матричного представления молекул: каждая молекула представлена в виде матрицы $n \times n$, где n - число атомов молекулы; каждый атом описывается дважды - в столбце и в строке матрицы.

- Матрица смежности

  Матрица смежности молекулы - квадратная матрица размера $n \times n$, показывающая все связи между атомами. Единица, стоящая на пересечении строки i и столбца j означает, что атомы с номерами i и j соединены. В противном случае на пересечении строк стоит 0. На диагонали матрицы всегда стоят нули.

  Данная матрица может рассматриваться как обобщение матрицы связей. Содержит дополнительную информацию о числе свободных валентных электронов на конкретном атоме диагональных элементов.

В целом матричное представление молекул обладает рядом плюсов и минусов. Из достоинств можно отметить, что молекулярный граф полностью кодирует молекулу, а также позволяет применять матричную алгебру. Существен-

ным недостатком является тот факт, что число значений возрастает пропорционально n^2, а также не учитывается стереохимия молекул.

3.1.4 Таблица связей

Для решения проблемы матричного представления и квадратичного возрастания значений при увеличении числа атомов в молекуле был предложен метод представления молекул, под названием таблица связей. Данная структура представляет собой два списка один из которых: список всех атомов молекулы и таблица из трех столбцов. Первые два столбца таблицы показывают какие атомы связаны между собой, а третий столбец - порядок связи (одинарная = 1, двойная = 2 и т.д.). Как и в большинстве других представлений молекул, атомы водорода в большинстве случаев опускаются и при визуализации молекул восстанавливаются по стандартной валентности атомов. Структура таблицы связей может быть дополнена различными списками, такими как списки свободных электронов или список зарядов атомов молекул.

Такой формат представления молекул имеет ряд существенных преимуществ над описанными выше, такие как: число значений растет линейно с увеличением числа атомов в молекулах, возможно добавление дополнительной информации о молекуле, помимо атомов и связей. Более того, данное представление широко используется в множестве различных пакетов для работы с химическими данными.

3.1.5 Форматы файлов для хранения химических данных

В настоящее время существует большое число форматов файлов для хранения разнообразных химических данный. В данной главе будут освещены основные форматы:

- Molfile

  Формат данных Molfille является одним из наиболее используемых форматов для хранения структуры химических молекул. Данный формат хранит

молекулы в виде таблицы связей. Основным недостатком этого формата является отсутствие единого формата таблиц связей. Расширение данного файла: *.mol

- SMILES

  Также широко распространенный формат файла. Данный формат хранит линейную нотацию молекул. Расширение файлов: *.smi

- PDB file

  Данный формат файла используется для хранения 3D структур биологических макромолекул. Расширение фойлов: *.pdb

- CIF

  Данный формат используется для хранения 3D структур молекул, полученных методом кристаллографии. Расширение файлов: *.cif

- JCAMP

  Данный формат используется для хранения информации о молекуле, полученной методом спектрометрии. Существует две модификации данного формата: первая, JCAMP-CS, содержит структурную информацию о молекулах и является аналогом Molfile; вторая, JCAMP-DX, непосредственно содержит спектроскопические данные о молекуле. Расширения фалов: *.jdx, *.dx, *.cs

- CML

  Данный формат файла является неким обобщением всех вышеперечисленных. В данном формате собрана вся химическая информация о молекуле, доступная на текущий момент. Расширение файлов: *.cml

3.1.6 Однозначное и единственное представление молекул

Все вышеперечисленные представления молекул обладают одной существенной проблемой, которая на практике затрудняет применение этих представлений: неоднозначность обратного преобразования молекулы в структурную формулу и множественность представления структурной формулы в виде, например, таблицы связей. Более того, для каждой молекулы, состоящей из n атомов, имеется $n!$ различных ее представлений в виде таблицы связей.

Для решения этой проблемы в большинстве случае применяют алгоритм Моргана. Этот алгоритм позволяет однозначно кодировать и декодировать химические вещества, а также учитывать их стереохимию.

Алгоритм Моргана состоит из двух основных частей: процесса релаксации, который классифицирует атомы в зависимости от их соседей, и присваивания уникальных и инвариантных номеров атомам. Далее более подробно будут описаны эти процессы.

- *Классификация атомов по их соседям (процесс релаксации)*
  На данном этапе происходит подсчет extended connectivity (EC), на основе которых атомам присваиваются уникальные номера.

- *Присваивание каждому атому уникального инвариантного номера*
  На этом этапе атомам присваиваются уникальные номера. Атому с наибольшим EC, полученным на последней итерации предыдущего этапа, присваивается номер 1. Номер 2 получает атом, связанный с атомом 1, с наибольшим значением EC. Дальнейшая нумерации строится аналогично. После нумерации всех атомов, связанных с первым, нумерация продолжается с еще не пронумерованных соседей атома 2 и так далее. Если при нумерации встречаются атомы с одинаковыми значениями EC, то дальнейшая нумерация подчиняется строгим правилам, которые учитывают тип атом, его связи, заряд и так далее.

3.1.7 Фрагментное кодирование молекул

Ещё одним распространенным способом представления молекул является фрагментное кодирование, а именно фингерпринты. Для этого молекула разбивается на отдельные функциональные группы, кольца и так далее. Структурные фингерпринты представляют собой бинарные последовательности нулей и единиц, которые показывают наличие (1) или отсутствие (0) какой-либо подструктуры в описываемой молекуле. Стандартный размер структурного фингерпринта - 150-2500 бит. Такое простое представление молекулы в виде строки позволяет легко находить сходства между молекулами и осуществлять поиск близких по структуре молекул.

Кроме того, структурные фингерпринты могут подвергаться хэшированию - преобразованию по специальному алгоритму в строку строго определенной длины. С хэшированными фингерпринтами удобнее работать, их быстрее сравнивать и быстрее производить поиск, так как их длина строго фиксирована. Более того, так как при появлении в молекуле каждый отдельный фрагмент активирует не одну позицию в фингерпринте, а сразу несколько, то вероятность ”коллизии”достаточна низкая, чтобы по такого рода фингерпринтам однозначно восстанавливать структурную формулу молекулы. На рисунке ниже приведен наглядный пример того, что было описано выше.

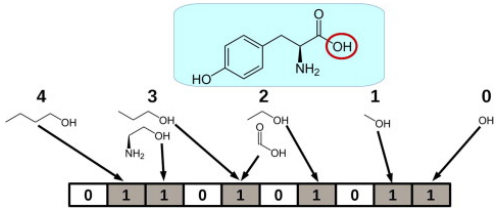


Рисунок 3.2 – Пример построения хэшированного фингерпринта

Подводя итог данной части, хотелось бы отметить, что существует множество разнообразных представлений молекул, каждое из которых подходит под строго определенные цели и задачи. В этой работе будут использованы такой формат файлов, как SMILES, а также фингерпринты Моргана для поиска

сходства молекул.

3.2 Топология биологических сетей

В настоящее время различные биологические сети получили широкое распространение. Существует несколько основных видов биологических сетей:

- сети белок-белковых взаимодействий
- генные сети
- сети метаболических путей

В данном разделе будут подробно рассмотрены все виды сетей, а также освещены основные метрики, используемые при работе с биологическими сетями.

3.2.1 Сети белок-белковых взаимодействий

Данный вид сетей представляет собой ненаправленный граф, вершинами которого являются белки. Две вершины соединены ребром, если два соответствующих белка взаимодействуют между собой в биологической системе. Сети такого рода являются по своему строению безмасштабными сетями, то есть степени вершин таких сетей распределены по степенному закону, то есть доля вершин со степенью k примерно или асимптотически пропорциональна $k^{-\gamma}$.

Для работы с такими сетями было предложено множество различных метрик центральности, таких как:

- степень вершины (degree) - число ребер графа, которым принадлежит эта вершина

$$d_i = \sum_j u_j,$$

где u_j - ребро графа, принадлежащее данной вершине.

- кратчайшее расстояние между вершинами (shortest distance) - минимальное число ребер невзвешанного графа, которое нужно пройти, чтобы попасть из одной вершины в другую.

- степень посредничества (betweenness) - это мера центральности в графе, основанная на кратчайших путях. Для любой пары вершин в связном графе существует по меньшей мере один (кратчайший) путь между вершинами, для которого минимально либо число рёбер, по которым путь проходит, (для невзвешенных графов), либо сумма весов этих рёбер (для взвешенных графов). Степень посредничества для каждой вершины равна числу этих кратчайших путей через вершину, нормированных на общее число кратчайших путей в графе.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

где σ_{st} равно общему числу кратчайших путей из узла s в узел t, а $\sigma_{st}(v)$ равно числу этих путей, проходящих через v.

- связанность (connectivity) - минимальное число элементов графа (вершин и ребер), которое необходимо удалить, чтобы разделить оставшиеся вершины на изолированные подграфы.

- степень близости узла (к другим узлам) (closeness) — это мера центральности в сети, вычисляемая как обратная величина суммы длин кратчайших путей между узлом и всеми другими узлами графа. Таким образом, чем более централен узел, тем ближе он ко всем другим узлам.

$$C(x) = \frac{1}{\sum_y d(y,x)},$$

где $d(x,y)$ - расстояние между вершинами x и y.

- степень влиятельности (eigenvector centrality) - метрика центральности гра-

фа, вычисляемая как собственные вектора матрицы смежности. Решается уравнение на поиск собственных векторов матрицы:

$$\mathbf{Ax} = \lambda\mathbf{x},$$

где **A** - матрица смежности, **x** - собственный вектор, а λ - собственное значение матрицы смежности, соответствующее данному собственному вектору. Также это векторное уравнение можно переписать в виде суммирования по всем вершинам:

$$x_v = \frac{1}{\lambda}\sum_{t\in M(v)} x_t = \frac{1}{\lambda}\sum_{t\in G} a_{v,t}x_t,$$

где $M(v)$ представляет собой множество соседей вершины v, а λ является константой.

- центральность графа Харари (Harary graph centrality) для вершины (v) определяется формулой:

$$C(v) = \frac{1}{max_u d(v,u)},$$

где $d(v,u)$ - кратчайшее расстояние между вершинами v и u.

- информационная центральность (information centrality) - средняя гармоническая длина путей, заканчивающихся в вершине s. Чем больше у вершины s коротких путей, соединяющих ее с другими вершинами, тем меньше данная метрика. Сначала определим понятие Матрица Кирхгофа:

$$\mathbf{L} = \mathbf{D} - \mathbf{A},$$

где **A** - взвешанная матрица смежности графа, **D** - диагональная матрица

такая, что

$$d_{i,j} := \begin{cases} \deg(v_i) & \text{при } i = j, \\ 0 & \text{в противном случае.} \end{cases}$$

Тогда

$$l_{i,j} := \begin{cases} \deg(v_i) & \text{при } i = j, \\ -1 & \text{при } (v_i,v_j) \in E(G), \\ 0 & \text{в противном случае.} \end{cases}$$

Определим матрицу **J** как матрицу всех вхождений, которые равны единице. Определим матрицу **B** как $\mathbf{B} = \mathbf{L} + \mathbf{J}$. Тогда определим информацию, проходящую между вершинами u и v как

$$I_{uv} = \frac{1}{\mathbf{B}^{-1}(u,u) + \mathbf{B}^{-1}(v,v) - 2\mathbf{B}^{-1}(u,v)}$$

Тогда информационной центральностью называется гармоническое среднее I_{uv} по всем вершинам u:

$$I_v = \frac{n}{\sum\limits_{u \in V} \frac{1}{I_{uv}}}.$$

- stress центральность (stress centrality) - простая сумма числа всех кратчайших путей, проходящих через вершины.

$$C_s(v) = \sum_{s \neq t \neq v \in V} \rho_{st}(v),$$

где $\rho_{st}(v)$ - число кратчайших путей, проходящих через вершину v.

- центральность близости вершин при случайном блуждании (random walk closeness) - это мера центральности сети, которая описывает среднюю скорость, с которой случайно идущие процессы достигают узла из других узлов сети.

Рассмотрим взвешенный граф (направленный или ненаправленный) с n вершинами, обозначенными $j = 1,\ldots, n$; и процесс случайного блуждания по этому графу с матрицей перехода M. Элемент m_{jk} в M описывает вероятность случайного блуждания из вершины i перейти непосредственно в вершину j . Эти вероятности определяются следующим образом.

$$M(i,j) = \frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}},$$

где a_{ij} - это (i, j)-й элемент матрицы весов графа A. Когда между двумя вершинами графа нет ребра, то соответствующий элемент матрицы A равен нулю.

