1 Connectivity Map

Connectivity Map - это ресурс, который использует клеточные ответы на возмущения для поиска взаимосвязей между болезнями, генами и терапией. База данных CMap содержит экспрессионные сигнатуры, возмущенные различными малыми молекулами или реагентами, вызывающие сверхэкспрессию или нокдаун генов, на разных клеточных линиях. Изменения в экспрессии генов (в совокупности называемые дифференциальной сигнатурой экспрессии), возникающие в результате заболевания или лечения, можно сравнить на предмет сходства со всеми возмущенными сигнатурами в базе данных. Возмущения, которые вызывают очень похожие сигнатуры, называются "связанными"; их сходные транскрипционные эффекты предполагают, что они оказывают связанные физиологические эффекты на клетку. Эти связи могут быть использованы для разработки различных методов лечения заболеваний.

Например, имея экспрессионную сигнатуру для какого-то заболевания, мы можем ее сравнить с сигнатурами, индуцированными малыми молекулами, из базы данных. Затем проранжируем сигнатуры по их сходству с сигнатурой запроса(в нашем случае, сигнатурой заболевания). Те малые молекулы, которые вызывают наиболее схожие изменения в экспрессии, можно назвать "положительно связанными". А малые молекулы, которые изменяют экспрессию противоположным образом, называются "отрицательно связанными". Именно они могут расматриваться в качестве кандидатов лекарственных препаратов для этого заболевания.

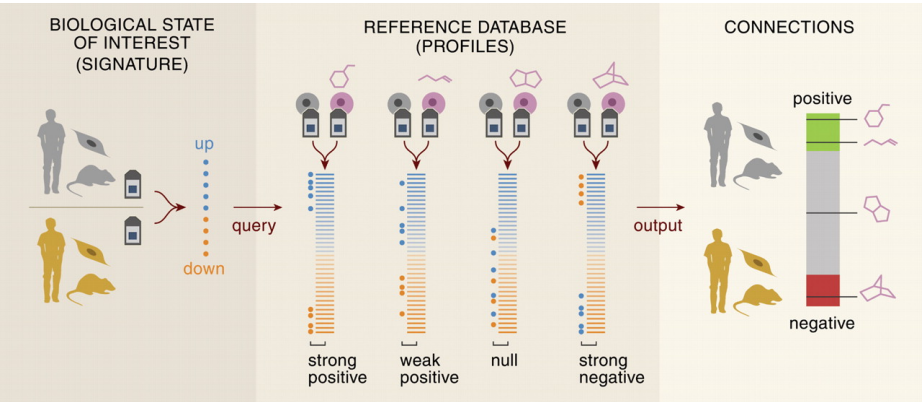

Рис. 1: Концепция Connectivity Map[1]

Центр транскриптомики LINCS в Broad Institute, используя технологию L1000, расширил ресурс CMap, чтобы охватить более 1 млн профилей [1].

2 LINCS L1000 data

A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles B рамках проекта NIH LINCS Центр транскриптомики Broad Institute LINCS создал более 1,3 миллиона транскриптомных профилей с использованием технологии L1000. Большая часть этих данных L1000 включает в себя лекарственные возмущения клеточных линий человека.

При разработке подхода было предположено, что можно было бы с небольшими затратами "поймать"любое клеточное состояние с помощью измерений сокращенного числа транскриптов. Были проанализированы 12031 профиль экспрессии Affymetrix HGU133A в Gene Expression Omnibus (GEO). Они же были использованы для определения оптимального количества информативных транскриптов, которые были названы 'landmark' транскрипты. Этот анализ показал, что 1000 landmarks было достаточно, чтобы восстановить 82% информации транскриптома.

Для измерения 1000 landmark транскриптов был адаптирован метод, включающий амплификацию, опосредованную лигированием (LMA), с последующим захватом продуктов амплификации на микросферы с флуоресцентной адресацией (Peck et al., 2006). Метод был расширен до 1000 параллельных реакций.

