Биомедицинская химия, 2020 том 66, вып. 1, с. 7-17.

МЕДИЦИНСКИЕ ПРЕДМЕТНЫЕ РУБРИКИ ДЛЯ АНАЛИЗА ЭВОЛЮЦИИ НАУЧНЫХ ГРУПП НА ПРИМЕРЕ НАУЧНОЙ ШКОЛЫ АКАДЕМИКА А.И. АРЧАКОВА

Е.В. Ильгисонис*, О.И. Киселева, А.В. Лисица, Е.В. Поверенная, М.Н. Топоркова, Е.А. Пономаренко

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; *эл. почта: ilgisonis.ev@gmail.com

Предложен метод сравнительного анализа научных траекторий ученых на основе библиографического профиля (т.н. “мешпринт”), который представляет собой перечень терминов MeSH (ключевые термины, используемые для индексации статей в библиотеке биомедицинских текстов PubMed) с указанием относительной частоты встречаемости каждого термина в статьях учёного. Сопоставление персонализированных библиографических профилей можно представить в виде семантической сети, где узлами являются фамилии учёных, а связи пропорциональны рассчитанным мерам сходства библиографических профилей. Данный метод был использован для анализа семантической сети ученых, объединённых научной школой академика А.И. Арчакова. Результаты работы позволили показать взаимосвязи между научными траекториями одной научной школы и соотнести результаты с мировыми трендами развития научных направлений.

Ключевые слова: семантические сети; анализ текстов; медицинские предметные рубрики; text-mining; MeSH; социология; research trajectory

DOI: 10.18097/PBMC20206601007

ВВЕДЕНИЕ

Цифровизация общества и развитие Интернета делает возможным перенести фокус исследования с классических объектов молекулярной биологии (генов, белков, заболеваний) на самих учёных. В этом контексте наряду с традиционной задачей биологии и биохимии исследования молекулярных сетей интересно также исследовать эволюцию социальных групп учёных, объединённых, например, одной научной школой.

Параллели между молекулярными и социальными сетями проводятся во многих исследованиях [1-3]. Анализ траектории фокуса внимания учёного представляет существенный интерес, поскольку позволяет проследить направление научного поиска, а систематизация и анализ такого рода примеров являются интереснейшей базой для развития методов компьютерного обучения и прогнозирования [4] .

Функциональные возможности библиотеки биомедицинских текстов PubMed/MEDLINE позволяют создавать алгоритмы, анализирующие в автоматическом режиме как содержание резюме научных публикаций [5, 6], так и сопутствующие мета-данные (ключевые слова, авторов, географическую принадлежность, финансирование и др. ([https://www.ncbi.nlm.nih.gov/books/NBK153385/]).

Принимая во внимание, что набор ключевых слов (терминов MeSH [7]) в сжатом виде отражает содержание статьи, мы использовали именно эту информацию для создания персонализированных библиографических профилей. За последние годы возросло количество работ, в которых объектом исследования являются не сами тексты научных публикаций, а термины MeSH. Их используют для анализа научных трендов [8, 9], лекарственных взаимодействий [10], научных журналов [11], механизмов развития заболеваний [12], оценки индивидуальной исследовательской траектории [4]. Создан целый ряд инструментов (так называемые “PubMed derivates” [13]), базирующихся на частотном анализе терминов MeSH для упрощения поиска релевантной информации в базе PubMed. Примерами таких систем являются GoPubMed [14] (не поддерживается с 2017 года), MeSHy [15] и др.

Перечень ключевых терминов в сочетании с “весом” каждого термина – долей статей учёного, которые проиндексированы с использованием данного термина – позволяют сгенерировать уникальный профиль исследователя MeSHPrint (по аналогии с FingerPrint – отпечатком пальца), отражающий его специфичную и исключительную область компетенций на научной карте. По аналогии с различиями в отпечатках пальцев у людей, мы предполагаем, что ситуации полностью совпадающих у учёных мешпринтов не встречаются. Даже в случае, если они долгое время работают над решением сходных научных задач, персональное публикационное портфолио будет различаться.

В данной работе мы предлагаем метод автоматического формирования мешпринтов учёных и метод расчёта близости персональных библиографических профилей. Визуализация результатов позволяет оценить дивергенцию научных путей исследователей, являющихся учениками одной научной школы. В качестве примера мы проанализировали семантическую карту, полученную при анализе библиографических профилей 69 ученых, защитившего докторскую или кандидатскую диссертации под руководством академика А.И. Арчакова.

Александр Иванович Арчаков создал научную школу в области изучения молекулярной организации и функционирования оксигеназных цитохром Р450-содержащих систем, исследования молекулярных механизмов структуры и функции мембран и биологического окисления. Он предложил схему молекулярной организации оксигеназной системы печени, разработал методы её реконструкции из изолированных белков и липидов. Под руководством А.И. Арчакова разработан принципиально новый лекарственный препарат с противовирусной активностью “Фосфоглив” для лечения заболеваний печени. В настоящее время этот препарат широко используется в практической фармакологии [16].

Современные научные интересы А.И. Арчакова связаны с исследованиями в области постгеномных технологий, нанобиотехнологий, развитием подходов к созданию персонализированной медицины. А.И. Арчаков является основоположником развития протеомики в России, одним из инициаторов крупнейшего международного проекта “Протеом человека” [17, 18]. Детализированная семантическая карта учёного представлена в работе [4].

МЕТОДИКА

Формирование персонализированных мешпринтов

В группу ученых, для которых формировали персонализированные библиографические профили (мешпринты), включены исследователи, защитившие докторские и/или кандидатские диссертации под руководством А.И. Арчакова с 1969 по 2017 гг. (полный список приведён в Дополнительных материалах), из которых 58 человек – кандидаты наук, 19 – доктора наук (из них 8 человек защитили кандидатскую диссертацию также под руководством А.И. Арчакова).

Фамилии из перечня в форме латинской транслитерации направляли в качестве запроса к библиотеке PubMed. Запрос состоял из фамилии и/или инициалов конкретного учёного (поскольку ряд статей не всегда содержит указание отчества авторов). Пример запроса для И.И. Карузиной: [*Karuzina II {Author} OR Karuzina I {Author}*]. Нерелевантные статьи удаляли в ходе экспертной проверки результатов поискового запроса. В результате запроса фамилию учёного ассоциировали с перечнем идентификаторов статей – PMID, для которых с использованием API PubMed загружали перечень соответствующих статье ключевых слов – терминов MeSH. На основе полученных данных формировали мешпринт – уникальный для каждого автора перечень ключевых терминов с указанием “веса” каждого термина. “Вес” термина рассчитывали путём нормирования количества статей каждого автора, проиндексированных данным термином, на общее количество статей автора. Мешпринт формировали на основе 100 терминов, для которых отношение частот встречаемости в публикациях автора более чем на порядок отличалось от частоты встречаемости этого же термина в библиотеке PubMed. Фрагменты сформированных мешпринтов представлены на рисунке 1.

