

## СКРИНИНГ ХИМИЧЕСКИХ СОЕДИНЕНИЙ, ПРИГОДНЫХ ДЛЯ ФУНКЦИОНАЛИЗАЦИИ ПОЛИОЛЕФИНОВ

**В.В. ЛИТВЯК, И.С. УСАЧЁВ**

*Всероссийский научно-исследовательский институт крахмалопродуктов – филиал ФНЦ  
пищевых систем им. В.М. Горбатова РАН,  
140051, Российская Федерация, Московская обл., пос. Красково, ул. Некрасова, 11,  
электронная почта: vniik@arrisp.ru*

Установлено, что для получения новых типов биodeградируемых полимерных материалов на основе полиолефинов и крахмала необходимо дополнительно ввести функционализированный полимер, позволяющий совместить гидрофобный полиолефин с гидрофильным крахмалом. Важными критериями органических веществ, прививаемых на полиолефин являются: доступность по цене, нетоксичность, нелетучесть, способность к полимеризации (наличие двойных связей), размер молекулы. В результате проделанной работы разработан алгоритм проведения скрининга органических веществ, пригодных для функционализации полиолефинов. Сформировано подмножество химических соединений из базы химической информации PubChem, удовлетворяющее критериям, отобранным для поиска соединения-кандидата для использования в качестве биоразлагающих добавок к полимерам. Отобранные химические соединения содержат в своем составе полярные группы (сульфонатную, карбоксильную) с большим значением дипольного момента и способностью участвовать в образовании водородных связей. Использованный в скрининге диапазон значений коэффициента разделения между гидрофобным растворителем и водой определил наличие у соединений гидрофобных группировок и, одновременно, потенциальную возможность растворения в воде. Выполненная работа представляет собой базис для проведения экспериментальных исследований по тестированию стимуляции биоразложения полимерных материалов микроорганизмами.

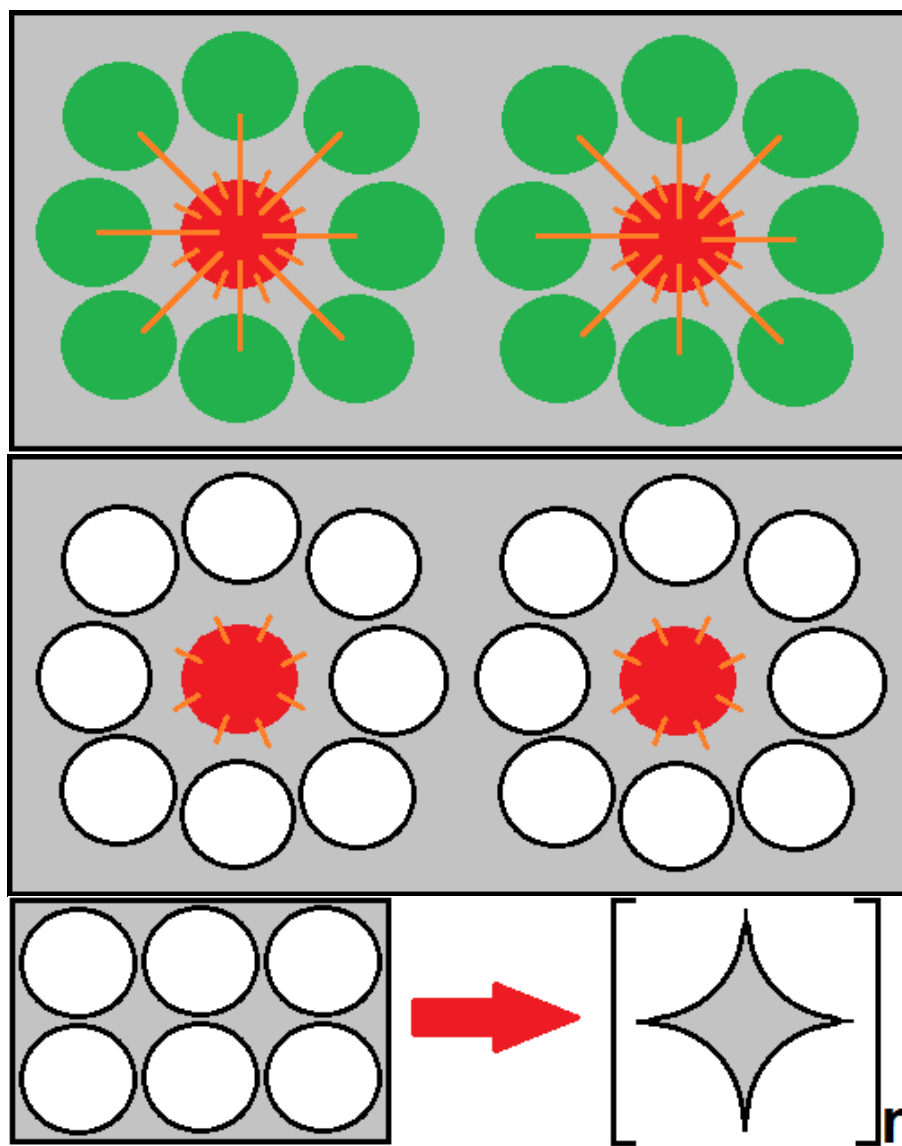
**Ключевые слова:** функционализированный полимер, биоразложение, биodeградируемый полимерный материал, базы химической информации, алгоритм.

В мире достаточно остро стоит проблема утилизации полиолефинов (полиэтилена, полипропилена и т.д.). Существует несколько путей создания биоразлагаемых полимерных материалов: создание новых типов полимеров (как синтетических, так и природных) или придание биodeградируемости уже освоенным крупнотоннажным полимерам путем их модификации или введения различных наполнителей [1–3].

Для получения новых типов биodeградируемых полимерных материалов на основе полиолефинов и крахмала необходимо введение  
<http://ntk.kubstu.ru/file/3072>

функционализованного полимера – сополимера, позволяющего совместить гидрофобный полиолефин с гидрофильным крахмалом. Особенности биоразложения композитного полимерного материала можно условно разделить на четыре этапа (рис.1).

Такая методика позволит получить полимерную пленку с регулируемым сроком биоразложения и хорошими физико-механическими свойствами, её можно использовать для создания инновационной упаковки (в т.ч. и для пищевых продуктов).



■ – основной трудноразрушаемый синтетический полимер, ● – добавка-1 (биоразрушаемый природный полимер), ● – добавка-2 (сополимер), | – связи  
 Рис. 1. – Схема биоразложения композитного полимерного материала

Функционизированным полимером может послужить полиолефин (полиэтилен и/или полипропилен) химически привитой:

-органическим соединением, содержащим одновременно функциональные группы, в которых разница электроотрицательности атомов составляет 0,4, и функциональные группы, состоящие из атомов, разница в электроотрицательности которых превышает 0,4 (табл. 1 и табл. 2), или

Таблица 1. Шкала электроотрицательности атомов

Атом	H	Li	Be	B	C	N	O	F	Br
Электроотрицательность	2,1	1,0	1,5	2,0	2,5	3,0	3,5	4,0	2,8
Атом	Na	Mg	Al	Si	P	S	Cl	J	–
Электроотрицательность	0,9	1,2	1,5	1,8	2,1	2,5	3,0	2,4	–

Таблица 2. Влияние типа связи на электроотрицательность

Тип связи	—C—	C=C	—C=C	C=	C≡	=C—	≡C—	O=	N≡
Электроотрицательность	2,5	2,3	2,7	2,1	2,0	2,8	2,8	4,71	4,208

-органическим соединением, дипольный момент которого  $\leq 1,84$  Д (табл. 3).

