УДК 004.891:[547.6.04:615.277.4]

Система, реализующая комбинаторно-численный подход к проблеме прогноза свойств химических соединений. Прогноз канцерогенности полициклических ароматических углеводородов (ПАУ)

М. В. Максин, Н. В. Харчевникова, В. Г. Блинова, Д. А. Добрынин, З. И. Жолдакова

Описываются математическая модель и программная реализация системы типа ДСМ, позволяющей включать числовые данные в ДСМ-анализ свойств химических соединений. Приведены результаты тестирования системы с использованием данных по канцерогенной активности полициклических ароматических углеводородов.

В наших предыдущих работах [1–3] описаны подходы к созданию программной системы, основанной на совместном использовании ДСМ-метода правдоподобных рассуждений [4,5] и анализа числовых параметров в задаче прогноза свойств химических соединений, в частности биологической активности. При этом числовые характеристики могут отражать "физику" явлений и процессов, в которых объект участвует (как, например, энергия активации в химических реакциях) и, таким образом, являться важным элементом настройки интеллектуальной системы анализа на конкретную предметную область.

Вопрос о включении анализа числовых параметров в ДСМ-систему обсуждался еще в работе [6]. Такие числовые параметры, как коэффициенты распределения октанол/вода (logP), характеризующие гидрофобность, т. е. способность соединения проникать через мембраны к рецептору, энергетические характеристики реакционной способности и т. д., не содержатся явно в структурном графе и дескрипторах фрагментарного кода суперпозиции подструктур (ФКСП), которые служат для представления соединений в ДСМ-системе. Числовые параметры могут быть использованы также как "границы" для отсекания лишних ветвей при построении дерева метаболизма.

В настоящей работе описывается метод поиска сходства на числовых параметрах, что позволяет включить их анализ в ДСМ-систему правдоподобных рассуждений, а также программная реализация гибридной интегрированной системы. В этой системе ДСМ-решатель дополнен квантово-химическим вычислителем, который используется для получения числовых характеристик соединений, и модулем, выполняющим операцию сходства над числовыми величинами. В основе математической модели этого модуля лежит алгебра интервалов.

АЛГЕБРА ИНТЕРВАЛОВ

Чтобы объект некой предметной области мог участвовать в ДСМ-рассуждениях [5], для него должна быть определена операция сходства, сопоставляющая двум объектам третий, который и выражает их сходство. Для корректной работы ДСМ-метода такая операция должна обладать определёнными алгебраическими свойствами, индуцирующими отношение толерантности (т. е. рефлексивное и симметричное отношение) на объектах. Более формально [4], операцией сходства на множестве объектов S, представляющем некоторую предметную область, будем называть операцию $\cap$, удовлетворяющую следующим свойствам (где s_i, s_j, s_k — произвольные объекты из S):

$$(1)\ s_i \cap s_i = s_i,$$

$$(2)\ s_i \cap s_j = s_j \cap s_i,$$

$$(3)\ (s_i \cap s_j) \cap s_k = s_i \cap (s_j \cap s_k),$$

$$(4)\ s_i \cap s_0 = s_0,$$

для некоторого s_0 из S, который называется пустым объектом.

Пустой объект представляет собой "неинформативное" сходство. Существует несколько вариантов задания операции сходства для численных величин [2,4]. Можно разбить всю область значений числовой величины на непересекающиеся интервалы и считать, что внутри одного интервала I значения сходны. Для таких значений сходство выражается как "интервал I", а для значений из разных интервалов — сходство пусто. Недостатком такого подхода является то, что предопределённое разбиение может оказаться неадекватным и в результате близкие "по смыслу" значения могут оказаться "несходными".