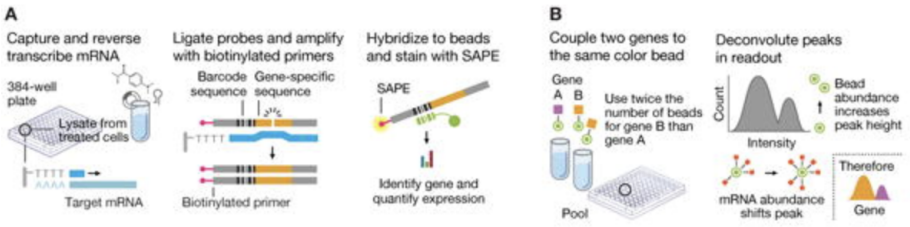

Рис. 2: Концепция L1000 assay[2]

3 iLINCS

Integrative LINCS (iLINCS) - это интегративная веб-платформа для анализа данных и сигнатур LINCS. Портал предоставляет удобный интерфейс для анализа omics (транскриптомных и протеомных) наборов данных LINCS.

3.1 Внутренняя база данных iLINCS

Внутренняя база данных iLINCS содержит более 10 000 обработанных наборов omics данных, более 220 000 omics сигнатур и более 10^9 статистически значимых "связей" между сигнатурами. Наборы данных Omics включают транскриптомные (RNA-seq и микрочипы) и протеомные (Reverse Phase Protein Arrays и LINCS targeted mass spectrometry proteomics) датасеты. Коллекции наборов данных включают транскриптомные и протеомные данные, сгенерированные проектом Cancer Genome Atlas (TCGA), наборы данных GEO GDS и полную коллекцию наборов данных GEO RNA-seq.

Библиотеки сигнатур iLINCS:

Библиотеки сигнатур LINCS L1000:

- Consensus gene knockdown signatures (CGS)
- Overexpression gene signatures
- Chemical perturbation signatures

Библиотеки LINCS L1000 содержат химически и генетически возмущенные сигнатуры LINCS, полученные технологией LINCS L1000.

Также стоит упомянуть следующие библиотеки сигнатур:

- LINCS targeted proteomics signatures : молекулярные сигнатуры на уровне протеома, генетически возмущенные или с помощью малых молекул на различных клеточных линий, полученные с помощь методов масс-спектрометрии
- Connectivity Map signatures : транскрипционные сигнатуры, построенные на второй версии первоначального датасета Connectivity Map с использованием метода анализа экспрессий Affymetrix
- Disease related signatures: транскрипционные сигнатуры, полученные при сравнении выборочных групп в GEO GDS
- DrugMatrix signatures: транскрипционные сигнатуры DrugMatrix.
- ENCODE transcription factor binding signatures
- Transcriptional signatures from EBI Expression Atlas

- Cancer Therapeutics Response Signatures: транскриптомные сигнатуры на основе данных проекта Cancer Therapeutics Response Portal (CTRP). Данные включают 860 линий раковых клеток и сочетают базовую (необработанную) экспрессию генов с измерениями чувствительности к 481 противораковому соединению.

- Pharmacogenomics transcriptional signatures : сигнатуры, полученные при анализе дифференциальной экспрессии генов между клеточными линиями, обработанными противораковыми препаратами, и соответствующими контролями в двух отдельных проектах: NCI Transcriptional Pharmacodynamics Workbench (NCI-TPW) и Plate-seq project dataset

Все предварительно вычисленные сигнатуры возмущений в iLINCS состоят из двух векторов: вектора значений logFC d=(d1,..., dN) и вектора p-value p=(p1,...,pN), где N-количество генов или белков в сигнатуре. Помимо такого представления, сигнатуры пользователя также могут состоять только из значений logFC без p-value, 2 списков генов: с повышенной и пониженной экспрессией, и только одного списка генов [3].

3.2 Анализ связности сигнатур

В зависимости от типа сигнатуры запроса используются различные метрики связности. Если сигнатура запроса создается из набора данных iLINCS или загружается пользователем с значениями logFC и p-value, то связность со всеми сигнатурами iLINCS вычисляется как взвешенная корреляция между двумя векторами значений logFC и вектором весов, равным [$-log_{10}$(p-value сигнатуры запроса) $- log_{10}$(p-value сигнатуры iLINCS)]. Когда загруженная пользователем сигнатура состоит только из значений logFC без p-value, вектор весов корреляции основан только на p-value сигнатур iLINCS [$-log_{10}$(p-value сигнатур iLINCS)] [3].

3.3 Анализ связности возмущений

Связь между сигнатурой запроса и возмущением устанавливается с помощью анализа обогащения оценок связности между сигнатурой запроса и набором всех сигнатур L1000 (для всех клеточных линий, временных точек и концентраций). Анализ устанавливает, являются ли оценки связности как набора достаточно высокими на основе метода Random Set [3].