Центральность близости случайного блуждания вершины i является обратной величиной среднего времени первого перехода к этому узлу:

$$C_i^{RWC} = \frac{n}{\sum_{j=1}^{n} H(j,i)}$$

Среднее время первого прохождения от вершины i к вершине j - это ожидаемое количество шагов, которое потребуется процессу, чтобы впервые достичь узла j из узла i:

$$H(i,j) = \sum_{r=1}^{\infty} rP(i,j,r),$$

где $P(i,j,r)$ - вероятность того, что требуется ровно r шагов, чтобы достичь j из i в первый раз. Чтобы вычислить эти вероятности достижения вершины в первый раз за r шагов, полезно ввести преобразование M путем удаления его j-ой строки и столбца; обозначим его как M_{-j}. Поскольку вероятность того, что процесс начнется с i и окажется в k после r-1 шагов,

просто является (i, k)-м элементом M_{-j}^{r-1}, $P(i,j,r)$ можно выразить как

$$P(i,j,r) = \sum_{k \neq j} ((M_{-j}^{r-1}))_{ik} m_{kj}$$

Подставляя это в выражение для среднего времени первого прохождения, получаем

$$H(i,j) = \sum_{r=1}^{\infty} r \sum_{k \neq j} ((M_{-j}^{r-1}))_{ik} m_{kj}$$

Используя формулу суммирования геометрических рядов для матриц, получаем

$$H(i,j) = \sum_{k \neq j} ((I - M_{-j})^{-2})_{ik} m_{kj}$$

где I - это n-1-мерная единичная матрица.

- центральность посредничества вершин при случайном блуждании (random walk betweenness) - по существу тоже самое, что и посредничество, однако вместо кратчайших путей используются случайные блуждания из одной вершину в другу.

$$C_i^{RWB} = \sum_{j \neq i \neq k} r_{jk},$$

где элемент r_{jk} матрицы R, который содержит вероятность случайного блуждания, начинающегося в узле j с поглощающим узлом k, проходящего через узел i.

- коэффициент кластеризации (clustering coefficient) - это мера того, в какой степени узлы в графе склонны группироваться вместе. Существуют две версии этой меры: глобальная и локальная. Глобальная версия была разработана для того, чтобы дать общее представление о кластеризации в сети, тогда как локальная показывает вовлечённость отдельных узлов.

  Предположим, что граф полностью описывается матрицей смежности A. Тогда локальный коэффициент кластеризации i-ой вершины графа можно

рассчитать следующим образом:

$$C_i = \frac{1}{k_i(k_i-1)} \sum_{j,k} A_{ij}A_{jk}A_{ki},$$

где $k_i = \sum_j A_{ij}$. Глобальный коэффициент кластеризации всего графа может быть рассчитан по следующей формуле:

$$C = \frac{\sum_{i,j,k} A_{ij}A_{jk}A_{ki}}{\sum_i k_i(k_i-1)},$$

где $k_i = \sum_j A_{ij}$.

- центральность подграфа (subgraph centrality) вершины показывает количество подграфов, в которых содержится данная вершина, нормализуя метрику на размеры подграфов. Определим локальный спектральный момент матрицы смежности **A** как i-ый диагональный элемент k-ой степени матрицы смежности:

$$\mu_k(i) = (\mathbf{A}^k)_{ii}$$

Тогда центральность подграфа будем называть следующее выражение:

$$SC(v) = \sum_{k=0}^{\infty} \frac{\mu_k(v)}{k!}$$

- центральность близости Фримена (Freeman closeness) - полное геодезическое расстояние от данной вершины до всех других . Данная метрика определяется следующей формулой:

$$C(v) = \frac{1}{\sum_{i \neq v} d(v,i)},$$

где $d(v,i)$ - кратчайшее расстояние между вершинами v и i.

Также довольно часто используются метрики консервативности белков, как мера их важности:

- эволюционное расстояние D определяется формулой

$$q = \ln \frac{1 + 2D}{2D},$$

где q - отношение числа консервативных сайтов в последовательности выравниваний пар белков

- ESC (excess sequence conservation) - средняя избыточная последовательность является мерой эволюционной консервативности белка. Определяется следующим соотношением:

$$\langle ESC_k \rangle = \frac{1}{N_k} \sum_{i}^{N_k} \frac{\langle D \rangle - D_i}{\langle D \rangle},$$

где N_k - количество белков в соответствующей группе k, а $\langle D \rangle = \frac{1}{N} \sum_{i}^{N} D_i$ - среднее эволюционное расстояние всех N белков

- ER (excess retension): В соответствии со степенью вершины k в базовой сети взаимодействия белков все белки группируются в ячейки по логарифмически увеличивающейся связности k. В каждой ячейке отношение $e_k^A = \frac{n_k^A}{N_k}$ представляет собой определенную характеристику A, где n_k^A - это количество белков, которые имеют характеристику A (например, являются незаменимыми или ортологичными в эталонном организме), а N_k- общее количество белков. В отсутствии корреляции между A и его положением в сети, e_k^A имеет общее значение, не зависящее от k, $e = \frac{n}{N}$, где $n = \sum_k n_k^A$ - общее количество белков организма, имеющих признак A, а $N = \sum_k N_k$ - общее количество белков в основной сети. Таким образом, для каждой группы k мы определяем эволюционное избыточное удержание признака A как $ER_k^A = \frac{e_k^A}{e}$, которое должно иметь независимое от k

значение $ER_k = 1$ при случайном присвоении A.

Вопрос выбора метрики для определения важности белков до сих пор остается нерешенным, поэтому актуальной является задача подбора оптимальных метрик для решения конкретных задач на конкретных данных. В литературе существует множество статей, которые описывают процесс подбора метрик для конкретного исследования, что подтверждает описанную выше проблему выбора метрики.

Jeong et al. [2] описали идею использования такой топологической метрики, как степень вершины графа белок-белковых взаимодействий для определения летальность мутации того или иного белка сети в дрожжах Saccharomyces cerevisiae. Мутации моделировались путем исключения случайной вершины из графа взаимодействий. Полученные результаты валидировались на списке известных летальных мутаций для дрожжей. Было показано, что белки с высокой степенью вершины в 3 раза вероятнее окажутся важными для выживаемости организма, нежели белки с низкой степенью вершины. Однако, существует противоположная точка зрения, опровергающая связь топологических метрик центральности и важности белков. Утверждается, что для важных белков накоплено больше данных, чем для редких белков, поэтому связь топологических метрик с важностью белка ставится под сомнение.

Hahn et al. [3] подробно описали использование трех метрик центральности, таких как связанность, степень посредничества и степень близости узла, для определения важности белков, а также их эволюционной консервативности в трех эукариотических организмах: Saccharomyces cerevisiae, Caenorhabditis elegans и Drosophila melanogaster. В датасет для построения сетей белок-белковых взаимодействий вошли только те белки, которые имеют ортологов во всех трех организмах. Полученные результаты валидировались на результатах более ранних исследований, в которых была выявлена летальность того или иного белка. Было показано, что белки с высокой степенью вершины в 3 раза вероятнее окажутся важными для выживаемости организма, нежели белки с низкой сте-

пенью вершины. Было показано, что положение белка в сети белок-белковых взаимодействий влияет как на скорость его эволюционных изменений, так и на вероятность оказаться важным для выживаемости. Также стоит отметить, что для необходимых генов все три метрики были выше, чем для остальных, что подтверждает наличие корреляции между центральностью белков в сети взаимодействий и их важностью для выживания организма.

Pržulj et al. [4] исследовали такие метрики, как связанность, длина кратчайшего пути и количество точек сочленения. Был проведен систематический анализ, основанный на теории графов сети белок-белковых взаимодействий для построения вычислительных моделей для описания и прогнозирования свойств летальных мутаций и белков, участвующие в генетических взаимодействиях, функциональных групп, белковых комплексов и сигнальных путей. Анализ показал, что летальные мутации не только сильно связаны в сети, но и обладают дополнительным свойством: их удаление вызывает нарушение целостности сети. Также были предоставлены доказательства существования альтернативных путей обхода жизнеспособных белков сети, в то время как подобных путей для летальных мутаций не существует. Кроме того, было установлено, что разные функциональные классы белков обладают разными характеристиками сети. Во время валидации была оценена весомость прогнозов путем их сравнения со случайной моделью, и оценена точность прогнозов за счет анализа их перекрытия с базой данных MIPS.

Joy et al. [5] исследовали такие метрики центральности графа, как степень посредничества и связанность для исследования структуры сетей белок-белковых взаимодействий. В проделанной работе была уточнено строение белковый сетей, а именно был найден такой вид вершин, как HBLC (high betweenness and low connectivity), то есть вершины с высоким значением степени посредничества и низким значением связанности. Однако ранее считалось, что существуют лишь два типа вершин графов белок-белковых взаимодействий: вершины с низкой степенью посредничества и низкой связанностью и наоборот -

вершины с высокими значениями обеих метрик.

Прежде чем описывать дальнейшие результаты, стоит описать вычислительные модели эволюции биологических сетей, использовавшиеся для валидации полученных данных.

- Модель Барабаши — Альберт (BA модель)

  Алгоритм генерации случайных безмасштабных сетей с использованием принципа предпочтительного присоединения. Сеть начинается с начальной сетки с m_0 узлами. $m_0 >= 2$ и степень каждого узла в начальной сети должна быть не меньше 1, иначе она всегда будет отделена от остальной части сети. В каждый момент времени в сеть добавляется новый узел. Каждый новый узел соединяется с существующими узлами с вероятностью, пропорциональной числу связей этих узлов. Формально, вероятность p_i того, что новый узел соединится с узлом i, равна:

  $$p_i = \frac{k_i}{\sum_j k_j},$$

  где k_i — степень i-го узла, а в знаменателе суммируются степени всех существующих узлов. Наиболее связанные узлы («хабы»), как правило, накапливают ещё больше связей, тогда как узлы с небольшим числом связей вряд ли будут выбраны для присоединения новых узлов. Новые узлы имеют «предпочтение» соединяться с наиболее связанными узлами. Такой принцип связывания узлов называется принципом предпочтительного соединения.

- Обобщенная модель Барабаши - Альберт (EBA модель)

  По сути является обобщением модели BA, где добавление соединений и их изменение происходит вместе с добавлением узлов с преимущественным присоединением.

- Модель Sole - Vazquez (SV модель)

Биологически ориентированная модель построения безмасштабных сетей. В этой модели существующие узлы (белки) копируются со всеми их существующими связями, за чем следует дивергенция дублированных узлов, вводимая путем изменения связей и/или добавления связей, имитируя мутации дублированных генов.

- Модель дупликация-мутация (DM модель)

  Биологически ориентированная модель построения безмасштабных сетей, учитывающая дупликации и мутации генов. Точечные мутации, которые влияют на способность белка участвовать в молекулярных взаимодействиях, моделируются как присоединение или отсоединение связей, в то время как количество узлов фиксировано («динамика связей»). Поскольку дублирование узлов в эволюционных временных масштабах происходит медленно, по сравнению с временной шкалой динамики связей, дублирование генов моделируется как добавление узлов без каких-либо связей, в то время как динамика связей происходит на каждом временном шаге. Это было оправдано наблюдением, что в дублированных генах полная диверсификация происходит почти сразу после дупликации. Обычно это расхождение является необъективным, так как один из белков сохраняет большую часть взаимодействий, в то время как другой сохраняет несколько или ни одного. Таким образом, для динамики связи в нашем моделировании новое присоединение устанавливается следующим образом: выбирается случайный узел и присоединяется к другому узлу с предпочтительным присоединением, то есть со скоростью, пропорциональной его связанности k, как в модели BA. Напротив, для отсоединения связь между двумя узлами выбирается со скоростью отсоединения, пропорциональной сумме инверсий их связностей. Это мотивировано наблюдением более высокой частоты мутаций для менее связанных белков.

При анализе четырех существующих вычислительных моделей эволюции биологических сетей (BA модель, EBA модель, SV модель и DM(duplication-

mutation) модель) было установлено, что только модель DM способна воспроизводить сети, содержащие HBLC белки. Сравнивая модели роста сети, было обнаружено, что мутации (изменения в сетевых связях из-за их добавления и удаления) играют центральную роль в механизме создания сети, вследствие чего появляются HBLC белки. Таким образом, предложенный алгоритм объясняет эту отличительную черту топологии сети без необходимости учета функциональной адаптации. В этом исследовании показано, что существование белков HBLC является неизбежным следствием определенных молекулярных механизмов роста сети, которые включают в себя случайные изменения схемы связи из-за мутаций. Это, вместе с открытием того, что узлы HBLC, по-видимому, не являются эволюционно более старыми белками, поддерживает идею о том, что присутствие белков HBLC обусловлено внутренними, структурными и механистическими ограничениями роста сети.

