БИОЛОГИЧЕСКИЙ АСПЕКТ ПРОБЛЕМЫ ПОСТРОЕНИЯ БАЗЫ ДАННЫХ БИОЛОГИЧЕСКИ АКТИВНЫХ СОЕДИНЕНИЙ

Шуваев А.В.
к.х.н., доцент
ФГБОУ ВО Сибирский государственный университет путей сообщения,
г. Новосибирск, Россия

Аннотация. Разработаны технологические основы обработки биологической и физической части патентной информации биологически активных химических соединений. Ввод и вывод информации осуществляется с помощью двух файлов: "Format", содержащего набор наименований характеристик в общем виде, и "Value" – конкретные данные для определенного химического соединения. На примере соединений с гербицидным видом активности показаны основные приемы процедуры предварительной обработки информации с последующим вводом в базу данных.

Ключевые слова: биоактивные соединения, химическая структура, обработка патентной биологической информации, база данных, генерация биологической информации.

Контакты автора: Шуваев А.В., shuvaev53@mail.ru

В предыдущих работах [6, 7] были изложены основные принципы построения базы структурных данных биологически активных химических соединений на основе обработки патентной информации с помощью семи отдельно создаваемых текстов-файлов. При этом основное внимание уделялось лишь только процедуре обработки химической структурной информации посредством подробного описания содержания следующих четырех файлов: "Atom" и "Radical" – символьное обозначение единичных атомов или их условных наборов; "Formula" – формулы Маркуша [1 – 3] в обобщенном виде для определенного структурного интервала; "Replace" – полный список операций замен конкретных радикалов на химические атомы или типы химических связей.

Следующим этапом является разработка основ технологии обработки и ввода в машинную память информации о биологических, физических и технологических характеристиках химических соединений. К ее основным требованиям следует отнести: создание по возможности однообразного характера представления информации, достаточного для понимания и работы с ней специалистов различного профиля, в сочетании с компактностью, емкостью, быстротой ввода/вывода информации. Решения этих взаимосвязанных задач можно достичь с помощью формирования двух файлов: "Format" – перечень наименований основных характеристик в общем виде; "Value" – конкретные данные для определенного химического соединения. В принципе для всего массива данных можно было бы создать один универсальный формат, однако сложность задачи заключается в том, что в зависимости от принадлежности к определенному виду активности, в патентах приводится разнообразная биологическая информация и, кроме того, даже в пределах одного вида имеется большая разнородность в представлении биологических данных разных авторов. В связи с этим имеет смысл обработку биологической информации проводить отдельно по видам и внутри каждого, в зависимости от его специфических особенностей, осуществлять попытку создания универсального формата.

В данной работе приводятся результаты обработки химических соединений с гербицидным видом активности.

Файл "Format"

Обработке биологической информации обычно предшествует процедура обработки химической структурной информации, в результате которой данные по конкретной химической структуре заносятся на отдельную часть магнитного диска с присвоением определенного номера. Как уже упоминалось ранее [7], патенты с гербицидным видом активности обычно содержат большой объем структурных данных – от нескольких сот до нескольких тысяч. Поэтому удобно структурные данные одного патента обрабатывать и заносить целиком на отдельную часть магнитного носителя. Что касается биологических данных, то по сравнению со структурными данными, их количество в патенте на 1-2 порядка меньше.

При изготовлении формата следует исходить из того, что его содержание можно условно разбить на две составные части: вспомогательную и основную. Вспомогательная часть включает в себя: номер формата (цифра); номер части магнитного носителя (ML) – условный номер одного или нескольких совместно обрабатываемых патентов; структурный интервал, охватываемый патентом (structure interval); химический класс соединений (chemical class); вид активности, название теста (test); описание метода (application mode); единицы шкалы биологической активности (numerical scale).