Так, разницу электроотрицательностей можно рассчитать по формуле:

$$\Delta X = X_A - X_B = 0,28\sqrt{\Delta E} = 0,28\sqrt{E_{A-B} - \frac{E_{A-A} + E_{B-B}}{2}},$$

где:  $X_A, X_B$  – электроотрицательность атомов А и В;

$\Delta X$  – разница электроотрицательностей атомов А и В;

$E_{A-A}, E_{B-B}, E_{A-B}$  – энергия связей.

Дипольный момент молекулы рассчитывают по формуле:

$$\mu_0 = q \cdot l,$$

где:  $\mu_0$  – постоянный дипольный момент, определяемый разностью

электроотрицательностей атомов А и В;

$q$  – разделенный заряд;

$l$  – расстояние разделения между центрами тяжести положительного и отрицательного заряда.

Таблица 3. – Дипольные моменты функциональных групп органических соединений

Функциональная группа		Дипольный момент, Д	Пример простейших соединений
Формула	Название		
1	2	3	4
Углеводородная цепь			
$-\text{SO}_2\text{O}^-$	Сульфонатная	4,507	Метансульфонат
$-\text{C} \equiv \text{N}$	Нитрил-	3,674	Ацетонитрил
	Карбамидная	3,630	Ацетамид
$-\text{NO}_2^-$	Нитро-	3,469	Нитрометан
	Хлорангидридная	3,229	Ацетилхлорид
	Фторангидридная	3,109	Ацетилфторид
	Карбоксилатная	2,941	Ацетат натрия
	Кето-	2,846	Ацетон
	Формил-	2,675	Ацетальдегид
	Карбоангидридная	2,633	Уксусный ангидрид
$-\text{N}^+ \equiv \text{N}$	Диазо-	2,621	Фенилдиазоний хлорид
$-\text{NO}$	Нитрозо-	2,334	Нитрозометан
$-\text{F}$	Фтор-	2,144	Фторметан
$-\text{Cl}$	Хлор-	2,101	Хлорметан
$=\text{NH}$	Имино-	2,037	Метанимин
$-\text{Br}$	Бром-	1,939	Бромметан
$-\text{OH}$	Гидрокси-	1,791	Метанол
$-\text{S}-$	Сульфидная	1,604	Диметилсульфид

—O—	О-эфир	1,538	Диметиловый эфир
—SH	Сульфгидрильная	1,536	Метилмеркаптан
$\begin{array}{c} \text{O} \\ \parallel \\ \text{—C—} \\ \diagdown \\ \text{O—R} \end{array}$	Сложноэфирная	1,480	Метилацетат
—NH <sub>2</sub>	Амино-	1,292	Метиламин
1	2	3	4
$\text{—}\overset{+}{\text{N}}\equiv\text{N}$	Диазо-	1,079	Диазометан
$\begin{array}{c}   \\ \text{—N—} \end{array}$	Третичная аминогруппа	0,692	Триметиламин
—NH—NH—	Гидразо-	0,413	1,1-диметилгидразин
—N=N—	Азо-	0,000	(E)-1,2-диметилдиазен
—NH—NH—	Гидразо-	0,000	1,2-диметилгидразин
5-членные гетероциклы			
=N—	Аза-	3,929	1H-имидазол
—S—	Тиа-	1,486	Тиазол
—O—	Окса-	1,538	Оксазол

Дипольные моменты молекул характеризуют распределение заряда по молекуле и определяют энергию межмолекулярного взаимодействия.

Дипольным момент также определяется интегралом (в атомной системе единиц):

$$\mu = \int r \left[ \sum Z_a \sigma(r - R_a) - \rho(r) \right] dr ,$$

где:  $\rho(r)$  – электронная плотность молекулы;

$Z_a$  и  $R_a$  – заряд и координата ядра;

$\delta(r - R_a)$  – дельта-функция Дирака.

Дипольный момент характеризует распределение электронной плотности относительно распределения зарядов ядер, то есть полярность молекулы. За направление дипольного момента принимают вектор, проведенный от центра тяжести положительных зарядов к центру тяжести отрицательных зарядов. Молекулы, обладающие центром симметрии, характеризуются нулевым дипольным моментом.

Абсолютные значения дипольных моментов могут быть определены методами микроволновой и ИК-спектроскопии, измерением комплексной диэлектрической проницаемости как функции частоты и температуры. Применение расширенных базисов в квантово-механических расчетах позволяет определить точное направление, знак и близкое к экспериментальному абсолютное значение для большинства молекул. Расчетные методы позволяют оценить величины дипольного момента, не прибегая к получению чистых образцов рассматриваемых соединений и не проводя трудоемких экспериментальных манипуляций. Таким образом, метод расчета значений дипольных моментов молекул служит выгодной альтернативой экспериментальным методикам, благодаря своей масштабируемости и воспроизводимости получаемых результатов, которые хорошо согласуются с эталонными экспериментальными данными.

Кроме расчета дипольного момента в приближении молекулярных орбиталей линейной комбинацией атомных орбиталей, применяется метод, основанный на квантово-топологической теории, в рамках которого учитывается величина межатомного переноса электронной плотности.

Расчет дипольного момента требует повышенной точности при построении распределения электронной плотности молекулы. Минимальный уровень расчетов, результаты которых согласуются с экспериментом, дают метод Хартри-Фока с учетом возможных возбужденных электронных конфигураций (теория возмущений Мёллера-Плессета) или метод Кона-Шема с усовершенствованным функционалом энергии (B3LYP и др.).

Нами был использован метод Хартри-Фока с TZ-набором базисных функций 6-311G(p)/MC (McLean/Chandler). Для учета возбужденных состояний использовалась теория Мёллера-Плессета второго порядка (MP2).

Набор 6-311G(p) предполагает наличие 6 орбиталей гауссового типа (ОГТ) для аппроксимации основных атомных орбиталей, 5 для каждой из

трех функций валентной оболочки тяжелого атома и дополнительно 3 р-функции для описания поляризации атомов водорода. Таким образом, для описания электронной структуры молекулы воды необходимо использовать 6 ОГТ для основной оболочки кислорода, 15 ОГТ для описания валентных оболочек кислорода, по 5 для валентных оболочек и по 3 поляризующие орбитали для атомов водорода. Итого для молекулы воды требуется параметризация  $6+15+2\cdot(5+3)=37$  базисных функций.

Вследствие того, что базисный набор MC, упомянутый выше, не подходит для описания молекул с атомом брома в составе, для расчета их свойств был использован валентно-расщепленный базисный набор MIDI.

Функциональные группы характеризуются электроотрицательностью и дипольным моментом. Замещение неполярными метильными радикалами позволяет произвести оценку параметров, присущих функциональным группам с присоединенным алифатическим заместителем.

Вычисления состояли из двух этапов. На первом проводилась оптимизация геометрии молекулы для достижения энергетического минимума. В ходе оптимизации положения атомов изменяются относительно друг друга, чтобы обеспечить равновесное их расположение, для которого получаемые физико-химические параметры молекулы имеют смысл. На втором производился расчет свойств молекулы в оптимальной конформации. Сравнение экспериментального значения дипольного момента молекулы воды ( $\Delta\chi_{\text{H}_2\text{O}} = 0,4$ ;  $\mu_{\text{H}_2\text{O}} = 1,84$  Д) с полученным с помощью расчета по выбранной схеме (1,822 Д, относительная ошибка 1%) показало, что расчетные значения с достаточной точностью воспроизводят величины, полученные в экспериментах.