Избежать произвола при разбиении численных значений на интервалы в рамках ДСМ-АПГ можно, например, введя следующую алгебру интервалов. Пусть (a_i, b_i) — интервалы, и $A \le a_i, b_i \le B$, где A и B — минимальное и максимальное возможные значения числового параметра. Операция сходства определяется как $(a_i, b_i) \cap (a_j, b_j) = (\min(a_i, a_j), \max(b_i, b_j))$, а (A, B) играет роль пустого объекта. Нетрудно проверить, что эта операция обладает свойствами (1)—(4). Исходные "точечные" значения параметра x_i представляются в виде интервалов (x_i, x_i). Свойством этого определения является то, что сходство практически никогда не бывает пустым; в ДСМ-выводе это может играть как положительную (порождая более полную систему гипотез), так и отрицательную роль (значительно увеличивая конечное число гипотез). Именно это и наблюдалось в ходе экспериментов: если при анализе учитывались числовые параметры, то число гипотез увеличивалось в десятки раз, но при этом удавалось спрогнозировать большее число свойств.

РЕАЛИЗАЦИЯ МОДУЛЯ

Модуль нахождения сходства числовых величин выполнен в виде COM-библиотеки с одним экспортируемым объектом, реализующим интерфейсы JSMSOLVERLib::IJSMFragment и NUMJSMOBJLib::INumJSMFragment. Интерфейс IJSMFragment определён в библиотеке JSMSOLVERLib [10] и содержит методы, общие для всех ДСМ-объектов, такие, как нахождение сходства, проверка вложимости, проверка равенства. Интерфейс INumJSMFragment определён локально и содержит методы, специфичные для данного ДСМ-объекта, а именно процедуры задания и извлечения интервала (в виде чисел — левой и правой границ интервала). Методы реализованы в соответствии с описанной выше моделью, при этом в качестве пустого объекта принимается интервал [–1.7E+308, 1.7E+308].

ПОЛЬЗОВАТЕЛЬСКИЙ ИНТЕРФЕЙС ИНТЕГРИРОВАННОЙ СИСТЕМЫ

Для проведения экспериментов с интегрированной системой потребовалось расширить описание входных данных. Теперь при вызове программы ей передаётся имя файла, содержащего описание эксперимента, который может содержать в себе ссылки на два других вспомогательных файла с описанием свойств объектов и числовыми характеристиками объектов. Все упомянутые файлы являются XML-файлами [11] приведенной ниже структуры и содержат иерархию (дерево) свойств, используемых программой. Тег <p> обозначает лист в таком дереве, а тег <set> — узел:

```
<!ELEMENT properties (p|set)+>
<!ELEMENT set (p|set)+>
<!ATTLIST   set
        id CDATA #REQUIRED>
<!ELEMENT p (#PCDATA)>
<!ATTLIST   p
        id CDATA #REQUIRED>
```

ФАЙЛ ОПИСАНИЯ ЭКСПЕРИМЕНТА

Этот файл определяет следующие свойства (параметры эксперимента):

- obj_data — маска для имени MOL-файлов, участвующих в эксперименте (относительно каталога, из которого запускается программа);
- obj_props — имя файла, содержащего описание свойств объектов (см. ниже);
- obj_params (необязательный) — имя файла, содержащего числовые характеристики объектов (см. ниже); если указан, все числовые характеристики будут включены в гибридный объект в качестве независимых компонентов;
- useMetabolit — индикатор использования структуры конечного метаболита в качестве одного из компонентов гибридного объекта;
- useStructure — индикатор использования структуры исходного соединения в качестве одного из компонентов гибридного объекта;
- useNumber — индикатор использования энергии образования конечного метаболита в качестве одного из компонентов гибридного объекта;
- allowInclusionInCongrObjects — индикатор проверки запрета на контрпримеры;
- printHasseDiagrams — индикатор вывода на экран диаграммы Хассе полученного множества гипотез;
- JSMEngine — имя набора параметров ДСМ-решателя; набор должен быть определён в этом же файле и содержать следующие параметры:

Learner — COM-имя объекта, выполняющего ДСМ-обучение, Solver — COM-имя объекта, выполняющего ДСМ-прогноз.