Park et al.[6] рассмотрели 40 различных метрик центральности, относящихся как к глобальным или локальным, так и метрики, ранее не рассматривавшиеся для оценки центральности в графах белок-белковых взаимодействий. Были рассмотрены две сети взаимодействий для белков дрожжей, полученных из разных баз данных. Результаты показали, что измерения центральности информации на основе маршрута и локализованной информации предсказывают важность белков в обеих сетях. И наоборот, предполагается, что меры глобальной центральности и меры, связанные с хабами (наиболее центральными белками сети), могут не подходить для выявления значимости белков. Кроме того, меры центральности локальной информации, охватывающие различные диапазоны, предоставляют релевантную информацию о важных узлах. Меры локализованной центральности, которые предполагают идеальные пути или случайные блуждания, показывают более слабую корреляцию со значимостью белков, чем меры информационной центральности. То есть те меры центральности, которые представляют сложность окружающей среды и учитывают локальную подсеть вокруг конкретного узла, являются лучшими мерами для прогнозиро-

вания важных узлов в сети белок-белковых взаимодействий. Основываясь на выводе о том, что меры центральности локализованной информации содержат наиболее важную информацию для прогнозирования существенности, был сделан вывод, что локальные плотные кластеры содержат важные узлы, поскольку влияние возмущения на кластеры может быть значительным на основаниях предположения, что сигнал проходит через несколько путей, использующих окружающую вершину среду, а не только кратчайший путь. Кроме того, результаты анализа кластеризации показывали, что определенные биологические процессы ассоциируются с определенными сетевыми кластерами, предполагая тесную взаимосвязь между конкретной топологией сети и биологической функцией. В заключение, было продемонстрировано, что клеточные функции, включая важность белков, тесно связаны с топологией сети.

Помимо топологических метрик центральности сетей белок-белковых взаимодействий Wuchty et al.[7] рассмотрели так называемые эволюционные метрики консервативности белков. В работе была предложена новая эволюционная метрика ER(evolutionary retention), которая позволяет выявить устойчивую и сильную корреляцию между консервативностью, значимостью и связностью белков дрожжей. Было показано, что сильно связанные белки с гораздо большей вероятностью будут важными и в то же время консервативными как ортологи у высших эукариот, чем менее связанные вершины графа. Сосредоточившись на независимой от эволюционного расстояния D мере ортологичного избыточного удерживания ER_k и подкорректировав безмасштабную статистику с помощью логарифмической группирования, был снижен уровень шума входных данных и обнаружена значительная корреляция между связностью и эволюционной консервативностью. Хотя более ранние подходы определения таких зависимостей сильно пострадали из-за используемых данных, предложенный метод в значительной степени нечувствителен к качеству входных данных, что также справедливо для несогласованности данных и шума.

3.2.2 Сети метаболических путей

В данном разделе будут кратко рассмотрены сети метаболических путей и топологические метрики, специфические для них. Сетью метаболических путей мы называем граф, следующего вида: вершинами графа являются метаболиты, участвующие в том или ином процессе и белки, которые используют эти метаболиты в качестве субстрата или продукта реакции. Метаболиты, участвующие в химической реакции в качестве субстрата для того или иного белка соединяются с вершиной, соответствующей этому белку, направленным ребром (от субстрата к белку), а продукты реакции - другим направленным ребром (от белка к продукту). По такому принципу выстраивается сложная разветвленная сеть метаболических путей.

Несмотря на значительные результаты по подбору оптимальных метрик центральности для сетей белок-белковых взаимодействий, использование тех же метрик для анализа метаболических сетей существенно проигрывало в точности такому методу, как анализ баланса потоков внутри клетки. Однако, Wunderlich et al. [8] предложили такую метрику, как синтетическая доступность для предсказания выживаемости штамма бактерии E.coli с определенными мутациями. Также было показано, что такие топологические метрики, как степень вершины, диаметр графа и степень посредничества не способны предсказывать летальность мутации на сети метаболических путей.

Топологическая метрика синтетическая доступность определяется следующим образом: рассмотрим метаболическую сеть, которая имеет доступ к определенным входам: субстратам, потребляемым из окружающей среды (например, сахару, кислороду и азоту), с целью производства определенных продуктов, таких как аминокислоты, нуклеотиды и другие компоненты, вместе называемые биомассой. Синтетическая доступность S_j выхода j определяется как минимальное количество метаболических реакций, необходимых для производства j из входных данных сети. $S_j = \infty$, если j не может быть синтезирован

из входных данных сети. Суммируя синтетическую доступность по всем компонентам биомассы, получаем общую синтетическую доступность биомассы $S = \sum_i S_i$. Мы предполагаем, что если нокаут фермента не изменяет S, то есть биомасса может быть произведена без дополнительных метаболических затрат, мутант является жизнеспособным. Если $S = \infty$, то есть по крайней мере один существенный компонент биомассы не может быть произведен из сетевых ресурсов, предсказывается летальный фенотип.

В этом исследовании было показано, что топология и функция метаболической сети тесно связаны. Введя новую меру, основанную на топологии, синтетическую доступность, удалось правильно предсказать жизнеспособность 443 из 598 мутантных штаммов E. coli на основе всеобъемлющего надежного датасета и 3477 из 4154 мутантных штаммов дрожжей, выращенных в нескольких условиях. Синтетическая доступность, S, по сути, представляет собой диаметр сети, специально предназначенный для транспортных сетей. Было показано, что увеличение S коррелирует с нежизнеспособным фенотипом. Значительное увеличение S при мутации предполагает увеличение метаболических затрат, что приводит к снижению скорости роста или смерти. Очевидный успех синтетической доступности можно объяснить только вкладом сетевой топологии, потому что никакая другая информация не использовалась в этих прогнозах.

3.3 Connectivity Map

3.3.1 Принцип метода

Прежде чем описывать метод Connectivity Map, стоит дать определение понятию сигнатура. Под сигнатурой мы понимаем два списка: список генов с повышенной экспрессией и список генов с пониженной экспрессией, полученных на основе анализа дифференциальной экспрессии.

Connectivity Map - это ресурс, позволяющий сравнивать полученные сигнатуры с сигнатурами из базы данных и находить определенные взаимосвязи [9]. Принцип работы данного метода изображен на рисунке ниже.

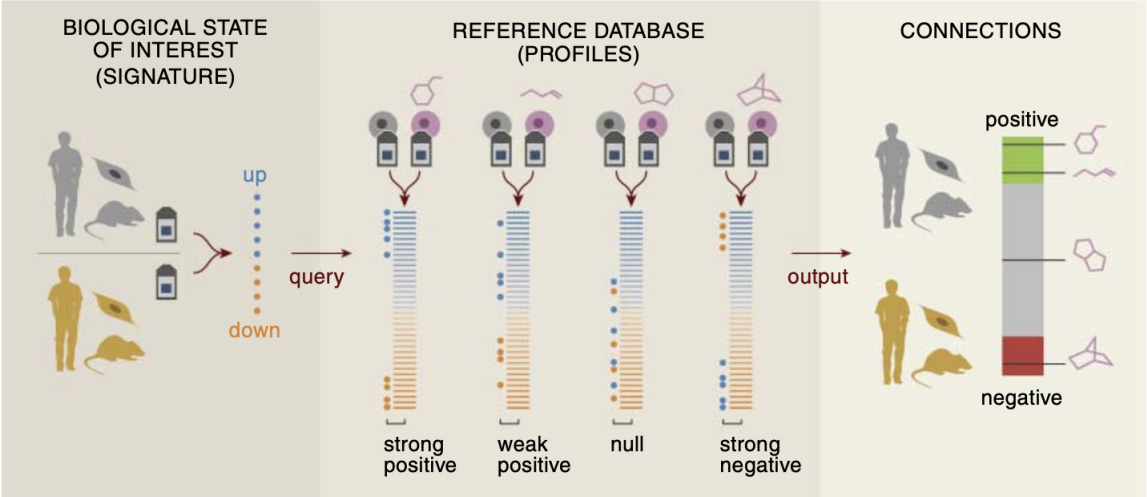


Рисунок 3.3 – Принцип метода Connectivity Map

На вход данному методу поступает список генов с повышенной экспрессии и список генов с пониженной экспрессией. Списки генов предварительно получают при анализе дифференциальной экспрессии генов между двумя состояниями, например, между здоровой и больной тканью. Далее выполняется запрос в референсную базу данных, где происходит сравнение сигнатуры запроса с известными сигнатурами. Для сигнатур, находящихся в базе данных, известны малые молекулы, способные обратить данную сигнатуру, то есть лекарства, которые могут быть применены для лечения того или иного заболевания, характеризующегося той или иной сигнатурой. После произведенного сравнения метод Connectivity Map выдает ранжированный список молекул с их метриками сходства, на основании которых можно судить о пригодности вещества для обращения сигнатуры.

Lamb et. al [9] описали метод Connectivity Map и провели ряд исследований для валидации предложенного метода. Было показано, что геномные сигнатуры можно использовать для распознавания лекарств с общими механизмами действия. Были найдены известные ингибиторы HDAC и модуляторы рецепторов эстрогена на основе сигнатур. Также было установлено, что метод хорошо работает для обнаружения неизвестных механизмов действия (была предсказана активность гедонина как ингибитора HSP90) и выявления по-

тенциальных новых терапевтических средств (обнаружена способность сиролимуса преодолевать резистентность к дексаметазону при остром лимфобластном лейкозе). Результаты также показывают, что сигнатуры часто сохраняются в различных типах клеток и в различных условиях (сигнатура устойчивости к дексаметазону была определена в образцах костного мозга, но поиск проводился по профилям из линии рака груди MCF7). В то же время результаты демонстрируют ограничения использования только нескольких клеточных линий (сигнатура эстрадиола не была обнаружена в клетках, лишенных рецепторов эстрогена) или только нескольких концентраций (хлорпромазин не распознавался как фенотиазин при 1 мМ).

3.3.2 Существующий инструмент

Duan et al. [10] представили инструмент, который называется L1000CDS2, с помощью которого можно произвести Connectivity Map и получить список малых молекул, которые могут обратить вспять или имитировать экспрессионную сигнатуру при болезни и других биологических условиях. Большой набор сигнатур, вычисленных методом CD (Characteristic Direction), предоставляется в виде современного веб-приложения. Помимо определения приоритета малых молекул для того, чтобы обратить или имитировать входную сигнатуру или предварительно вычисленные сигнатуры для 670 заболеваний и набора эндогенных лигандов, веб-инструмент поисковой машины L1000CDS2 также предсказывает попарные комбинации малых молекул, выполняет подструктурный анализ обогащения и вычисляет прогнозируемые цели на основе внешнего набора сигнатур. Так для валидации метода было выполнено предсказание лекарственной молекулы для лечения вируса Эбола на ранних стадиях. Наиболее вероятная молекула, обращающая сигнатуру клеток, инфицированных вирусом Эбола, кенпауллон, как было показано, ослабляет инфекцию дозозависимым образом, не вызывая клеточной токсичности в двух линиях клеток. Были предсказаны задействованные гены-мишени и сигнальные пути клеток,

которые указывали на гены иммунного ответа, управляемые ингибированием путей CDK1-2 и GSK3B, и потенциально активирующие передачу STAT сигналов.

4 Материалы и методы

4.1 Сырые данные и анализ дифференциальной экспрессии

В качестве входных данных инструмента используется нормализованная матрица количества прочтений, полученная из ресурса ARCHS 4 [11] для каждого рассматриваемого перехода. С помощью широко используемого инструмента edgeR [12] был выполнен анализ дифференциальной экспрессии и получены данные анализа дифференциальной экспрессии.

4.2 Построение генных сетей и подсчет метрик центральности

4.2.1 Получение данных о взаимодействии

Для получения информации о взаимодействии генов внутри сигнатуры проводятся API запросы в базы данных STRING [13] и BioGRID [14]. API запросы производились с помощью пакета requests языка Python. В этих базах данных содержится информация о белок-белковых взаимодействиях, как физических и регуляторных (воздействия транскрипционных факторов на гены), так и предсказанных взаимодействиях. Помимо информации о взаимодействии, в базе данных STRING [13] содержатся так называемые коэффициенты достоверности, показывающие насколько вероятно взаимодействие между белками или генами.

4.2.2 Построение генных сетей

С помощью эффективного и действенного пакета под названием Graph-Tool [15] языка Python, строятся генные сети, а также вычисляются метрики центральности. Более конкретно, метрики центральности, такие как pagerank centrality, betweenness centrality, eigenvector centrality, closeness centrality, Katz centrality, Hits centrality, eigentrust, вычисляются с помощью соответствующих функций пакета на основании реконструированного неориентированного графа белок-

белковых и ген-генных взаимодействий.

4.3 Нормализация

Стандартная нормализация, использовавшаяся для уменьшения мат. ожидания и дисперсии log FC была взята из пакета scikit-learn [16] языка Python. Была использована функция StandardScaler, которая центрирует и нормализуют входную матрицу (делает мат.ожидание равным нулю и стандартное отклонение равным единице).

4.4 Статистический анализ

Для оценки статистической значимости получаемых результатов был использован тест Колмогорова-Смирнова, реализованный средствами пакета scipy языка Python в виде функции ks_2samp. Данная функция вычисляет статистику Колмогорова-Смирнова для двух выборок. Это двусторонняя проверка нулевой гипотезы о том, что две независимые выборки взяты из одного и того же непрерывного распределения.