В основной части формата содержится следующая информация: структурный номер испытуемого соединения (structure number); температура плавления твердых веществ (°C) (m.p.(grad C)); для жидких индивидуальных соединений температура кипения при определенном значении давления (b.p.(grad C)/P(torr)); показатель преломления при определенной температуре (ND/(grad C)); спектральные характеристики (IR(1/SM)); доза (Rate(KG/HA)); стадия развития растения – до и после всхожести (stage) и, наконец, полный перечень названий растений, на которых изучалось воздействие испытуемого соединения. Отдельные элементы формата отделяются друг от друга с помощью

амперсанда (&), в конце формата ставится звездочка: *.

Все тестируемые в разных патентах растительные культуры и различные сорные травы можно было бы выписать в виде единого ряда, и тогда получился бы один формат для всех испытуемых соединений с гербицидной активности. Однако в этом случае сильно затруднится процедура обработки, т. к. разные авторы не придерживаются какого-либо одного порядка представления наименований растений. С другой стороны, создание отдельного формата для каждого набора данных приведет к очень большому числу форматов, что ухудшит компактность базы данных. В качестве поиска наиболее оптимального решения можно рекомендовать следующее: при обработке нового патента необходимо просматривать картотеку форматов базы данных и в случае точного или близкого совпадения как набора характеристик, так и последовательности их расположения, с одним из имеющихся форматов файла "Format", использовать его, при больших различиях – создавать новый формат.

Ниже приведены некоторые разработанные форматы, наиболее типичные для тест-испытаний соединений на гербицидный вид активности, использованы данные работ [4, 5, 8 – 11]. Всего таких форматов в базе данных насчитывается около 60, в файле "Format" они записаны под следующими номерами: 40-47, 51-68, 70, 200-219, 221-225, 227-234, 242-245, 247-256.

*Format:

5&ML=&structure interval=&structure number=&m.p.(grad C)=*

38&ML=&structure interval=&structure number=&m.p.(grad C)=&ND/(grad C)=&f.p. (grad C) = & b.p.(grad C)/P(torr)=*

40&ML=&structure interval=&structure number=&Rate(KG/HA)=&stage=&Bushbean=& Cotton=&Morningglory=&Cocklebur=&Cassia=&Nutsedge=&Crabgrass=&Barnyardgrass=&Wild Oats=&Wheat=& Corn=&Soybean=&Rice=&Sorghum=*

200&ML=&structure interval=&chemical class=&test=&Application mode=&numerical scale=&Rate(g/ha)=&stage=&structure number=&bush bean=&cotton=&morningglory=&cocklebur =&sicklepod=& nutsedge=&crabgrass=&barnyardgrass=&wild oats=&wheat=&corn=&soybean =& rice=&sorghum=&sugar beet= &velvetleaf=&cheatgrass=*

210&ML=&structure interval=&application mode=&stage=&rate(g/ha)=&structure number =&soybean=&velvetleaf=&sesbania=&cassia=&cotton=&morningglory=&alfalfa=&jimsonweed=&

cocklebur=&corn=&crabgrass=&rice=&nutsedge=&barnyardgrass=&wheat=&giant foxtail=&wild oats=&sorghum=*

230&ML=&structure interval=&test=&application mode=&numerical scale=&stage=& rate (g/ha)=& structure number=&bush bean=&cotton=&sorghum =&corn=&soybean=&wheat=&wild oats=&rice=& barnyardgrass=&crabgrass=&morningglory=&coklebur=&cassia=&nutsedge=*

244&ML=&structure interval=&chemical class=&test=&application mode=&numerical scale=&stage=& time (weeks)=&rate (KG/HA)=&structure number=&Canada thistle=&cocklebur= &velvetleaf=& morningglory=& lambsquarters=&smartweed=&yellow nutsedge=&quackgrass=& johnsongrass=&downy brome=&barnyardgrass=&soybean=&sugar beet=&wheat=&rice=&sorghum= &wild buckwheat=&hemp sesbania=&panicum spp=&crabgrass=*

Файл "Value"

После изготовления формата осуществляется обработка биологических данных патента. Строго по позициям формата записываются цифровые значения или текстовый материал. При обработке информации в рамках одного патента вспомогательная часть формата будет одинакова для всего структурного интервала и поэтому позиции этой части формата заполняются данными только один раз. В основной же части вся информация носит переменный характер – каждому испытуемому соединению на определенных тест-объектах соответствуют свои физико-химические и биологические характеристики. Более трудоемкой представляется работа над второй частью формата.

