нерешённые задачи для молодых учёных

x+y=?

Павлов Дмитрий Алексеевич

НАВИГАЦИЯ В МИРЕ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

Аннотация

В статье в несколько упрощённой форме изложен ряд проблем химической информатики, решение которых рано или поздно позволит создать систему поиска химических соединений и научных статей, которая бы отвечала нуждам современной фармакологии. По каждой проблеме перечислены основные достижения отрасли на настоящий момент и указаны недостатки имеющихся решений.

Ключевые слова: фармацевтика, фармакология, органическая химия, химическая информатика.

Сколько органических соединений вы знаете? А сколько вы знаете лекарств? Каждое лекарство, не считая тех, что производятся из растений, представляет собой комплект из действующего вещества и оболочки/растворителя, в которой/ом оно проходит свой путь до усваивания организмом пациента. Действующее вещество в лекарственном препарате – это одно конкретное органическое соединение[1]. Количество известных органических соединений, которые можно добыть или синтезировать, превышает 30 миллионов, а количество лекарств на рынке – всего несколько тысяч. Создание любого нового лекарства занимает от 10 до 15 лет и является очень дорогостоящим. Расходы на программное обеспечение составляют в этой индустрии (как и почти в любой другой) весьма скромную долю.

Для программистов это не беда, а большая удача: средства на разработку программ выделяются фармакологическими компаниями так щедро, как только возможно, ибо если случится так, что какая-нибудь программа, созданная например за два года, сократит 10–15-летний цикл создания лекарства *хотя бы* на две недели, то траты окажутся оправданы.

Роль компьютеров в этом процессе за последние два десятилетия стремительно возросла, и вот почему. Производство лекарственного средства – комплексная задача, в которой есть место пробам и ошибкам.

Представим, что усилиями биологов в организме выявлен «нездоровый» белок, вызывающий болезнь или болезненные ощущения. Дело за малым – найти вещество, которое разрушит или заблокирует белок, не причинив в процессе вреда организму. Затруднение состоит в том, что на эту роль может годиться одна молекула из 30 миллионов или ещё не открытая молекула. Современные технологии массового синтеза (так называемая комбинаторная, или *сочетательная химия* [1]) и массовых биохимических опытов *high-throughput screening, HTS* [2] позволяют за короткие сроки

[1] В редких лекарствах их два, например в бисептоле.

получать сотни тысяч новых молекул и гигабайты экспериментальных данных.

Опишем карьеру лекарственного вещества от конца к началу. До того как попасть на прилавки аптек, лекарство должно пройти клинические испытания (clinical trials) на пациентах, под присмотром врачей. Это представляет определённый риск, поэтому до пациентов доходят лишь немногие вещества, прошедшие доклинические тесты (preclinical testing) на животных. Их тоже берегут, поэтому доклиническим испытаниям предшествуют массовые опыты на отдельных живых клетках (in vitro, лат. «в стекле», то есть в пробирке). Но и в пробирки не бросаются все молекулы подряд. Люди должны выбирать нужные (перспективные для лечения данной болезни) вещества и отбрасывать заведомо не подходящие, а без компьютера им этого не сделать[1]. На самом деле, и с ним не очень удобно. Эффективная система навигации по химическим соединениям пока ещё не создана, и о перспективах создания таковой сейчас и пойдёт речь[2].

ПОИСК ХИМИЧЕСКИХ СОЕДИНЕНИЙ В БАЗАХ ДАННЫХ

Состояние систем поиска химических соединений в наши дни примерно соответствует состоянию поисковых систем во всемирной паутине в 90-е годы[3]. Да, именно так. Примитивные алгоритмы поиска (и поиск ведётся далеко не по всем имеющимся источникам), весьма вялая поддержка языковой грамматики, никакого ранжирования результатов.

Допустим, стало известно какое-то вещество, эффективно подавляющее проблемный белок[4]. Пилюлю с веществом скормили крысе, та пошла зелёными пятнами и спустя час сдохла. Есть основания полагать, что данное вещество токсично, и людям его давать нельзя. Но можно попробовать найти вещества, близкие ему по структуре. Если повезёт, они окажутся менее токсичными при той же эффективности.