(а)

#	MeSH_AAI	Qty	Freq
1	Oxidation-Reduction	82	18.4
2	Kinetics	75	16.9
3	Rabbits	71	16
4	Cytochrome P-450 Enzyme System/metabolism	70	15.7
5	Electron Transport	38	8.5
6	Microsomes, Liver/enzymology	47	10.6
7	Amino Acid Sequence	35	7.9
8	In Vitro Techniques	36	8.1
9	Protein Binding	37	8.3
10	Binding Sites	24	5.4
11	Catalysis	27	6.1
12	Protein Conformation	25	5.6
13	Hydroxylation	23	5.2
14	Liposomes	22	4.9
15	Substrate Specificity	24	5.4
16	Biosensing Techniques	20	4.5
17	Cytochrome P-450 Enzyme Inhibitors	20	4.5
18	Electrodes	19	4.3
19	Microsomes, Liver/metabolism	20	4.5
20	Models, Molecular	19	4.3
21	Proteomics/methods	22	4.9
22	Biosensing Techniques/methods	14	3.1
23	Computer Simulation	13	2.9
24	Electrophoresis, Gel, Two-Dimensional	12	2.7
25	Electrophoresis, Polyacrylamide Gel	16	3.6
26	Liver/cytology	12	2.7
27	Liver/enzymology	13	2.9
28	Liver/metabolism	11	2.5
29	Macromolecular Substances	12	2.7
30	Mass Spectrometry/methods	12	2.7
31	NAD/metabolism	18	4
32	Proteins/analysis	12	2.7
33	Proteins/chemistry	11	2.5
34	Proteomics	15	3.4
35	Spectrometry, Mass, Matrix-Assisted Laser Desorption-Ionization	19	4.3

(б)

#	MeSH_LAV	Qty
1	Proteomics	15
2	Proteomics/methods	12
3	Spectrometry, Mass, Matrix-Assisted Laser Desorption-Ionization	11
4	Hep G2 Cells	8
5	Tandem Mass Spectrometry	8
6	Microsomes, Liver/enzymology	8
7	Amino Acid Sequence	7
8	Cytochrome P-450 Enzyme System/metabolism	7
9	Electrophoresis, Gel, Two-Dimensional	7
10	Molecular Sequence Data	6
11	Transcriptome	6
12	Electrophoresis, Polyacrylamide Gel	6
13	Software	5
14	Protein Binding	5
15	Databases, Protein	5
16	Blood Proteins/analysis	4
17	Chromatography, Liquid	4
18	Rabbits	4
19	Oxidation-Reduction	4
20	Mass Spectrometry	4
21	Proteins/analysis	4
22	Metal Nanoparticles/chemistry	3
23	Limit of Detection	3
24	Computational Biology	3
25	Gold/chemistry	3
26	Linear Models	3
27	Proteome	3
28	Liver/metabolism	3
29	Biomarkers/blood	3
30	Polymorphism, Single Nucleotide	3
31	Gene Library	3
32	Gene Expression Profiling	3
33	Sensitivity and Specificity	3
34	Mass Spectrometry/methods	3
35	Alternative Splicing	3

Рисунок 1. Пример мешпринта (**а**) А.И. Арчакова (MeSH_AAI) (**б**) А.В. Лисицы (MeSH_LAV). Общее количество доступных в библиотеке PubMed работ составляет 444 и 85 соответственно. *Qty* – количество публикаций, проиндексированных термином MeSH, *Freq* – отношение *Qty* к общему количеству публикаций автора (“вес”).

Расчёт расстояния между мешпринтами и визуализация

Отличия между мешпринтами *D(a,b)* рассчитывали как сумму дистанций между *n* терминами MeSH (см. формулу 1), где *n* – количество уникальных терминов MeSH в составе обоих мешпринтов, *Mi* – меш-термин, *Freq(a)* – частота встречаемости термина *M* в профиле *(a)*, *Freq(b)* – частота встречаемости термина *M* в профиле *(b)*:

$$D(a,b) = \sum_{i=1}^{n} Mi \cdot [Freq(a) - Freq(b)] \quad (1).$$

На основе полученных данных визуализировали общую семантическую карту, узлами которой являются имена ученых, а длина связи между узлами пропорциональна значению *D(a,b)*.

Аннотацию полученной семантической сети проводили с использованием данных о терминах MeSH, совпадающих между узлами сети. Полученные результаты сопоставляли с информацией о частоте встречаемости данного термина в библиотеке PubMed, аффилиацией учёного.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

За полувековой период от момента защиты первой кандидатской диссертации под руководством А.И. Арчакова – с 1969 по 2019 год – в библиотеке Pubmed присутствуют 1,6 тыс. статей, найденных по именам авторов из сформированного перечня участников научной школы (см. раздел “МЕТОДИКА”). На рисунке 2а представлено распределение количества защит по годам. Видно, что распределение количества кандидатских и докторских диссертаций носит “волнообразный” характер; при этом амплитуда колебаний докторских защит существенно меньше и максимум сдвинут на несколько лет во времени. Очевидно, что выбранный временной период и эволюция тематик научной школы отражают естественные закономерности научного поиска, при котором цикл от научной идеи до её реализации занимает несколько десятилетий [19].

Приведённое на рисунке 2а распределение согласуется с результатами анализа индивидуальной научной траектории А.И. Арчакова. Наиболее “продуктивный” по количеству подготовленных учеников период – это 1980-1984 годы; в это время в фокусе исследований научной школы впервые появляется семейство цитохромов Р450. Результаты работы в этой области стали основой защищённых в период 1985-1989 гг. докторских диссертаций (А.В. Карякина, Е.А. Бородина и А.Н. Арипова), посвященных исследованию процессов при повреждении и восстановлении биологических мембран при различных заболеваниях.

Несмотря на то, что количество защищаемых под руководством А.И. Арчакова кандидатских диссертаций снижается, а докторских – остаётся на стабильном уровне, количество статей, публикуемых участниками его научной школы, неизменно растёт (см. рис. 2б). Таким образом, было показано, что люди, защитившие диссертации под руководством Александра Ивановича продолжают активную научную деятельность.