4 L1000FWD

L1000 fireworks display (L1000FWD) - это веб-приложение, которое обеспечивает интерактивную визуализацию экспрессионных сигнатур из набора данных LINCS L1000. Большая часть данных L1000 получена возмущением клеточных линий человека с помощью лекарственных препаратов. Эти профили экспрессии генов могут быть использованы для генерации сотен тысяч лекарственных экспрсессионных сигнатур. Такой набор сигнатур является ценным каталогом для поиска связей между лекарствами, генами и болезнями; и для открытия механизмов действия (MOA) для менее изученных лекарств и малых молекул. Также это предоставляет возможность перепрофилирования лекарств [4].

4.1 Подготовка сигнатур и построение сети

Подготовка сигнатур начинается с нормализации данных 3 уровня LINCS L1000. Затем вычисляют сигнатуры, используя метод характерного направления. Для построения матрицы смежности сигнатур было использовано 34 502 сигнатур. Эта матрица смежности содержит попарное косинусное сходство между сигнатурами в пространстве 978 landmark генов. Часть ребер с низким уровнем сходства была удалена. Таким образом был получен взвешенный неориентированный граф, состоящий из 18 082 узлов (сигнатур) и 595 177 ребер. Связанные компоненты с менее чем 10 узлами были удалены из графа, в результате чего граф состоял из 16 848 узлов и 594 372 ребер. Окончательный граф сигнатур охватывает 68 различных клеточных линий, 3237 препаратов/соединений, 3 временных точки и 132 дозы.

Как было упомянуто ранее, в L1000FWD каждый узел представляет собой экспрессионную сигнатуру, индуцированную препаратом. Интерфейс L1000FWD позволяет менять форму и цвет узлов в соответствии со связанными с ними атрибутами, такими как клеточная линия, MOA, временная точка и batch. Сигнатуры кластеризованны по сходству. Многие кластеры содержат сигнатуры различных типов клеток, разделяя при этом общие MOA. Кроме того, карта L1000FWD позволяет пользователям проецировать свои собственные сигнатуры на карту для определения того, где находится сигнатура в глобальном пространстве экспрессий. Поиск на основе сходства позволяет находить сигнатуры, имитирующие или обращающие сигнатуру запроса, представленную 2 списками генов: генов с повышенной и пониженной экспрессией [4].

4.2 Вычисление уровня сходства сигнатур

Оценка сходства считается как пересечение списков генов с повышенной и пониженной экспрессией сигнатуры запроса с аналогичными списками генов для сигнатуры из базы данных, деленное на "эффективный ввод". "Эффективный ввод"расчитывается как число общих генов для входных списков и генов из L1000. Для оценки статистической значимости пересечения списков генов сигнатуры запроса со списками сигнатуры базы данных используется точный тест Фишера. Для расчета z-score моделируются 10 000 пар случайных списков генов с повышенной и пониженной экспрессией в качестве входных данных. Для каждой сигнатуры запроса вычисяется уровень сходства с сигнатурой из базы данных. Затем уже для значения сходсва исходной сигнатуры запроса с сигнатурой из базы данных вычисляется z-score на основе полученной выборки оценок сходства. Рассчитанные метрики комбинируются следующим образом:

$$c = z * log_{10}(p) \tag{1}$$

Для пользователя доступны оценка сходтсва сигнатур, p-value, Z-score, и комбинированная оценка [4].

5 CLUE

CLUE - это платформа, созданная Broad Institute для анализа сингатур с помощью CMap и последующего ранжирования малых молекул. В состав платформы входят несколько инструментов: Query app, TouchStone app, Morpheus app и некоторые другие. Рассмотрим функции, которые выполняет каждое отдельное приложение:

- TouchStone ссылается на набор данных о химических соединениях и генетических пертурбагенах, которые хорошо изучены, а также генерируют стабильные экспрессионные сигнатуры в клетках. Таким образом, набор данных Touchstone служит эталоном для оценки связей между пертурбагенами. Использование этого приложения помогает узнать больше о пертурбагенах и изучить их взаимосвязь.

- Приложение Query используется для того, чтобы найти положительные и отрицатель-

ные связи между интересующей экспрессионной сигнатурой и всеми сигнатурами в CMap.

- Morpheus - это приложение, которое позволяет аннотировать существующие наборы данных и извлекать больше информации о полученных химических соединениях.