Например, амфетамин является подструктурой мезокарба, и оба препарата подавляют реакцию обратного захвата дофамина (dopamin reuptake) в мозге, что ведёт к повышению активности. Но мезокарб, в отличие от амфетамина, не вызывает тахикардии и повышения артериального давления (см. рис 1).

Вообще говоря, нередко случается так, что добавление или удаление небольшого фрагмента идёт молекуле (точнее, пациентам) на пользу. Чем добавлять и удалять всевозможные фрагменты вручную, проще запустить поиск по базе данных и найти все молекулы, содержащие данную как подструктуру, или все молекулы, содержащиеся в данной. Соответствующие виды поиска называются «подструктурные поиски» (substructure search) (см. рис. 2) и «надструктурные поиски» (superstructure search).

Более общий критерий структурного сходства молекул основан на количестве различных фрагментов, которые присутствуют одновременно в обеих молекулах. Поиск молекул по такому критерию называ-

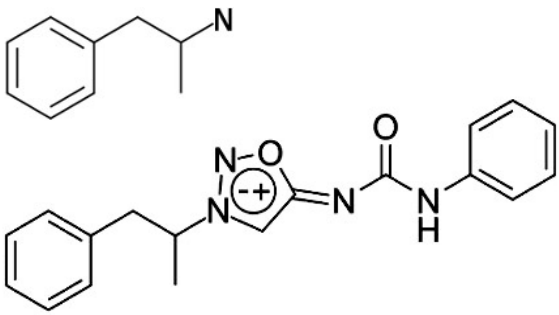

Рис. 1. Амфетамин и мезокарб

[1] Следует заметить, что без компьютеров не появилось бы такое количество данных, которое не под силу обработать вручную; возможно, компьютеры в этой истории продвигают сами себя.

[2] На прочих стадиях роль вычислительных машин не менее важна, однако возникающие там задачи лежат за рамками данной статьи.

[3] (До появления Google). Тогдашних «королей» информационного поиска (Altavista, Lycos, Rambler) мало кто помнит, в том числе потому, что они были практически бесполезны.

[4] Взаимодействие молекулы с белком тоже можно моделировать на компьютере. Этому посвящена отдельная область вычислительной химии под английским названием «docking».

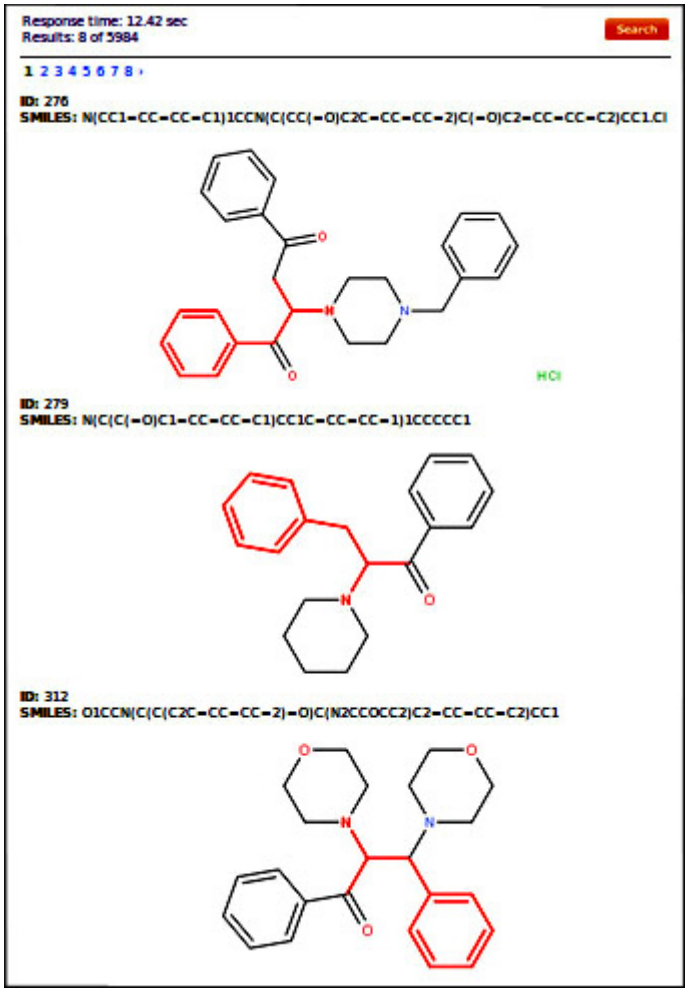

Рис. 2. Подструктурный поиск в сервисе Bingo с амфетамином в качестве запроса

ется *поиском по сходству* [4] (similarity search) (см. рис. 3).

