МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)"
ФИЗТЕХ-ШКОЛА БИОЛОГИЧЕСКОЙ И МЕДИЦИНСКОЙ ФИЗИКИ
КАФЕДРА БИОИНФОРМАТИКИ И СИСТЕМНОЙ БИОЛОГИИ

---

Выпускная квалификационная работа по направлению

03.03.01

прикладная математика и физика

на тему:

# Разработка нового метода Connectivity Map основанного на топологических метриках вершин генных сетей

Студент _____ Минаева М.В.

Научный руководитель _____ Муртазалиева Х.А.

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

МОСКВА, 2020

# 1 Аннотация

Здесь будет аннотация к моей работе.

# Содержание

# 2 Обозначения, сокращения, основные определения

Здесь будут важные понятия и определения.

# 3 Введение

## 3.1 Представление химических веществ

Проблема представления химических молекул в виде, пригодном для анализа и обработки на компьютере, стоит довольно остро. 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.

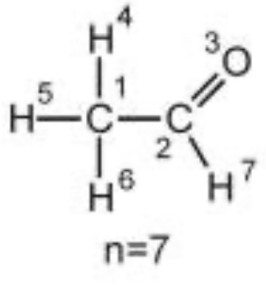
В настоящее время существует множество разновидностей форматов SMILES. Наиболее интересным для текущей задачи среди всего этого многообразия является формат USMILES (Unique SMILES), позволяющий однозначно восстанавливать химическую структуру молекул после конвертации из формата USMILES.

### 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 органи-

ческой 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 представляемся 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 |

Рис. 3.1: Представление молекулярного графа в виде матрицы[1]

### 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 возможно применение матричной алгебры. Существенным недостатком является то 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 квадратичным возрастанием 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 - показывает 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. Данных 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 $n!$ раз-личных 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 тд. Если при нумерации встречаются атомы с одинаковыми значениями EC, то дальнейшая нумерация подчиняется строгим правилам, которые учитывают тип атом, его связи, заряд и тд.

### 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 тд. 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 сход-

ства 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, как 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.

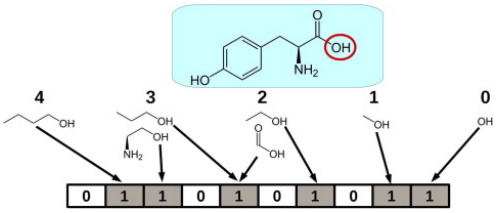



Рис. 3.2: Пример построения хэшированного фингерпринта[1]

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 $\sigma_{st}$ 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 графа, вычисляемая как собственные вектора матрицы смежности. Решается уравнение на поиск собственных векторов матрицы:

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

где $\mathbf{A}$ - матрица смежности, $\mathbf{x}$ - собственный вектор, а $\lambda$ - 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 $\lambda$ 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. Данная метрика тем меньше, чем больше у вершины s коротких путей, соединяющих ее с другими вершинами. Сначала определим понятие Матрица Кирхгофа:

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

где $\mathbf{A}$ - взвешанная матрица смежности графа, $\mathbf{D}$ - 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 $\mathbf{J}$ как матрицу всех вхождений, которые равны единице. Определим матрицу $\mathbf{B}$ 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, достигшего 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 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 рассматривать целевую вершину как поглощающую и 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 $\mathbf{A}$ 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 отсутствие 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 присвоения A.

Вопрос выбора метрики для решения установления 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 белками, 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 установления 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 взаимодейцствий вошли только те белки, которые имеют ортологи во всех трех организмах. Полученные результаты

валидировались на результатах более ранних исследований, в которых была выявлена летальность того или иного белка. 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 показывал, что летальные мутации не только сильно связаны в сети, но они также обладают дополнительным свойством: их удаление вызывает нарушение целостности сети. Мы также предоставляем 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, вероятностью $p_i$ 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] было рассмотрели 40 различных метрик центральности, относящиеся 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, полученные 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 прогнозирования важных узлов в сети белок-белковых взаимодействий. Основываясь 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 эволюцион-

ные метрики консервативности белков. В работе был 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 мета-

болических 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, т.е. биомасса может быть произведена без дополнительных метаболических затрат, мутант является жизнеспособным. Если $S = \infty$, т.е. 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 метабо-

лической 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, и 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 (эти списки получаются 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.