Кроме того, важными критериями органических веществ, прививаемых на полиолефин являются: доступность по цене, нетоксичность, нелетучесть, способность к полимеризации (наличие двойных связей), размер молекулы.

Характеристика сополимеров или смешанных полимеров и реакции сополимеризации представлены в таблице 4.

Цель – проведение скрининга, т.е. отбор химических соединений, пригодных для функционализации полиолефинов.

Таблица 4. – Характеристика сополимеров или смешанных полимеров и реакции сополимеризации

№	1. Типы сополимеров (полимерные цепи содержат несколько типов мономерных звеньев)
1	2
1.	<p><i>Статистический или нерегулярный сополимер:</i></p> $[M_1]_n + [M_2]_n \longrightarrow \left[ -M_1-M_1-M_1-M_2-M_1-M_2-M_1-M_1-M_2- \right]_n$ <p>где: M<sub>1</sub> и M<sub>2</sub> – разные момеры мономерные звенья располагаются случайно, т.е. нерегулярно или хаотично</p>
2.	<p><i>Чередующийся или регулярный сополимер:</i></p> $[M_1]_n + [M_2]_n \longrightarrow \left[ -M_2-M_1-M_2-M_1-M_2-M_1- \right]_n$ <p>мономерные звенья располагаются в определенной последовательности</p>
3.	<p><i>Блок-сополимер:</i></p> $[M_1]_n + [M_2]_n \longrightarrow \left[ -M_1-M_1-M_1-M_1-M_2-M_2-M_2-M_1-M_1- \right]_n$ <p>мономерные звенья образуют длинные непрерывные однородные последовательности (блоки)</p>
4.	<p><i>Привитой сополимер:</i></p> $[M_1]_n + [M_2]_n \longrightarrow \left[ -M_1-\overset{M_2}{\underset{ }{M_1}}-M_1-M_1-\overset{M_2}{\underset{ }{M_1}}-M_1- \right]_n$ <p>разветвленные полимеры главная цепь которых состоит из одних типов мономеров, а боковые ответвления из других</p>
2. Возможные варианты сополимеризации	
1.	Сополимеризацию органического соединения, способного полимеризоваться с гидрофильными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофобные группы
2.	Сополимеризацию органического соединения, способного полимеризоваться с гидрофобными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофильные группы
3.	Сополимеризацию органического соединения, способного полимеризоваться с гидрофобными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофобные группы
4.	Сополимеризацию органического соединения, способного полимеризоваться с



	гидрофильными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофильные группы
5.	Сополимеризацию органического соединения, способного полимеризоваться с гидрофильными и гидрофобными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофильные группы
1	2
6.	Сополимеризацию органического соединения, способного полимеризоваться с гидрофильными и гидрофобными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофобные группы
7.	Сополимеризацию органического соединения, способного полимеризоваться с гидрофильными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофобные и гидрофильные группы
8.	Сополимеризацию органического соединения, способного полимеризоваться с гидрофобными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофобные и гидрофильные группы
9.	Сополимеризацию органического соединения, способного полимеризоваться с гидрофобными и гидрофильными группировками (мономер-1), проводят совместно с органическим соединением, способным полимеризоваться (мономер-2) и содержащим гидрофобные и гидрофильные группы

Методика проведения скрининга. База данных PubChem [<http://pubchem.ncbi.nlm.nih.gov>] является хранилищем экспериментальных данных о биологической активности малых органических молекул. Проект PubChem является составной частью Инициативы медицинских исследований Национальных институтов здоровья (National Institutes of Health, NIH) США. В PubChem сохраняются данные о молекулярной структуре и биологической активности, получаемые от MLSCN и других организаций. PubChem предоставляет инструментарий для поиска, получения информации и ее анализа. В PubChem интегрирована информация из различных публично доступных источников и других баз знаний NIH.

Основной целью PubChem является предоставление свободно доступного он-лайн ресурса, содержащего всестороннюю информацию о биологической активности малых молекул, для молекулярных биологов, а

также специалистов в медицинской и вычислительной химии. Проект запущен в сентябре 2004. PubChem повторяет подход GenBank [4] по наполнению данными, когда исследователи могут напрямую предоставлять экспериментальную информацию для включения в банк. Таким образом, PubChem зависит от исследователей-участников, которые вносят свой вклад в наполнение базы и поддержание ее информации в актуальном состоянии. PubChem характеризуется высокоавтоматизированной обработкой информации, что делает поддержание базы недорогим, в противоположность проектам, которые используют подход по ручной аннотации.

Базы, подобные PubChem, существуют уже более 25 лет в биологии, например, GenBank, SwissProt [5], PDB [6] и др. Однако для химических баз данных подобный подход реализован впервые. PubChem тесно интегрирована с другими сервисами Национального центра по биотехнологической информации (National Center for Biotechnology Information, NCBI) [7] и доступна через поисковую систему NCBI Entrez [8].

PubChem подразделяется на 3 раздела: PubChem Substance, PubChem Compound, и PubChem BioAssay. PubChem Substance содержит описания химических образцов, предоставленные участниками, и ссылки на их биологическую активность. Описание включает в себя идентификаторы из базы PubChem Compound в случае, если химическая структура входящих в состав соединений известна. Ссылки включают в себя цитирование в базе литературных источников PubMed [9], указание белковых кристаллографических структур, имеющих отношение к данному веществу, ссылки на веб-сайты участника, депонировавшего информацию, и ссылки на результаты биологического тестирования из PubChem BioAssay.

PubChem Compound содержит уникальную информацию о химической структуре из PubChem Substance. Химические соединения могут найдены по рассчитанным химическим свойствам и сравнению по структуре с помощью

алгоритмов кластеризации. Где это возможно, даны ссылки на биологическую активность.

В PubChem BioAssay хранятся результаты тестирования биологической активности веществ сложного состава и химических соединений из различных источников. Данный раздел PubChem предоставляет описания каждого биологического испытания, включая условия и полученные результаты. PubChem BioAssay предоставляет ссылки на ассоциированные с исследованием записи в PubChem Substance. Таким образом, формируется полный биологический профиль отдельных химических соединений.

Выделить уникальную химическую информацию из PubChem Substance, для того, чтобы сформировать базу PubChem Compound, – нетривиальная задача. Не существует единых общепринятых норм представления информации о химической структуре, и каждый участник, вносящий свои данные в базу PubChem, представляет ее в том виде, который считает удобным для него. Вследствие этого, одна и та же химическая структура может быть добавлена в базу в различном виде. Создание алгоритма, который бы распознавал и унифицировал виды представления химической структуры соединений, до сих пор представляет значительные трудности.

В PubChem реализован метод приведения химических структур к единому виду представления с помощью алгоритма «стандартизации». Он включает в себя следующие этапы: проверка отнесения каждого атома к известному элементу периодической системы, расстановка водородных атомов в соответствии с нереализованной валентностью в органических соединениях, приведение к общему виду представления функциональных групп, проверка валентностей и формальных зарядов на атомах молекулы, проверка возможности существования резонансных структур, дополнительное определение ароматических структур, нахождение стереохимических центров. Дополнительная обработка применяется в случае

декларирования существования смеси, в которой могут протекать реакции обмена ионов и нейтрализации. Прошедшая стандартизацию химическая структура используется для создания двумерной диаграммы с координатами атомов, расчета основных химических свойств (молекулярная масса, брутто-формула и т.д.) и дескрипторов (каноническое представление SMILES, InChI, наименование согласно номенклатуре IUPAC и т.д.).

Дескрипторы веществ не создаются для химических соединений, для которых процедура стандартизации не была завершена успешно.