4.5 Байесовская оптимизация

Байесовская оптимизация проводилась с помощью пакета skopt языка Python функцией gp_minimize. Так как в ходе реализации инструмента стояла задача максимизации статистики Колмогорова-Смирнова, а функция gp_minimize предназначена для минимизации оптимизируемой функции, то для оптимизации использовалась следующая функция: - статистика Колмогорова-Смирнова.

4.6 Оценка качества модели

Метрика GSEA enrichment score [17] использовалась для оценки качества ранжирования модели. Она показывает насколько представлен выбранный список в ранжированном списке. Метрика GSEA рассчитывается следующим образом:

- Упорядочивается N генов в D, чтобы сформировать $L = \{g_1, ..., g_N\}$ в соответствии с корреляцией $r(g_j) = r_j$ их профилей экспрессии с C.

- Оценивается доля генов в S («совпадениях»), взвешенная по их корреляция и доля генов, не входящих в S («промахи»), присутствующих до данной позиции i в L.

$$P_{hit}(S, i) = \sum_{\substack{g_j \in S \\ j \leqslant i}} \frac{|r_j|^p}{N_R}, \qquad \text{where} \quad N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{miss}(S, i) = \sum_{\substack{g_j \notin S \\ j \leqslant i}} \frac{1}{(N - N_H)}$$

ES - это максимальное отклонение $P_{hit} - P_{miss}$ от нуля. Для случайно распределенная S, ES (S) будет относительно небольшой, но если она сосредоточены в верхней или нижней части списка или иным образом распределены неслучайно, тогда ES (S) будет соответственно высоким. Когда $p = 0$, ES (S) сводится к стандартной статистике Колмогорова – Смирнова; когда $p = 1$, гены в S взвешены по их корреляции с C, нормированным на сумму корреляций по всем генам из S.

5 Полученные результаты

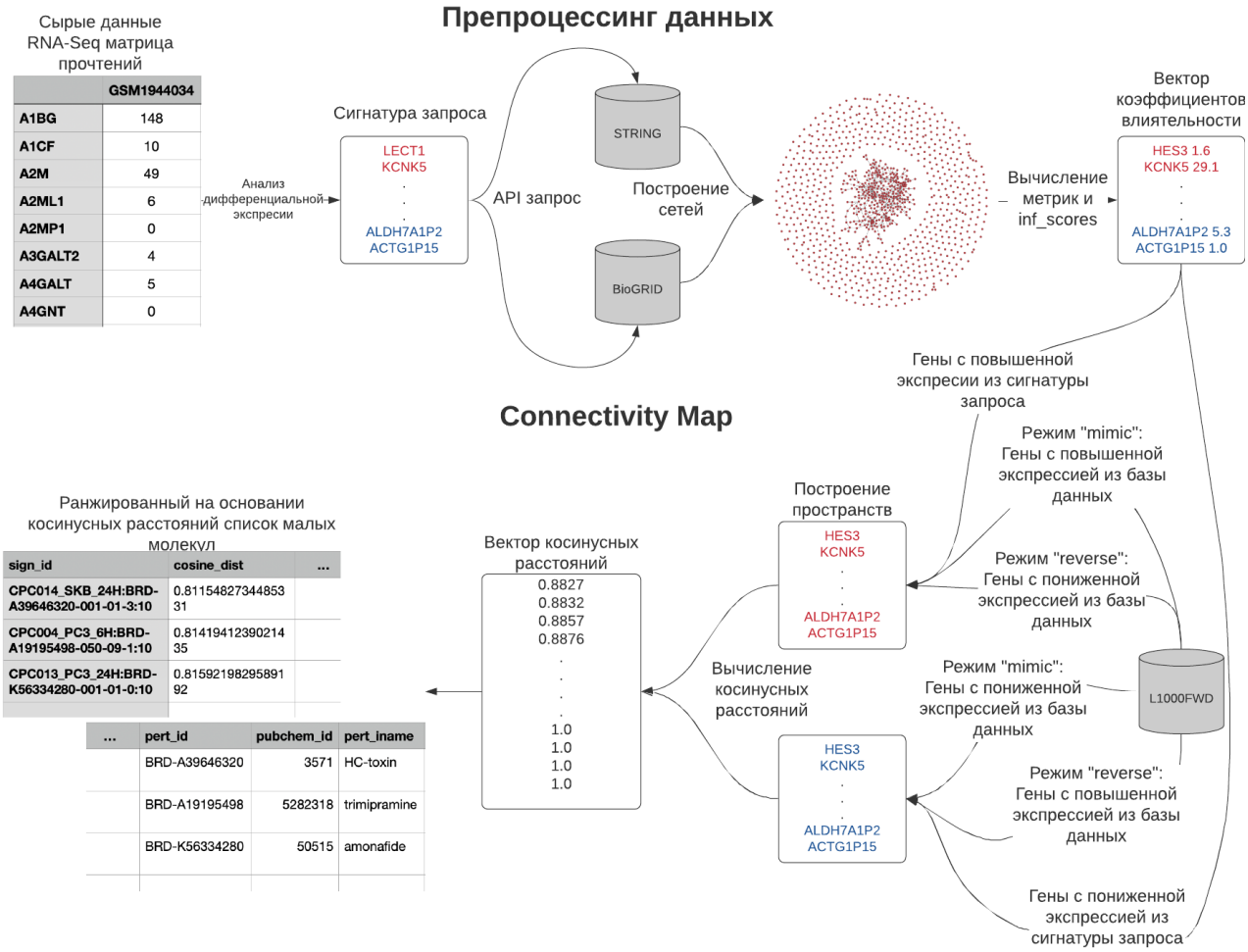


Рисунок 5.1 – Схема инструмента TopoCMap

5.1 Структура подхода

TopoCMap - инструмент, разработанный, для проведения анализа Connectivity Map, который учитывает биологическую значимость генов в экспрессионной сигнатуре путем расчета так называемого коэффициента влиятельности для каждого дифференциально экспрессированного гена. В основе инструмента лежат различные этапы обработки данных, которые представлены на рисунке выше. Далее в этом разделе будет подробно описан каждый из них.

5.1.1 Сырые данные и анализ дифференциальной экспрессии

Используя фильтрацию генов по полученным в ходе анализа дифференциальной экспрессии значениям $\log FC$ и p_value, была получена экспрессионная сигнатура, состоящая из двух списков генов: генов с повышенной экспрессией (up) и генов с пониженной экспрессией (down). На этом этапе использовались следующие пороги по $\log FC$ и p_value: $|\log FC| \geqslant 1.5$ и $p\_value \leqslant 10^{-3}$. Полученную сигнатуру будем в дальнейшем называть сигнатурой запроса.

5.1.2 Построение генных сетей и подсчет коэффициентов влиятельности

Для оценки биологической значимости генов для определенного клеточного состояния, используется анализ сетей белок-белковых и ген-генных взаимодействий.

5.1.3 Построение генных сетей

На основании информации, полученной из STRING и BioGRID строятся генные сети. На основании построенных сетей рассчитываются топологические метрики центральности, такие как pagerank centrality (pr), betweenness centrality (bw), eigenvector centrality (ev), closeness centrality (cl), Katz centrality (kz), Hits centrality (hs), eigentrust (et). Метрики центральности позволяют определить так называемые белки-хабы, которые по сути являются наиболее связанными вершинами в сети белок-белковых взаимодействий. Эти метрики были выбраны на основе анализа ряда статей [2—8], в которых подробно описывается корреляция между этими метриками и важностью белков в сети белок-белковых взаимодействий. В перечисленных статьях также подробно описывалась летальность мутаций для клетки в одном из белков-хабов.

5.1.4 Вычисление коэффициентов влиятельности

Затем метрики центральности используются для расчета коэффициентов влиятельности для каждого гена в сигнатуре запроса. Также при вычислении коэффициентов влиятельности используется кратность изменения (FC), поскольку этот параметр количественно выражает степень изменения экспрессии гена. Предварительно FC были прологарифмированы, для уменьшения среднеквадратичного отклонения, а также нормализованы. Нормализация была проведена для унификации масштабов топологических метрик центральности, которые находятся в диапазоне от 0 до 1, и FC. Коэффициент влиятельности для гена i вычисляется как комбинация метрик:

$$inf\_score_i = \begin{cases} (a_1 \cdot \log FC + b_1) \cdot (a_2 \cdot \text{pr} + b_2) \cdot (a_3 \cdot \text{bw} + b_3) \\ \cdot (a_4 \cdot \text{ev} + b_4) \cdot (a_5 \cdot \text{cl} + b_5) \cdot (a_6 \cdot \text{kz} + b_6) \\ \cdot (a_7 \cdot \text{hs} + b_7) \cdot (a_8 \cdot \text{et} + b_8) + a_9, \quad \text{if gene in STRING} \\ 1, \quad \text{otherwise} \end{cases}$$

где a_1, a_2, b_1, b_2 и тд. - числовые коэффициенты. Эти коэффициенты были подобраны на основании масштабной валидации на базе данных CFM [18], которая будет подробно описана ниже.

5.1.5 Вычисление уровня сходства сигнатур

В качестве сигнатур базы данных были взяты сигнатуры LINCS, использовавшиеся в сходном инструменте L1000FWD [19]. Оценка сходства сигнатуры запроса с сигнатурами базы данных рассчитывается как косинусное расстояние между генными векторами сигнатуры запроса и аналогичными векторами для сигнатуры из базы данных, усредненное для каждой пары. Генные вектора составляются следующим образом: генный вектор состоит из генов с повышенной (или пониженной) экспрессией, где на месте генов стоят коэффициенты

влиятельности. Каждый генный вектор раскладывается по следующему пространству: гены из сигнатуры запроса + гены из сигнатуры базы данных. Таким образом, в позициях генного вектора могу стоять следующие величины:

- inf_score_i, если ген есть в раскладываемом списке и для него доступна информация о взаимодействиях в базах данных STRING или BioGRID
- 1, если ген есть в раскладываемом списке, но для него нет информации о взаимодействиях в базах данных STRING или BioGRID
- 0, во всех остальных случаях

На основании вычисленных косинусных расстояний ранжируются сигнатуры базы данных и, как следствие, малые молекулы, вызывающие такого рода изменения экспрессии.

5.1.6 Подбор коэффициентов для расчёта коэффициента влиятельности

В этой главе будут подробно описаны все этапы валидации, приведенные на рисунке ниже.

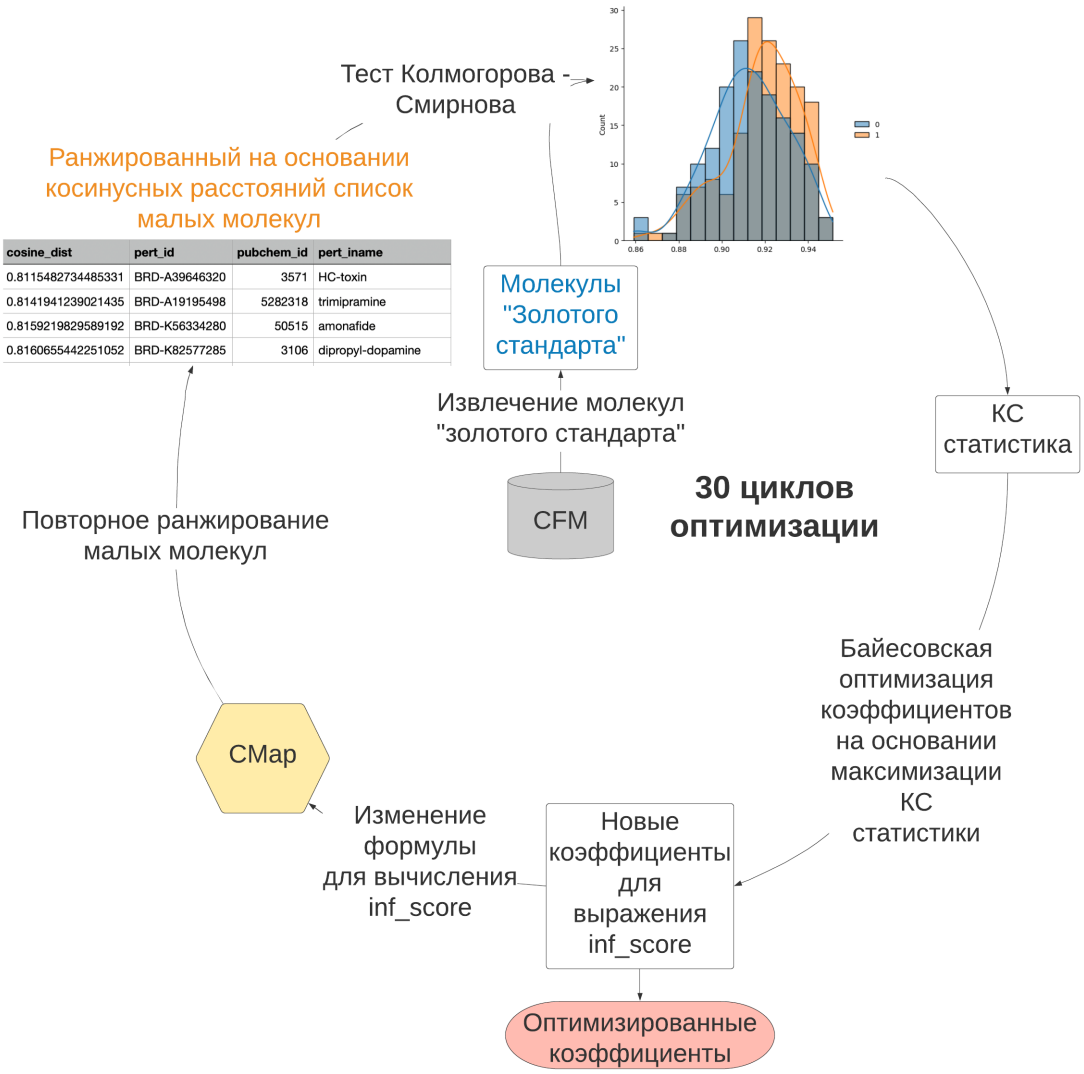


Рисунок 5.2 – Схема инструмента ТороCMap

5.1.7 Извлечение молекул ”золотого стандарта”

