УДК 543:001.891 (01)

ГРУППИРОВАНИЕ ПРИРОДНЫХ СОЕДИНЕНИЙ С ИСПОЛЬЗОВАНИЕМ АССОЦИАТИВНОГО БИБЛИОМЕТРИЧЕСКОГО АНАЛИЗА

А.М. МАЛЬКО, В.Ю. КИСТАНОВА, С.А. ГУСЕВ

Библиометрический анализ позволяет алгоритмическим способом сопоставлять объекты, имеющие индивидуальные обозначения в текстах научных статей. В качестве таких объектов в работе рассматривали наименования 1812 химических соединений природного происхождения. Ассоциативные связи между объектами устанавливали путем сопоставления библиографических списков, полученных для каждого соединения в результате поиска соответствующих наименований в реферативной библиотеке PubMed. Результат обработки >500 тыс. рефератов отображали в виде сетевой диаграммы, графически представляющий рассчитанные смысловые взаимосвязи между объектами. В составе диаграммы были выявлены обособленные кластеры, содержащие соединения, сходные по химической структуре, проявляемым биологическим или химическим свойствам. В работе показано, что ассоциативный анализ позволяет обобщать информацию о строении и свойствах природных соединений.

Ключевые слова: природное соединение, библиография, ассоциативные связи, классификация.

Количество научных работ, посвященных получению и исследованию свойств новых химических соединений из природного сырья, постоянно увеличивается. Это обусловлено, прежде всего, поиском природных веществ, обладающих лекарственным действием [1]. Ферментативные системы растений катализируют сложные биохимические реакции, многие из которых невозможно повторить в лабораторных условиях. В свою очередь, по мере совершенствования аналитических методик будет появляться все больше информации об уникальных компонентах природных экстрактов.

Информация о физико-химических и биологических свойствах природных экстрактов содержится в научных публикациях в описательной форме. Публикации довольно разнородны: в одних работах рассматриваются вопросы, связанные с выделением и очисткой веществ, в других — анализируется их структура, в-третьих — исследуется биологическая или терапевтическая активность. Возникает задача обобщения разнородных сведений в форме, удобной для анализа основных тенденций в области практического использования компонентов природных экстрактов.

Автоматические средства интерпретации сведений, содержащихся в научных статьях, активно разрабатывались в применении к информации о функциональных свойствах белковых молекул [2]. Например, в одних работах в результате автоматического распознавания наименований биомолекул были установлены связи между 5 тыс. белками и 1,7 тыс. раз-

ФГУ «Россельхозцентр», Москва; Общество с ограниченной ответственностью «КуБ», Москва.

личными заболеваниями [3]. В другой работе тематическая декомпозиция научных статей позволила выявить группы белков, вовлеченных в процесс программированной клеточной гибели [4] или участвующих в обеспечении клеточного метаболизма [5]. В данной работе предлагается применить подходы для анализа природных химических соединений, сходные с процитированными выше.

Для решения поставленной задачи в данной работе использовали возможности электронной библиотеки PubMed [6]. Эта система предоставляет доступ к 19 млн. рефератов статей в области биомедицины и биотехнологии. В систему PubMed встроен алгоритм определения ассоциативных связей между публикациями [7]. Публикации считаются родственными по смыслу, если у них совпадают частотные характеристики употребления ключевых терминов. С использованием встроенного в PubMed алгоритма оценки родственности статей ассоциативные связи между природными соединениями графически отображали в виде сетевой диаграммы, для которой характерно: а) наличие связности между объектами одного кластера; б) объединение объектов в кластеры при наличии сходства химической структуры (отнесение к одному классу природных соединений), сходного типа биологической активности, сходности проявляемых химических свойств. Кластеры, образованные группами соединений на диаграмме, были охарактеризованы по принадлежности к классам химических соединений и по биологической активности.

Материалы и методы исследования

Исходную выборку для выполнения работы получили из базы данных «Библиография природных соединений» [8]. В выборку вошли наименования 1812 химических соединений, каждому из которых соответствовал регистрационный номер CAS. Это уникальный численный идентификатор химических соединений, внесённых в реестр Chemical Abstracts Service.