Организация связей между данными в трех базах PubChem является довольно простой. Имеется соответствие между идентификаторами веществ (SIDs) и идентификаторами соединений. Каждое «вещество», если была проведена процедура стандартизации, будет содержать CID, дающий ссылку на запись, представляющую всю стандартизованную структуру. В случае, если «вещество» представляет собой смесь соединений, запись будет содержать CID-ы на каждый независимый компонент смеси. По возможности, каждому соединению, вне зависимости от степени ионизации и замены протонов на какие-либо другие ионы, дается соответствие с «родительским» химическим соединением в типичной форме представления. Особенно важно это для кислот и щелочей. Идентификаторы PubChem Assay (AIDs) содержат ссылки на записи о веществах (их SID-ы).

Для расширенного использования базы PubChem необходимо прояснить взаимоотношения между различными идентификаторами, которые в терминологии реляционных баз данных называются ключами.

Существует взаимоотношение «один ко многим» между стандартизованными CID, с одной стороны, и SID, с другой. Причиной этому является то, что более чем один участник формирования базы данных может отправить вещество, содержащее в своем составе стандартизованное химическое соединение. Более того, данные, присланные одним участником,

также могут быть избыточными, так как могут содержать информацию об оптических изомерах либо различных формах при различном pH. Нет прямых взаимоотношений между идентификаторами BioAssay (AID) и Compound (CID). Эти связи можно проследить через записи о веществах (SID).

По состоянию на март 2008 года, PubChem содержал информацию о более чем 1000 типах опытов на биологическую активность, 28 миллионов результатов экспериментов, 40 миллионов веществ и 19 миллионов уникальных химических соединений. Информация поставляется более чем 70 организациями-участниками. Основная часть информации по скринингу биологической активности поступает от центров, финансируемых Национальными институтами здоровья США в рамках сети MLSC. PubChem содержит также результаты таких исследований, как скрининг 60 опухолевых клеточных линий от DTP/NCI, данные о токсичности от проекта DSSTox в EPA, данные о биологической активности, извлеченные из специальной литературы в рамках проекта BindingDB

Доступ к интерфейсу PubChem обеспечивает NCBI Entrez. Он является достаточно простым в использовании и одновременно позволяет формировать сложные поисковые запросы. Имеется возможность поиска точного совпадения индекса путем добавления имени индекса в квадратных скобках после поискового запроса. К примеру, “aspirin[CompleteSynonym]” возвращает одну запись, соответствующую только аспирину, так как существует лишь одна запись, которая имеет полное соответствие поисковому запросу.

Некоторые индексы, вводимые в Entrez, настроены таким образом, что требуют полного совпадения всего запроса, в то время как другие индексы допускают частичное совпадение. Для числовых индексов имеется возможность задавать условия поиска в виде диапазона: минимальное и максимальное значения разделяются в запросе двоеточием, после чего в

квадратных скобках указывается наименование индекса. К примеру, чтобы найти все химические структуры в PubChem Compound с количеством доноров водородной связи от 0 до 5, поисковый запрос должен иметь вид “0:5[HydrogenBondDonorCount]”. При указании диапазона можно использовать дробные числа (в качестве разделителя целой и дробной частей используется точка). Так, для нахождения соединений по диапазону молекулярной массы, можно использовать запрос “214.31456:215 [MolecularWeight]”.

В одном запросе может быть использован поиск по нескольким индексам одновременно, объединенным булевыми операторами И, ИЛИ, НЕ. К примеру, запрос вида “Br[Element] AND 1[CovalentUnitCount]” позволит найти все химические структуры с бромом, которые не являются частью смеси различных молекул или ионов.

Булевы операторы в Entrez обязательно пишутся заглавными буквами (“AND”, “OR”, “NOT”). По умолчанию, Entrez убирает из запроса лишние пробелы, пунктуацию и нерегламентированные символы. Чтобы быть уверенным, что поисковый алгоритм воспримет запрос в виде единого слова либо фразы, его необходимо взять в кавычки. К примеру, чтобы найти в базе данных PubChem Compound аспирин, используя строку InChI, необходимо использовать

```
"InChI=1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)/f/h11H"[InChI]"
```

в качестве запроса.

Использование разнообразных индексов позволяет повысить эффективность работы с базой данных. Полный список индексов доступен в закладке “Preview/Index” в Entrez, а также в открывающемся меню внизу слева. Также имеется интерфейс для конструирования сложных поисковых запросов, комбинирующих поиск по нескольким индексам.

Фильтры представляют собой булевы функции, возвращающие значение «да» или «нет» в зависимости от того, соответствует ли заданным условиям записи в базе данных. Например, фильтр “pcsubstance\_pcassay” возвращает значение «да» для каждого вещества, которое имеет ассоциированные с ним записи в PubChem BioAssay. Комбинированный запрос “100:200[MolecularWeight] AND pcsubstance\_pcassay[Filter]” в PubChem Substance возвращает список веществ (субстанций), содержащих соединения с молекулярной массой от 100 до 200 Да, имеющие ассоциированные с ними записи в PubChem BioAssay. Фильтр “has\_pharm” в PubChem Compound позволяет определить, имеют ли данные химические соединения фармакологическую активность.

Структурный поиск, реализованный в PubChem, позволяет запрашивать подмножества PubChem Compound с помощью различных типов структурных запросов и опциональных фильтров. Поддерживаемые форматы для структурного поиска включают в себя SMILES, SMARTS, InChI, CID (идентификатор PubChem Compound), молекулярную формулу и SDF. Также имеется он-лайн редактор на основе JavaScript, в котором химическая структура может быть нарисована вручную. Поиск похожих соединений основан на метрике, которая учитывает набор связей в молекуле, наличие информации об изотопах и оптически активных центрах. Поиск по подобию определяет процент совпадения, используя уравнение Танимото, а также набор словарных ключей. Имеется возможность осуществлять поиск по подмножеству, которое было создано предыдущим поисковым запросом или загруженным файлом, содержащим список идентификаторов химических соединений. Поисковые запросы обрабатываются серверами NCBI. Как правило, выделяется достаточно времени для завершения любого запроса, однако могут применяться ограничения на время выполнения и количество выдаваемых результатов.

Учитывая тот факт, что в PubChem применяется стандартизация химических структур, перед использованием базы данных важно представлять, какие изменения вносит эта процедура. Для этого создан веб-инструмент, позволяющий использовать алгоритм стандартизации для произвольного пользовательского файла. На ввод принимаются форматы представления химических структур SMILES, InChI или SDF.

После работы с поисковым сервисом PubChem зачастую бывает необходимо экспортировать результаты для последующего локального анализа. Вся информация, хранимая в базе данных, доступна через протокол FTP (URL <ftp://ftp.ncbi.nlm.nih.gov/pubchem/>). Средство для скачивания данных доступно по адресу [http://pubchem.ncbi.nlm.nih.gov/pc\\_fetch/](http://pubchem.ncbi.nlm.nih.gov/pc_fetch/).

Важно отметить, что база PubChem Substance содержит оригинальную информацию, переданную участником наполнения базы, тогда как PubChem Compound предоставляет данные, прошедшие процедуру стандартизации.

Для экспорта доступно множество форматов данных. Они включают в себя SDF, форматы изображений, SMILES, InChI, XML, текстовый и бинарный ASN.1. Собственным форматом хранения данных PubChem является ASN.1. Файлы в формате XML по содержанию полностью соответствуют ASN.1. SDF представляет собой промышленный стандарт обмена химической информацией, поэтому он поддерживается большинством химических программных пакетов. К сожалению, SDF не способен хранить всю информацию из ASN.1, следовательно, часть информации не будет включаться в файлы SDF.

