MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 27, NO. 2, 2020
journal homepage: www.mais-journal.ru

COMPUTING METHODOLOGIES AND APPLICATIONS

Method of the Joint Clustering in Network and Correlation Spaces

A. N. Gainullina[1], A. A. Shalyto[1], A. A. Sergushichev[1]

DOI: 10.18255/1818-1015-2020-2-180-193

[1]ITMO University, 49 Kronverkskiy Prospekt, Saint Petersburg 197101, Russia.

MSC2020: 68P99, 92B99
Research article
Full text in Russian

Received June 8, 2020
After revision June 17, 2020
Accepted June 17, 2020

Network algorithms are often used to analyze and interpret the biological data. One of the widely used approaches is to solve the problem of identifying an active module, where a connected subnetwork of a biological network is selected which best reflects the difference between the two considered biological conditions. In this work this approach is extended to the case of a larger number of biological conditions and the problem of the joint clustering in network and correlation spaces is formulated.

To solve this problem, an iterative method is proposed that takes as the input graph G and matrix X, in which the rows correspond to the vertices of the graph. As the output, the algorithm produces a set of subgraphs of the graph G so that each subgraph is connected and the rows corresponding to its vertices have a high pairwise correlation. The efficiency of the method is confirmed by an experimental study on the simulated data.

Keywords: active module; clustring; gene expression; biological networks.

INFORMATION ABOUT THE AUTHORS

Anastasiia N. Gainullina	orcid.org/0000-0003-3796-2337. E-mail: anastasiia.gainullina@gmail.com PhD student.
Anatoly A. Shalyto	orcid.org/0000-0002-2723-2077. E-mail: shalyto@mail.ifmo.ru Chief researcher, professor, Doctor of Sciences.
Alexey A. Sergushichev correspondence author	orcid.org/0000-0003-1159-7220. E-mail: alserg@itmo.ru Associate professor, PhD.

Funding: Government of Russian Federation, grant 08-08.

For citation: A. N. Gainullina, A. A. Shalyto, and A. A. Sergushichev, “Method of the Joint Clustering in Network and Correlation Spaces”, *Modeling and analysis of information systems*, vol. 27, no. 2, pp. 180-193, 2020.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 27, № 2, 2020
сайт журнала: www.mais-journal.ru

COMPUTING METHODOLOGIES AND APPLICATIONS

Метод совместной кластеризации в графовом и корреляционном пространствах

А. Н. Гайнуллина[1], А. А. Шалыто[1], А. А. Сергушичев[1]

DOI: 10.18255/1818-1015-2020-2-180-193

[1]Университет ИТМО, Кронверкский пр. 49, г. Санкт-Петербург, 197101 Россия.

УДК 519.1
Научная статья
Полный текст на русском языке

Получена 8 июня 2020 г.
После доработки 17 июня 2020 г.
Принята к публикации 17 июня 2020 г.

Алгоритмы на графах часто используются для анализа и интерпретации биологических данных. Одним из широко используемых подходов является решение задачи поиска активного модуля, в которой в графе биологических взаимодействий выделяется связный подграф, лучше всего отражающий разницу между двумя рассматриваемыми биологическими состояниями. В настоящей работе этот подход расширяется на случай большего числа биологических состояний и формулируется задача совместной кластеризации в графовом и корреляционном пространстве.

Для решения этой задачи предлагается итеративный метод, принимающий на вход граф G и матрицу X, в которой строки соответствуют вершинам графа. На выходе алгоритм выдает набор подграфов графа G так, что каждый подграф является связным и строки, соответствующие его вершинам, обладают высокой попарной корреляцией. Эффективность метода подтверждается экспериментальным исследованием на смоделированных данных.

Ключевые слова: активный модуль; кластеризация; экспрессия генов; биологические графы.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Анастасия Наильевна Гайнуллина	orcid.org/0000-0003-3796-2337. E-mail: anastasiia.gainullina@gmail.com Аспирант.
Анатолий Абрамович Шалыто	orcid.org/0000-0002-2723-2077. E-mail: shalyto@mail.ifmo.ru Гл. науч. сотр., профессор, докт. техн. наук.
Алексей Александрович Сергушичев автор для корреспонденции	orcid.org/0000-0003-1159-7220. E-mail: alserg@itmo.ru Доцент, канд. техн. наук.

Финансирование: Правительство Российской Федерации, субсидия 08-08.

Для цитирования: A. N. Gainullina, A. A. Shalyto, and A. A. Sergushichev, “Method of the Joint Clustering in Network and Correlation Spaces”, *Modeling and analysis of information systems*, vol. 27, no. 2, pp. 180-193, 2020.

Введение

Методы с использованием графов часто применяются в биоинформатике для интерпретации экспериментальных данных [1]. Такие методы применяются в разных контекстах: в исследованиях полногеномного поиска ассоциаций [2], метаболических процессов [3], соматических мутаций в раке [4] и других. Главная идея этих методов состоит в том, что учет известных биологических связей (например, между белками, метаболитами или другими сущностями) может повлечь более глубокое понимание данных и соответствующих биологических процессов.

Одним из часто применяемых подходов является выделение в графе биологических связей некоторого связного подграфа, лучше всего соответствующего так называемому *активному модулю*. Эта концепция впервые была представлена в работе [5], в которой авторы предложили метрику для оценки подграфов на основе данных экспрессии генов и эвристический метод *jActiveModules* для поиска наиболее оптимального подграфа. В дальнейшем эта концепция была развита в методе *BioNet* [6], в котором был предложен другой способ оценки подграфов, оптимизация которого соответствовала решению задачи поиска связного подграфа максимального веса (*Maximum Weight Connected Subgraph*, *MWCS*). Хотя эта задача является NP-полной, для нее существуют практические программы-решатели, позволяющие находить оптимальные или хорошие субоптимальные решения за небольшое время.