Найденные в базе данных молекулы не очень интересны сами по себе, они интересны в контексте. В каких медицинских и химических статьях упомянуто данное соединение? Есть ли на него патент? Представлено ли оно в коммерческих каталогах? Известны ли его свойства, такие как растворимость, кислотность, токсичность, внутренняя абсорбция и другие? Известна ли химическая реакция его синтеза? Доступны ли исходные компоненты этой химической реакции? На эти вопросы и на многие другие должна давать ответ система поиска.

Перечислим наиболее популярные поисковики химических соединений:

- *PubChem* [5] – база данных из 27 миллионов соединений с богатыми возможностями для поиска: по номеру, по названию, по структурной формуле, по подструктуре и по сходству. Химические свойства также можно задавать в качестве дополнительных критериев поиска (например, ограничиться только молекулами, молекулярная масса которых не превышает 120). Базу постоянно пополняют более 80 организаций.
- *ChemSpider* [6] содержит 25 миллионов соединений и имеет важное отличие от PubChem: добавлять молекулы и обновлять информацию о них здесь могут не только избранные организации, но и простые пользователи. Вместе с последними список источников ChemSpider составляет почти 300 пунктов. Поиск соединений в ChemSpider не имеет такого количества опций, как в PubChem, в частности, отсутствует поиск по сходству.
- *eMolecules* [7] – компиляция из 7 миллионов соединений, собранных в 150 коммерческих каталогах. Возможности поиска минимальны, никакой информации о соединениях, кроме ссылок на каталоги, сайт не показывает. Это, скорее, платформа для продавцов химических веществ, нежели поисковая система для исследователей.

ПОИСК ХИМИЧЕСКИХ СОЕДИНЕНИЙ ПО НАУЧНЫМ ПУБЛИКАЦИЯМ

Пионерами поиска в научных работах по химии были создатели химической реферативной службы (*Chemical Abstracts Service*, *CAS* [8]), существующей с 1907 года. В этой службе ведётся учёт всех известных химических соединений. Тысячи людей в течение десятков лет вручную составляют библиографические справки и заполняют

базу данных *SciFinder* [9] (см. рис. 4) – отдельного продукта CAS для поиска публикаций. Аналогичная база данных, поддерживаемая издательством Elsevier, называется *«Crossfire Beilstein»* [10]. Сервисы PubChem и ChemSpider также выдают пользователю вместе с каждой найденной молекулой список публикаций, к которым данная молекула может иметь отношение, но возможности для поиска собственно публикаций в этих сервисах не очень развиты.

Как же, наконец, отправить на отдых тысячи «индексаторов», от рассвета до заката читающих статьи и заполняющих библиографические базы? Эта задача несколько сложнее, чем найти слово «парацетамол» по текстам статей. Во-первых, само вещество может иметь несколько названий (пример альтернативного названия парацетамола – «N-(4-гидроксифенил)ацетамид»). Во-вторых, лекарство, содержащие это вещество, может упоминаться под разными торговыми марками (в данном случае «Панадол», «Эффералган» и десяток других). В третьих, вещество может быть не написано, а *нарисовано* в статье в растровом виде (в старых отсканированных статьях) или в векторном (начиная с 90-х годов). Программы по автоматическому распознаванию рисунков с молекулами сегодня находятся в плачевном состоянии. Вот наиболее известные из них:

- *CLiDE* [11] канадской фирмы SimBioSys,
- *OSRA* [12] – проект с открытыми исходниками нашего соотечественника, работающего в США,
- *ChemoCR* [13] – проект Марка Циммермана из института Фраунгофера в Германии