Спецификации формата PubChem ASN.1, схема XML и описание структуры SDF доступны на FTP-ресурсе NCBI. ASN.1 представляет собой двоичный формат. NCBI использует текстовые дескрипторы в ASN.1, которые могут быть прочитаны как программными средствами, так и пользователем. Однако это не соответствует в строгом смысле формату ASN.1 Это означает,



что сторонние, не NCBI, библиотеки для парсинга будут непригодны для чтения PubChem ASN.1. Экспорт структурных дескрипторов SMILES и InChI производится в текстовом формате, содержащими CID и SID, с полями, разделенными символами табуляции. Формат SMILES позволяет записывать информацию только о ковалентных связях, в то время как PubChem поддерживает также описание ионных, координационных связей и металлокомплексов. Файлы дополнительно могут быть сжаты в форматах gzip (.gz) или bzip2 (.bz2). Скачивание структур ограничено максимальным количеством в 250000 записей на один запрос. Максимальное количество изображений составляет 50000 штук на запрос.

Несмотря на то, что PubChem предоставляет множество веб-сервисов для формирования поисковых запросов и анализа данных, подобные интерфейсы не предназначены для выполнения многократно повторяемых действий, требующих автоматизации. Без средств программирования осуществление интенсивного анализа большого количества информации из базы данных PubChem оказывается невозможным.

Индексы PubChem описаны в языке терминов, значение которых приводится ниже.

AID: идентификатор PubChem BioAssay, положительное целое число.

CID: идентификатор химического соединения PubChem, положительное целое число.

Сложность: рейтинг сложности химических соединений представляет собой приблизительную оценку того, насколько сложна их химическая структура с позиции разнообразия входящих в состав химических элементов и топологии строения, включая явление хиральности. Значение сложности рассчитывается по формуле Бергца/Хендриксона/Игленфелда (Bertz/Hendrickson/Ihlenfeldt). Для ароматических соединений применяется приводящий коэффициент, вследствие чего сложность молекул бензола и

циклогексана является равной. Параметр сложности является, в общем случае дробным числом от 0 до нескольких тысяч. Общей тенденцией является то, что большие по размеру соединения оказываются «сложнее» меньших. Однако, у высокосимметричных соединений, а также тех, что состоят из малого количества типов атомов или элементов, получают меньшую величину сложности при оценке. Показано, что величина сложности слабо коррелирует с доступностью при искусственном синтезе. Наиболее сложное соединение в базе PubChem имеет идентификатор CID 6338588 (C<sub>124</sub>H<sub>185</sub>N<sub>9</sub>O<sub>20</sub>S<sub>36</sub>) с величиной сложности 18425. Средняя сложность по всем химическим соединениям в PubChem составляет около 551.

Комментарии в записях содержат комментарии от участника, выславшего информацию о соединении, а также некоторую дополнительную информацию, не подпадающую ни под одну из категорий формата хранения данных.

Компонент: в составе сложных субстанций, представляющих собой смеси соединений, компонент соответствует одному типу молекул.

Соединение: химическое соединение из состава субстанций, депонированных в PubChem Substance, прошедшее процедуру стандартизации и депонированное в PubChem Compound.

Рассчитанные дескрипторы: информация, описывающая химическое соединение, в различных форматах (SMILES, InChI, наименование согласно номенклатуре IUPAC).

Рассчитанные свойства: информация, полученная в результате расчетов на основании данных о химическом строении соединения (молекулярная масса, брутто-формула, XLogP и т.д.).

Категория участника: дополнительная категория данных, введенная участником наполнения базы данных при депонировании.

Исключенное соединение: Идентификатор соединения (CID) без ссылки

на какую-либо субстанцию (вещество). Исключенные соединения не доступны из интерфейса Entrez.

НВА: количество акцепторов водорода в структуре. Классификация атомов согласно работе [10]. НВД: количество доноров водорода в структуре. Классификация атомов согласно работе [10].

Тяжелые атомы: все атомы за исключением атомов водорода.

InChI: международный химический идентификатор IUPAC. Поиск по строке InChI доступен из Entrez PubChem.

Запись о субстанции старой версии: запись о субстанции считается «старой», если доступна обновленная информация о субстанции от участника.

Молекулярная масса: молекулярная масса является суммой масс атомов, входящих в состав химического соединения, выраженная в дальтонах или г/моль. В случае, если не был явно указан изотопный состав или наличие изотопных меток, молекулярная масса рассчитывается по усредненной массе элемента с учетом природной частоты встречаемости изотопов. Если указана изотопная метка, то для расчета молекулярной массы используется точная масса изотопа даже в том случае, если этот изотоп является короткоживущим, и рассчитанная таким образом масса не имеет физического смысла. В настоящее время в PubChem нет возможности сохранять данные о детальном соотношении изотопов в веществе. Псевдоатомам присваивается нулевая атомная масса.

SID: идентификатор PubChem Substance identifier, положительное целое число.

SMILES: Simplified Molecular Input Line Entry System, упрощенная молекулярная линейная система записи, система линейной нотации молекулярной структуры.

SMARTS: язык, используемый для обозначения подструктур химических соединений, согласно правилам, основанным на SMILES.

Субстанция (вещество): индивидуальная запись, полученная от участника наполнения базы PubChem, представляющая собой информацию о субстанции, использовавшейся в анализе биологической активности (BioAssay).

Synonyms: все названия, в том числе тривиальные, синонимы, часто используемые идентификаторы химического соединения, All names, trivial names, synonyms, frequently used IDs, and other names collected from depositors. In the compound summary page, synonyms are distinct synonyms from all corresponding substances.

TPSA: Topological Polar Surface Area, топологическая площадь полярной поверхности. Представляет собой оценку площади молекулярной поверхности (в квадратных ангстремах), которая является полярной согласно работе [11]. Метод является достаточно простым – учитываются только N и O, трехмерные координаты не используются, расчет ведется с использованием большого числа предварительно полученных параметров для различных типов гибридизации, зарядов и ароматических систем.

Версия: версия субстанции в PubChem, которая увеличивается по мере обновления участником информации о нем.

Xref: внешние ссылки из записей базы данных PubChem.

XLogP: коэффициент разделения, отображающий логарифм соотношения растворимости соединения в различных растворителях. С февраля 2009 года, в PubChem используется версия 3 алгоритма по расчету XlogP [12].

Скрининг химических соединений, пригодных для использования в качестве биоразлагающей добавки к полимерам, основан на применении ряда критериев, которые позволяют отобрать наилучшие соединения-кандидаты из многообразия веществ, информация про которых хранится в химической базе данных.

Для поиска в базе данных PubChem воспользуемся механизмом поиска

по индексам, который реализован в поисковой системе Entrez от NCBI.

Искомые соединения ограничим простыми по композиции субстанциями, для чего установим значение индекса 1[CUC]. Элементный состав соединений укажем выражением (O[EL] OR N[EL] OR C[EL] OR S[EL] OR Cl[EL]). Молекулярная масса интересующих соединений лежит в пределах от 32 (молярная масс метанола) до 400 Да: 0:400[MW]. Комбинируя указанные выше запросы по индексам, получаем результат поиска (ноябрь 2011). Всего найдено 14 920 779 соединений.