Продолжительность публикационной активности представителей научной школы

Продолжительность публикационной активности исследователя в какой-то мере является функцией от его возраста. Для всех диссертантов Александра Ивановича были загружены сведения о датах публикации их статей. В таблице перечислены 25 диссертантов с наибольшим количеством статей.

На основе полученных данных была построена тепловая карта, отражающая продолжительность и интенсивность публикационной активности представителей научной школы А.И. Арчакова (рис. 3). Анализ тепловой карты показывает, что на 2019 г. публикационно-активными являются 20 исследователей (более 30% учеников А.И. Арчакова).

На рисунке 4 представлена гистограмма распределения продолжительности публикационной активности для представителей научной школы А.И. Арчакова.

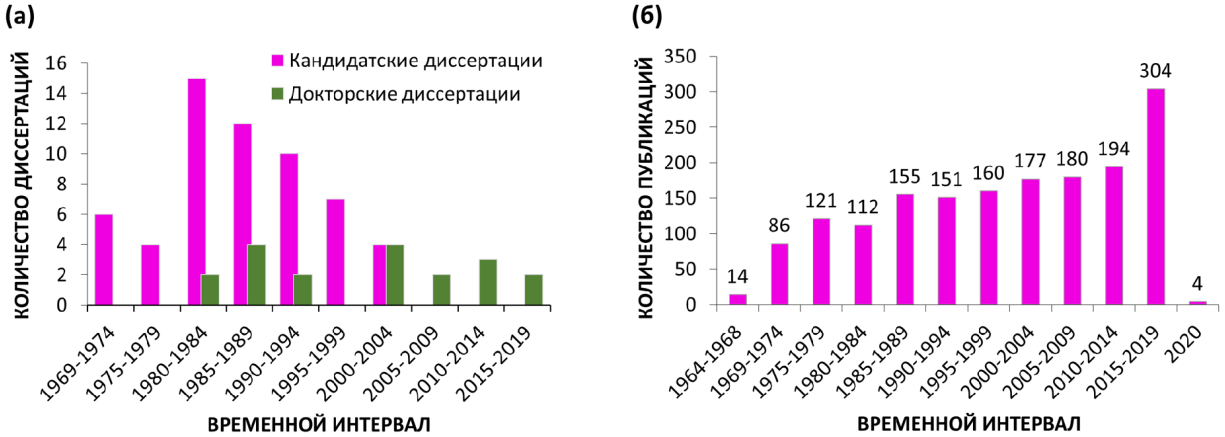


Рисунок 2. (**а**) Гистограмма распределения количества защищённых в рамках научной школы (под руководством А.И. Арчакова) кандидатских и докторских диссертаций в период с 1969 по 2019 гг. (**б**) Количество публикаций участников научной школы в те же временные интервалы (только для участников научной школы и А.И. Арчакова).

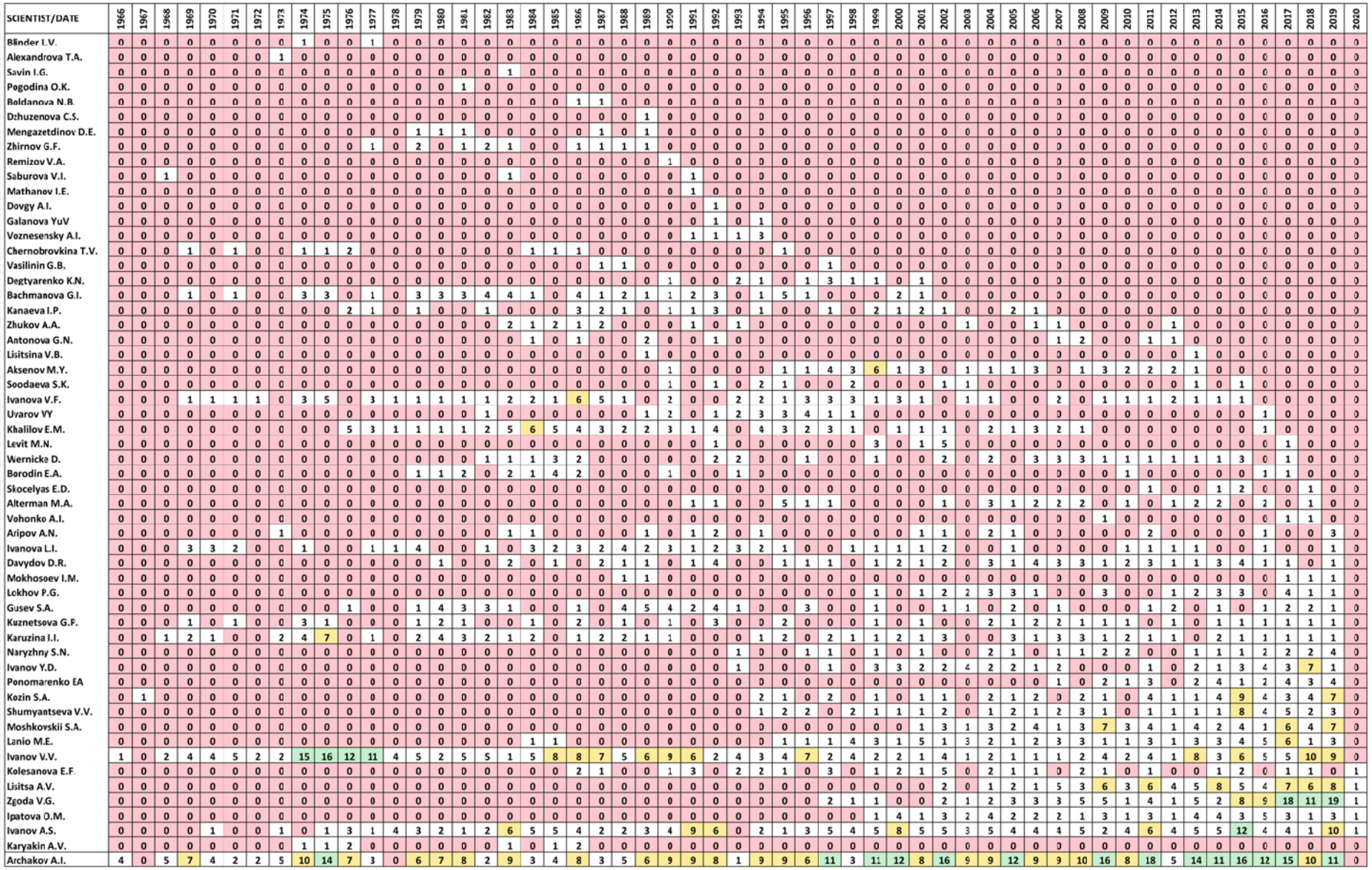


Рисунок 3. Тепловая карта, отражающая продолжительность и интенсивность публикационной активности представителей научной школы А.И. Арчакова. Каждый столбец соответствует одному году существования научной школы А.И. Арчакова, каждая строка – одному диссертанту. Красным цветом обозначено отсутствие публикаций, белым – 1-5 публикаций, жёлтым – 6-10 публикаций, зелёным – более 10 публикаций за указанный период.