При составлении текста данных в пределах одного формата возможно перемещение как в прямом, так и в обратном направлении с пропуском отдельных позиций. При этом, если необходимо пропустить ряд позиций, например при отсутствии информации в патенте, то после последней обработанной информации ставится амперсанд: *&* и знак: #+*n*, где *n* – натуральное число, его значение показывает сколько позиций надо в формате пропустить при перемещении вправо. Если в патенте для одного и того же соединения имеется набор данных при разных дозах, стадиях всхожести или других условиях, а также при необходимости перехода к обработке данных по другим соединениям, осуществляется перемещение по формату назад с помощью операции: #-*m*, при этом перенос влево происходит на *m*-1 позицию.

Таким образом, после ввода информации для вспомогательной части

формата, заносится в виде непрерывного текста полные данные об испытанных соединениях и тест-объектах с пропуском позиций при отсутствии информации или с возвращением на определенные позиции при обработке данных следующего соединения из патента. По окончании обработки данных патента по определенному формату в конце текста ставится знак: *. Далее записывается номер нового формата и все необходимые данные в соответствии с его позициями. Так в конечном итоге происходит формирование текста-файла "Value", в котором, в частности, накапливается информация по биологической активности.

В качестве примера ниже показан фрагмент файла "Value" для соединений с гербицидным видом активности.

*Value:

3-40&437-821&446&0.05&Post/Pre&6C,9G,6Y&4C,9G&6C,9G/9G&9C/9H&6C,9G/9G&8G/10E&2C,8G/3C,8G&6C,9G/9H&3C,9G/2C,8G&3C,9G/9H&2U,9G/9G&9C/9H&6C,9G/10E&2C,9G/6C,9G&#-17&447&0.05&Post/Pre&5C,9G,6Y&5C,9G&9C/9G&3C,9G/9H&6C,9G/9G&2C,8G/10E&2C,8G/4C,8G&9C/5C,9H&3C,7G/2C,8G&3C,8G/1C,9H&5C,9G/9G&6C,9G/9H&9C/10E& 2C,9G/5C,9H&...

5-200&1355-4214,7137-7154&Pyrimidine sulfonylureas&Herbicidal activity&seeds were planted and in pre-phase treated with the chemicals dissolved in non-phytotoxic solvent. In post-phase when plants had a few leafs from 2 to 5 they were treated too. Treated plants and controls were maintained in greenhouse for 16 days and compared visually rated for response to treatment&0=no injury&10=complete kill. Symbols: C=chlorosis or necrosis; D=defoliation; E=emergence inhibition; G=growth retardation; H=formative effects; S=albinism; U= unusual pigmentations; X=axillary stimulation; 6Y=abscised buds or flowers&50&Post-emergence&1355&5C,9G,6Y&2C,3H,6G&1C,5G&2C,8G&2C,5G&1C,3G&5G&2C,9H&2C,8H&2C,8H&5U,9C&2C,7G&6C,9G&4U,9C&#-15&1356&9C&6C,9G&2C,5G&9C&5C,9G&3C,9G&4C,9G&9C&3U,9G&9C&6C,9G&6C,9G& 5U,9C&...