Чтобы исключить из результатов соединения, которые проявили себя в тестах в качестве биологически активных в отношении клеток, тканей или органов человека, определим круг исследований, у которых определялась подобная активность. Для этого откроем «Расширенный поиск» по базе PubChem BioAssay. В условиях определим, что нас интересуют тесты, проведенные в отношении метаболизма человека, а в тестах были использованы химические соединения (не интерферирующая РНК). Результат поиска – 67119 тестов. Сохраним результаты поиска в созданном пользовательском аккаунте, а также в текстовом файле. Таким образом, в файле сохранен список исследований активности химических соединений в отношении клеток и тканей человека.

Соединения, показавшие активность в биологических тестах, исключаются применением поиска с указанием индекса 0[AC]. Результат уточняется, таким образом, до 713 676 соединений.

Искомый показатель гидрофобности (XlogP) заключим в диапазон от – 0.69 (метанол) до 1.88 (бензойная кислота), для чего используем предусмотренный разработчиками поисковой системы индекс – 0.69:1.88[XLGP]. Исключим из результатов изотопно-меченные химические соединения (0[IAC]). Электрический заряд химических соединений задается запросом [TFC]. Сложность химических соединений ограничим средним значением по базе данных PubChem – 551. Использование сформированного

поискового запроса, указывающего базовые ограничения на химические структуры (приведен ниже) позволяет получить подмножество из 140832 соединений.

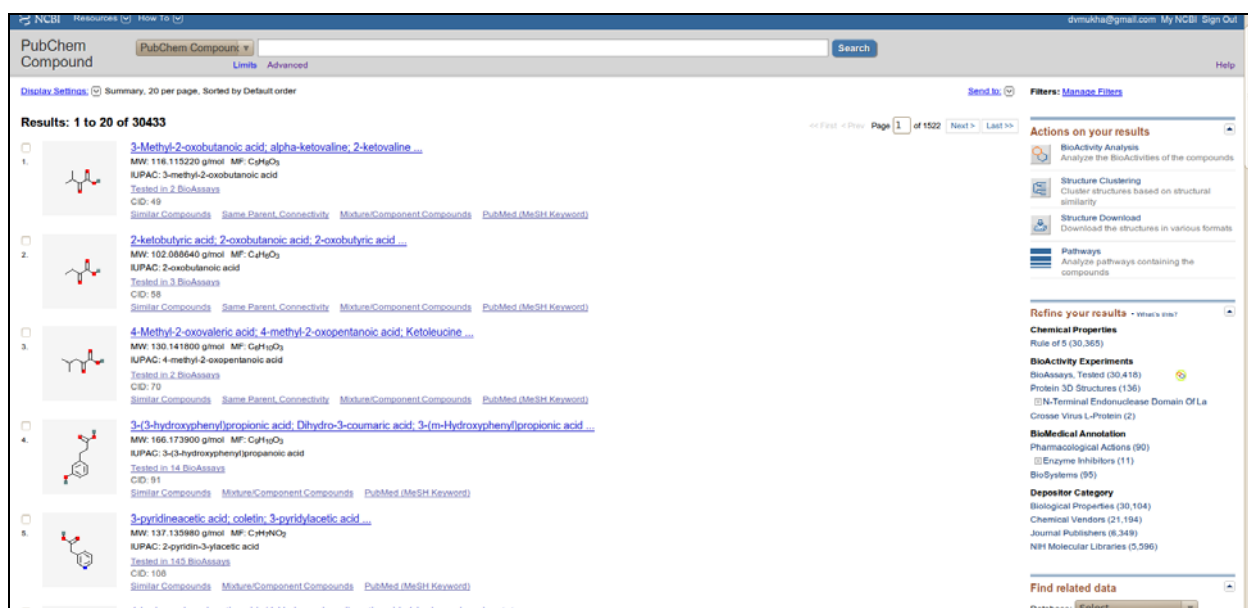
```
1[CUC] AND (O[EL] OR N[EL] OR C[EL] OR S[EL] OR Cl[EL]) AND 32:400[MW] AND 0[AC] AND -0.69:1.88[XLGP] AND 0[IAC] AND 0[TFC] AND 0:551[CPLX]
```

Поисковые запросы по мере выполнения добавляются в «Историю запросов», что позволяет использовать их в качестве подмножеств для дальнейшей работы. Для поиска одной из подструктур интереса этого оказывается вполне достаточно. Но если работа с подмножеством будет заключаться в поиске нескольких альтернативных подструктур, то средства, представленные на сайте PubChem окажутся непригодными.

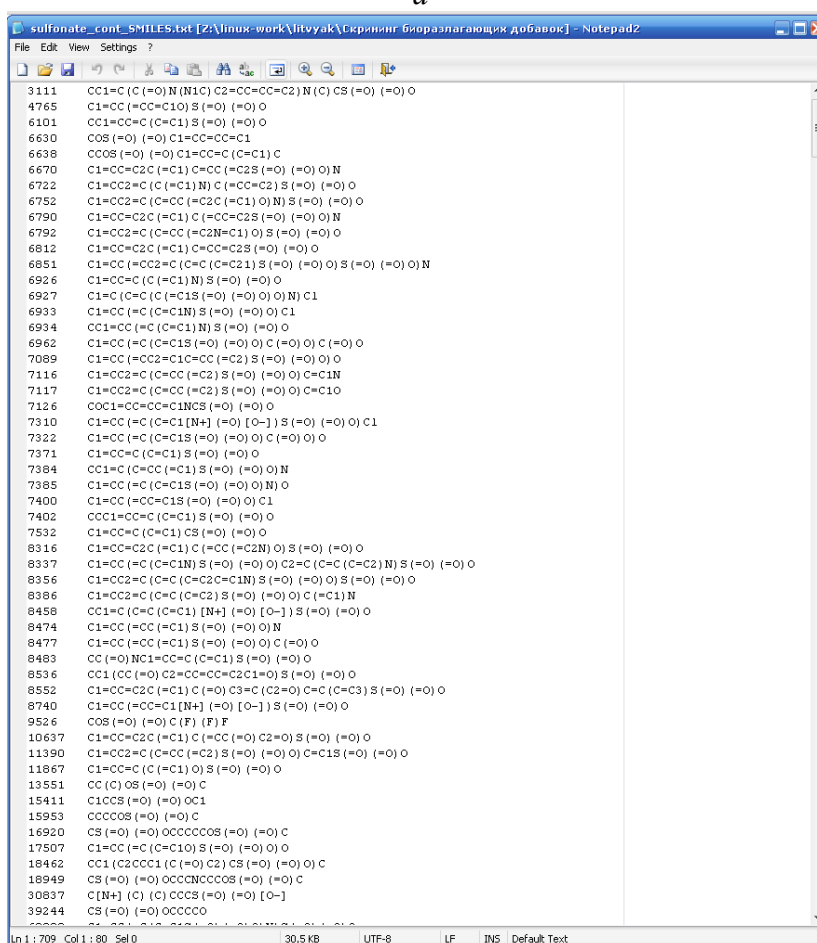
Воспользуемся поиском подструктур для нахождения среди сформированного подмножества тех соединений, которые содержат сульфонатную функциональную группу. Результат поиска состоит из 708 химических соединений. Для возможности дальнейшей обработки результатов как пользователем, так и программным средством в автоматическом режиме, сохраним результаты в виде аннотаций, сортированного списка идентификаторов и XML-файла. Последний представляет собой коллекцию записей, извлеченных без потерь информации из базы данных PubChem.

Аналогичный поиск по карбоксильной группе дает 30433 химических соединений (рис. 1а).