Таблица. Топ-25 учеников А.И. Арчакова с наибольшей публикационной активностью

№	ФИО	Число публикаций
1	Ivanov A.S.	188
2	Zgoda V.G.	109
3	Lisitsa A.V.	79
4	Khalilov E.M.	77
5	Karuzina I.I.	76
6	Lanio M.E.	66
7	Moshkovskii S.A.	63
8	Kozin S.A.	56
9	Davydov D.R.	56
10	Ipatova O.M.	56
11	Gusev S.A.	55
12	Bachmanova G.I.	53
13	Shumyantseva V.V.	50
14	Ivanov Y.D.	47
15	Kuznetsova G.P.	41
16	Kolesanova E.F.	40
17	Aksenov M.Y.	39
18	Wernicke D.	39
19	Lokhov P.G.	33
20	Alterman M.A.	31
21	Kanaeva I.P.	29
22	Ponomarenko E.A.	29
23	Naryzhny S.N.	29
24	Uvarov V.Y.	22
25	Borodin E.A.	18

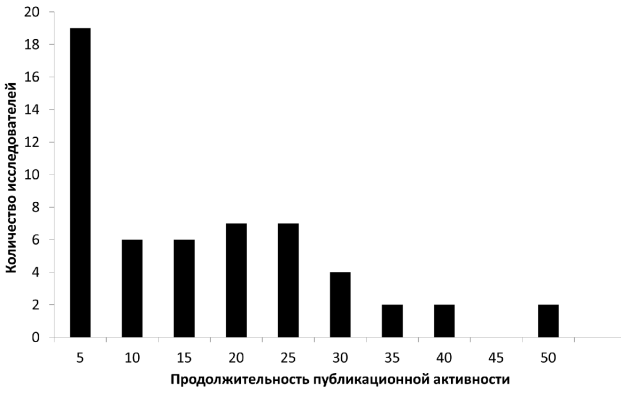


Рисунок 4. Гистограмма распределения продолжительности публикационной активности для представителей научной школы А.И. Арчакова.

На рисунке 4 видно, что 30% завершили свою публикационную активность в течение 5 лет, по-видимому, необходимых для подготовки диссертационной работы. Возможными причинами прекращения публикационной активности также могут быть смена сферы профессиональной деятельности, смена фамилии и т.п.

В среднем публикационная активность исследователей составляла 15 лет. На протяжении всего существования научной школы (50 лет) продолжают активно публиковаться два диссертанта А.И. Арчакова: И.И. Карузина и А.С. Иванов. Интересно, что большая часть учеников А.И. Арчакова, которые продолжают научную деятельность, являются сотрудниками Института биомедицинской химии имени В.Н. Ореховича (ИБМХ) и возглавляют собственные тематики исследований.

Количество публикаций, продолжительность активного публикационного периода, как и принятый в наукометрии индекс Хирша, не отражают в полной мере ценность научного вклада учёного, являясь своеобразной функцией от возраста учёного. При этом отношение количества цитирований работ к общему количеству работ является более-менее стабильным показателем. С использованием пакета Scopus 2.0.9 для языка программирования python (https://scopus.readthedocs.io/) для всех исследователей было загружено количество их цитирований. Было показано, что, в независимости от возраста, продолжительности публикационной активности, тематик исследований публикации всех представителей научной школы А.И. Арчакова в среднем цитируются 10-20 раз.

Согласно предложенному в работе алгоритму формирования мешпринтов (см. раздел “МЕТОДИКА”), для каждого учёного – участника исследуемой научной школы – загружали перечень публикаций с его авторством и соответствующие публикациям термины MeSH. Мешпринты были сформированы для 55 учёных из числа участников научной школы и руководителя – А.И. Арчакова. В среднем для анализируемой группы учёных мешпринт был построен на основе 35 публикаций

На рисунке 5а представлено облако TOP100 встречающихся для исследуемой научной школы MeSH-терминов. Облако наглядно показывает “чемпионов” по частоте исследований внутри научной школы, к числу которых относятся термины “*цитохром P450*”, “*кинетика*”, “*микросомы печени*”, “*белки и аминокислоты*” (отфильтрованы малоинформативные термины, например, описывающие объект исследования: “*Human*”, “*Rabbit*” и т.п.). Взвешенный список тегов (рис. 5б) отражает направления исследований научной школы А.И. Арчакова, значительно обогащённые (более чем в 5 раз) по сравнению со всем массивом научных исследований из базы Pubmed. Такая визуализация позволяет сформировать представление о том, чем научная школа А.И. Арчакова отличается от “усреднённой” научной школы или отдельного учёного, и оценить специфичность научного поля. Список терминов с наибольшим обогащением во многом повторяет наиболее активно исследуемые направления школы: как и в случае самых популярных тематик внутри школы, прочные позиции лидеров по обогащению занимают тематики цитохромов P450 и микросом печени. Любопытно пронаблюдать то, чем различаются облака на рисунке 5: облако (а) на первом плане содержит глобальные термины, облако (б) – более конкретные и специфичные, тем самым даёт представление не только о сфере интересов научной школы, но и об объектах (HepG2, эритроциты) и методах исследований (масс-спектрометрия MALDI, тандемная масс-спектрометрия, жидкостная хроматография, двумерный гель-электрофорез, применение биосенсоров).