CLiDE – наиболее развитая программа из перечисленных, но она нередко ошибается, требует вмешательства человека и «не знает» многих особенностей молекул. OSRA активно развивается, но обладает на данный момент не лучшим качеством распознавания. ChemoCR, похоже, находится в перманентной закрытой разработке: эту программу никто никогда не видел, тем не менее, доступно немалое количество публикаций по алгоритмам, используемым в ней. Указанные программы ещё менее пригодны к распознаванию более сложных химических объектов, как-то: химических реакций, таблиц с заместителями. Комбинированный семантический анализ текста и рисунков (например, «молекула на рис. 10а имеет показатель LD50, равный 5.6 г/кг для взрослых крыс») вообще нигде не реализован.

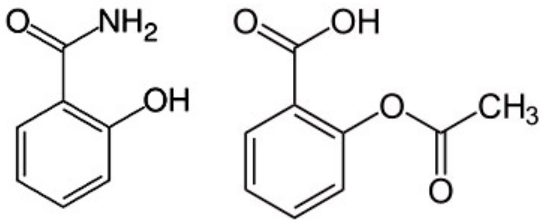

Рис. 3. Салициламид и ацетилсалициловая кислота – схожие по химическим свойствам соединения, использующиеся в медицине (последнее более известно под названием «аспирин»)

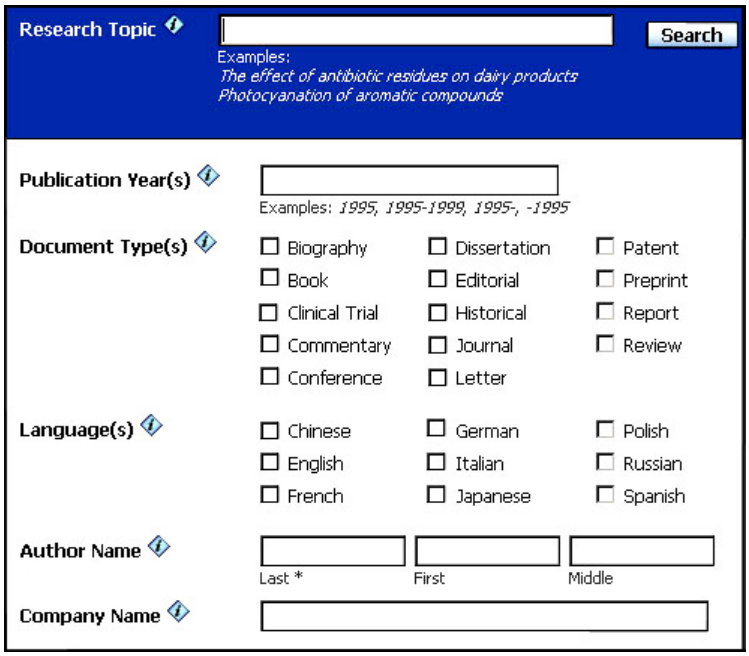

Рис. 4. SciFinder

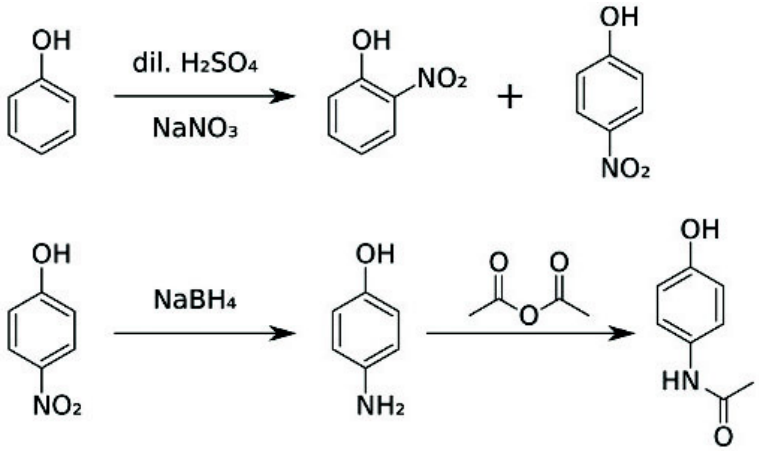

Рис. 5. Многоступенчатая реакция синтеза парацетамола

ПЛАНИРОВАНИЕ СИНТЕЗА

Схема синтеза химического соединения представляет собой цепочку химических реакций (см. рис. 5).