23-244&193-213&N-carbobenzoxy-N-phosphonomethylglycine thioesters&Herbicidal activity&Alumunium pans having holes in the bottom and compacted to a depth of 0.95 to 1.27 cm from the top were filled with soil, seeds of plants, covered with soil and placed in greenhouse. Some time after when plants reach the desired age, each pan, except control pans, were treated with suspensions of chemicals. Then plants returned to the greenhouse, watered and 2 or 4 weeks late injury to the plants was compared to the control&plant response (% control) = Index:0-24=0;25-49=1;50-74=2;75-99=3;100=4&Post&2&11.2&193&l&0&0&0&l&l&0&0&0&0&1&#-14&4&11.2

&194&1&2&1&1&2&2&1&2&3&2&2&#-13&5.6&194&1&1&1&1&2&1&0&1&1&1&2&#-13&11.2&195&1&1&1&1&1&2&1&2&3&2& 2&...

5-5&1355-4214,7137-7154&1355&150-153&#-2&1356&121-125&...

17-5&206-662,761-770&762&155-159&...

23-38&193-213&194&&1.5580/25&...

Содержимое файлов "Format" и "Value" хранится на магнитных носителях. С помощью специальной обрабатывающей программы для каждого химического соединения базы данных возможна генерация всей имеющейся информации из отдельных файлов "Patent", "Atom", "Radical", "Formula", "Replace", "Format", "Value" в виде набора данных о литературном источнике, структурной формуле соединения, характеристиках физических и биологических свойств. Демонстрация последнего в режиме "Activity" показана ниже на ряде примеров.

*Activity:

ML=5&structure interval=1355-4214,7134-7154&chemical class=Pyrimidine sulfonylureas& test =Herbicidal activity&Application mode=seeds were planted and in pre-phase treated with the chemicals dissolved in a non-phytotoxic solvent. In post-phase when plants had a few leafs from 2 to 5 they were treated too. Treated plants and controls were maintained in greenhouse for 16 days and compared visually rated for response to treatment&numerical scale=0=no injury; 10=complete kill. Symbols: C=chlorosis or necrosis; D=defolation; E=emergence inhibition; G=growth retardation; H=formative effects; S=albinism; U=unusual pigmentations; X=axillary stimulation; 6Y=abscised buds or flowers& Rate(g/ha)= 50&stage= Post-emergence&structure number=1355&bushbean=5C, 9G,6Y&cotton=2C,3H,6G&morningglory=1C,5G&cucklebur=2C,8G&sicklepod=2C,5G&nutsedge= 1C,3G&crabgrass=5G&barnyardgrass=2C,9H&wildoats=2C,8H&wheat=2C,8H&corn=5U,9C&soybean =2C,7G&rice=6C,9G&sorghum=4U,9C*

*Activity:

ML=5&structure interval=1355-4214,7137-7154&structure number=1355&m.p.(grad C)= 150-153*

*Activity:

ML=23&structure interval=193-213&chemical class=N-carbobenzoxy-N-phosphonomethyl glycine thioesters&test=herbicidal activity&application mode=aluminium pans having holes in the bottom and compacted to a depth of 0,95 to 1,27cm from the top were filled with soil, seeds of plants, covered with soil and placed in greenhouse. Some time after when plants reach the desired age each pan, except control pans, were treated with suspensions of chemicals. Then pans returned to the greenhouse, watered and 2 or 4 weeks late injury to the plants was compared to the control&

numerical scale=plant response (%control)= Index:0-25=0,25-49=1,50-74=2,79-99=3,100=4&stage =post&time(weeks)=2&rate(KG/HA)=11.2&structure number=194&Canada thistle=1&cocklebur=2& velvetleaf=1&morningglory=1&lambsquarters=2&smartweed=2&yellow nutsedge=1&quackgrass= 2&johnsongrass=3&downy brome=2&barnyagrass=2&rate(KG/HA)=5.6&structure number=194& Canada thistle=1&cocklebur=1&velvetleaf=1&morningglory=1&lambsquarters=2&smartweed=1& yellow nutsedge=0&quackgrass=1& johnsongrass=1& downy brome=1& barnyagrass=2*

*Activity:

ML=23&structure interval=193-213&structure number=194&m.p.(grad C)=&ND/(grad C)= 1.5580/25*

В заключении необходимо отметить, что изложенная в этой работе методика обработки и хранения биологически активной информации химических соединений на магнитных носителях в виде создаваемой базы данных целиком себя оправдывает. Запись наиболее важной части патентной информации осуществляется в концентрированной и компактной форме. Определенные сложности, как в стадии создания, так и в последующем использовании единого универсального формата, приводят к выводу о целесообразности производить обработку биологической информации по отдельным видам независимо. В дальнейшем мы намерены продолжить исследования форм патентного представления биологической информации с другими видами активности с последующим созданием их компьютерного аналога.

Список литературы

1. Блэдуц Г.Э., Гейвандов Э.А. Автоматизированные информационные системы для химии. – М.: Наука, 1974 – 312 с.

2. Мартиросов А.К. Разработка автоматизированных подсистем регистрации и классификации химико-структурных данных с использованием формул Маркуша в системе информационного обеспечения по проблемам химической безопасности: Дис. … канд. тех. наук. Москва, 2005. 168 с.

3. Пирс Т., Хони Б. Искусственный интеллект: применение в химии. – М.: Мир, 1988 – С. 238-259.

4. Фишер А., Рор В. Гербицидный состав. Патент СССР № 573115. Заявл. 15.05.74: опубл. 15.09.77, 10 с.

5. Штурм Е., Целлариус Х.Е., Бредов Б., Фогель К. Гербицидное средство. Патент СССР № 572175. Заявл. 01.02.72: опубл. 05.09.77, 8 с.

6. Шуваев А.В. Основы построения базы данных биологически активных химических соединений для решения экологических вопросов: Вопросы строительства и инженерного оборудования объектов железнодорожного транспорта: материалы науч. – практ. конф. Новосибирск: Изд-во СГУПСа, 2017. С. 147 – 157.

7. Шуваев А.В. Подготовка патентной химической структурной информации к вводу в базу данных биологически активных химических соединений // Аэкономика: экономика и сельское хозяйство. 2018. № 2 (26). С. 5. URL: http://aeconomy.ru/science/agro/podgotovka-patentnoy-khimicheskoy-s/

8. Brookes R.F., Godson D.H., Greenwood D., Tulley M., Wakerley S.B. 1-carbamoyl-1,2,4-triazole herbicidal agents. Patent UK 31922/70. Filed 01.07.70: publ. 31.03.71, 40 p.

9. Chupp J.P., Worley J.W. Herbicidal composition and method of use. Patent US 4242123 A. Filed 30.11.78: publ. 30.12.80, 5 c.

10. Takahashi R., Fujikawa K., Yokomichi I., Someya S., Sakashita N. Herbicidal compound, herbicidal composition containing the same, and methods of use thereof. Patent Japan 49-82403. Filed 17.07.74: publ. 14.10.74, 28 p.

11. Zimal S.D. Herbicidal and plant-growth-regulating N-substituted-N-(2,5-dialkylpyrrol-1-yl) haloacetamides. Patent US 4282028 A. Filed 25.07.79: publ. 04.08.81, 10 p.

THE BIOLOGICAL ASPECT OF THE PROBLEM TO CONSTRACT A DATABASE OF BIOLOGICALLY ACTIVE CHEMICAL COMPOUNDS

Shuvaev A.V.
PhD in Chemistry, docent,
Siberian State University of Railway Engineering,
Novosibirsk, Russia

Annotation. The technological bases of processing of the biological and physical patent parts information about biologic active chemical compounds have been developed. Input and output of data carried out by means of two files: "Format", containing a set of characteristics of items in general, and "Value" – specific data for a particular chemical compound. The basic techniques of the procedure for preliminary processing of information with subsequent entry into the database are shown on the example of compounds with an herbicidal activity.

Keywords: bioactive compounds, chemical structure, biological processing of patent information, the database, the generation of biological information.