Пример файла описания эксперимента приведен ниже:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE properties
[
    <!ELEMENT properties (p|set)+>
    <!ELEMENT set (p|set)+>
    <!ATTLIST   set
          id CDATA #REQUIRED>
    <!ELEMENT p (#PCDATA)>
    <!ATTLIST   p
          id CDATA #REQUIRED>
]
>
<properties>
    <p id="obj_data">.\data\all\*mol.</p>
    <p id="obj_props">.\data\all\props.xml</p>
    <p id="obj_params">..\data\params.xml</p>
    <p id="useMetabolit">false</p>
    <p id="useStructure">true</p>
    <p id="useNumber">true</p>
    <p id="JSMEngine">JSMEngineSimple</p>
    <p id="allowInclusionInContrObjects">false</p>
    <p id="printHasseDiagrams">false</p>
    <set id="JSMEngineSimple">
        <p
id="Learner">JSMSolver.JSMLearnerSimple.1</p>
        <p
id="Solver">JSMSolver.JSMSolverSimple.1</p>
    </set>
    <set id="JSMEngineGeneralized">
        <p
id="Learner">JSMSolver.JSMLearnerGeneralized.1</p>
        <p
id="Solver">JSMSolver.JSMSolverGeneralized.1</p>
    </set>
</properties>
```

ФАЙЛ ОПИСАНИЯ СВОЙСТВ ОБЪЕКТА

Файл содержит не более одной записи для каждого из объектов, участвующих в эксперименте. Запись идентифицируется именем MOL-файла, и

содержит признак наличия для одного или нескольких из no_of_props свойств. Свойства идентифицируются порядковым номером, начиная с 0. Признак может принимать одно из трех значений: 0 — если объект обладает свойством, 1 — если не обладает и ? — если для данного свойства требуется прогноз. Если для какого-либо объекта запись не найдена или она не содержит значения для какого-либо из свойств, то этот объект считается τ-объектом по всем ненайденным свойствам.

Пример файла описания свойств объекта приведен ниже:

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE properties
[
    <!ELEMENT properties (p|set)+>
    <!ELEMENT set (p|set)+>
    <!ATTLIST   set
          id CDATA #REQUIRED>
    <!ELEMENT p (#PCDATA)>
    <!ATTLIST   p
          id CDATA #REQUIRED>
]
>
<properties>
    <p id="no_of_props"> 2 </p>
    <set id="j01.mol"><p id="0"> 1 </p><p id ="1"> 0
</p></set>
    <set id="j02.mol"><p id="0"> 0 </p><p id ="1"> 0
</p></set>
...
    <set id="j35.mol"><p id="0"> ? </p></set>
    <set id="j36.mol"><p id="0"> 1 </p><p id="1">?
</p></set>
<properties>
```

ФАЙЛ ЧИСЛОВЫХ ХАРАКТЕРИСТИК ОБЪЕКТА

Как и файл описания свойств, этот файл содержит не более одной записи для каждого из объектов, участвующих в эксперименте. Запись идентифицируется именем MOL-файла и содержит значения для одного или нескольких из no_of_params числовых параметров. Параметры идентифицируются порядковым номером, начиная с 0. Если для какого-либо объекта запись не найдена или не содержит значения для какого-либо из параметров, то значения для всех ненайденных параметров принимаются равными –1.7E+308.

Пример файла числовых характеристик объекта приведен ниже:

```
<?xml version="1.0" endoding="iso-8859-1"?>
<!DOCTYPE properties
[
    <!ELEMENT properties (p|set)+>
    <!ELEMENT set (p|set)+>
    <!ATTLIST   set
          id CDATA #REQUIRED>
    <!ELEMENT p (#PCDATA)>
    <!ATTLIST   p
          id CDATA #REQUIRED>
]
>
<properties>
    <p id="no_of_params"> 2 </p>
    <set id="j01.mol"><p id="0">  .125  </p><p
id="1"> 42.4 </p></set>
    <set id="j02.mol"><p id="0">  .155  </p><p
id="1"> 49.6 </p></set>
...
    <set id="j35.mol"><p id="0">  .000  </p><p
id="1"> 44.4 </p></set>
    <set id="j36.mol"><p id="0">  .321  </p><p
id="1"> 46.8 </p></set>
    </properties>
```

Система предоставляет возможность работы в различных режимах:

1) прогноз с поиском сходства на структурах исходных соединений без учета числовых параметров;

2) прогноз с поиском сходства на структурах исходных соединений с учетом числовых параметров. Причем в случае прогноза для ПАУ числовой параметр рассчитывается системой, а в случае прогноза для соединений других структурных рядов числовые параметры считываются из входного файла;

3) прогноз с поиском сходства на структурах конечных метаболитов, которые генерируются системой, и учетом числовых параметров.