Стало быть, если соединение нельзя заказать через каталоги, можно попытаться осуществить синтез самостоятельно, имея схему реакции, где в правой части стоит искомое вещество. Исходные компоненты реакции (прекурсоры, или предшественники) придётся всё же раздобыть, или вновь синтезировать.

Базы данных с органическими реакциями имеют размер, на три порядка меньший, чем с молекулами, однако многие записи в них задают на самом деле не одну реакцию, а группу реакций, объединённых некоей неподвижной частью, на месте которой может быть всё, что угодно (см. рис. 6, 7).

Благодаря этому обстоятельству, значительно увеличивается разнообразие синтезируемых веществ и в то же время усложняется поиск. Вкупе с «многоступенчатостью», планирование синтеза становится сложной задачей. В некоторой степени эта задача решена в сервисе *Reaxys* [14] (см. рис. 8), который доступен только по подписке.

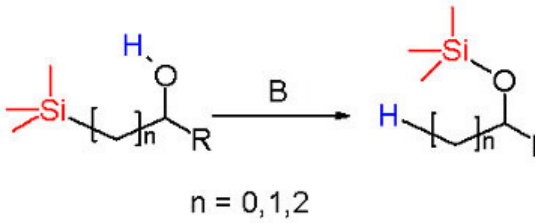

Рис. 6. Перегруппировка Брука

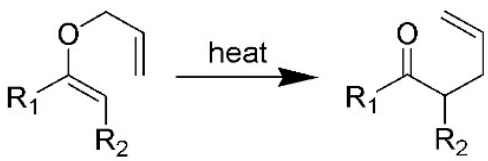

Рис. 7. Перегруппировка Кляйзена

ЗАКЛЮЧЕНИЕ

Положение дел с поиском органических соединений не отвечает современным нуждам. Да, есть отдельные полезные сайты, успешно решающие отдельные части задачи, но не существует пока ни химического аналога Google в сфере поиска, ни аналога Википедии, который позволил бы учёным со всего мира объединить свои усилия по описанию свойств миллионов органических веществ.

В настоящий момент наиболее перспективная система, которая может в будущем стать «Гуглом и Википедией» химиков – это ChemSpider. Её создатели явно принимают в расчёт ключевые элементы успеха глобальных сервисов: кросс-платформенность (работает через браузер и даже с мобильных устройств), доступность для каждого, богатые возможности, «дружба» с многими другими сервисами (включая PubChem), право пользователей публиковать свой контент. ChemSpider имеет недостатки, как общие для индустрии, так и «свои собственные», но движется в правильном направлении (см рис. 9).