Однако с развитием технологий получения биологических данных эксперименты все чаще стали включать множество разнородных биологических образцов. Для таких типов экспериментов разработанные ранее методы, основанные на выделении одного активного модуля, отличающегося между двумя группами образцов, стали все менее применимы. Для них становится актуальна задача выделения нескольких активных модулей, где каждый модуль характеризуется своим профилем активности в исследуемых биологических образцах [7].

В настоящей работе рассматривается задача поиска нескольких активных модулей. Для простоты в качестве графа биологических связей используется граф белок-белковых взаимодействий, в котором вершинами являются гены, и между генами существуют ребро, если соответствующие генам белки могут взаимодействовать друг с другом в клетке. В качестве экспериментальных данных рассматриваются данные экспрессии генов, в которых каждому гену в каждом биологическом образце сопоставлено некоторое число – значение экспрессии этого гена. В таком контексте задачу поиска активных модулей можно сформулировать как задачу совместной кластеризации в графовом (на основе графа белок-белковых взаимодействий) и корреляционном (на основе таблицы экспрессии генов) пространствах.

1. Задача совместной кластеризации в графовом и корреляционном пространствах

Пусть дан граф $G = (V, E)$ порядка $n = |V|$, $V = \{1, \dots, n\}$. Пусть также дана матрица X размерности $n \times m$, где i-ая строчка матрицы соответствует i-й вершине графа. Будем называть i-ю строчку матрицы *профилем* вершины i.

Сформулируем задачу совместной кластеризации в графовом и корреляционном пространствах как поиск набора связных подграфов $S = \{S_i = (V(S_i), E(S_i))\}$ графа G таких, что:

- для каждого подграфа S_i высока попарная корреляция профилей вершин $V(S_i)_j$;
- для каждой пары подграфов S^i и S^k $(i \neq k)$ низка корреляция профилей вершин $V(S^i)_j$ и $V(S^k)_l$.

Обратим внимание, что в данном определении не требуется, чтобы подграфы S^i не пересекались. Также это определение допускает различные метрики оценки качества кластеризации, которые будут зависеть от конкретного приложения.

В настоящей работе мы будем рассматривать следующую модель. Пусть связные подграфы $A = \{A_1, \dots, A_K\}$ – истинные *активные модули*. Активность модулей в каждом из *m образцов* задается

матрицей P размерности $K \times m$. Вклад i-го гена в j-й модуль задается неотрицательной матрицей B размерности $n \times K$, при этом $B_{i,j} = 0$ как только $i \notin V(A_j)$. Тогда матрица X может быть представлена как:

$$X = B \cdot P + \varepsilon,$$

где ε – случайная матрица, соответствующая шуму.

В такой модели, задачей является по графу G и матрице X как можно более точно восстановить активные модули A. При этом K – число активных модулей – неизвестно.

2. Итеративный алгоритм совместной кластеризации

Для решения задачи совместной кластеризации в графовом и корреляционном пространствах мы предлагаем следующий алгоритм итеративной кластеризации. Алгоритм основан на идеях кластеризации с помощью алгоритма *k-means* и *EM*-алгоритма. Псевдокод алгоритма приведен на рисунке 1.

Algorithm: Network clustering

Input: Graph $G = (V, E)$ of order $n = |V|$, matrix X of size $n \times m$, initial module profiles approximation $P^{(1)}$ of size $k^{(1)} \times m$, value of *base*.

Result: Final approximation of profiles P^* of size $k^* \times m$ and a set of connected subgraphs A_i^* for $i \in 1, \ldots, k^*$ as a final approximation of active modules

for $i \in \{1, 2, \ldots\}$ **do**

- $k^{(i)} \leftarrow$ number of rows in $P^{(i)}$;
- $d_{x,y} \leftarrow 1 - \mathrm{corr}(X_x, P_y^{(i)})$ for $x \in \{1, \ldots, n\}, y \in \{1, \ldots, k^{(i)}\}$;
- $d_{x,0} \leftarrow base$ for $x \in \{1, \ldots, n\}$;
- $d'_{x,y} \leftarrow \min_{z \in \{0, \ldots, k^{(i)}\}, z \neq y} d_{x,z}$ for $x \in \{1, \ldots, n\}, y \in \{1, \ldots, k^{(i)}\}$;
- **for** $j \in \{1, \ldots, k^{(i)}\}$ **do**
  - $w_x \leftarrow -\log \frac{d_{x,y}}{d'_{x,y}}$ for $x \in \{1, \ldots, n\}$;
  - $A_j^{(i)} \leftarrow$ connected subgraph of G with maximum sum of vertex weights w;
  - $P^{(i+1)} \leftarrow$ coordinate-wise average of X_x, for $x \in V(A_j^{(i)})$ if $w_x > 0$;
- **end**
- **if** $P^{(i+1)}$ *substantially differs from* $P^{(j)}$ *for* $j \leq i$ **then**
  - **continue**
- **end**
- **if** *there are very small modules in* $A^{(i)}$ **then**
  - remove one row from $P^{(i+1)}$ that corresponds to the smallest module;
  - **continue**
- **end**
- **break**

end

Fig. 1. Proposed algorithm of clustering in network and correlation spaces

Рис. 1. Предлагаемый алгоритм совместной кластеризации в графовом и корреляционном пространствах

На i-ой итерации алгоритма выполняется два основных шага:

1. По приближению матрицы активностей модулей на предыдущей итерации $P^{(i)}$ выполняется поиск потенциальных активных модулей – подграфов $A^{(i)}$.