Программа `gper` из стандартного набора утилит Linux позволяет отобрать из результатов те записи, которые содержат нежелательные химические элементы, кроме тех, что были указаны в поисковом запросе. Воспользуемся командой, учитывающей синтаксис файлов с записями SMILES:



*a*



*b*

Рис. 1. – *a* – результат выполнения полного поискового запроса в Entrez; *b* – химическая информация о сульфатсодержащих соединениях в формате SMILES с указанием идентификаторов CID

```
egrep -v "^[0-9]+.+(\(|\)|\[|\]|O|C|l|N|H|S|\-  
|\=|1|2|3|4|5|6|7|8|9|0|@|\+|\#|\|/)+$" имя_файла
```

Таким образом, в результатах с сульфонат-содержащими соединениями оказалось 36 соединений с «нежелательными» химическими элементами. Сформируем новый файл без этих химических соединений командой (рис. 1б)

```
egrep "^[0-9]+.+(\(|\)|\[|\]|O|C|l|N|H|S|\-  
|\=|1|2|3|4|5|6|7|8|9|0|@|\+|\#|\|/)+$"   
sulfonate_cont_SMILES.txt > sulfonate_cont_SMILES_el-  
pure.txt
```

Такой же запрос по отношению файла с записями о химических соединениях, содержащих карбоксильную функциональную группу, позволяет уменьшить размер списка с 30433 до 29289 химических соединений.

```
grep "^[0-9]+.+(\(|\)|\[|\]|O|C|l|N|H|S|\-  
|\=|1|2|3|4|5|6|7|8|9|0|@|\+|\#|\|/)+$"   
carboxylic_cont_SMILES.txt > carboxylic_cont_SMILES_el-  
pure.txt
```

Выполнив следующий набор команд, пользователь получает файлы, содержащие строчный список идентификаторов соединений, который можно использовать для запроса сводных данных по биологической активности.

```
awk '{print $1"  "'} sulfonate_cont_SMILES_el-  
pure.txt > sulfonate_cont_CID_el-pure.txt
```

```
for i in `cat sulfonate_cont_CID_el-pure.txt`; do  
echo -n "$i ">>sulfonate_cont_CID_el-pure_in-line.txt;  
done
```

```
awk '{print $1"  "'} carboxylic_cont_SMILES_el-  
pure.txt > carboxylic_cont_CID_el-pure.txt
```

```
for i in `cat carboxylic_cont_CID_el-pure.txt`; do
```



```
echo -n "$i ">>carboxylic_cont_CID_el-pure_in-line.txt;  
done
```

Данные, полученные в результате скрининга базы данных PubMed, в формате SMILES пригодны для поиска по каталогам производителей базы ZINC (рис. 2) [13], а также некоторых других, в том числе коммерческих, баз данных. Найденные записи содержат информацию о потенциальном производителе химического соединения, наличия его на складе или под заказ, реквизитах для заказа.

Установлено, что для получения новых типов биodeградируемых полимерных материалов на основе полиолефинов и крахмала необходимо дополнительно ввести функционализированный полимер, позволяющий совместить гидрофобный полиолефин с гидрофильным крахмалом. Важными критериями органических веществ, прививаемых на полиолефин являются: доступность по цене, нетоксичность, нелетучесть, способность к полимеризации (наличие двойных связей), размер молекулы.

В результате проделанной работы разработан алгоритм проведения скрининга органических веществ, пригодных для функционализации полиолефинов. Сформировано подмножество химических соединений из базы химической информации PubChem, удовлетворяющее критериям, отобранным для поиска соединения-кандидата для использования в качестве биоразлагающих добавок к полимерам. Отобранные соединения прошли предварительный скрининг на отсутствие специфического воздействия и токсичности в отношении человеческих клеток, тканей и органов, согласно информации в PubChem BioAssay.

Сформирован общий вид поискового запроса, основанный на реализованных в базе PubChem поисковых механизмах на основе индексов,

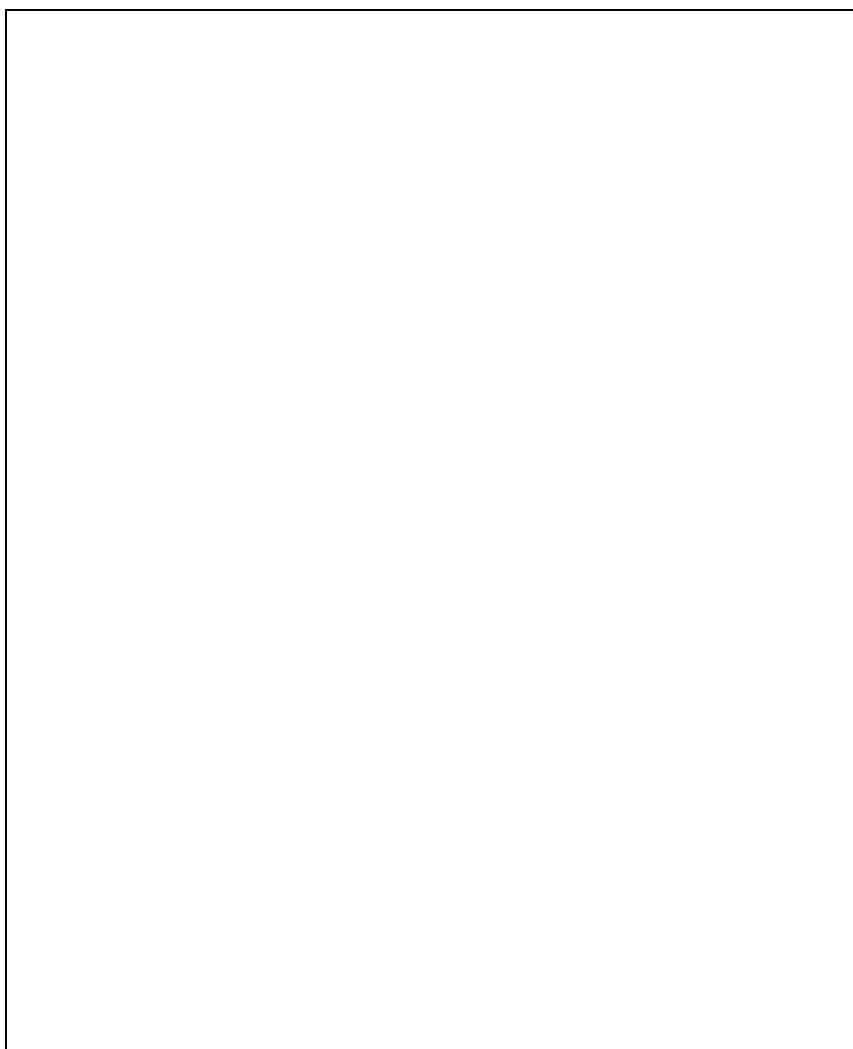


Рис. 2. – Подмножества базы данных ZINC

описывающих экспериментальные и рассчитанные физико-химические свойства химических соединений. В полученном подмножестве, состоящем из 140 тысяч химических соединений, отобраны молекулы, содержащие сульфатные и карбоксильные группы. Полученные группы веществ имеют размер 708 и 30433 соединения, соответственно. Далее размер этих групп был уменьшен путем удаления соединений, содержащих «нежелательные» химические элементы, модификация которыми способна затруднить процесс биоразложения органических соединений, а также потенциально повысить их токсичность. Данные о целевых соединениях-кандидатах переведены в несколько различных форматов представления химической информации

(XML, SMILES), позволяющие вести дальнейшую их обработку как пользователем в ручном режиме, так и автоматически с помощью программных средств и исполняемых сценариев.