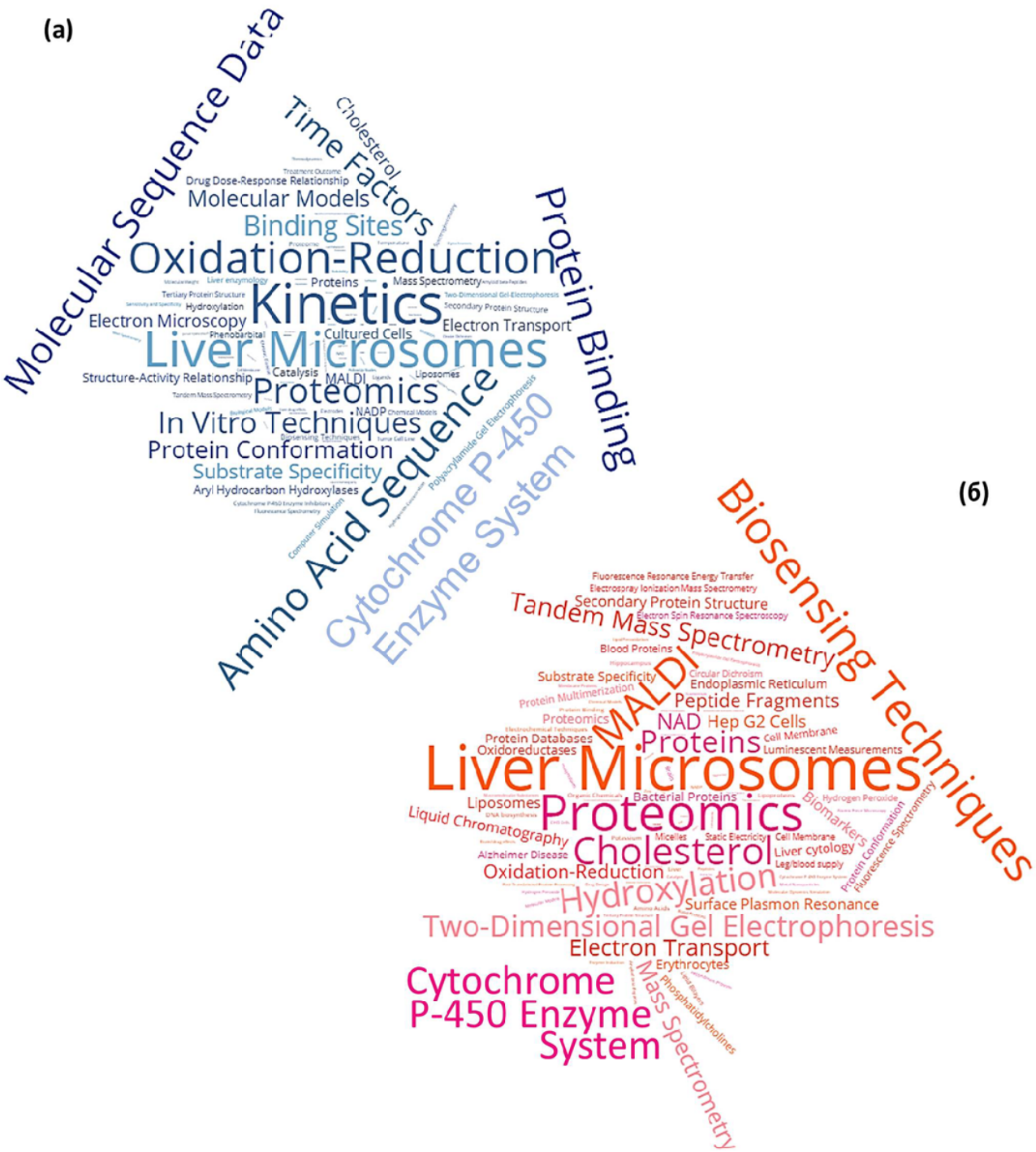
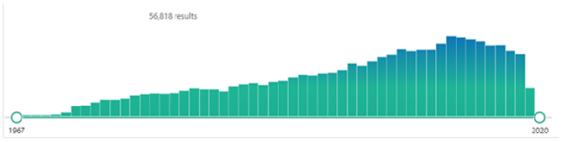


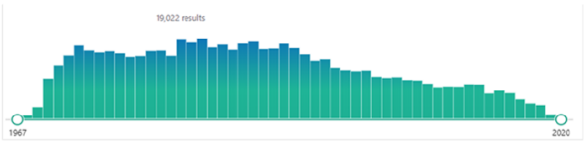
Рисунок 5. Облака выборки из TOP100 (**а**) наиболее часто встречающихся ключевых терминов научной школы и (**б**) характеризующихся максимальным обогащением по сравнению с базой PubMed. Размер шрифта коррелирует с частотой встречаемости термина и степенью обогащения.

Построение временных трендов наиболее специфичных терминов школы А.И. Арчакова позволяет по количеству публикаций в базе Pubmed оценить периоды интереса к ним всего мирового научного сообщества. На рисунке 6 приведены примеры тематик, которые набирают популярность и активно развиваются (в частности, Blood proteins/analysis и Biosensing Techniques), а также областей с тенденцией на спад – например, как в случае с двумерным гель-электрофорезом, который замещается новыми, более совершенными технологиями. Интересно рассмотреть тренды встречаемости термина Enzyme kinetics. В соответствии с данными базы PubMed, частота атрибутирования научных статей этим термином вышла на плато и даже снижается. Тем не менее, кинетика ферментов – это тема, определяющая существование целых индустрий и неизменно привлекающая внимание учёных из разных областей естественных наук. Работы по кинетике

(а) *Cytochrome P450 Enzyme System/metabolism*



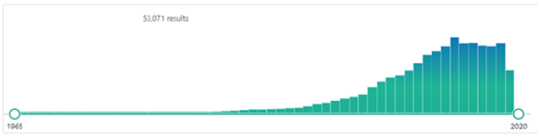
(б) *Microsomes, Liver/enzymology (*)*



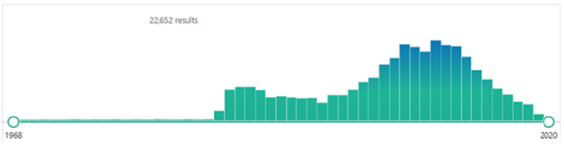
(в) *Proteomics/methods*

(г) *Spectrometry, Mass, Matrix-Assisted Laser Desorption-Ionization (*)*

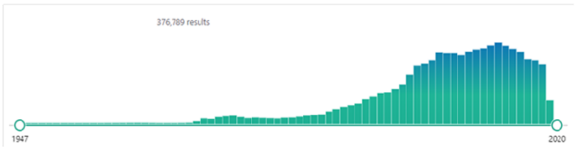
(д) *Biosensing Techniques*



(е) *Electrophoresis, Gel, Two-Dimensional (*)*



(ж) *Protein Conformation*



(з) *Blood Proteins/analysis*

(и) *Surface Plasmon Resonance(*)*

(к) *Enzyme kinetics(*)*

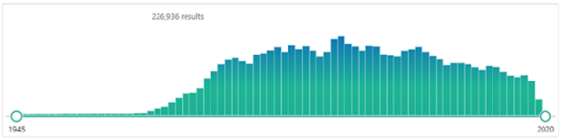


Рисунок 6. Публикационные тренды в библиотеке PubMed для наиболее специфичных для научной школы терминов (частота встречаемости терминов в статьях научной школы более, чем в пять раз превосходит частоту встречаемости этих же терминов в среднем в библиотеке PubMed); * – тенденция к спаду.