Система тестирована в ходе прогноза канцерогенной активности полициклических ароматических углеводородов. Согласно биохимической и квантово-химической модели [7–9] канцерогенная активность ПАУ определяется их биоактивацией с образованием метаболитов — диолэпоксидов, способных связываться с ДНК. Диолэпоксид атакует критические нуклеофильные позиции ДНК по механизму нуклеофильного замещения первого порядка (S_N1) с образованием в качестве интермедиатов триолкарбокатионов. Скорость процесса биоактивации ПАУ лимитируется образованием этих интермедиатов из диолэпоксидов. Отсюда следует, что биологическое действие ПАУ — мутагенное или канцерогенное — будет тем сильнее, чем меньше требуется энергии для образования карбокатиона из диолэпоксида. В [3] описан программный модуль, реализующий нахождение наиболее активного конечного метаболита с помощью квантово-химических расчетов. При этом рассчитывается энергетический параметр, характеризующий энергию активации реакции образования карбокатиона из диолэпоксида. На числовых параметрах, соответствующих наиболее активному конечному метаболиту, был осуществлен поиск сходства и анализ этих параметров был включен в ДСМ-систему правдоподобных рассуждений.

Компьютерный эксперимент проводили на массиве 25 ПАУ. Канцерогенная активность характеризовалась индексами Бэджера. В качестве структурной компоненты использовали химический граф генерированного системой конечного метаболита. Полученная в результате эксперимента система гипотез покрывала ВСЕ рассмотренные соединения (т. е. критерий достаточности оснований принятия гипотез выполнился полностью), чего не удавалось достигнуть, рассматривая только структуру соединений.

Для кодирования ПАУ был усовершенствован язык ФКСП [12]. Был предложен новый способ кодирования ПАУ, основанный на их представлении в виде цепочек циклов. Усовершенствованный кодировщик ФКСП входит в состав комбинированной системы прогноза.

Канцерогенную активность трех соединений необходимо было предсказать, так как экспериментальные данные для них противоречивы. По результатам одних экспериментов соединения были отнесены к одной группе — по Бэджеру, а по результатам других экспериментов — к другой. Такие результаты описываются, например, как +/–. В соответствии с индексами Бэджера ДСМ-эксперимент проводился по четырем свойствам:

1. Соединение (по крайней мере) слабоактивно.

2. Соединение (по крайней мере) обладает средней активностью.

3. Соединение (по крайней мере) сильноактивно.

4. Соединение сверхактивно.

В табл. 1 приведены экспериментальные оценки канцерогенной активности соединений (значения свойств) и рассчитанные значения числового параметра ΔE.

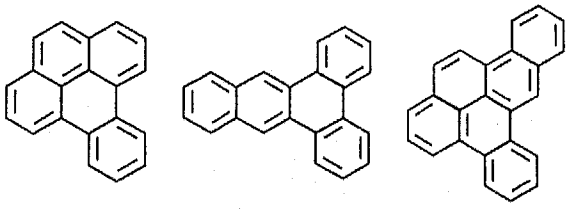
Таблица 1

Экспериментальные оценки канцерогенной активности соединений (значения свойств) и значения числового параметра ΔE (в единицах β)

Соединение	1	2	3	4	ΔE
Бензо(е)пирен	+/–	–	–	–	0.826
Дибенз(а,с)антрацен	+/–	–	–	–	0.833
Дибензо(а,е)пирен	+	+	+/–	+/–	0.846

Структуры соединений приведены на рисунке.

Бензо(е)пирен Дибензо(а,с)антрацен Дибензо(а,е)пирен



Структура соединений, канцерогенную активность которых необходимо было предсказать

Результаты компьютерных экспериментов при работе системы в различных режимах приведены в табл. 2–4. Знак вопроса означает, что значение свойства не может быть доопределено с использованием сгенерированных гипотез.