Отобранные химические соединения содержат в своем составе полярные группы (сульфонатную, карбоксильную) с большим значением дипольного момента и способностью участвовать в образовании водородных связей. Использованный в скрининге диапазон значений коэффициента разделения между гидрофобным растворителем и водой определил наличие у соединений гидрофобных группировок и, одновременно, потенциальную возможность растворения в воде.

Выполненная работа представляет собой базис для проведения экспериментальных исследований по тестированию стимулирования биоразложения полимерных материалов микроорганизмами.

## ЛИТЕРАТУРА

1. Усачёв, И.С. Применение ультразвукового воздействия для модификации физико-механических свойств биоразлагаемых полимерных композиций с термопластичным крахмалом / И.С. Усачёв, В.В. Колпакова, А.С. Сарджвеладзе, Д.А. Соломин, Н.Д. Лукин, И.Ю. Васильев, В.В. Ананьев // Пищевая промышленность. – 2019. – №8. – С. 48–51.

2. Лукин, Н.Д. Биоразлагаемые съедобные плёнки на основе крахмала / Н.Д. Лукин, В.В. Ананских, Л.Д. Шлеина, А.В. Родионова // Пищевая пром-сть. – 2019. – №11. – С. 13–15.

3. Литвяк, В.В. Перспективы производства современных упаковочных материалов с применением биоразлагаемых полимерных композиций Перспективы производства современных упаковочных материалов с применением биоразлагаемых полимерных композиций / В.В. Литвяк // Экология / журнал Белорусского государственного университета. –

2019. – №2. – С. 84–94.

4. Benson, D.A. GenBank / D.A. Benson [et al.] // *Nucleic Acids Res.* – 2007. – Vol. 35. – P. D21–D25.
5. Boeckmann, B. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. // A. Bairoch [et al.] / *Nucleic Acids Res.* – 2003. – Vol. 31. – P. 365–370.
6. Berman, H. The Protein Data Bank. / H. Berman [et al.] // *Nucleic Acids Res.* – 2000. – Vol. 28. – P. 235–242.
7. Electronic source: <http://www.ncbi.nlm.nih.gov>, accessed 16.11.2011.
8. Electronic source: <http://www.ncbi.nlm.nih.gov/sites/gquery>, accessed 16.11.2011.
9. Electronic source: <http://pubmed.gov>, accessed 16.11.2011.
10. Wang, R. A New Atom-Additive Method for Calculating Partition Coefficients // R. Wang, Y. Fu, L. Lai / *J. Chem. Inf. Comput. Sci.* – 1997. – Vol. 37. – P. 615–621.
11. Ertl, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. / P. Ertl, B. Rohde, P. Selzer // *J Med. Chem.* – 2000. – Vol. 43. – P. 3714–3717.
12. Cheng, T. Computation of Octanol–Water Partition Coefficients by Guiding an Additive Model with Knowledge. // T. Cheng [et. al.] / *J. Chem. Inf. Model.* 2007, 47. – P. 2140–2148.
13. Irwin, J. Virtual Screening against Metalloenzymes for Inhibitors and Substrates / J. Irwin, F. Raushel, B. Shoichet // *J. Chem. Inf. Model.* – 2005; Vol. 45, No. 1. – P. 177–182.

## REFERENCES

1. Usachev, I.S. *Primenenie ultrazvukovogo vozdeystviya dlya*  
<http://ntk.kubstu.ru/file/3072>

modifikatsii fiziko-mekhanicheskikh svoystv biorazlagaemykh polimernykh kompozitsiy s termoplastichnym krakhmalom / I.S. Usachev, V.V. Kolpakova, A.S. Sardzhveladze, D.A. Solomin, N.D. Lukin, I.Yu. Vasilev, V.V. Ananov // Pishchevaya promyshlennost. – 2019. – №8. – S. 48–51.

2. Lukin, N.D. Biorazlagaemye sedobnye plenki na osnove krakhmala / N.D. Lukin, V.V. Ananskikh, L.D. Shleina, A.V. Rodionova // Pishchevaya prom-st. – 2019. – №11. – S. 13–15.

3. Litvyak, V.V. Perspektivy proizvodstva sovremennykh upakovochnykh materialov s primeneniem biorazlagaemykh polimernykh kompozitsiy Perspektivy proizvodstva sovremennykh upakovochnykh materialov s primeneniem biorazlagaemykh polimernykh kompozitsiy / V.V. Litvyak // Ekologiya / zhurnal Belorusskogo gosudarstvennogo universiteta. – 2019. – №2. – S. 84–94.

4. Benson, D.A. GenBank / D.A. Benson [et al.] // Nucleic Acids Res. – 2007. – Vol. 35. – P. D21–D25.

5. Boeckmann, B. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. // A. Bairoch [et al.] / Nucleic Acids Res. – 2003. – Vol. 31. – P. 365–370.

6. Berman, H. The Protein Data Bank. / H. Berman [et al.] // Nucleic Acids Res. – 2000. – Vol. 28. – P. 235–242.

7. Electronic source: <http://www.ncbi.nlm.nih.gov>, accessed 16.11.2011.

8. Electronic source: <http://www.ncbi.nlm.nih.gov/sites/gquery>, accessed 16.11.2011.

9. Electronic source: <http://pubmed.gov>, accessed 16.11.2011.

10. Wang, R. A New Atom-Additive Method for Calculating Partition Coefficients // R. Wang, Y. Fu, L. Lai / J. Chem. Inf. Comput. Sci. – 1997. – Vol. 37. – P. 615–621.

11. Ertl, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug

Transport Properties. / P. Ertl, B. Rohde, P. Selzer // J Med. Chem. – 2000. – Vol. 43. – P. 3714–3717.

12. Cheng, T. Computation of Octanol–Water Partition Coefficients by Guiding an Additive Model with Knowledge. // T. Cheng [et. al.] / J. Chem. Inf. Model. 2007, 47. – P. 2140–2148.

13. Irwin, J. Virtual Screening against Metalloenzymes for Inhibitors and Substrates / J. Irwin, F. Raushel, B. Shoichet // J. Chem. Inf. Model. – 2005; Vol. 45, No. 1. – P. 177–182.

## *SCREENING OF CHEMICAL COMPOUNDS, SUITABLE POLYMERS FOR ACTIVATION*

**V.V. LITVYAK, I.S. USACHEV**

*All-Russian research Institute of starch products – branch of the V. M. Gorbatov Federal research center for food systems of the Russian Academy of Sciences,  
11, Nekrasova str., Kraskovo, Moscow region, Russian Federation, 140051;  
e-mail: vniik@arrisp.ru*

It was established that in order to obtain new types of biodegradable polymer materials based on polyolefins and starch, it is necessary to additionally introduce a functionalized polymer, which allows combining a hydrophobic polyolefin with hydrophilic starch. Important criteria for organic substances grafted onto a polyolefin are: affordability, nontoxicity, non-volatility, polymerization ability (presence of double bonds), and molecule size. As a result of the work done, an algorithm for screening organic substances suitable for the functionalization of polyolefins was developed. A subset of chemical compounds has been formed from the PubChem chemical information database that meets the criteria selected for the search for a candidate compound for use as biodegradable additives to polymers. The selected chemical compounds contain polar groups (sulfonate, carboxyl) with a large dipole moment and the ability to participate in the formation of hydrogen bonds. The range of values of the separation coefficient between the hydrophobic solvent and water used in the screening determined the presence of hydrophobic groups in the compounds and, at the same time, the potential for dissolution in water. The work performed constitutes the basis for experimental research on testing the stimulation of biodegradation of polymeric materials by microorganisms.

**Key words:** functionalized polymer, biodegradation, biodegradable polymer material, chemical information bases, algorithm.