ферментов отмечены престижными премиями, в том числе Нобелевской (2018 г., Френсис Арнольд (Frances Arnold) и её исследования направленной эволюции ферментов). Можно предположить, что популярность этого (и, возможно, других) терминов имеет тенденцию к снижению потому, что в ходе разработки проблематики появляются более специфичные теги, точнее характеризующие работу.

Среди относительно молодых, но уверенно входящих в TOP15 наиболее активно изучаемых направлений исследуемой научной школы – протеомика и постгеномные исследования. А.И. Арчаков является ведущим российским учёным в области протеомики в России, одним из инициаторов крупнейшего международного проекта “Протеом человека” [17, 18]. На момент написания статьи, в библиотеке

PubMed по запросу *"(("proteomics"[MeSH Terms] OR "proteomics"[All Fields])"* возвращалось 92,3 тыс. статей, из которых российские учёные опубликовали 0,5 тыс. статей (см. рис. 5). Таким образом, доля публикаций в области протеомики российских авторов составляет около 0,5% всех публикаций данного направления.

На рисунке 7 представлено сопоставление долей публикаций российских авторов среди всего количества опубликованных в области работ. Графики опубликованных работ российскими учёными в области геномики, протеомики и метаболомики в целом повторяют мировые тренды как по данным WoS, так и PubMed. Количество работ в области геномики значительно превышает долю работ по протеомике и метаболомике, периодом "отрыва" являются 2008-2010 года, когда стоимость секвенирования генома стала резко снижаться благодаря появлению новых методов прочтения нуклеиновых молекул (https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost). С этого периода наблюдается и рост работ в области протеомики и метаболомики, что объясняется возрастающим интересом к понимаю взаимосвязи генетических событий с постгеномными данными. Согласно приведённым графикам, 2010 год стал переломным и для российских учёных; так, наблюдается рост доли отечественных работ по отношению ко всем публикациям в соответствующих областях (геномика ~ 1,5%, протеомика ~ 1,8% и метаболомика ~ 0,7%). Рост доли протеомных работ имеет наибольший темп, что можно объяснить участием России в проекте "Протеом человека" – втором крупнейшем примере международного сотрудничества в области биологии.

В 2018 году, по данным WoS, наблюдается уменьшение количества публикаций, относящихся к геномным и постгеномным исследованиям, выполненным как в мире, так и в России. Соответствующая мировая статистика по ресурсу PubMed также показывает снижение количества работ в области протеомики и метаболомики, а также резкий спад роста геномных работ. При этом для России наблюдается незначительное снижение в области метаболомики и видимый рост работ в области протеомики.

В ранее опубликованной работе, посвященной основным научным направлениям исследовательской работы А.И. Арчакова [4], было показано, что тематика исследований Александра Ивановича является доминирующей в рамках ИБМХ.

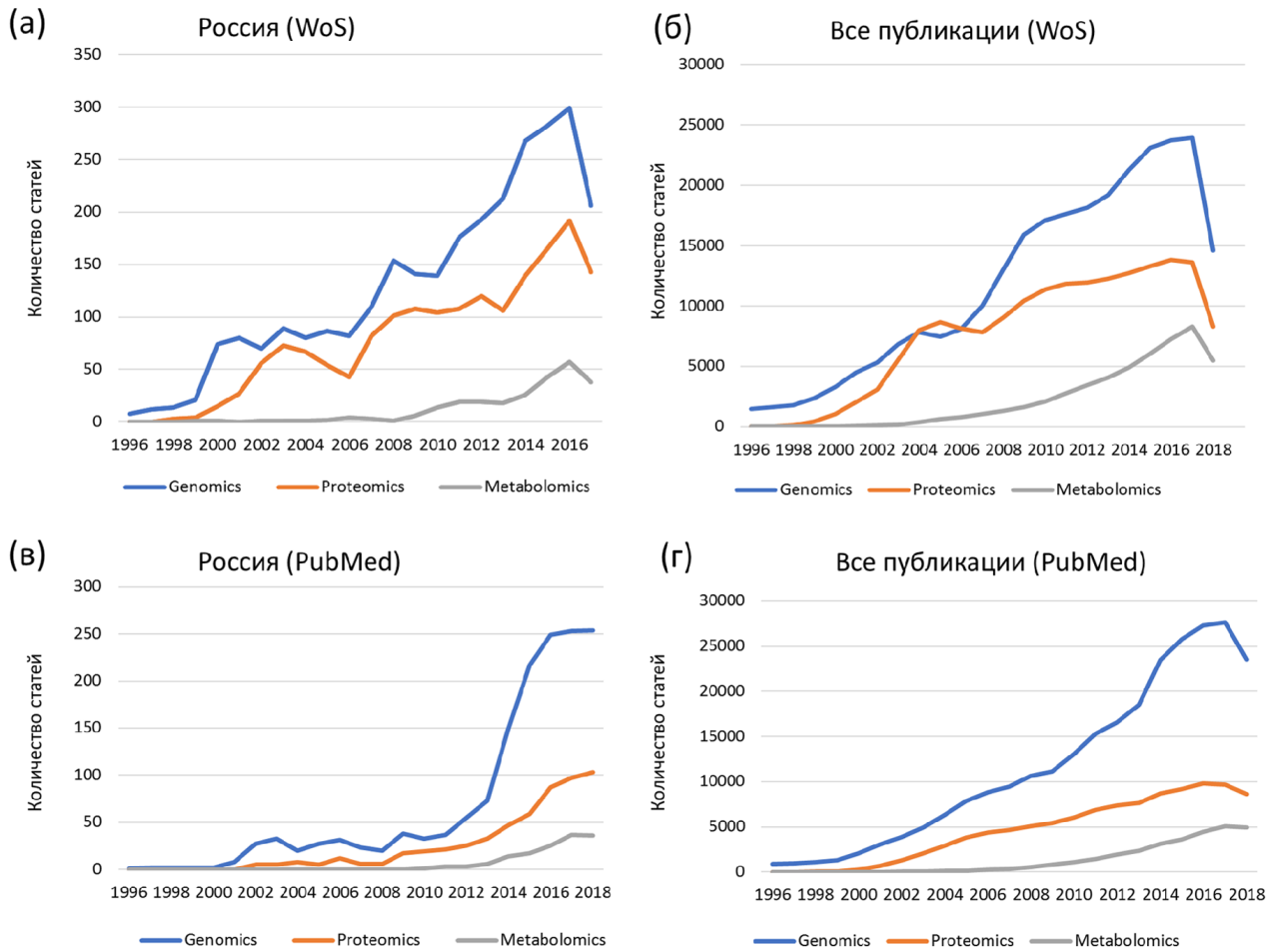


Рисунок 7. Число публикаций в России и в мире по "Омиксным" наукам (геномные, протеомные и метаболомные исследования) за 1996-2019 г. **(а)** выполненных российскими учеными, согласно ресурсу WoS; **(б)** выполненных суммарно в мире, согласно ресурсу WoS; **(в)** выполненных российскими ученными, согласно ресурсу PubMed; **(г)** выполненных суммарно в мире, согласно ресурсу PubMed.