Для оптимизации необходимо сначала извлечь из доступных ресурсов некий ”золотой стандарт”, чтобы при последующей оптимизации была возможность оценивать точность модели. Для этих целей была использована база данных CFM [18], в которой содержится информация о малых молекулах, вызывающих трансдифференцировку между различными клеточными типами, например, фибробластами и кардиомиоцитами. В CFM содержится список экспериментальных протоколов из различных опубликованных исследований по химическому перепрограммированию клеток. Для каждого клеточного перехода был составлен список соединений ”золотого стандарта”, а также туда были до-

бавлены молекулы похожие на соединения ”золотого стандарта”. Поиск таких соединений был осуществлен на основе данных, содержащихся в L1000FWD [19]. Сходство молекул вычислялось с помощью коэффициента Танимото, также известного как мера Жаккара.

5.1.8 Оптимизация коэффициентов для расчета коэффициента влиятельности

Для оптимизации коэффициентов в выражении для расчета коэффициента влиятельности была выбрана Байесовская оптимизация. Данный вид оптимизации был выбран, так как он наиболее хорошо подходит для поиска минимума аналитически неизвестной функции. В качестве функций минимизации была выбрана функция статистика Колмогорова-Смирнова, которая показывает расстояние между распределениями. В качестве анализируемых распределений рассматривались распределение косинусных расстояний случайного набора малых молекул, не являющихся ”золотым стандартом”, а также распределение косинусных расстояний для соединений из ”золотого стандарта” для конкретного перехода. Метрика GSEA enrichment score [17] использовалась для оценки качества ранжирования модели. В нашем случае мы использовали в качестве ранжированного списка - список всех малых молекул, ранжированный на основании косинусных расстояний, а в качестве выбранного - список молекул ”золотого стандарта” для интересующего клеточного перехода. Байесовская оптимизация состояла из 30 итераций на 15 начальных точках. Результаты оптимизаций на различных клеточных переходах были усреднены и приняты за финальные коэффициенты влиятельности.

5.1.9 Результаты оптимизации

Описанная в методах оптимизация производилась на шести трансдифференцировках, а именно:

- Фибробласты (fb) → Индуцированные кардиомиоциты (heart)

- Фибробласты → Индуцированные нейроны (neuron)
- Фибробласты → Индуцированные нейральные стволовые клетки (neural)
- Фибробласты → Индуцированные панкреатические бета клетки (beta)
- Фибробласты → Индуцированные плюрипотентные стволовые клетки (ips)
- Мезенхимальные стволовые клетки (mes) → Индуцированные нейроны

В таблице 5.1 представлена статистика по каждому из переходов. Видно, что больше всего соединений и сигнатур было доступно для переходов из фибробластов в кардиомиоциты и нейроны. Для каждого перехода, представленного в таблице 5.1, в результате оптимизации был получен вектор оптимальных коэффициентов. В таблице 5.2 представлены оптимальные коэффициенты в зависимости от перехода.

После оптимизации для каждой клеточной конверсии был получен ранжированный список малых молекул, которые предположительно должны вызывать такой переход. Статистика по каждому переходу представлена в таблице 5.3. Из этой таблице можно сделать вывод, что оптимизация оказалась наиболее удачной для перехода из фибробластов в индуцированные плюрипотентные стволовые клетки, так как большое количество соединений из ”золотого стандарта” было приоритезировано методом, а также первое вхождение соединения ”золотого стандарта” было на 20-ой позиции (tretinoin).

Таблица 5.1 – Статистика по трансдифференцировкам

Переход	Протоколы	Количество соединений из ”золотого стандарта”	Количество сигнатур для соединений из ”золотого стандарта”
Фибробласты ↓ Индуцированные кардиомиоциты	19	42	3195
Фибробласты ↓ Индуцированные нейроны	8	27	1249
Фибробласты ↓ Индуцированные нейральные стволовые клетки	6	17	1278
Фибробласты ↓ Индуцированные панкреатические бета клетки	6	18	394
Фибробласты ↓ Индуцированные плюрипотентные стволовые клетки	4	13	953
Мезенхимальные стволовые клетки ↓ Индуцированные нейроны	5	14	548

Таблица 5.2 – Оптимальные коэффициенты для каждого перехода

Метрика	fb to heart	fb to neuron	fb to beta	fb to ips	fb to neural	mes to neuron
logFC	8,332	9,981	3,589	1,725	6,713	1,0
pagerank	1,000	4,821	7,958	4,682	4,527	7,125
betweenness	5,873	8,342	7,639	7,977	2,543	9,464
eigenvector	7,344	10,0	10,0	3,511	6,146	4,890
closeness	1,0	10,0	10,0	6,253	1,625	9,919
katz	4,073	1,102	10,0	2,196	2,170	6,324
hits	1,060	10,0	9,438	1,178	3,481	9,475
eigentrust	5,939	8,487	5,641	9,412	7,990	3,988
Свободный член	1,0	0,544	0,551	0,722	0,809	0,665

Таблица 5.3 – Результаты валидации для оптимальных коэффициентов для каждого перехода

Метрика качества	fb to heart	fb to neuron	fb to beta	fb to ips	fb to neural	mes to neuron
среднее значение p_value	0,262	0,531	0,164	0,095	0,370	0,298
p_value методом GSEA	0,271	0,0543	0,038	0,061	0,292	0,096
Первое вхождение соединения из ”золотого стандарта”	207	587	121	20	137	526
Количество соединений ”золотого стандарта” в топ 5% списка	10	10	9	15	8	4

5.1.10 Результаты усреднения коэффициентов

После проведенной оптимизации оптимальные для каждого перехода коэффициенты были усредененны. В результате получились следующие коэффициенты (таблица 5.4):

Таблица 5.4 – Усредненные оптимальные коэффициенты

logFC	pr	bw	ev	cl	kz	hs	et	free coeff
5,233	5,019	6,973	6,982	6,466	4,311	5,772	6,909	0,715

Исходя из полученных значений можно сделать следующие выводы: оптимальные коэффициенты, полученные для различных переходов слишком вариабельны - сложно сделать вывод, какая из метрик центральности вносит наибольший вклад; несмотря на большой разброс значений метрик, более приоритетными всё же являются метрики: eigenvector, betweenness, eigentrust, closeness.

Также была проведена валидация инструмента с усредненными коэффициентами для рассматриваемых переходов. Результаты валидации представлены в таблице 5.5.

Таблица 5.5 – Результаты валидации для усредненных коэффициентов для каждого перехода

Метрика качества	fb to heart	fb to neuron	fb to beta	fb to ips	fb to neural	mes to neuron
среднее значение p_value	0,749	0,552	0,623	0,142	0,484	0,583
p_value методом GSEA	0,395	0,153	0,116	0,029	0,182	0,329
Первое вхождение соединения из ”золотого стандарта”	97	258	120	64	84	1029
Количество соединений ”золотого стандарта” в топ 5% списка	9	6	7	11	7	3

Видно, что после усреднения результаты немного ухудшились: возросло p_value, уменьшилась статистическая значимость появления соединений ”золотого стандарта” в топе ранжированного списка (p_value, определенное методом GSEA уменьшилось), уменьшилось число соединений ”золотого стандарта” в топ 5%, однако примечательно, что ранг первого вхождения соединения из ”золотого стандарта” возрос.

В целом, можно сделать вывод, что усреднение коэффициентов привело к ухудшению качества предсказаний модели, однако это не слишком искажает общую картину.

5.2 Валидация инструмента

5.2.1 Анализ полученных списков для оптимальных коэффициентов

Из-за того, что база данных CFM не на 100% пересекается с базой сигнатур в ответ на малые молекулы L1000FWD, часть соединений ”золотого стандарта” не могут быть отранжированы. Более того, каждое соединение встречается в списке более одного раза, так как в действительности инструмент ранжирует не вещества, а сигнатуры, вызванные этим веществом. Исходя из этого, полезно ввести метрику, такую как отношение количества соединений из пересечения CFM и L1000FWD в топ 5% списка к общему числу молекул на пересечении CFM и L1000FWD для конкретного перехода. В таблице 5.6 представлены результаты расчета данной метрики.

Таблица 5.6 – Отношение числа соединений из ”золотого стандарта” в топ 5% к общему числу соединений из ”золотого стандарта”

fb to heart	fb to neuron	fb to beta	fb to ips	fb to neural	mes to neuron
4 из 19 (21,1 %)	5 из 10 (50%)	5 из 11 (45,5%)	4 из 8 (50%)	2 из 5 (40%)	3 из 8 (37,5%)

5.2.2 Поиск соединений по механизму действия

Помимо этого, так как в качестве ”золотого стандарта” используются соединения из протоколов статей по клеточным конверсия, то стоит предположить, что не все соединения с такого рода активностью вошли в наш ”золотой стандарт”. Для этого был проведен анализ механизмов действия для соединений ”золотого стандарта” для каждого клеточного перехода. В таблицах 5.7 - 5.12 приведены механизмы действия соединений ”золотого стандарта”. Из этих таблиц можно

сделать вывод, что наиболее распространенными механизмами действия для переходов являются:

- Агонисты рецепторов ретиноевой кислоты и ингибиторы пути TGF-β для перехода из фибробластов в кардиомиоциты
- Ингибиторы ДНК метилтрансфераз для перехода из фибробластов в панкреатические бета клетки
- Ингибиторы пути TGF-β для перехода из фибробластов в нейроны
- Агонисты рецепторов ретиноевой кислоты для перехода из фибробластов в индуцированные плюрипотентные стволовые клетки
- Ингибиторы пути TGF-β для перехода из фибробластов в нейральные стволовые клетки
- Ингибиторы пути TGF-β для перехода из мезенхимальных стволовых клеток в нейроны

Также был проведен анализ механизмов действия соединений в топ 50 для каждого клеточного перехода. Таблицы анализа представлены в приложениях 9.1 - 9.6. Для большинства переходов, таких как: фибробласты - кардиомиоциты, фибробласты - панкреатические бета клетки, фибробласты - индуцированные плюрипотентные стволовые клетки, в топ 50 были найдены соединения с аналогичной активностью, что и активность некоторых соединений ”золотого стандарта”. Далее немного подробнее будут рассмотрены эти соединения.

Для перехода из фибробластов в бета клетки в топ 50 соединений были найдены вещества со следующими активностями: ингибиторы HDAC (HC-toxin, vorinostat, trichostatin-a). В топ 50 они встречаются 5 раз. Для этого перехода в ”золотом стандарте” содержится ингибитор HDAC - romidepsin. Кроме того, в ”золотом стандарте” есть ингибиторы GSK-3α/β, ингибитор моно-

амин оксидазы, агонисты глюкокортикоидного рецептора и ингибиторы ДНК-метилтрансферазы. В топ 50 были найдены вещества с такой же активностью: TWS-119 (ингибитор GSK-3β), salsolinol (ингибитор моноамин оксидазы), beta-methasone (агонист глюкокортикоидного рецептора), temozolomide(ингибиторы ДНК - метилтрансферазы).

Для перехода из фибробластов в кардиомиоциты в топ 50 соединений были найдены вещества со следующими активностями: ингибиторы GSK-3β (NSC-693868, indirubin). Для этого перехода в ”золотом стандарте” содержится ингибитор GSK-3β - CHIR-99021. Кроме того, в ”золотом стандарте” есть ингибиторы сигнального пути TGF-β. Хотя в топ 50 не нашлось ингибиторов TGF-β и его рецепторов, в топ 50 нашлось большое количество ингибиторов белков-трансдукторов сигнала от TGF-β-рецептора. Например, TGF-β запускает сигнальный путь NF$\varkappa$B; в топ 50 был найден parthenolide, который является ингибитором сигнального пути NF$\varkappa$B. Также TGF-β запускает MAPK-киназный каскад посредством активации малой ГТФазы RAF, инструментом также были найдены 2 ингибитора RAF ГТФазы (SB-590885, GW-5074).