2. По полученным на предыдущем шаге подграфам $A^{(i)}$ выполняется коррекция матрицы активностей и получается приближение $P^{(i+1)}$ для следующего шага.

Для получения начального приближения матрицы активностей модулей $P^{(1)}$ будем использовать алгоритм кластеризации *k-medoids* для некоторого значения k, являющегося параметром. На вход этому алгоритму передается матрица d корреляционных расстояний между профилями (строчками X_i):

$$d(V_i, V_j) = 1 - \mathrm{corr}(X_i, X_j), \tag{1}$$

где corr – корреляция Пирсона. В результате работы алгоритма получается разбиение всех вершин графа на кластеры с хорошей внутренней корреляцией. Однако эти кластеры не соответствуют связным подграфам в графе G. В то же время при достаточно большом значении k, большем числа истинных активных модулей K, для всех строчек матрицы активностей модулей P в полученных кластерах будет хотя бы один медоид (вершина – центр кластера) с высокой корреляцией профиля с активностью модуля. Таким образом, в качестве начального приближения матрицы активностей модулей $P^{(1)}$ можно взять матрицу размерности $k \times m$, в которой i-ая строчка равна профилю i-го медоида.

Теперь рассмотрим шаг поиска потенциальных активных модулей по приближению матрицы активностей $P^{(i)}$ размерности $k^{(i)} \times m$.

Во-первых, определим вес, отражающий насколько хорошо вершина графа подходит к тому или иному профилю. Сначала введем корреляционное расстояние между вершиной и профилем:

$$d(V_x, P^{(i)}_y) = 1 - \mathrm{corr}(X_x, P^{(i)}_y), x \in \{1, \dots, n\}, y \in \{1, \dots, k^{(i)}\}.$$

Далее введем фиктивный нулевой профиль, расстояние до которого по определению будет всегда равно некоторой константе *base*:

$$d(V_x, P^{(i)}_0) \equiv base, x \in \{1, \dots, n\}.$$

Затем определим «референсное» расстояние до ближайшего профиля, отличного от рассматриваемого:

$$d'(V_x, P^{(i)}_y) = \min_{z \in \{0,\dots,k^{(i)}\}, z \neq y} d(V_x, P^{(i)}_z), x \in \{1, \dots, n\}, y \in \{1, \dots, k^{(i)}\}.$$

Наконец, определим искомый вес как:

$$w(V_x, P^{(i)}_y) = -\log \frac{d(V_x, P^{(i)}_y)}{d'(V_x, P^{(i)}_y)}. \tag{2}$$

Введенный вес обладает следующими свойствами:

1. Чем выше корреляция профиля вершины V_x с профилем модуля $P^{(i)}_y$, тем выше вес $w(V_x, P^{(i)}_y)$.
2. Вес $w(V_x, P^{(i)}_y)$ может быть положительным только если профиль $P^{(i)}_y$ является ближайшим к профилю вершины V_x. В частности, для заданной вершины V_x только для одного y вес $w(V_x, P^{(i)}_y)$ может быть положительным.
3. Вес $w(V_x, P^{(i)}_y)$ может быть положительным, только если корреляция между профилем вершины V_x и профилем модуля $P^{(i)}_y$ больше $1 - base$.

Таким образом, чем более положительным является вес $w(V_x, P^{(i)}_y)$, тем увереннее мы можем сказать, что вершина V_x должна принадлежать модулю для профиля $P^{(i)}_y$.

Теперь, определив вес $w(V_x, P^{(i)}_y)$ для всех вершин V_x и некоторого профиля $P^{(i)}_y$, мы можем найти такой связный подграф $A^{(i)}_y$, что суммарный вес его вершин максимален:

$$\sum_{v \in V(A^{(i)}_y)} w(v, P^{(i)}_y) \longrightarrow \max.$$

Задача поиска связного подграфа максимального веса (maximum weight connected subgraph, MWCS) является NP-полной. Однако для нее существуют несколько практических программ решателей, в том числе точных [6, 8]. Кроме того, для этой задачи есть быстрый эвристический решатель, часто находящий оптимальные или близкие к оптимальным решения [9]. В настоящей работе мы будем использовать именно этот метод, реализованный в программном пакете *mwcsr* для языка R (https://github.com/ctlab/mwcsr).

После того, как были получены подграфы $A^{(i)}$, по ним можно построить новые профили $P^{(i+1)}$. Для этого для каждого модуля усредним нормализованные (центрированные и деленные на дисперсию) значения профилей вершин модулей $A^{(i)}$ с положительным весом w. В случае, если в каком-то модуле мало положительных вершин (на практике – если одна или две), то в качестве профиля будем использовать значение с предыдущей итерации. Это позволяет не сойтись в локальный оптимум модуля.

Полученные профили $P^{(i+1)}$ затем сравниваются с профилями на предыдущих итерациях. Если профиль $P^{(i+1)}$ не совпадает ни с одним предыдущим профилем, то итерации продолжаются дальше.

В случае, если профиль $P^{(i+1)}$ совпал с каким-то из предыдущих профилей, то выполняется проверка, все ли найденные модули $A^{(i)}$ являются достаточно большими. В случае, если хотя бы один модуль имеет небольшой порядок (на практике – четыре или меньше) или небольшой диаметр (на практике – два или меньше), то выполняется процедура удаления одного модуля из рассмотрения. В качестве модуля для удаления выбирается модуль с минимальным числом вершин. Если таких модулей несколько, то выбирается один из пары наиболее скоррелированных модулей. При удалении модуля удаляется соответствующая строчка в $P^{(i+1)}$ и на единицу уменьшается число модулей $k^{(i+1)}$.