Целью данной работы было оценить сходство научных тематик, определяющих круг интересов представителей научной школы А.И. Арчакова. Для этого в соответствии с формулой 1 проводили попарное сравнение мешпринтов А.И. Арчакова и его учеников. Результаты сравнения визуализировали в виде семантической карты (рис. 8).

Наибольшее совпадение наблюдалось для мешпринтов А.И. Арчакова и И.И. Карузиной. Решающее значение в этом играют их совместные работы в области исследований цитохрома Р450, микросомального окисления, а также протеомики. Эти же тематики обеспечили сходство мешпринта Г.П. Кузнецовой. Исследования цитохрома Р450 и атомно-силовой микроскопии объединяют профили А.И. Арчакова и Ю.Д. Иванова. Исследования в области протеомики и масс-спектрометрии обеспечили сходство мешпринтам А.В. Лисицы, В.Г. Згоды и А.И. Арчакова.

Узкие места методики

В ходе загрузки публикационных профилей и создания персональных мешпринтов мы столкнулись со следующими проблемами:

1. Разные варианты транслитерирования фамилий затрудняют поиск публикаций.

2. По правилам некоторых журналов отчество не используется для идентификации автора. При использовании автоматического анализа публикационных профилей это приводит к возникновению интерферирующих данных, требующих ручной обработки экспертом. На сегодняшний день многие журналы стали использовать уникальные идентификаторы авторов, однако большая часть накопленных в базе данных PubMed публикаций (более 30000000 статей) не снабжены ими.

3. Для диссертантов, обладающих часто встречающимися фамилиями, составить публикационный профиль в автоматическом режиме не представляется возможным, поэтому для них проводился контроль загружаемых публикаций в ручном режиме. Наибольшее количество публикаций (252) в автоматическом режиме было загружено для В.В. Иванова. При экспертной проверке выяснилось, что эти публикации суммарно принадлежат четырём исследователям.

4. Для семи исследователей, работавших над диссертациями до 1990 года, в базе данных PubMed не было обнаружено ни одной публикации. В некоторых случаях это связано с тем, что журнал не имеет переводной версии и из-за этого не индексируется Pubmed. Кроме того, журналы некоторых издательских домов также не индексируются в PubMed. Для остальных же поиск публикаций с использованием поисковых систем Google, Yandex, а также ресурса E-library не увенчался успехом.

Помимо вышесказанного, на результаты исследования влияет человеческий фактор: решение о присвоении термина MeSH публикации принимается экспертами. За счёт того, что словарь MeSH имеет сложную разветвленную структуру, статьи, относящиеся к одной тематике, могут индексироваться различными (хоть и близкими по значению) терминами MeSH. Несмотря на описанные ограничения, база данных PubMed, термины MeSH и предложенный нами алгоритм их анализа являются удобными инструментами для описания объектов и взаимосвязей между ними.

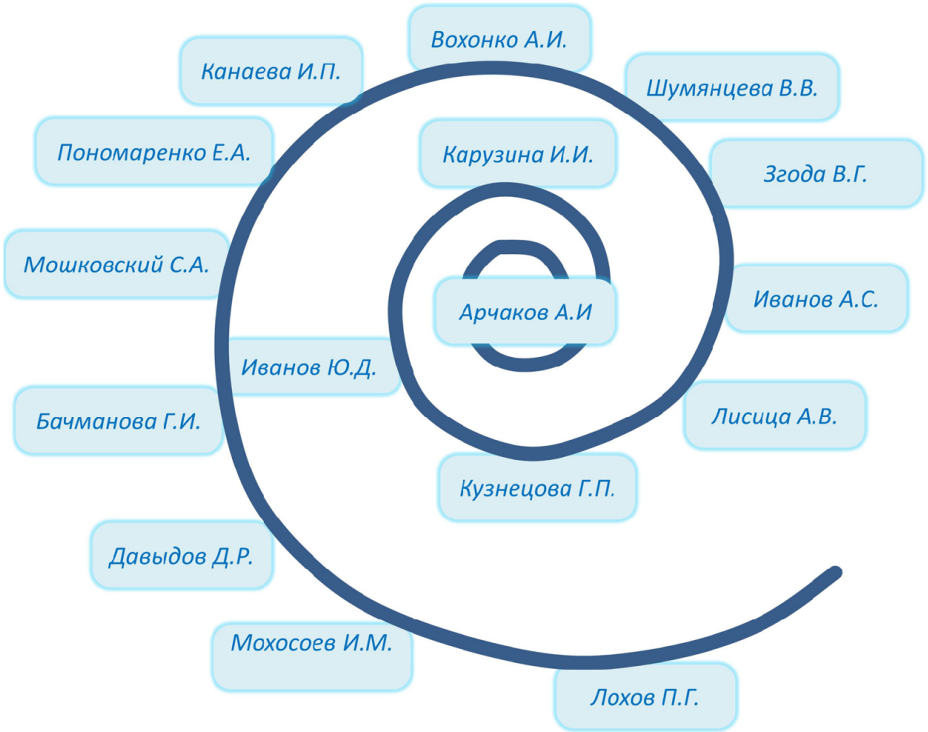


Рисунок 8. Семантическая карта научной школы академика А.И. Арчакова.

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

В данной работе на примере анализа публикационной активности представителей научной школы академика А.И. Арчакова показана возможность использования автоматизированной обработки научных публикаций для выявления трендов развития научных групп и поиска перспективных партнёров для сотрудничества. На примере А.И. Арчакова было продемонстрировано влияние руководителя на круг научных интересов и перспективы эволюции научных групп. Практическое использование полученных данных напрямую связано с возможностью предсказания потенциала той или иной тематики исследований.