Таблица 5.7 – Механизм действия соединений ”золотого стандарта” для перехода из фибробластов в кардиомиоциты

Ранг	CID	Название	Механизм действия
93	444795	Retinoic acid	агонист рецептора ретиноевой кислоты
587	4521392	SB-431542	ингибитор пути Activin/BMP/TGF-β
951	11238147	ICG-001	ингибитор Wnt
987	5289501	TTNPB	агонист рецептора ретиноевой кислоты
1316	25150857	BIX-01294	ингибитор G9a гистон метилтрансферазы
1498	5092	rolipram	ингибитор фофсодиестеразы типа 4 (PDE4)
2810	11524144	dorsomorphin	ингибитор ALK2, ALK6 и AMPK
3073	448042	Y-27632	ингибитор ROCK1 и ROCK2
4460	447966	LY-364947	TGF-b сигнальный путь
4718	3121	valproic-acid	ингибитор гистон деацетилазы
4951	47936	forskolin	CAMP агонист
7073	9826528	PD-0325901	ингибитор MEK/ERK сигнального пути
9236	9956119	CHIR-99021	ингибитор GSK3β

Для перехода из фибробластов в индуцированные плюрипотентные стволовые клетки в топ 50 соединений были найдены вещества со следующими активностями: агонисты рецептора ретиноевой кислоты (fenretinide, isotretinoin, tretinoin). Для этого перехода в ”золотом стандарте” содержатся агонисты рецептора ретиноевой кислоты - AM580, Retinoic acid. Кроме того, в ”золотом стандарте” есть ингибиторы сигнального пути TGF-β. Хотя в топ 50 не нашлось ингибиторов TGF-β и его рецепторов, в топ 50 нашлось большое количество ингибиторов белков-трансдукторов сигнала от TGF-β-рецептора. Например, TGF-β запускает сигнальный путь NF$\varkappa$B; в топ 50 был найден IKK-2-inhibitor-V, который является ингибитором сигнального пути NF$\varkappa$B. Также TGF-β запускает PI3K-Akt-киназный каскад посредством активации PI3K киназы, инструментом также были найдены 2 ингибитора PI3K киназы (wortmannin, AS-605240) и ингибитор Akt киназы (triciribine). В свою очередь Akt киназа активирует белковый комплекс mTORC, ингибитор которого также был найден ин-

струментом (sirolimus). Аналогичные результаты были получены и для других трансдифференцировок.

Таблица 5.8 – Механизм действия соединений ”золотого стандарта” для перехода из фибробластов в бета клетки

Ранг	CID	Название	Механизм действия
13128	9444	Azacitidine	ингибитор ДНК (цитозин-5)-метилтрансферазы 1
	5352062	Romidepsin	селективный ингибитор гистон деацетилазы
	5222465	Sodium butyrate	холинэстераза
	19493	Parnate	ингибитор моноамин оксидазы
601	702558	RG-108	RG108 ингибитор ДНК метилтрансферазы
2150	9956119	CHIR-99021	ингибитор GSK-3α и GSK-3β; ингибитор киназы гликоген синтазы
5091	222786	Cortisone	агонист глюкокортикоидного рецептора; индуктор аннексина 1
	24776445	Vismodegib	ингибитор трансмембранного белка-гомолога Smoothened (SMO) в сигнальном пути Hedgehog
1448	25195294	LDN-193189	селективный ингибитор ALK2 и ALK3
15909	936	Nicotinamide	продукт АДФ-рибозил циклазы 2

2652	25150857	BIX-01294	ингибитор гистон метилтрансферазы
889	444795	Tretinoin	природная производная витамина А
86	24775005	Erismodegib	ингибитор сигнального пути Hedgehog
	10990876	Sodium ascorbyl phosphate	синтетическая форма витамина С
1892	5743	Dexamethasone	агонист глюкокортикоидного рецептора; стимулятор ядерного рецептора; агонист аннексина А1; индуцибельный отрицательный модулятор синтазы оксида азота

Таблица 5.9 – Механизм действия соединений ”золотого стандарта” для перехода из фибробластов в нейроны

Ранг	CID	Название	Механизм действия
	52912189	I-BET151	Inhibitor of the BET family
1366	448042	Y-27632	Inhibitor of ROCK1 and ROCK2
	16047442	KC7f2	Inhibitor of HIF 1-alpha
5935	4521392	SB431542	Sonic hedgehog
742	25195294	LDN193189	Inhibitor of TGF-beta/Smad signaling
1533	4878	PP2	активатор SIRT1 и ингибитор аминоресвератолсульфата и Src киназы
	449054	RepSox	ингибитор ALK5 (рецептор TGF-β типа 1)
556	11524144	Dorsomorphin	ингибитор ALK2, ALK6 и AMPK
1547	19582717	ISX9	Нейрогенный модулятор
9517	47936	Forskolin	CAMP агонист

Таблица 5.10 – Механизм действия соединений ”золотого стандарта” для перехода из фибробластов в индуцированные плюрипотентные стволовые клетки

Ранг	CID	Название	Механизм действия
9839	5743	Dexamethasone	активатор глюкокортикоидного пути
	56962337	SGC0946	ингибитор DOT1L метилтрансферазы
	19493	Parnate	ингибитор лизин-специфической деметилазы 1 (LSD1)
	449054	RepSox	ингибитор ALK5 (рецептор TGF-β типа 1)
1150	47936	Forskolin	агонист CAMP
334	2126	AM580	агонист рецептора ретиноевой кислоты
1092	451668	5-Aza-2’-deoxycytidine	ингибитор метилирования ДНК
20	444795	Retinoic acid	агонист рецептора ретиноевой кислоты

Таблица 5.11 – Механизм действия соединений ”золотого стандарта” для перехода из фибробластов в индуцированные нейральные стволовые клетки

Ранг	CID	Название	Механизм действия
1190	25150857	BIX01294	ингибитор G9a гистон метилтрансферазы
3070	702558	RG108	ингибитор ДНК метилтрансферазы
	46209426	Thiazovivin	ингибитор RHO/ROCK сигнального пути
137	25195294	LDN193189	Ингибитор TGF-beta/Smad сигналинга
	449054	RepSox	Ингибитор ALK5 (рецептор TGF-β типа 1)

Таблица 5.12 – Механизм действия соединений ”золотого стандарта” для перехода из мезенхимальных стволовых клеток в индуцированные нейроны

Ранг	CID	Название	Механизм действия
	52912189	I-BET151	ингибитор семейства BET
1124	47936	Forskolin	агонист CAMP
4230	4521392	SB431542	ингибитор Activin/BMP/TGF-β пути
526	448042	Y-27632	ингибитор ROCK1 и ROCK2
	5288382	Geldanamycin	ингибитор белка теплового шока 90
	9687	Bucladesine	активатор пути CAMP
4437	11524144	Dorsomorphin	ингибитор ALK2, ALK3, ALK6 и AMPK
	449054	RepSox	ингибитор ALK5 (рецептор TGF-β типа 1)

5.2.3 Сравнение инструмента с инструментом Connectivity Map

В ходе проделанной работы были проведены тесты инструмента на сигнатурах из статьи Lamb et al.[9]. Инструмент был опробован на 11 следующих сигнатурах:

- Сигнатура S1 - клетки T24 (мочевой пузырь), MDA 435 (карцинома груди) и MDA 468 (карцинома груди), обработанные тремя ингибиторами гистон-деацетилазы (HDAC): вориностатом (также известным как субероиланилидгидроксамовая кислота или SAHA), MS -27-275 и трихостатином A
- Сигнатура S2 - клетки MCF7 обрабатывали лигандом природного рецептора эстрогена (ER), 17β-эстрадиолом (E2)
- Сигнатура S3 - из пяти сигнатур в ответ на 5 различных фенотиазинов была создана одна сигнатура из генов, регулируемых всеми пятью соединениями (thioridazine, chlorpromazine, fluphenazine, trifluoperazine и prochlorperazine)
- Сигнатура S4 - сигнатура S6 - альтернативные внутренние фенотиазиновые сигнатуры
- Сигнатура S7 - сигнатура гедонина при лечении этим соединением клеток рака простаты LNCaP в течение 6 часов

- Сигнатура S8 - агонисты PPARγ, связанные с ожирением, вызванным специфическим питанием у крыс
- Сигнатура S9 - Сигнатура S10 - DAPH как потенциальное терапевтическое средство против AD
- Сигнатура S11 - чувствительности к дексаметазону, полученная путем сравнения лейкозных клеток костного мозга пациентов, проявляющих либо чувствительность к дексаметазону, либо резистентность к нему in vitro

Инструмент ToroCMap показал высокое качество предсказаний на данных сигнатурах. Так для сигнатуры S1 топ 100 молекул полученного списка имели активность ингибиторов HDAC, для сигнатуры S7 gendanamycin, который обладает аналогичной активностью, что и gedunin, встречается в топ 100 списка 8 раз. Аналогичные результаты были получены и для остальных сигнатур.

На рисунке 5.3 схематически представлены 2 ранжированных списка малых молекул, полученных для сигнатур S1 (слева) и S7 (справа), описанные выше.

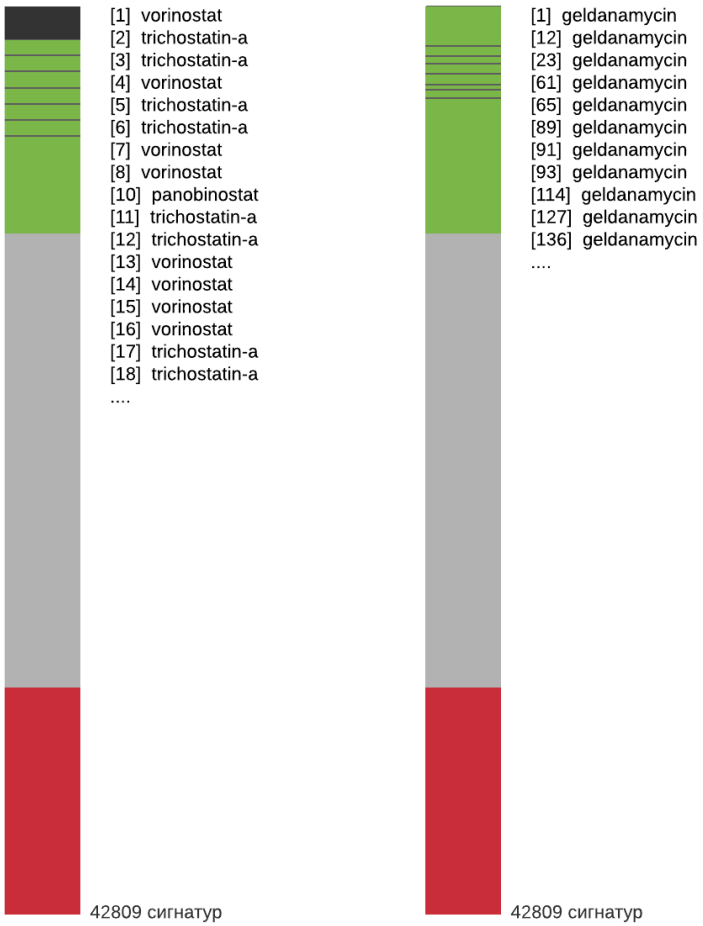


Рисунок 5.3 – Схематичное представление ранжированных списков малых молекул для сигнатур S1 (слева) и S7 (справа)

В базе сигнатур, применяющейся Lamb et al., было 453 сигнатур, большинство из которых были сигнатурами в ответ на соединения интереса из статьи. Также в этой базе сигнатур содержалось всего 3 клеточные линии, которые были весьма схожи по своей биологии, и одно время обработки.

В базе сигнатур ТороСМар содержится 42809 сигнатур в ответ на более чем 4000 соединений. Также в этой базе анализируется большое количество клеточных линий, различное время обработки химическими агентами, а также большое количество биологических копий экспериментов.

Из-за столь существенных различий между базами сигнатур инструмента ТороСМар и инструмента, описанного в статье Lamb et al., количественное сравнение подходов затруднено.

6 Заключение. План дальнейших исследований

В ходе работы была реализована модель инструмента, основанного на подходе CMap, а также анализе топологических метрик центральности сетей белок-белковых и ген-генных взаимодействий.

В процессе создания была проведена широкомасштабная оптимизация коэффициентов в выражении для коэффициентов влиятельности. После нахождения оптимальных коэффициентов для каждого клеточного перехода коэффициенты в выражении для inf_score были усреденены.