5.1 Подготовка сигнатур

Подготовка образцов заключалась в расширении числа сигнатур баз данных предыдущего проекта. Расширение было выполнено по нескольким направлениям. Во-первых, было увеличено количество малых молекул с 164 до 19811 низкомолекулярных лекарств, инструментальных соединений и соединений из библиотеки для скрининга, включая те, которые имеют клиническую применимость, известный механизм действия или номинацию от NIH Molecular Libraries Program. Каждое соединение было профилировано в трех экземплярах либо через 6, либо через 24 часа после обработки.

Во-вторых, были расширены масштабы генетических нарушений с помощью нокдаунов или оверэкспрессии 5 075 генов, отобранных на основе их связи с заболеваниями человека или принадлежности к биологическим путям. Каждое генетическое нарушение было профилировано в трех экземплярах через 96 часов после заражения. Для исследований оверэкспрессии использовали один клон кДНК, тогда как были профилированы три различных shРНК, нацеленных на каждый ген.

В-третьих, было вовлечено больше клеточных линий. Хорошо аннотированные генетические и низкомолекулярные пертурбагены были профилированы в основном наборе из 9 клеточных линий, в результате чего был получен эталонный набор данных, который был назван Touchstone v1. Неохарактеризованные малые молекулы без известного механизма действия (MOA) были по-разному профилированы в пределах от 3 до 77 клеточных линий, что дало набор данных, который был назван Discovery v1. В общей сложности были сгенерированы 1319138 L1000 профилей из 42080 пертурбагенов (19811 низкомолекулярных соединений, 18493 shРНК, 3462 кДНК и 314 биопрепаратов), что соответствует 25200 биологическим объектам (19811 соединений, shРНК и / или кДНК против 5075 генов и 314 биопрепаратов)всего для 473647 сигнатур(включая реплики), что более чем в 1000 раз больше, чем в пилотном наборе данных CMap.

5.2 Вычисление уровня сходства сигнатур

5.2.1 Взвешенная оценка связанности

Взвешенная оценка связности (WTCS) представляет собой непараметрическую меру сходства, основанную на взвешенной статистике обогащения (ES) Колмогорова-Смирнова. WTCS - это составная двунаправленная версия ES. Для данной пары набора генов запроса (q_{up}, q_{down}) и сигнатуры r базы данных WTCS вычисляется следующим образом:

$$w_{q,r} = \begin{cases} \frac{ES_{up} - ES_{down}}{2}, & if \;\; sgn(ES_{up}) \neq sgn(ES_{down}) \\ 0, & otherwise \end{cases}$$

Где ES_{up} - это обогащение q_{up} в r, а ES_{down} - это обогащение q_{down} в r. WTCS находится в диапазоне от -1 до 1. Он будет положительным для сигнатур, которые связаны положительно, и отрицательным для сигнатур, которые связаны обратно, и близким к нулю для сигнатур, которые не связаны друг с другом. Ноль присваивается в том случае, когда и ES_up, и ES_down имеют один и тот же знак.

5.2.2 Нормализация оценки связанности

Чтобы обеспечить возможность сравнения оценок связности между клеточными типами и типами возмущений, оценки нормализуются для учета глобальных различий в связности, которые могут возникать по этим ковариатам. Учитывая вектор значений WTCS w, полученных в результате запроса, мы нормализуем значения в каждой клеточной линии и типе возмущения, чтобы получить нормализованные оценки связности (NCS) следующим образом:

$$NCS_{c,t} = \begin{cases} \frac{w_{c,t}}{\mu^{+}_{c,t}}, & if \;\; sgn(w_{c,t}) \;\; > \;\; 0 \\ \frac{w_{c,t}}{\mu^{-}_{c,t}}, & otherwise \end{cases}$$

где $NCS_{C,t}, w_{C,t}, \mu^{+}_{c,t} и \mu^{c,t}$ - нормализованные оценки связности, сырые взвешенные оценки связи и средние со знаком исходных взвешенных оценок связности (среднее положительных и отрицательных значений, оцениваемых отдельно) в подмножестве сигнатур Touchstone, соответствующих клеточной линии c и пертурбагену типа t, соответственно.