Если профиль $P^{(i+1)}$ совпал с каким-то из предыдущих профилей и все найденные модули удовлетворяют критериям по размеру, то алгоритм завершается. Последним шагом является пересчет модулей $A^{(i+1)}$ для профиля $P^{(i+1)}$. Таким образом, на выходе алгоритма получается приближение матрицы профилей активности модулей $P^* = P^{(i+1)}$ и приближение набора активных модулей $A^* = A^{(i+1)}$.

3. Экспериментальное исследование работы алгоритма

3.1. Описание симулированных данных

В настоящей работе мы будем рассматривать три типа матрицы профилей активности модулей P, соответствующие различным дизайнам биологических экспериментов. Для всех типов все биологические состояния представлены в трех повторностях – типичное число для биологических экспериментов, в которых анализируется экспрессия генов. Для упрощения сравнения между разными типами матриц во всех типах в матрицах представлены шесть биологических состояний и десять модулей.

В первой матрице P^S (рисунок 2) рассматривается простой эксперимент из шести биологических состояний. Первое состояние является контрольным – в нем активности всех модулей равны нулю. Кроме этого, есть пять других независимых состояний, и для каждого состояния есть по два независимых модуля: один активирующийся (значение активности равно единице), другой – подавляемый (значение активности равно минус одному).

В следующей матрице P^C (рисунок 3) рассматривается более сложная ситуация. В этой матрице рассматривается шесть биологических состояний и десять модулей. Каждый модуль при этом может быть активен в нескольких из биологических состояний и подавлен в остальных. Состояния, в которых активны модули, были выбраны случайным образом.

В последней матрице P^T (рисунок 4) рассматривается тип эксперимента, в котором некоторый процесс исследуется в нескольких временных точках. Как и в матрице P^S, в нем присутствует кон-

$$P^S = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1
\end{pmatrix}$$

Fig. 2. Matrix P for a simple experiment design with five biological conditions (excluding the control), with each condition having two corresponding modules: one that gets activated and one that gets inhibited

Рис. 2. Матрица P для простого типа эксперимента с пятью биологическими состояниями, не считая контрольного, где каждому состоянию соответствуют два модуля: активирующийся и подавляющийся

$$P^C = \begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0
\end{pmatrix}$$

Fig. 3. Matrix P for a complex experiment design with six biological conditions, with each module being active in a certain subset of the conditions

Рис. 3. Матрица P для сложного типа эксперимента с шестью биологическими состояниями, где каждый модуль активен в некотором подмножестве состояний

трольное состояние. Еще пять состояний соответствуют последовательным временным точкам. Для каждой временной точки присутствует модуль, который в ней начинает активироваться, и модуль, который начинает подавляться. Для такого профиля активности может усложниться разделение модулей, так как модули соответствующие близким временным точкам имеют высокую корреляцию профилей.

В качестве графа G во всех экспериментах рассматривался граф белок-белковых взаимодействий, используемый в пакете *BioNet*. Граф состоит из 2034 вершин и 7756 ребер.

Активные модули A_i генерировались следующим образом. Во всех случаях число вершин в модуле выбиралось случайно перестановкой множества $\{20, 40, \dots, 200\}$. После того, как порядок модуля был выбран, модуль выбирался равномерным случайным образом из всех связных подграфов такого порядка. Для генерации случайных подграфов использовался пакет *mcmcRanking*.

После того, как выбраны модули A_i, задавалась матрица B вклада модуля в профили вершин. Для всех ненулевых элементов $B_{i,j}$ (соответствующих вершинам i, входящим в модуль A_j) значение случайно выбиралось из экспоненциального распределения $\mathrm{Exp}(\lambda)$ с параметром $\lambda = 1$.

$$P^T = \begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1
\end{pmatrix}$$

Fig. 4. Matrix P for a time course experiment design where each module either gets activated or inhibited at certain time point and keeps its state till the end of the experiment

Рис. 4. Матрица P для эксперимента с несколькими временными точками, в котором каждый модуль либо активируется, либо подавляется в некоторый момент времени и сохраняет свое состояние до конца эксперимента

Наконец, шум ε генерировался из нормального распределения $\mathcal{N}(0, \sigma^2)$. Значение среднеквадратичного отклонения σ являлось варьируемым параметром – большие значения среднеквадратичного отклонения усложняют задачу поиска модулей.

3.2. Базовые методы для сравнения

Для сравнения использовалось три базовых метода.

Первый метод выполняет кластеризацию *k-medoids* по матрице корреляционных расстояний (1). Значение k может варьироваться. Результатом является набор из k непересекающихся множеств – найденных кластеров.

Второй метод также выполняет кластеризацию, но методом *WGCNA* (weighted gene correlation network analysis) [10]. Аналогично предыдущему, для кластеризации этот метод использует корреляционное расстояние. Одной из особенностей метода является то, что в этом методе заранее не задается число кластеров, а его работа может регулироваться другими параметрами. Другой особенностью является то, что он способен выделить «мусорный» кластер, состоящий из вершин, которые не относятся ни к какому из «настоящих» кластеров с хорошей внутренней корреляцией. В качестве результата для сравнения используется набор из выделенных методом *WGCNA* непересекающихся кластеров, за исключением «мусорного».

Третьему методу (будем называть его *nearest*) передается истинная матрица профилей P и параметр *base*. Метод вычисляет вес $w(V_x, P_y^{(i)})$ по формуле (2). Метод возвращает K кластеров, где в i-й кластер входят все вершины V_x, для которых $w(V_x, P_y^{(i)}) > 0$.