После оптимизации была произведена валидация полученного инструмента с помощью нашего ”золотого стандарта”, извлеченного из базы данных CFM, как для оптимальных коэффициентов для каждого перехода, так и для усредненных значений.

Помимо валидации, также была проверена применимость ToroCMap для поиска соединений исходя из их механизма действия на клетки. Было установлено, что инструмент весьма успешно приоритизирует соединения исходя из механизма действия на уровне транскриптома. Однако не все механизмы действия успешно приоритизируются нашим инструментом. Это может быть связано с нестабильностью транскриптомных сигнатур при воздействии одним и тем же химическим агентом на различные клеточные линии, при различном времени инкубирования клеточной линии в химическом агенте, а также вариабельностью транскриптомных данных от эксперимента к эксперименту.

Вдобавок, было проведено сравнение нашего инструмента с инструментом, предложенным в статье Lamb et al. [9]. Несмотря на ряд сложностей при сравнении инструментов, вызванных различиями баз данных сигнатур этих инструментов, было показано, инструмент ToroCMap показывает результаты не хуже тех, что представлены в статье. В некоторых случаях, наш инструмент показывает более высокие результаты за счет того, что вероятность случайно ранжировать вещества верным образом для нашего инструмента на порядок

ниже, чем для инструмента Lamb et al. Такое понижение вероятности случайности ранжирования вызвано существенным увеличением базы данных сигнатур нашего инструмента по сравнению с инструментом Lamb et al.

В дальнейшем планируется улучшить наш инструмент. Также планируется добиться статистической значимости предсказаний при валидации на клеточных переходах, описанных выше. Кроме того, планируется объединение ToroCMap с инструментом предсказания синергического эффекта на основе подхода CMap, также разрабатываемого в нашей лаборатории.

7 Благодарности

Хочу выразить огромную благодарность моему научному руководителю, Муртазалиевой Халимат Асадулаевне, за неисчерпаемый вклад в мою научную работу. Также хочу выразить благодарность заведующий лабораторией биоинформатики клеточных технологий, к. б. н. Медведевой Юлии Анатольевне и заместителю заведующего лабораторией, PhD, Ступникову Алексею Ильичу за содействие в проведении научной работы и написании дипломной работы.

8 Список используемых источников

[1] Thomas Engel Johann Gasteiger. *Chemoinformatics*. WILEY-VCH GmbH Co. KGaA, 2003.

[2] H.Jeong и др. «Lethality and centrality in protein networks». В: *Nature* (2001).

[3] Matthew W. Hahn и Andrew D.Kern. «Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks». В: *Molecular Biology and Evolution* (2005).

[4] N Pržulj, Dennis A Wigle и Igor Jurisica. «Functional topology in a network of protein interactions». В: *Bioinformatics* 20.3 (2004), с. 340—348.

[5] Maliackal Poulo Joy и др. «High-Betweenness Proteins in the Yeast Protein Interaction Network». В: *Journal of Biomedicine and Biotechnology* (2005).

[6] Keunwan Park и Dongsup Kim. «Localized network centrality and essentiality in the yeast–protein interaction network». В: *Proteomics* 9.22 (2009), с. 5143—5154.

[7] S Wuchty. «Topology and evolution in yeast interaction networks». В: *Genome Res* 14 (2004), с. 1310—1314.

[8] Zeba Wunderlich и Leonid A. Mirny. «Using the topology of metabolic networks to predict viability of mutant strains». В: *Biophysical journal* (2006).

[9] Justin Lamb и др. «The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease». В: *Science* (2006).

[10] 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.

[11] Alexander Lachmann и др. «Massive mining of publicly available RNA-seq data from human and mouse». В: *Nature Communications* (2018).

[12] Mark D. Robinson, Davis J. McCarthy и Gordon K. Smyth. «edgeR: a Bioconductor package for differential expression analysis of digital gene expression data». В: *Bioinformatics* (2010).

[13] Damian Szklarczyk и др. «STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets». В: *Nuclear acids research* (2018).

[14] Chris Stark и др. «BioGRID: a general repository for interaction datasets». В: *Nucleic Acids Research* (2006).

[15] Tiago P. Peixoto. «The graph-tool python library». В: *figshare* (2014). DOI: 10.6084/m9.figshare.1164194. URL: http://figshare.com/articles/graph_tool/1164194 (дата обр. 10.09.2014).

[16] F. Pedregosa и др. «Scikit-learn: Machine Learning in Python». В: *Journal of Machine Learning Research* 12 (2011), с. 2825—2830.

[17] Aravind Subramanian и др. «Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles». В: *PNAS* (2005).

[18] Sizykh A и др. «CFM: a database of experimentally validated protocols for chemical compound-based direct reprogramming and transdifferentiation». В: *F1000Research* (2021).

[19] Zichen Wang и др. «L1000FWD: fireworks visualization of drug-induced transcriptom signatures». В: *Bioinformatics* (2018).

9 Приложение

Анализ механизма действия топ 50 соединений для перехода из фибробластов в кардиомиоциты

Таблица 9.1 – Анализ механизма действия топ 50 соединений для перехода из фибробластов в кардиомиоциты

position	cid	name	moa
1	4792	phorbol-12-myristate-13-acetate	activates PKC
2	438981	NSC-693868	Cdk inhibitor; also inhibits GSK-3
3	60167550	serdemetan	MDM(nuclear-localized E3 ubiquitin ligase) inhibitor
4	25226483	HG-5-113-01	LOK(Serine/Threonine Kinase 10 Lymphocyte-Oriented Kinase), LTK (Leukocyte Receptor Tyrosine Kinase), TRCB, ABL(T315I)(ABL Proto-Oncogene 1, Non-Receptor Tyrosine Kinase)
5	2812	clotrimazole	Intermediate conductance calcium-activated potassium channel protein 4 inhibitor; Nuclear receptor subfamily 1 group I member 2 activator; Hydroxycarboxylic acid receptor 2 partial agonist; cytochrome P450 inhibitor, imidazoline receptor ligand

6	41684	nitazoxanide	pyruvate ferredoxin oxidoreductase inhibitor
7	3689416	K784-3188	-
8	3921152	SA-792574	microtubule inhibitor
9	5284329	purmorphamine	smoothened receptor agonist
10	2415	BML-190	cannabinoid receptor inverse agonist
11	3083928	fraxidin	-
12	44632017	YK-4279	apoptosis inhibitor
13	3899	leflunomide	dihydroorotate dehydrogenase inhibitor, PDGFR tyrosine kinase receptor inhibitor
14	6710658	estrone	Estrogen receptor agonist
15	4381125	K784-3131	-
16	42623900	QS-11	-
17	9806489	ZK-164015	estrogen receptor antagonist
18	5326739	indirubin	CDK inhibitor, glycogen synthase kinase inhibitor
19	4792	phorbol-12-myristate-13-acetate	activates PKC
20	2777391	T-0070907	PPAR receptor antagonist
21	3854666	SB-225002	CC chemokine receptor antagonist
22	680502	AMG-9810	TRPV(Transient Receptor Potential Cation Channel Subfamily V Member) antagonist
23	5757	estradiol	Estrogen receptor agonist

24	3396	fluspirilene	Dopamine D2 receptor antagonist; 5-hydroxytryptamine receptor 2A antagonist; Voltage-dependent calcium channel gamma-1 subunit inhibitor
25	3035714	ABT-751	tubulin polymerization inhibitor
26	40854	oxfendazole	anthelmintic agent
27	262093	SA-792541	CDC (cell division cycle) inhibitor
28	6450551	axitinib	VEGFR and kinase inhibitor
29	25195294	LDN-193189	bone morphogenic protein inhibitor
30	2378	bifonazole	sterol demethylase inhibitor
31	44142132	BRD-K51377689	-
32	5310849	KF-38789	P selectin inhibitor
33	9886086	RO-28-1675	glucokinase activator
34	3478439	fulvestrant	estrogen receptor antagonist
35	439654	scoulerine	antagonist in vitro at the α2-adrenoceptor, α1D-adrenoceptor and 5-HT receptor. It has also been found to be a GABAA receptor agonist in vitro.
36	107967	FK-888	tachykinin antagonist
37	24747177	BRD-K70415218	-
38	10000456	RS-102895	CCR (chemokine receptor family) antagonist
39	327653	pifithrin-mu	HSP inhibitor
40	46907806	BRD-K98896788	-
41	7251185	parthenolide	NFkB pathway inhibitor

42	1242010	BRD-K42419294	-
43	44142088	BRD-K96704648	-
44	5702107	prednisolone	glucocorticoid receptor agonist
45	11316960	SB-590885	RAF inhibitor
46	5757	estradiol	Estrogen receptor agonist
47	3685735	BRD-K10484463	-
48	5924208	GW-5074	RAF inhibitor, leucine rich repeat kinase inhibitor
49	44506645	BRD-K51556300	-
50	51003718	BRD-A59145032	-

Анализ механизма действия топ 50 соединений для перехода из фибробластов в нейроны

Таблица 9.2 – Анализ механизма действия топ 50 соединений для перехода из фибробластов в нейроны

position	cid	name	moa
1	9927531	BIBR-1532	telomerase inhibitor
2	1637653	BRD-K19499941	
3	7326481	BRD-K86574132	
4	3235086	BRD-K63569039	
5	5326739	indirubin	CDK inhibitor, glycogen synthase kinase inhibitor
6	16007391	barasertib	Aurora kinase inhibitor
7	23648036	BRD-K53653395	
8	73707396	artesunate	DNA synthesis inhibitor
9	73707438	YL-54	
10	6710659	fludrocortisone	Mineralocorticoid receptor agonist; Glucocorticoid receptor agonist
11	13017911	geldanamycin	HSP inhibitor
12	60138075	BRD-K59222562	
13	4911	probenecid	uricosuric blocker
14	10367662	KI-16425	lysophosphatidic acid receptor antagonist
15	4728	penicillic-acid	

16	6708801	austricine	A hypolipidemic, antiatherosclerotic, anti-inflammatory agent, with angioprotective and hepatoprotective activity; antiulcerous.
17	2734756	hydroquinine	sodium channel blockade, beta-adrenergic blockade, repolarization prolongation, or calcium channel blockade.
18	151170	cilomilast	phosphodiesterase inhibitor
19	44489862	BRD-K00664012	
20	24180719	PLX-4720	RAF inhibitor
21	42623900	QS-11	
22	4065	meprylcaine	local anesthetic
23	2432214	BRD-K39345836	
24	60138116	BRD-K05979026	
25	3062316	dasatinib	Bcr-Abl kinase inhibitor, ephrin inhibitor, KIT inhibitor, PDGFR tyrosine kinase receptor inhibitor, src inhibitor, tyrosine kinase inhibitor
26	2442307	BRD-K79390395	
27	374536	RITA	MDM inhibitor
28	3796	KIN001-055	
29	5494449	tozasertib	Aurora kinase inhibitor, Bcr-Abl kinase inhibitor, FLT3 inhibitor, JAK inhibitor

30	5288209	fenretinide	apoptosis stimulant, retinoid receptor agonist
31	11054313	KU-C103867	
32	5691	wortmannin	PI3K inhibitor
33	44506424	BRD-K86027709	
34	5702239	naltrexone	opioid receptor antagonist
35	5337366	SA-25547	
36	60749	gemcitabine	ribonucleotide reductase inhibitor
37	3476986	blebbistatin	ATPase inhibitor
38	2850562	fatostatin	SREBP inhibitor
39	5877	depomedrol	Glucocorticoid receptor agonist; Annexin A1 agonist
40	9926999	LY-456236	glutamate receptor antagonist
41	24039271	BRD-A95820578	
42	702558	RG-108	DNA methyltransferase inhibitor
43	24180719	PLX-4720	RAF inhibitor
44	2775426	BRD-K85275009	
45	10042240	GYKI-52466	glutamate receptor antagonist
46	9549297	SU-11274	hepatocyte growth factor receptor inhibitor, tyrosine kinase inhibitor
47	9840076	AG-14361	PARP inhibitor
48	5691	wortmannin	PI3K inhibitor
49	60923	quiflapon	leukotriene synthesis inhibitor
50		trequinsin	phosphodiesterase inhibitor

Анализ механизма действия топ 50 соединений для перехода из фибробластов в бета клетки

Таблица 9.3 – Анализ механизма действия топ 50 соединений для перехода из фибробластов в бета клетки

position	cid	name	moa
1	3571	HC-toxin	an HDAC inhibitor
2	5282318	trimipramine	decreasing the reuptake of norepinephrine and serotonin (5-HT)
3	50515	amonafide	inhibitor of topoisomerase II
4	3106	dipropyl-dopamine	Dopamine receptor agonist.
5	6839	phensuximide	Phensuximide’s mechanism of action not understood, but may act in inhibitory neuronal systems that are important in the generation of the three per second rhythm. It’s effects may be related to its ability to inhibit depolarization-induced accumulation of cyclic AMP and cyclic GMP in brain tissue.
6	13017911	geldanamycin	Heat shock protein HSP 90-beta; Heat shock protein HSP 90-alpha; Endoplasmin inhibitor
7	262093	SA-792541	CDC inhibitor
8	20862025	BRD-K69676861	

9	19385000	L-755507	potent and selective β-3 adrenergic receptor agonist
10	45359153	BRD-K59460069	inhibits LPS- or concanavalin A-induced proliferation of isolated mouse splenic lymphocytes. Sydowinin B inhibits superoxide generation in, and elastase release from, isolated human neutrophils induced by cytochalasin B
11	9906875	LFM-A13	Bruton’s tyrosine kinase (BTK) inhibitor
12	6419979	salsolinol	Type A monoamine oxidase
13	19582717	CHEMBL-1222381	
14	5284590	montelukast	leukotriene receptor antagonist
15	5344054	PAC-1	caspase activator
16	1400	PP-1	Tyrosine-protein kinase HCK; Proto-oncogene tyrosine-protein kinase receptor
17	45006158	betamethasone	Glucocorticoid receptor agonist
18	44552569	BRD-K64979116	
19	9705	AY-9944	hedgehog pathway modulator
20	5394	temozolomide	methylates DNA at the N7 position of guanine (N7-MeG, 70%), the N3 position of adenine (N3-MeA, 9%), and the O6 position of guanine (O6-MeG, 6%).