5.2.3 Connectivity Map скоры

Тау τ сравнивает наблюдаемую оценку обогащения со всеми остальными в референсной базе данных. В принципе, τ можно вычислить путем сравнения с оценками из любой базы данных сигнатур, и наиболее распространенным подходом является создание нулевого распределения путем случайной перестановки. Однако более строгий тест, позволяющий избежать необходимости делать предположения относительно сложной корреляционной структуры данных экспрессии генов, заключается в использовании сборника разнообразных, биологически значимых пертурбационных сигнатур, таких как те, что находятся в CMap-L100Ovl, поскольку именно с этими сигнатурами любая новая связь должна конкурировать. Таким образом, результаты запроса оцениваются с помощью τ в качестве стандартизированной меры в диапазоне от -100 до 100; Значение τ, равное 90, указывает на то, что только 10% возмущенных сигнатур показали более сильную связь с запросом.

5.2.4 Тонкости вычисления τ

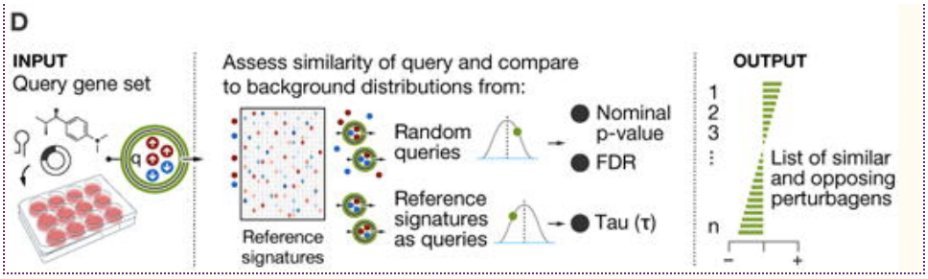

Рис. 3: Концепция выполнения запроса и рассчета τ скора [2]

Поскольку сигнатура фиксирована, τ можно использовать для сравнения результатов по запросам - соединение со значительным p-value и FDR, но низким τ может указывать на очень неразборчивые отношения, связи которых не уникальны. Хотя значимые сравнения могут быть выполнены между значениями NCS сигнатур базы данных относительно запроса q, также полезно оценить, существенно ли отличается связь между q и конкретной сигнатурой r от той, которая наблюдается между r и другими запросами. Это выполняется путем сравнения каждого наблюдаемого значения NCS ($ncs_{q,r}$) между запросом q и сигнатурой r базы данных с распределением значений NCS, представляющих сходство между эталонным сборником запросов (Q_{ref}) и r. Результатом этой процедуры является стандартизированная мера, которую мы называем Тау (τ), которая представляет процент запросов в Q_{ref} с

меньшим значением $|NCS|$ чем $|ncs_{q,r}|$, с поправкой на сохранение знака $ncs_{q,r}$:

$$\tau_{q,r} = sign\,(ncs_{q,r}) \frac{100}{N} \sum_{i=1}^{N} [|ncs_{i,r}| < |ncs_{q,r}|]$$

где $ncs_{q,r}$ - нормализованная оценка связности для сигнатуры r по запросу q, $ncs_{i,r}$ - нормализованная оценка связности для сигнатуры r относительно i-го запроса в Q_{ref}, а N - количество запросов в Q_{ref}.

5.2.5 Обобщение по клеточным линиям

При изучении результатов запроса часто бывает удобно получить ориентированную на возмущения меру связности, которая суммирует результаты, наблюдаемые в отдельных типах клеток. Это может быть особенно полезно при поиске соединений, которые сохраняются между клеточными линиями, или когда вы не уверены, какую клеточную линию исследовать. Учитывая вектор нормализованных оценок связности для perturbagen p, относительно запроса q, по всем линиям клеток, в которых был профилирован p, суммарная оценка связности клеток получается с использованием максимальной квантильной статистики:

$$NCS_{c,t} = \begin{cases} Q_{hi}(ncs_{p,c}, & if \;\; |Q_{hi}(ncs_{p,c}| \;\; >= \;\; |Q_{lo}(ncs_{p,c}| \\ Q_{lo}(ncs_{p,c}, & otherwise \end{cases}$$

где $ncs_{p,c}$ - вектор нормализованных оценок связности для perturbagen p относительно запроса q по всем линиям клеток, в которых был профилирован p, а Q_{hi} и Q_{lo} - это верхний и нижний квантили соответственно. Эта процедура сравнивает квантили Q_{hi} и Q_{lo} $ncs_{p,c}$ и сохраняет то, что имеет более высокую абсолютную величину. Таким образом, максимальный квантиль более чувствителен к сигналу в подмножестве клеточных линий, чем меры центральной тенденции, такие как среднее значение или медиана. В представленных здесь анализах были использованы $Q_{hi} = 67$, $Q_{lo} = 33$.