3.3. Сравнение методов получения стартовых приближений

В первом эксперименте было проведено сравнение способов получения стартовых приближений. В сравнении участвовало четыре метода:

1. Метод *k-means* со значениями k из множества $\{16, 24, 32, 64, 96\}$. В качестве расстояния использовалось евклидово расстояние. Дополнительная модификация значений экспрессии не проводилась.
2. Метод *k-medoids* со значениями k из множества $\{16, 24, 32, 64, 96\}$. В качестве метрики использовалось корреляционное расстояние.
3. Метод *WGCNA*. В качестве финального профиля использовалось покоординатное усреднение профилей, входящих в соответствующий кластер.

4. Метод *all* – метод для сравнения, который возвращал все строки исходной матрицы X.

Все методы были запущены на всех трех типах исходных матриц P. Среднеквадратичное отклонение для шума σ выбиралось из множества $\{0.25, 0.5, 1\}$. Каждый эксперимент проводился пять раз для разных исходных значений состояния генератора случайных чисел.

Для оценки качества получаемых стартовых приближений использовалась следующая процедура. Вычислялись попарные корреляции между строчками истиной матрицы P и стартовыми приближениями S, полученными некоторым методом. Затем для каждой строки матрицы P выбиралось максимальное значение корреляции, таким образом, получалось значение того, как хорошо соответствующая строка P представлена в матрице S. Наконец, вычислялись две суммирующие метрики: *corAvg* – усредненное значение корреляции для всех строк P, и *corMin* – минимальное значение корреляции.

На рисунке 5 приведены результаты сравнения. Для упрощения визуализации данные по разным типам матриц P объединены, так как показывают похожее поведение. В соответствии с ожиданиями, с увеличением среднеквадратичного шума все методы хуже справляются с восстановлением исходных профилей. Методы *k-means* и *k-medoids* показывают похожие значения для одинаковых значений k, и эти результаты улучшаются с увеличением k, однако на значении $k = 64$ уже наступает насыщение. Метод $WGCNA$ почти всегда проигрывает методам *k-medoids* и *k-means* с $k \geq 32$, кроме случая большого шума ($\sigma = 1$), где результаты похожи.

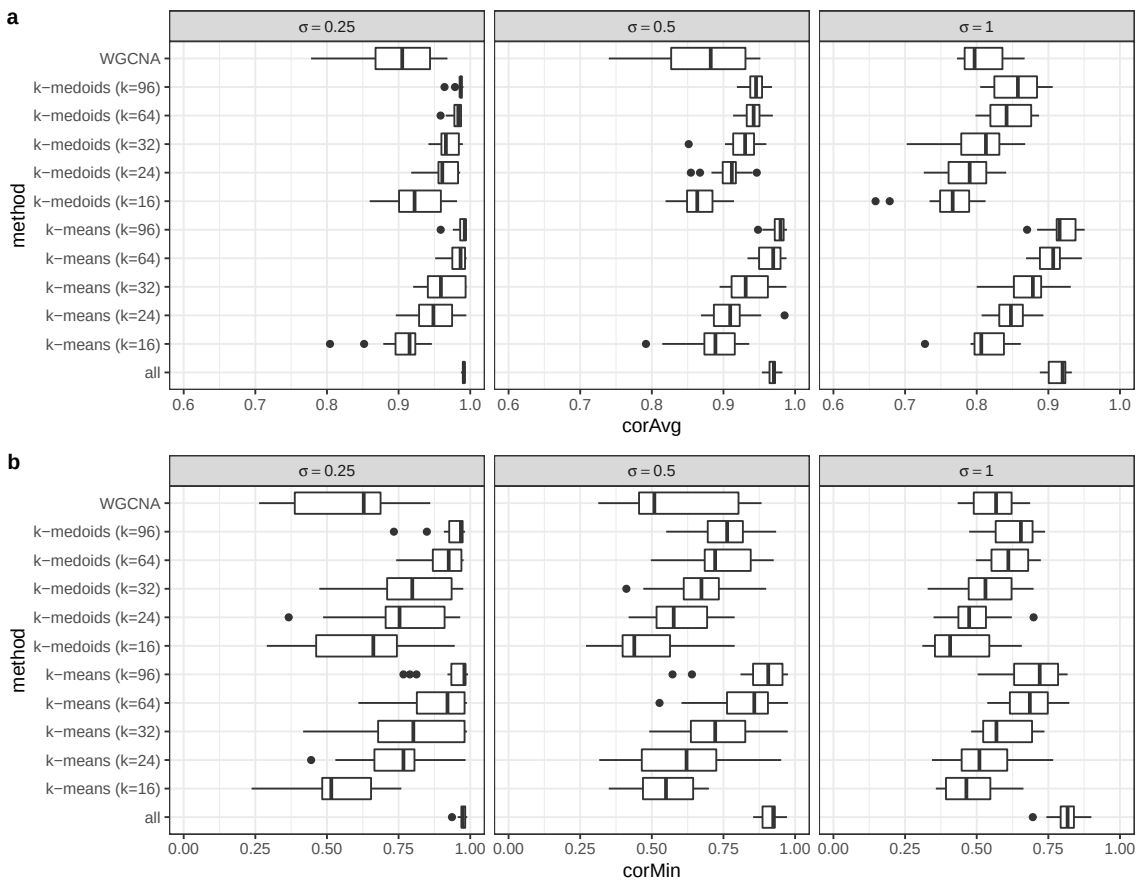

Fig. 5. Comparison of different start profiles generation methods by correlation with true values of P

Рис. 5. Сравнение различных способов получения стартовых профилей по корреляции с истинными профилями P

Также отметим, что для методов *k-means* и *k-medoids* для достижения хороших показателей требуются значения k в несколько раз больше истинного числа модулей (десяти). Таким образом, актуальной является задача разработки метода для восстановления матрицы P даже без ограничений на связность модулей в графе G.