21	5311	vorinostat	an HDAC inhibitor
22	6376322	trichostatin-a	an HDAC inhibitor
23	9549289	TWS-119	a GSK-3β inhibitor ; glycogen synthase kinase inhibitor
24	3101	diphenyleneiodonium	nitric oxide synthase inhibitor
25	5691	wortmannin	inhibitor of phosphoinositide 3-kinase enzymes
26	54911	nefazodone	nefazodone acts as an antagonist at type 2 serotonin (5-HT2) post-synaptic receptors and, like fluoxetine-type antidepressants, inhibits pre-synaptic serotonin (5-HT) reuptake.
27	2928239	BRD-A36059655	
28	9982218	NTNCB	neuropeptide receptor antagonist
29	3062316	dasatinib	tyrosine kinase inhibitor
30	3689416	K784-3188	
31	69923936	JW-7-24-1	target: Lymphocyte kinase (Lck)
32	44507790	BRD-K41668190	
33	3006531	U0126	selective inhibitor of both MEK1 and MEK2, a type of MAPK/ERK kinase
34	13017911	geldanamycin	Heat shock protein HSP 90-beta; Heat shock protein HSP 90-alpha; Endoplasmin inhibitor
35	73055061	BRD-K64024097	

36	4114	methoxsalen	inhibits the synthesis of deoxyribonucleic acid (DNA)
37	2977478	BRD-A07614565	
38	2736301	TBEP	
39	134551	ergocryptine	Dopamine receptor agonist
40	44498245	BRD-K78596368	
41	44499001	BRD-K32485462	
42	14709	dehydroisoandosterone	DHEA can be understood as a prohormone for the sex steroids
43	4216	ML-7	potently inhibits MLCK; also inhibits YAP/TAZ
44	68617	sertraline	selective serotonin reuptake inhibitor (SSRI)
45	6376322	trichostatin-a	an HDAC inhibitor
46	2819152	KM-00927	
47	6376322	trichostatin-a	an HDAC inhibitor
48	3038522	tandutinib	inhibit type III receptor tyrosine kinases
49	5702164	puromycin	antibiotic that prevents bacterial protein translation
50	5757	estradiol	Estrogen receptor agonist

Анализ механизма действия топ 50 соединений для перехода из фибробластов в индуцированные плюрипотентные стволовые клетки

Таблица 9.4 – Анализ механизма действия топ 50 соединений для перехода из фибробластов в индуцированные плюрипотентные стволовые клетки

position	cid	name	moa
1	310557	EXO-1	ARF inhibitor
2	44142147	BRD-K62466453	
3	4792	phorbol-12-myristate-13-acetate	activates protein kinase C
4	44142091	BRD-K49712247	
5	3342390	BRD-K97274161	
6	44497312	SD-6-035-B3	
7	4792	phorbol-12-myristate-13-acetate	activates protein kinase C
8	44142139	BRD-K44366801	
9	5282379	isotretinoin	retinoid receptor agonist
10	10077147	BX-795	IKK inhibitor
11	22434552	BRD-K51774655	
12	60138086	VU-0415533-1	
13	454217	prostratin	protein kinase C activator
14	5225086	ingenol	Protein kinase C delta type ligand
15	5353342	fenretinide	apoptosis stimulant, retinoid receptor agonist
16	322968	PRIMA1	TP53 inhibitor
17	446378	thapsigargin	non-competitive inhibitor of the sarco/endoplasmic reticulum Ca2+ ATPase (SERCA)

18	9821459	J-104129	
19	5022426	vincristine	tubulin polymerization inhibitor
20	2853908	SA-3676	
21	444795	tretinoin	retinoid receptor agonist, retinoid receptor ligand
22	44484589	BRD-K81249836	
23	4055	menadione	mitochondrial DNA polymerase inhibitor, phosphatase inhibitor
24	5757	estradiol	Estrogen receptor agonist
25	5081913	IKK-2-inhibitor-V	IKK inhibitor, NFkB pathway inhibitor
26	24747339	BRD-K63157263	
27	46943339	BRD-A38793261	
28	5667	vinblastine	microtubule inhibitor, tubulin polymerization inhibitor
29	650228	BRD-A58482795	
30	4792	phorbol-12-myristate-13-acetate	activates protein kinase C
31	5326739	indirubin	CDK inhibitor, glycogen synthase kinase inhibitor
32	5374464	sirolimus	mTOR inhibitor
33	2788	clioquinol	chelating agent
34	64715	mevastatin	HMGCR inhibitor
35	6605258	CGK-733	ATM kinase inhibitor, ATR kinase inhibitor
36	5225086	ingenol	Protein kinase C delta type ligand
37	5691	wortmannin	PI3K inhibitor

38	11957524	BRD-K07079548	
39	1502520	BRD-K22215695	
40	16759157	triciribine	AKT inhibitor
41	33741	tramadol	norepinephrine reputake inhibitor, opioid receptor agonist, serotonin reuptake inhibitor
42	15993715	BRD-K66381707	
43	23648023	BRD-K19117583	
44	10377751	AS-605240	potent and selective PI3Kγ inhibitor
45	5225086	ingenol	Protein kinase C delta type ligand
46	5702295	benzamil	sodium channel blocker
47	5225086	ingenol	Protein kinase C delta type ligand
48	1242560	BRD-K05649647	
49	443375	devazepide	CCK1 (CCK-A) receptor antagonist and a CCK8 antagonist.
50	4792	phorbol-12-myristate-13-acetate	activates protein kinase C

Анализ механизма действия топ 50 соединений для перехода из фибробластов в индуцированные нейральные стволовые клетки

Таблица 9.5 – Анализ механизма действия топ 50 соединений для перехода из фибробластов в индуцированные нейральные стволовые клетки

position	cid	name	moa
1	44616501	BRD-K80786583	
2	454217	prostratin	protein kinase C activator
3	5225086	ingenol	Protein kinase C delta type ligand
4	4792	phorbol-12-myristate-13-acetate	activates protein kinase C
5	12425537	BRD-K12238169	
6	5225086	ingenol	Protein kinase C delta type ligand
7	195165	SR-95639A	acetylcholine receptor agonist
8	2432214	BRD-K39345836	
9	3114	disopyramide	sodium channel blocker
10	44142148	BRD-K66037923	
11	17757274	BRD-K91691979	
12	24180719	PLX-4720	RAF inhibitor
13	566661	BRD-K48576794	
14	44507145	BRD-K80094086	
15	45281797	VU-0365117-1	
16	16747683	AZD-5438	CDK inhibitor
17	10125107	lawsone	red-orange dye
18	25109911	BRD-A11095214	
19	105078	BRD-K49061529	

20	20279	cladribine	adenosine deaminase inhibitor, ribonucleotide reductase inhibitor
21	3689416	K784-3188	
22	265237	withaferin-a	inhibits NF-κB
23	2416356	BRD-K62459624	
24	13200033	vinorelbine	tubulin polymerization inhibitor
25	44485781	BRD-K51454562	
26	6450551	axitinib	PDGFR tyrosine kinase receptor inhibitor, VEGFR inhibitor
27	5225086	ingenol	Protein kinase C delta type ligand
28	11626927	IKK3-inhibitor-IX	a potent ATP-competitive inhibitor of IKK-3
29	40854	oxfendazole	anthelmintic agent
30	4879301	BRD-A06909528	
31	46907727	BRD-K11778076	
32	4813	piceatannol	SYK inhibitor
33	4792	phorbol-12-myristate-13-acetate	activates protein kinase C
34	4261	entinostat	HDAC inhibitor
35	44142134	BRD-K28366633	
36	379077	NSC-663284	CDC inhibitor
37	5353609	methyl-2,5-dihydroxycinnamate	
38	16747683	AZD-5438	CDK inhibitor
39	44484589	BRD-K81249836	

40	44506577	BRD-K76938712	
41	3387354	SB-218078	CHK inhibitor
42	2327953	BRD-K70831943	
43	9830392	WAY-170523	metalloproteinase inhibitor
44	6710662	fluocinonide	glucocorticoid receptor agonist
45	24747231	JAS07-009	
46	11507802	LY-2183240	FAAH inhibitor, FAAH reuptake inhibitor
47	6711154	CGP-53353	EGFR inhibitor, PKC inhibitor
48	73265211	XMD-1499	
49	5447130	nitrofural	bacterial DNA inhibitor
50	5225086	ingenol	Protein kinase C delta type ligand

Анализ механизма действия топ 50 соединений для перехода из мезенхимальных стволовых клеток в индуцированные нейроны

Таблица 9.6 – Анализ механизма действия топ 50 соединений для перехода из мезенхимальных стволовых клеток в индуцированные нейроны

position	cid	name	moa
1	6708778	teniposide	topoisomerase inhibitor
2	44142351	BRD-A24021119	
3	46931017	BRD-K96147838	
4	16667695	selamectin	nematocide
5	4369491	SA-792987	PKC inhibitor
6	11404337	obatoclax	BCL inhibitor
7	44497912	BRD-K68038686	
8	5225086	ingenol	Protein kinase C delta type ligand
9	4477	niclosamide	DNA replication inhibitor, STAT inhibitor
10	5081913	IKK-2-inhibitor-V	IKK inhibitor, NFkB pathway inhibitor
11	16014494	BRD-K09661167	
12	44640183	KU-C104486	
13	56643196	BRD-K86492010	
14	5330286	palbociclib	CDK inhibitor
15	4122	nocodazole	tubulin polymerization inhibitor
16	56643206	BRD-K78385490	
17	1285940	BRD-K68548958	
18	2853908	SA-3676	
19	300471	elesclomol	oxidative stress inducer
20	23631972	BRD-K83194053	

21	5081913	IKK-2-inhibitor-V	IKK inhibitor, pathway inhibitor
22	44142144	BRD-K24156250	
23	44631784	VU-0410183-2	
24	659101	BRD-K92317137	
25	3003565	wortmannin	PI3K inhibitor
26	9888590	sunitinib	PLK inhibitor
27	44507260	BRD-K49010888	
28	6918369	tegaserod	serotonin receptor partial agonist
29	5757	estradiol	Estrogen receptor agonist
30	44634693	sirolimus	mTOR inhibitor
31	5225086	ingenol	Protein kinase C delta type ligand
32	5022426	vincristine	tubulin polymerization inhibitor
33	4677798	simvastatin	HMGCR inhibitor
34	659101	BRD-K92317137	
35	56588033	CAY-10594	
36	5281847	rottlerin	It is a protein kinase Cδ (PKCδ) inhibitor. Rottlerin acts as an uncoupler of mitochondrial respiration from oxidative phosphorylation. It has antitumor, autophagy, anti-proliferative, anti-metastasis and anti-invasive properties.
37	5081913	IKK-2-inhibitor-V	IKK inhibitor, NFkB pathway inhibitor

38	53338822	BRD-K14788918	
39	135411	CD-437	retinoid receptor agonist
40	3426979	BRD-K73610817	
41	4817	pifithrin-alpha	TP53 inhibitor
42	2777391	T-0070907	PPAR receptor antagonist
43	73707429	oligomycin-a	ATP synthase inhibitor, ATPase inhibitor
44	5691	wortmannin	PI3K inhibitor
45	4792	phorbol-12-myristate-13-acetate	activates protein kinase C
46	16759157	triciribine	AKT inhibitor
47	354624	clofarabine	ribonucleotide reductase inhibitor
48		QL-XI-92	
49	5691	wortmannin	PI3K inhibitor
50	6610310	MBCQ	