ФИНАНСИРОВАНИЕ

Исследование выполнено при поддержке гранта РФФИ №19-29-01138\19 “Оценка состояния здоровья человека путём сопоставления персонального молекулярного профиля с текущим уровнем знаний, накопленных в форме научных публикаций в библиотеке PubMed/MEDLINE” (научные проекты междисциплинарных фундаментальных исследований по теме “Информационные технологии для анализа больших массивов данных в задачах превентивной и персонализированной медицины” (26-901)).

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая статья не содержит каких-либо исследований с использованием людей или с использованием животных в качестве объектов.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

Дополнительные материалы доступны в электронной версии статьи на сайте журнала (pbmc.ibmc.msk.ru).

ЛИТЕРАТУРА

1. *Wang M., Carver J.J., Phelan V.V., Sanchez L.M., Garg N., Peng Y., Nguyen D.D., Watrous J., Kapono C.A. et al.* (2016) Nature Biotechnol., **34**(8), 828-837.
2. *Vasylyeva T.I., Friedman S.R., Paraskevis D., Magiorkinis G.* (2016) Infection, Genetics and Evolution, **46**, 248-255.
3. *Robinson C.V., Sali A., Baumeister W.* (2007) Nature, **450**(7172), 973-982.
4. *Ilgisonis E., Lisitsa A., Kudryavtseva V., Ponomarenko E.* (2018) Advances Bioinformatics, **2018**, DOI: 10.1155/2018/4625394.
5. *Ponomarenko E.A., Lisitsa A.V., Il'gisonis E.V., Archakov A.I.* (2010) Molekuliarnaia biologiia, **44**(1), 152-161.
6. *Skusa A., Rüegg A., Köhler J.* (2005) Briefings Bioinformatics, **6**(3), 263-276.
7. *Restrepo M.I., McGrath M.C., Sarkar I.N., Chen E.S.* (2019) Studies Health Technology Informatics, **264**, 1490-1491.
8. *Zhao F., Shi B., Liu R., Zhou W., Shi D., Zhang J.* (2018) BMC Ophthalmology, **18**(1), 1-11.
9. *Gan J., Cai Q., Galer P., Ma D., Chen X., Huang J., Bao S., Luo R., Zhang Q.* (2019) Medicine (United States), **98**(32), DOI: 10.1097/MD.0000000000016782.
10. *Lu Y., Figler B., Huang H., Tu Y.C., Wang J., Cheng F.* (2017) PLoS ONE, **12**(4), 1-13. DOI: 10.1371/journal.pone.0173548.
11. *Irwin A.N., Rackham D.* (2017) Research Social Administrative Pharmacy, **13**(2), 389-393.
12. *Joppich M., Weber C., Zimmer R.* (2019) Thrombosis Haemostasis, **119**(8), 1247-1264.
13. *Lu Z.* (2011) Database, **2011**, 1-13. DOI: 10.1093/database/baq036.
14. *Doms A., Schroeder M.* (2005) Nucl. Acids Res., **33**, 783-786.
15. *Theodosiou T., Vizirianakis I.S., Angelis L., Tsaftaris A., Darzentas N.* (2011) J. Biomed. Informatics, **44**(6), 919-926.
16. *Ивашкин В.Т., Бакулин И.Г., Богомолов П.О., Мациевич М.В., Гейвандова Н.И., Корой П.В., Недогода С.В., Саблин О.А., Ленская Л.Г., Белобородова Е.В., Багрецова А.А., Абдулхаков Р.А., Осипенко М.Ф., Осипова И.В.* (2017) Гепатология, **27**(2), 34-43. [*Ivashkin V.T., Bakulin I.G., Bogomolov P.O., Matsiyevich M.V., Geyvandova N.I., Koroy P.V., Nedogoda S.V., Sablin O.A., Lenskaya L.G., Beloborodova Y.V., Bagretsova A.A., Abdulkhakov R.A., Osipenko M.F., Osipova I.V., Pocheptsov D.A., Chumachek Y.V., Khromtsova O.M., Kuzmicheva Y.V.* (2017) Russ. J. Gastroenterology, Hepatology, Coloproctology, **27**(2), 34-43.]
17. *Legrain P., Aebersold R., Archakov A., Bairoch A., Bala K., Beretta L., Bergeron J., Borchers C.H., Corthals G.L., Costello C.E., Deutsch E.W., Domon B., Hancock W., He F., Hochstrasser D., Marko-Varga G., Salekdeh G.H., Sechi S., Snyder M., Srivastava S., Uhlen M., Wu C.H., Yamamoto T., Paik Y.-K., Omenn G.S.* (2011) Molecular Cellular Proteomics: MCP, **10**(7), M111.009993. DOI: 10.1074/mcp.M111.009993.
18. *Archakov A., Aseev A., Bykov V., Grigoriev A., Govorun V., Ivanov V, Khlunov A., Lisitsa A., Mazurenko S., Makarov A.A., Ponomarenko E., Sagdeev R., Skryabin K.* (2011) Proteomics, **11**(10), 1853-1856.
19. *Vandenbroucke J.P., Pearce N.* (2018) Clinical Epidemiology, **10**, 253-264.

Поступила в редакцию: 03. 02. 2020.
После доработки: 15. 02. 2020.
Принята к печати: 17. 02. 2020.

MEDICAL SUBJECT HEADINGS FOR THE SCIENTIFIC GROUPS EVOLUTION ANALYSIS ON THE EXAMPLE OF ACADEMICIAN A.I. ARCHAKOV'S SCIENTIFIC SCHOOL

E.V. Ilgisonis*, O.I. Kiseleva, A.V. Lisitsa, E.V. Poverennaya, M.N. Toporkova, E.A. Ponomarenko

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; *e-mail: ilgisonis.ev@gmail.com

This paper proposes a method of comparative analysis of scientific trajectories based on bibliographic profiles. The bibliographic profile ("meshprint") is a list of MeSH terms (key terms used to index articles in the PubMed), indicating the relative frequency of occurrence of each term in the scientist's articles. Comparison of personalized bibliographic profiles can be represented in the form of a semantic network, where the nodes are the names of scientists, and the relationships are proportional to the calculated measures of similarity of bibliographic profiles. The proposed method was used to analyze the semantic network of scientists united by the academic school of the academician A.I. Archakov. The results of the work allowed us to show the relationship between the scientific trajectories of one scientific school and to correlate the results with world trends.

Key words: semantic networks; text analysis; medical subject headings; Text-mining, MeSH, sociology; research trajectory

Funding. The study was supported by the RFBR grant No. 19-29-01138\19 "Human health risk assessment based on the comparison between personal molecular profile and current level of knowledge accumulated in scientific publications in the PubMed / MEDLINE library".

Received: 03.02.2020, revised: 15.02.2020, accepted: 17.02.2020.