3.4. Старт с истинных значений

В следующем эксперименте было проведено исследование качества работы предлагаемого метода кластеризации при старте с истинных значений P. Как и в предыдущем эксперименте, запуск проводился на всех трех матрицах P^S, P^C и P^T. Среднеквадратичное отклонение для шума σ выбиралось из множества $\{0.25, 0.5, 1\}$. Каждый эксперимент проводился пять раз для разных исходных значений состояния генератора случайных чисел.

Для каждого сгенерированного набора предлагаемый метод (*net-clust*) запускался со значениями *base* из множества $\{0.2, 0.3, 0.4, 0.5, 0.6\}$. Текущая матрица P передавалась в качестве стартового значения $P^{(1)}$. Для сравнения использовался метод *nearest*, описанный выше, которому также передавалась матрица P и значение *base*.

Оценка качества результатов выполнялась следующим образом. Во-первых, задача рассматривалась как задача классификации вершин на те, которые принадлежат хотя бы одному модулю, и те, которые не принадлежат. В этом случае можно вычислить метрики точности (*precision*) и полноты (*recall*). Однако эти метрики не отражают, насколько хорошо вершины разделяются на отдельные модули. Чтобы это учесть, для каждого найденного модуля вычислялась максимальная доля его вершин, полностью совпадающих с одним из истинных модулей A. Усредненное значение этих долей обозначалось как метрика *average module consistency* – чем ближе ее значение к единице, тем лучше.

Результаты анализа метрик точности и полноты представлены на рисунке 6 (панели *a* и *b*, соответственно). Как и в предыдущем эксперименте, поведение метрик не сильно зависело от типа матрицы P, поэтому данные приведены в агрегированном по всем типам виде. Результаты показывают, что добавление ограничения на связность влияет на результаты, но не очень сильно. Для больших значений *base* метод *net-clust* дает большую точность, но меньшую полноту.

Также на рисунке 6*c* приведен анализ метрики *average module consistency*. В целом, поведение этой метрики достаточно хорошо повторяет поведение метрики точности, однако на ней наблюдается выход предлагаемого метода *net-clust* на плато при значениях $base \leq 0.4$ и «провале» относительно базового метода *nearest*.

3.5. Исследование работы алгоритма с разных способов получения начальных приближений

Затем было исследовано, насколько влияет способ получения начальных приближений матрицы P на конечный результат работы алгоритма. Для получения начальных приближений использовались методы *k-means* и *k-medoids* со значениями $k = 32$ и $k = 64$. Исследование проводилось только для $P = P^S$. Среднеквадратичное отклонение для шума σ выбиралось из множества $\{0.25, 0.5, 1\}$. Для каждой пары матрицы P и выбранного уровня шума σ генерировалось по три набора данных для разных исходных значений состояния генератора случайных чисел.

Результаты исследования показали, что результат работы алгоритма не сильно зависит от того, как выбирался начальный набор. Так как время работы алгоритма было значительно больше для $k = 64$ по сравнению с $k = 32$ и метод *k-medoids* при $k = 32$ показал чуть лучшие результаты по сравнению с методом *k-means* для того же k, в дальнейшем сравнении для начального приближения использовался только метод *k-medoids* с $k = 32$.