6 TopoCMap

TopoCMap - инструмент, разработанный в нашей лаборатории, для проведения анализа Connectivity Map, который учитывает биологическую значимость геннов в экспрессионной

сигнатуре путем расчета так называемого скора влиятельности для каждого дифференциально экспрессированного гена.

6.1 Построение генных сети и подсчет скоров влиятельности

Подсчету скоров влиятельности для генов сигнатуры предшествуют несколько важных шагов.

6.1.1 Получение данных взаимодействий

Для получения информации о взаимодействии генов внутри сигнатуры проводятся запросы в базу данных STRING [5]. В этой базе данных содержится информациия о белок-белковых взаимодействиях, как физических и регуляционных (воздействия транскрипционных факторов на гены), так и предсказанных взаимодействиях.

6.1.2 Построение генных сетей

На информации, полученной из STRING [5] строятся генные сети, где вериной в сети является ген, а ребра, выходящие из вершины, соединяют взаимодествующие между собой гены. Генные сети строятся отдельно для генов сигнатуры запроса с повышенной экспрессией и пониженной экспрессией. На основании построинных сетей рассчитываются топологические метрики центральности, такие как pagerank centrality, betweenness, eigenvector centrality, closeness, Katz centrality, Hits centrality, eigentrust Эти метрики были выбраны, поскольку есть несколько статей, в которых корреляция между этими метриками и важностью белков в сети белок-белковых взаимодействий была показаны, и они включены в эффективный и действенный пакет под названием Graph-Tool [6].

6.1.3 Вычисление скоров влиятельности

Затем эти метрики центральности использовались для расчета скоров влиятельности для каждого гена в сигнатуре запроса. Мы также использовали кратное изменение (FC) при вычислении скоров влиятельности, поскольку это показывает значимость изменения выражения. Предварительно FC были прологарифмированы, для уменьшения среднеквадратичного отклонения (σ), а также нормализованы с помощью z-score. Такого рода нормализации бы-

ли проведены для унификации масштабов топологических метрик центральности, которые находятся в диапазоне от 0 до 1, и FC. Скор влиятельности для гена i вычисляется как комбинация метрик:

$$inf\_score_i = \begin{cases} (a_1 \cdot \log FC + b_1) \cdot (a_2 \cdot \text{pagerank} + b_2) \cdot (a_3 \cdot \text{betweenness} + b_3) \\ \quad \cdot (a_4 \cdot \text{eigenvector} + b_4) \cdot (a_5 \cdot \text{closeness} + b_5) \cdot (a_6 \cdot \text{Katz} + b_6) \\ \cdot (a_7 \cdot \text{Hits} + b_7) \cdot (a_8 \cdot \text{eigentrust} + b_8) + a_9, \quad \text{if gene in STRING} \\ 1, \quad otherwise \end{cases}$$

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 [7].

6.2 Вычисление уровня сходства сигнатур

Оценка сходства считается как косинусное расстояние между генными векторами сигнатуры запроса с 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 скоры 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 есть в раскладываемом списке и для него доступна информация о взаимодействиях в базе данных STRING [5]
- 1, если если есть в раскладываемом списке, но для него нет информации о взаимодействиях в базе данных STRING [5]
- 0, во всех остальных случаях

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

[1] Justin Lamb и др. “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease”. в: *science* 313.5795 (2006), с. 1929—1935.

[2] Aravind Subramanian и др. “A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles”. в: *Cell* (2018), с. 1437—1452.

[3] Marcin Pilarczyk и др. “Connecting omics signatures of diseases, drugs, and mechanisms of actions with iLINCS”. в: *bioRxiv* (2019), с. 826271.

[4] Zichen Wang и др. “L1000FWD: fireworks visualization of drug-induced transcriptomic signatures”. в: *Bioinformatics* 34.12 (2018), с. 2150—2152.

[5] 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).

[6] 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).

[7] Sizykh A и др. “CFM: a database of experimentally validated protocols for chemical compound-based direct reprogramming and transdifferentiation”. в: *F1000Research* (2021).