Попытки создания химических поисковиков также предпринимались и предпри-

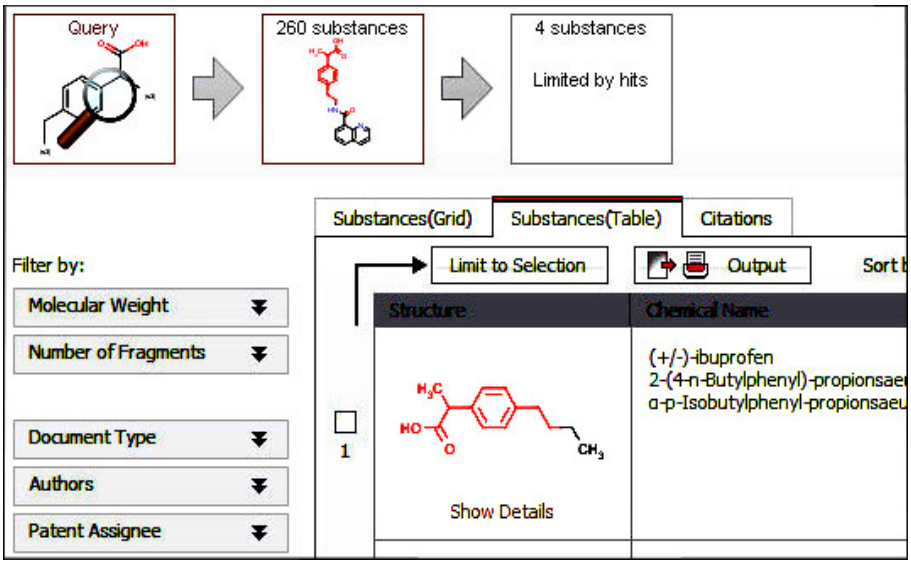

Рис. 8. Reaxys

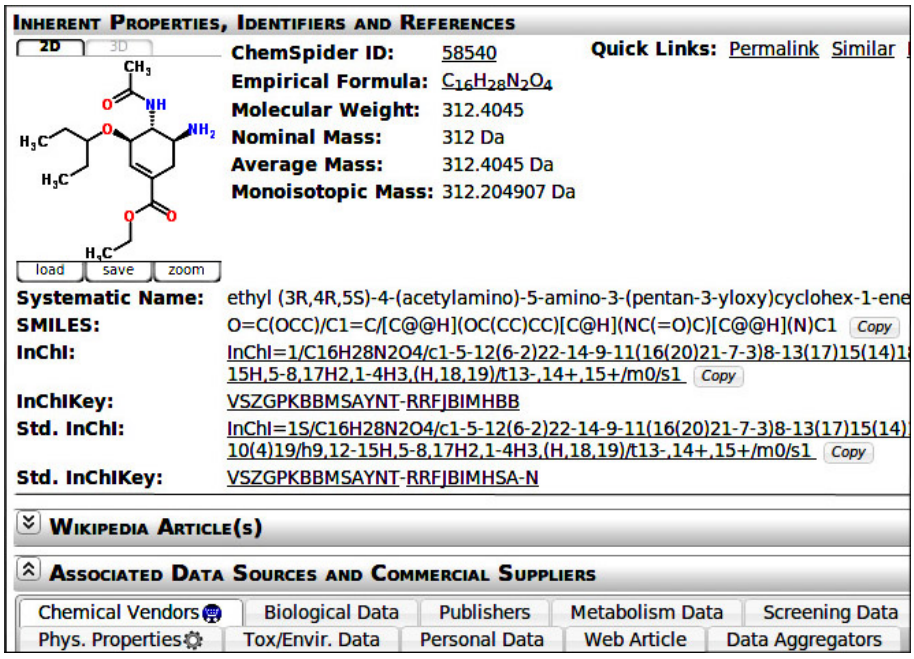

Рис. 9. ChemSpider

нимаются в стенах фармацевтических компаний для внутреннего пользования. Сказать о них нечего, кроме того, что любая ценная информация рано или поздно выходит на свет и там, находясь в общем доступе, постепенно очищается и повышается в качестве, а информация «только для своих» обречена стать бесполезной. Когда вы в последний раз находили что-нибудь дельное в локальной сети своей организации?

Автор признателен Н. Велецкому и Д. Лушникову за ценные замечания по тексту статьи.

Литература

1. http://wsyachina.narod.ru/chemistry/compatible_chemistry.html
2. http://en.wikipedia.org/wiki/High-throughput_screening
3. http://bingo-demo.scitouch.net/
4. http://ru.wikipedia.org/wiki/Молекулярное_подобие
5. http://pubchem.ncbi.nlm.nih.gov/
6. http://chemspider.com
7. http://emolecules.com
8. http://cas.org/
9. http://www.cas.org/products/scifindr/index.html
10. http://www.info.crossfirebeilstein.com/
11. http://www.simbiosys.ca/clide/
12. http://cactus.nci.nih.gov/osra/
13. http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/products/chemocr.html
14. http://reaxys.com/

Abstract

The article briefly describes a number of open problems in cheminformatics, solving which will allow sooner or later to work out a proper chemical compounds and articles search engine, suitable for the needs of present-day pharmacology. On each problem, the most well-known results are listed, and the weaknesses of the available services are mentioned.

Key words: pharmaceutics, pharmacology, organic chemistry, cheminformatics.

Павлов Дмитрий Алексеевич,
выпускник кафедры прикладной математики ФМФ СПбГПУ,

dmitry.pavlov@gmail.com