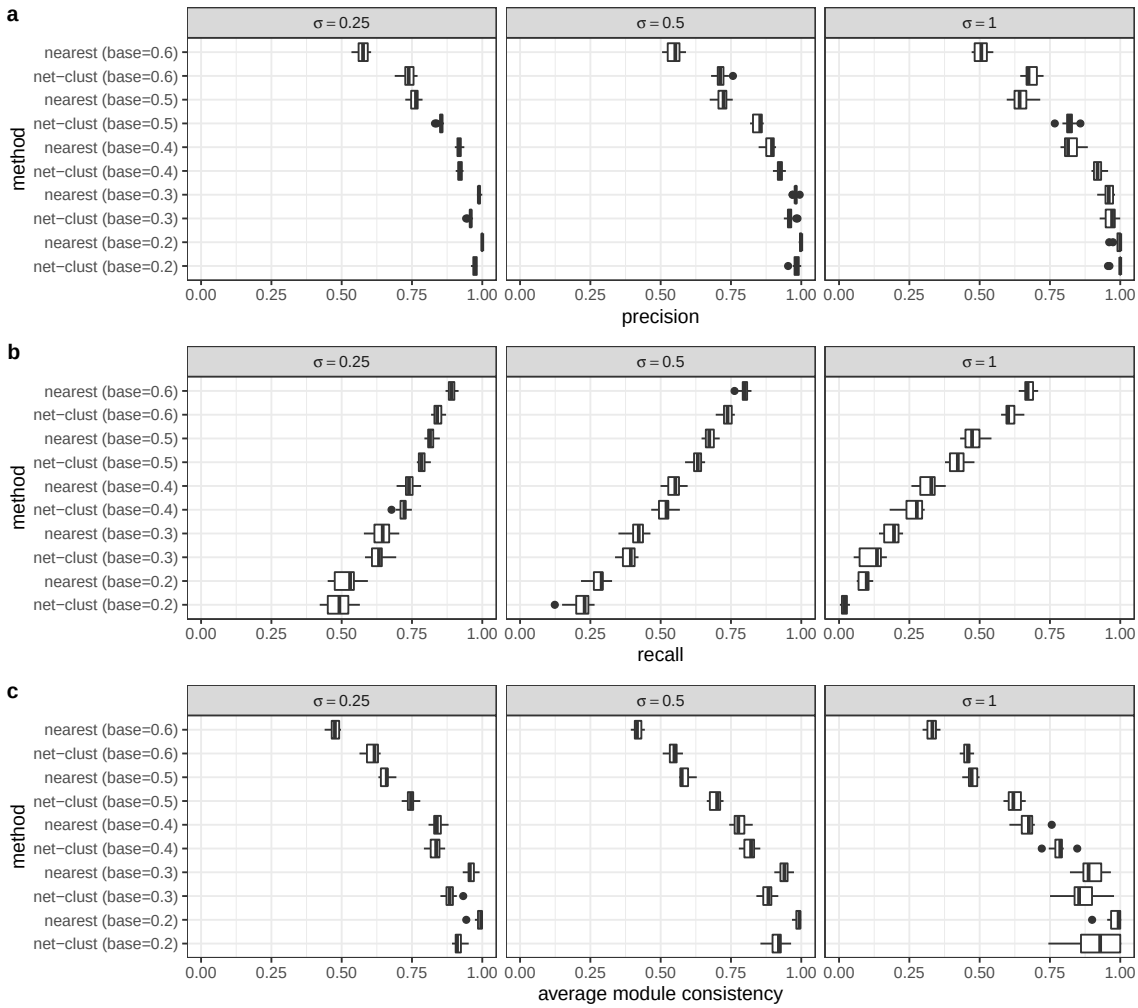

Fig. 6. Analysis of results of method *net-clust* as run with true values of P as a start approximation and compared with *nearest* method as a baseline

Рис. 6. Анализ результатов предлагаемого метода *net-clust* по сравнению с базовым методом *nearest* при старте с истинного значения матрицы P

3.6. Итоговое сравнение алгоритма с базовыми методами

Последний эксперимент заключался в запуске предлагаемого алгоритма от начала до конца и сравнении его с базовыми методами. Исходные данные для эксперимента генерировались для всех трех матриц P^S, P^C и P^T. Как и в предыдущих экспериментах, среднеквадратичное отклонение для шума σ выбиралось из множества $\{0.25, 0.5, 1\}$. Для каждой пары матрицы P и выбранного уровня шума σ генерировалось по три набора данных для разных исходных значений состояния генератора случайных чисел.

Предлагаемый метод запускался со стартовых значений, полученных методом *k-medoids* с параметром $k = 32$. Параметр *base* выбирался из множества $\{0.3, 0.4, 0.5\}$. Для сравнения использовался метод *k-medoids*, также с параметром $k = 32$, метод *WGCNA* и метод *nearest* с таким же выбором параметра *base*.

Результаты оценки качества алгоритмов представлены на рисунке 7. Как и ранее, результаты для разных типов матрицы P отличаются незначительно, поэтому приведены в агрегированном виде.

Из этого эксперимента видно, что предлагаемый метод позволяет достигать высокой точности в определении модулей при достаточно неплохой полноте.

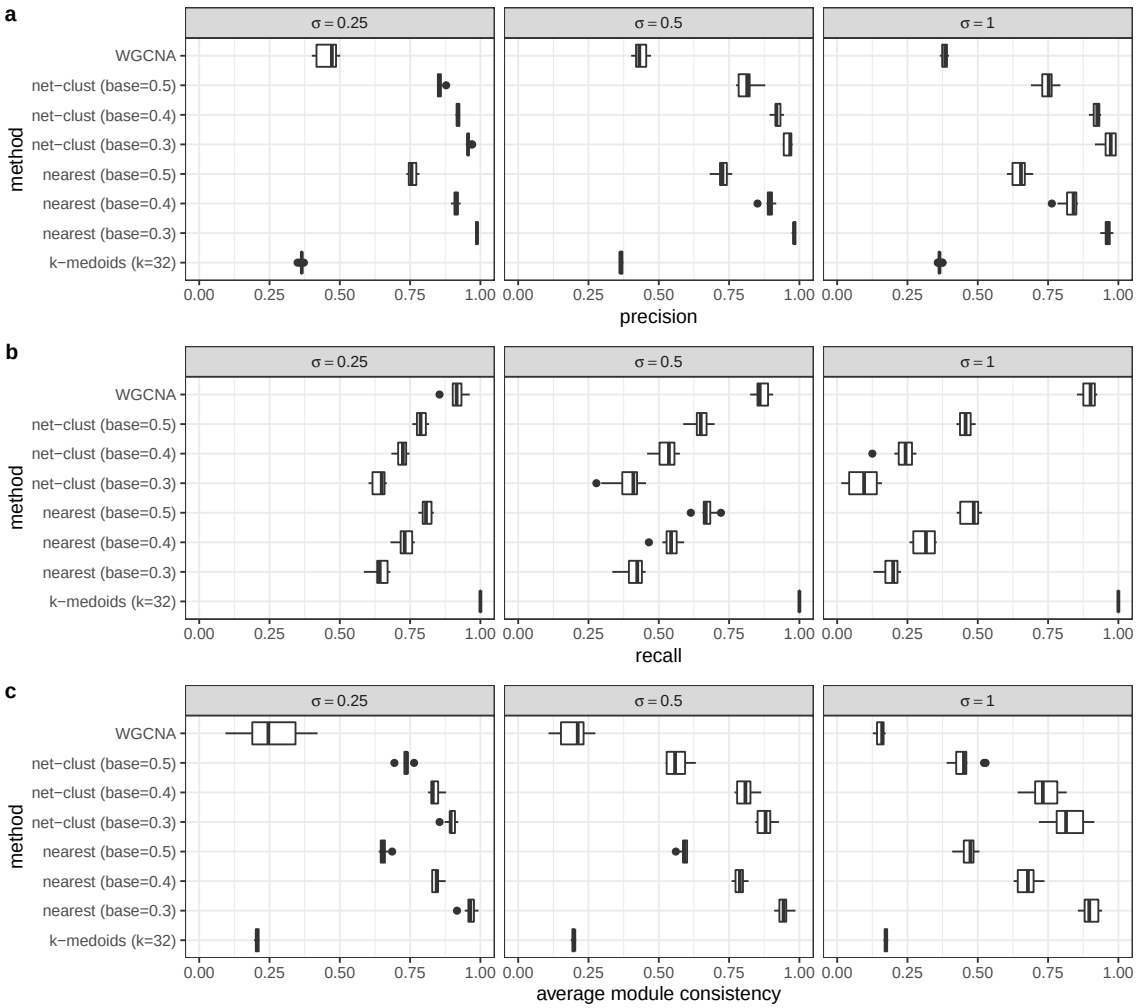

Fig. 7. Analysis of results of method *net-clust* compared with baseline methods

Рис. 7. Анализ результатов предлагаемого метода *net-clust* по сравнению с базовыми методами