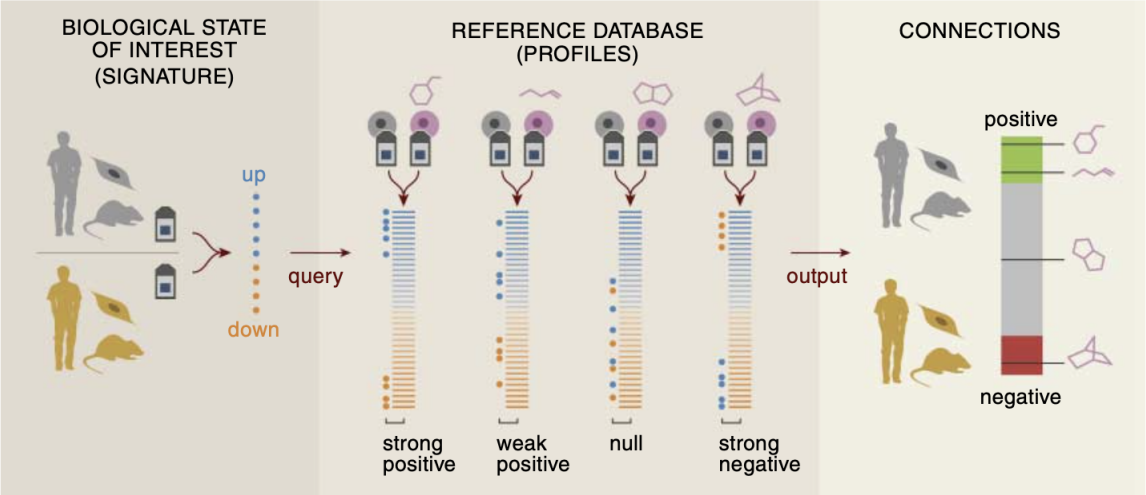



Рис. 3.3: Принцип метода Connectivity Map

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 индуцирующие данную сигнатуру. После произведенного сравнения метод Connectivity Map выдает ранжированный список молекул с их скорами, 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 пердсказана активность гедунин как ингибитора

HSP90) и выявления потенциальных 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 L1000CDS2

Duan et al. [10] представили подход, который называется L1000CDS2, с помощью которого можно произвести Connectivity Map и получить список малых молекул, которые могут обратить вспять или имитировать сигнатуру болезни и других биологических состояний. Метод доступен в виде веб-ресурса и позволяет сравнить сигнатуры пользователей с референсной базой даных сигнатур вычисленных методом CD (Characteristic Direction)[11]. Помимо определения приоритета малых молекул для того, чтобы обратить или имитировать входную сигнатуру или предварительно вычисленные сигнатуры для 670 заболеваний и набора эндогенных лигандов, L1000CDS2 также можно применять для предсказания попарные комбинации малых молекул, подструктурного анализа обогащения и других задач. Так для валидации метода было выполнено предсказание лекарственной молекулы для раннего лечения вируса Эбола. Кенпауллон - молекула, предсказанная с наивысшим рангом для обращающения экспрессионной сигнатуры клеток человека, инфицированных вирусом

Эбола, как было показано при экспериментальных испытаниях, 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

Здесь будут описаны все методы,

# 5 Полученные результаты

Здесь будут полученные результаты. Когда-нибудь я их получу)

# 6 Заключение. План дальнейших исследований

Здесь будут заключения из когда-нибудь проделанной работы.

# 7 Благодарности

Это тоже когда-нибудь будет написано и доведено до ума.

# 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] Neil R Clark и др. «The characteristic direction: a geometrical approach to identify differentially expressed genes». в: *BMC bioinformatics* 15.1 (2014), с. 1—16.