Регистрационные номера CAS использовали в качестве ключей, по которым проводили поиск в системе PubChem [9] для получения информации о наименованиях химических соединений. В соответствии с методикой, предложенной в работе [5], наименования химических соединений направляли по HTTP-запросу в систему PubMed. В ответ на запрос система PubMed предоставляет перечень релевантных статей, в которых упоминается данное химическое соединение. Для каждого химического соединения идентификаторы таких статей объединяли в состав релевантного библиографического профиля. Затем, для каждой релевантной публикации определяли 5 родственных по смыслу документов, используя поле «Related Articles» системы PubMed, и включали их в состав родственного библиографического профиля. При этом в число родственных публикаций не включали тривиальные случаи, когда в тексте статьи наименования двух любых соединений встречаются совместно. Техническое описание методики и подпрограммы для создания библиографических профилей доступны на интернет-сайте [11].

Ассоциативные связи рассчитывали между каждой парой химических соединений. Принимали за ***к*** количество ссылок, совпадающих в библиографических профилях родственных публикаций двух соединений, индекс сходства г вычисляли по формуле [12]

$$r = k \,/\, (m + n - k), \qquad (1)$$

где ***т и п*** — количество статей в библиографических профилях одного и другого соединения.

Для визуализации результатов использовали программу построения

сетевых диаграмм GVEdit [13]. Программа размещает на диаграмме регистрационные номера соединений так, чтобы наиболее оптимальным образом отразить существующие между этими соединениями попарные ассоциативные взаимосвязи. Построение диаграмм проводили при разных значениях индекса г, при этом, чем ниже задавалось значение индекса г, тем а) менее специфичные связи получали отображение на диаграмме; б) большее количество объектов входило в состав диаграммы; в) меньшее количество дискретных кластеров отображалось на сетевой диаграмме вследствие объединения отдельных кластеров в один мажорный. Последовательное построение сетевых диаграмм при разных индексах сходства привело к определению оптимального значения индекса сходства г > 0,032, при котором сетевая диаграмма содержала наибольшее количество дискретных кластеров, содержащих наибольшее количество объектов, характеризующихся наличием специфичных взаимосвязей между каждой парой, вошедшей в состав диаграммы.

Результаты и их обсуждение

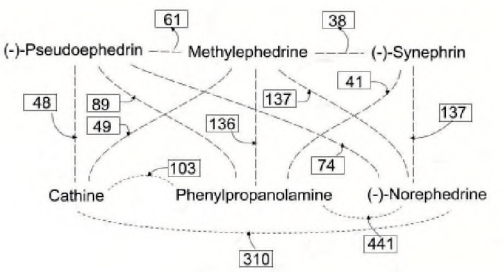
Многообразие природных соединений крайне широко, поэтому практически невозможно в литературном обзоре или монографии полностью описать их специфические свойства. При высоком темпе исследований в области получения и анализа природных компонентов, а также в настоящий момент характерному данной области науки активному совершенствованию новых высокоэффективных инструментальных методов выделения, очистки, а также идентификации природных химических соединений, информация, изложенная в обзорах, быстро теряет свою актуальность. Поэтому задача обработки и анализа больших объёмов получаемой информации становится крайне сложной и длительной. В данной работе для обобщения наиболее актуальных опубликованных сведений о строении и свойствах природных соединений предлагается полностью автоматический алгоритм установления ассоциативных связей. При этом фактическая визуализация актуальной информации позволяет в достаточно краткий период времени ознакомиться с конкретным интересующим исследователя, а также, возможно, ранее не известным, объектом, наблюдая в то же самое время его взаимосвязи с другими объектами, информация о которых уже имеется и широко известна.

Наименования большинства химических соединений представляют собой уникальные обозначения. Это позволяет средствами поиска соотнести каждому соединению список релевантной литературы, в которой упоминается наименование определенного вещества. Ассоциативная связь между двумя соединениями оценивается в зависимости от количества совпадений между соответствующими библиографическими списками.

Попарные ассоциативные связи были вычислены по формуле (1) для 1620 соединений, вследствие чего только для 820 соединений значение индекса сходства превысило установленный порог 0,032, при котором в составе сетевой диаграммы было представлено наибольшее количество дискретных кластеров, характеризующихся наличием наибольшего количества объектов. Отобранные 820 соединений были отображены на сетевой диаграмме, пример одного из кластеров, присутствующих на диаграмме, показан на рисунке. Узлами представленной сетевой диаграммы являются наименования химических соединений, а показанные пунктирами ребра в соответствии с количеством общих публикаций попарно соединяют химические соединения. Количество общих для двух объектов релевантных публикаций указано

стрелкой для каждой отображённой взаимосвязи. Так, соединение Cathine имеет 103 общие релевантные публикации с Phenylpropanolamine, 48 — с (-)- Pseudoephedrin и 49 — с Methylephedrine.