Отдельно важно отметить, что качество работы предлагаемого алгоритма сравнимо с качеством базового метода *nearest*, запускаемого с теми же значениями *base*. При этом методу *nearest* на вход передается истинная матрица P, а предлагаемый метод *net-clust* выводит ее автоматически.

Заключение

В работе рассматривается задача поиска активных модулей в графах белок-белковых взаимодействий по данным экспрессии генов. Математически, эта задача была сформулирована как задача совместной кластеризации в графовом и корреляционном пространствах.

В настоящей работе впервые сформулирована математическая модель этой задачи и предложены метрики оценки качества решений. В рамках этой формулировки для решения задачи был предложен метод *net-clust* основанный, с одной стороны, на идеях решения задачи поиска одного активного модуля, с другой – на идеях алгоритма *k-means*. В методе последовательно приближается

набор профилей искомых кластеров (центров) и находятся связные подграфы, профили вершин которых хорошо коррелируют с центрами.

Так как для решения этой задачи отсутствуют точные аналоги, для оценки качества метода были сформулированы несколько базовых методов. В качестве базовых, были рассмотрены как методы кластеризации, применяющиеся для анализа экспрессии генов: *k-medoids* и *WGCNA*, так и метод, использующий информацию об истинных значения центров кластеров – не применимый на практике, но позволяющий оценить качество оптимального решения. Все базовые методы не требуют связности кластеров в заданном графе биологических связей, в отличие от предлагаемого метода *net-clust*.

Экспериментальное исследование было проведено на симулированных данных, для которых известен правильный ответ и, таким образом, на которых можно сравнить качество предлагаемого метода с базовыми методами. Исследование показало, что для получения начальных приближений профилей кластеров подходят методы *k-medoids* и *k-means* при значениях k в несколько раз превышающих истинное число модулей. При этом дальнейшее увеличение k не приводит к улучшению результатов, но увеличивает время работы метода.

Сравнение с практическими методами *k-means* и *WGCNA* показало, что метод *net-clust* имеет значительно более высокую точность. С другой стороны, предлагаемый метод *net-clust* достигает параметров качества сравнимых с методом, принимающим на вход истинные значения центров кластеров, что означает, что качество метода *net-clust* близко к оптимально возможному для этой задачи.

Таким образом, было показано, что задача поиска активных модулей может быть сформулирована как задача совместной кластеризации в графовом и корреляционном пространствах, которая, в свою очередь, может быть решена близко к оптимальности с помощью предлагаемого метода *net-clust*.

References

[1] K. Mitra, A. R. Carvunis, S. K. Ramesh, and T. Ideker, “Integrative approaches for finding modular structure in biological networks”, *Nat. Rev. Genet.*, vol. 14, no. 10, pp. 719–732, 2013.

[2] E. J. Rossin *et al.*, “Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology”, *PLoS Genet.*, vol. 7, no. 1, e1001273, 2011.

[3] A. K. Jha, S. C. Huang, A. Sergushichev, V. Lampropoulou, Y. Ivanova, E. Loginicheva, K. Chmielewski, K. M. Stewart, J. Ashall, B. Everts, E. J. Pearce, E. M. Driggers, and M. N. Artyomov, “Network integration of parallel metabolic and transcriptional data reveals metabolic modules that regulate macrophage polarization”, *Immunity*, vol. 42, no. 3, pp. 419–430, 2015.

[4] M. D. Leiserson *et al.*, “Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes”, *Nat. Genet.*, vol. 47, no. 2, pp. 106–114, 2015.

[5] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, “Discovering regulatory and signalling circuits in molecular interaction networks”, *Bioinformatics (Oxford, England)*, vol. 18 Suppl 1, S233–S240, 2002.

[6] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Müller, “Identifying functional modules in protein-protein interaction networks: an integrated exact approach.”, *Bioinformatics (Oxford, England)*, vol. 24, no. 13, pp. i223–31, 2008, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btn161.

[7] M. N. Artyomov, A. Sergushichev, and J. D. Schilling, “Integrating immunometabolism and macrophage diversity”, *Semin. Immunol.*, vol. 28, no. 5, pp. 417–424, Oct. 2016.

[8] A. A. Loboda, M. N. Artyomov, and A. A. Sergushichev, "Solving Generalized Maximum-Weight Connected Subgraph Problem for Network Enrichment Analysis", in *Algorithms in Bioinformatics: 16th International Workshop, WABI 2016, Aarhus, Denmark, August 22-24, 2016. Proceedings*. Cham: Springer International Publishing, 2016, pp. 210–221, ISBN: 978-3-319-43681-4.

[9] E. Álvarez-Miranda and M. Sinnl, "A Relax-and-Cut framework for large-scale maximum weight connected subgraph problems", *Computers & Operations Research*, vol. 87, pp. 63–82, 2017.

[10] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis", *BMC Bioinformatics*, vol. 9, p. 559, 2008.