При выбранном уровне сходства r в составе сетевой диаграммы образовался доминирующий кластер из 380 объектов, а также относительно меньшие кластеры, сведения о которых представлены в таблице. Крупный кластер объединил разнородные по своей химической структуре,



Фрагмент сетевой диаграммы, относящийся к кластеру №8 (см. таблицу). Пунктирными линиями обозначены связи между химическими соединениями; для каждой связи указано количество общих для двух соединений родственных публикаций

способам получения и биологическим свойствам вещества. Для более детального анализа этого кластера необходимо использовать более высокие пороговые значения индекса сходства r, так как при повышении индекса r соединения, вошедшие в состав этого мажорного кластера, образуют дискретные кластеры небольших размеров, однако при этом общее число отображаемых соединений становится меньше 380. Для рассматриваемого мажорного кластера можно, в целом, отметить высокую встречаемость и частоту статей, относящихся к описанию свойств алкалоидов, в том числе опиатов.

Для остальных кластеров, включавших от 13 до 89 объектов, также был проведен анализ публикаций, характерных для большинства ассоциативных связей. Выраженные ассоциации, обусловленные сходством химической природы соединений, наблюдали в 9 кластерах из 14. Например, в кластере №6 сгруппированы соединения, относящиеся к группе природных гликозинолатов. В крупный кластер №8 вошли 89 соединений, которые образуют 42 ассоциативные связи между микотоксинами и афлотоксинами, индикация которых

Характеристики кластеров в составе семантической диаграммы химических соединений природного происхождения

№ кластера	Среднее значение индекса *r*	Количество веществ	Примечание
1	0,061	18	Пирролизидиновые алкалоиды
2	0,058	82	Эфирные масла
3	0,047	42	Флавоноиды/антиоксиданты
4	0,043	17	–
5	0,041	40	–
6	0,041	25	Гликозинолаты
7	0,039	13	Токоферолы
8	0,039	89	Афлотоксины/микотоксины
9	0,036	21	Углеводы
10	0,035	13	Желчные кислоты и их конъюгаты
11	0,034	16	Эфедрины и их производные
12	0,033	14	Экдистероиды
13	0,033	15	Производные дисахаридов
14	0,032	13	Индукторы опухолей

важна в пищевой отрасли. Кластер №3 содержит 22 родственные связи между соединениями — флавоноидами, которые объединились в силу проявляемых ими антиоксидантных свойств.

Для двух кластеров в составе полученной диаграммы (№4 и №5) не удалось выявить превалирующих публикаций, которые позволили бы соотнести вошедшие в эти кластеры соединения с определенными классами. В данном случае кластеры сформировались по общности препаративных методик, т.е. составляют набор химических соединений, выделение и очистка которых проводится с использованием определённого набора методов. Например, в кластере №4 ассоциативные связи обусловлены публикациями по теме хроматографической очистки компонентов природной биомассы.

В состав кластера №10 вошли природные вещества — продукты биосинтеза млекопитающих. Эта группа представлена желчными кислотами и их конъюгатами, анализ которых проводится в рамках исследования механизмов гепатотоксичности. Отдельную группу составляют представители кластера №14, куда вошли соединения, индуцирующие опухолевый рост клеток. Такие соединения, к которым, например, относится ***12-О-тетрадеканоил-форбол-13-а.цетат*** (индуктор опухоли кожных покровов), применяются в экспериментальных моделях для исследования механизмов онкогенеза [14].

Кластеры, приведенные в таблице, могут быть также охарактеризованы по своей плотности. Каждая связь, отображённая на диаграмме, имеет индекс сходства выше 0,032. Средние значения индекса г для кластера приведены в таблице. Представленные данные указывают, что кластеры в составе сетевой диаграммы отличаются средними значениями индекса ***г.*** Наибольшее значение г = 0,061 получено для кластера №2. Это означает, что кластер обладает компактной структурой, и вершины в его составе образуют много перекрестных ребер, т.е. химические соединения, вошедшие в состав этого кластера, являются очень тесно связанными, при этом каждое соединение связано более чем с 2 соседними. Другие кластеры, например № 11, 12, 13, наоборот, обладают низкой плотностью связей (каждое соединение редко образует более 3 взаимосвязей с другими), и поэтому соответствующие значения индекса находятся практически на уровне выбранного порога отсечения. Можно предположить, что различия усредненных индексов связаны с интенсивностью выполнения исследований в области соответствующих групп соединений.

Сетевая диаграмма позволяет получить общее представление о дифференцированном распределении объектов исследования по семантически родственным группам (кластерам). Варьирование индекса ассоциативного сходства позволяет регулировать степень обобщения опубликованного материала. В интересных участках сетевой диаграммы структура ассоциативных связей между химическими соединениями может быть детализирована. Каждая ассоциативная связь, появляющаяся на диаграмме, может быть интерпретирована с привлечением обеспечивающих эту связь статей. Преимуществом данного подхода является также быстрота построения представленной информационной модели, которая может быть использована всякий раз при возникновении необходимости обобщения поступившей информации. Ранее сходный подход использовали применительно к анализу генов [2] и белков [3,4], однако при этом связи устанавливали только для тех объектов, названия которых вместе встречались в одном

реферате. В данной работе мы показали, что для природных компонентов сетевые диаграммы могут быть получены на основе анализа смысловой родственности публикаций, т.е., когда названия объектов совместно в тексте документа не упоминаются.

Вывод

Ассоциативные связи химических соединений, установленные путем сопоставления библиографических профилей, отражают распределение природных компонентов в соответствии с общностью их химического строения, источников получения и методов экстракции.

Библиографический список

1. *Rishton G.M.* Natural products as a robust source of new drugs and drug leads: past successes and present day issues // Bioinformatics, 2007. 8.
2. *Stapley B.J., Benoit G.* Biobibliometrics information retrieval and visualization from co-occurrence of gene names in Medline abstracts //Pac symp Biocomput, 2000. 529-540.
3. *Bundschus M. et al.* Extraction of semantic biomedical relations from text using conditional random fields// BMC Bioinformatics, 2008.9.
4. *Zheng B., Lu* 1 Novel metrics for evaluating the functional coherence of protein groups via protein semantic network// Genome Biol., 2007.8.
5. *Пономаренко Е.А. и соавт.* Создание семантических сетей белков с использованием Pubmed/Medline //Молекулярная биология (принято в печать), 2009.
6. www.pubmed.org
7. *Lin J, WilburW.J.* PubMed related articles: a probabilistic topic-based model for content similarity // BMC Bioinformatics, 2007.8.
8. *Кудрявцев А.М. и соавт.* Библиография природных соединений (www.ookub.ru\ npbb\) //Свидетельство о регистрации базы данныхдля ЭВМ, № 2009620326, ФИПС. М., 2009.
9. http://pubchem.ncbi.nlm.nih.gov/
11. http: //www.ookub.ru/upload/fckeditor/SOP.rar
12. *Rogers D.J.,Tanimoto T.T.* A Computer Program for Classifying Plants //Science, 1960. 132, 1115-1118.
13. www.graphviz.org
14.Wattenberg E.V.// Physiol Cell Physiol., 2007. 292(1). C. 24-32.

Рецензент — д. с.-х. н. А.Н. Березкин

SUMMARY

Analysis of semantic associations enables the algorithmic comparison of the objects, which are individually entitled in the documents' texts. The designations of 1812 chemical compounds of natural origin were treated as such objects in the course of this work. Associative links are created by matching the bibliography lists, attributed to each compound via searching the PubMed for respective designations. Results of processing >500 thousand abstracts are displayed as a network diagram. Within this network the isolated clusters have been indicated, which comprise the compounds, sharing structural similarity and biological activity. It has been shown that associative semantics can be used to generalize the information on structure and properties of natural compounds.

Key words: natural compound, bibliography, associative links, classification

Малько Александр Михайлович — д. с.-х. н., ФГУ «Россельхозцент».
Кистанова Валерия Юрьевна — асп. ФГУ «Россельхозцент».
Эл. почта: vkistanova@mail.ru
Гусев Семен Александрович — к. б. н., Общество с ограниченной ответственностью «КуБ».