Таблица 2

Результат прогноза при работе системы в режиме поиска сходства на структурах исходных соединений без учета числовых параметров

Соединение	1	2	3	4
Бензо(е)пирен	+	?	?	?
Дибенз(а,с)антрацен	+	?	?	?
Дибензо(а,е)пирен	+	+	?	?

Таблица 3

Результат прогноза при работе системы в режиме поиска сходства на структурах исходных соединений с учетом числового параметра

Соединение	1	2	3	4
Бензо(е)пирен	?	–	–	–
Дибенз(а,с)антрацен	+	?	–	–
Дибензо(а,е)пирен	+	+	–	–

Таблица 4

Результат прогноза при работе системы в режиме поиска сходства на структурах конечных метаболитов и учетом числового параметра

Соединение	1	2	3	4
Бензо(е)пирен	?	–	–	–
Дибенз(а,с)антрацен	+	?	–	–
Дибензо(а,е)пирен	+	+	–	–

Анализ результатов свидетельствует, что учет числового параметра позволил уточнить прогноз канцерогенной активности соединений, были доопределены свойства сильной и сверхсильной активности.

Прогноз на структурах исходных соединений и структурах конечных метаболитов дает одинаковые результаты. Возможно, это связано с несовершенством дескрипторов ФКСП, применяемых для описания карбокатионов.

Второй эксперимент был проведен на массиве из 63-х ПАУ [11]. С использованием гипотез, генерированных в режиме поиска сходства на структурах исходных соединений и числовых параметрах, была правильно определена группа канцерогенности двух соединений, исключенных из обучающей выборки. Затем прогнозировалась активность 14-ти неизученных соединений. Полностью доопределены девять соединений. Прогноз группы канцерогенности этих соединений совпал с результатами, полученными в [11], на основе системы правил, предложенных экспертами.

Разработанная система может быть применена для прогноза биологической активности соединений любых структурных рядов, в частности канцерогенной активности непрямых канцерогенов из рядов диалкилнитрозаминов, ароматических аминов, замещенных и гетероциклических ПАУ. Такие компьютерные эксперименты в настоящее время проводятся.

СПИСОК ЛИТЕРАТУРЫ

1. Дьячков П. Н., Маневич С. И. Автоматизированная система прогнозирования канцерогенности полициклических углеводородов и их производных методом Хюккеля // НТИ. Сер. 2 .— 1996 .— № 7.

2. Маневич С. И., Харчевникова Н. В., Дьячков П. Н. Прогнозирование контрпродуктивных свойств химических соединений при комбинированном использовании структурных формул и численных энергетических параметров // НТИ. Сер. 2 .— 2000.— № 5.

3. Максин М. В., Харчевникова Н. В. Квантово-механический модуль системы, реализующей комбинаторно-численный подход к проблеме прогнозирования свойств химических соединений // НТИ. Сер. 2 .— 2002 .— № 6 .— С. 25–31.

4. Кузнецов С. О. ДСМ-метод как система автоматического обучения // Итоги науки и техники. Сер. Информатика. Т. 15 .— М.: ВИНИТИ.— 1991 .— С. 17–53.

5. Финн В. К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. Сер. Информатика. Т. 15 .— М.: ВИНИТИ, 1991 .— С. 54–101.

6. Забежайло М. И. Интеллектуальные системы и задача восстановления эмпирических зависимостей структурно-числового характера // Итоги науки и техники. Сер. Информатика. Т. 15 .— М.: ВИНИТИ, 1991 .— С. 54–101.

7. Jerina D. M., Lehr R. E., Yagi H. Mutagenicity of benz(a)pyrene derivatives and the description of a quantum mechanical model which predicts the case of carbonium ion formation from diol epoxides // In vivo metabolic activation and mutagenesis testing / Ed. De Serres F.— N. Y., 1976 .— P. 159–177.

8. Lehr R. E., Kumar S., Levin W., Jerina D. M. The bay region theory for polycyclic aromatic hydrocarbons — induced carcinogenesis // ACS Symposium Series.— 1985 .— Vol. 283 .— P. 63–84.

9. Дьячков П. Н. Квантово-химические расчеты в изучении механизма действия и прогнозе токсичности чужеродных соединений. // Итоги науки и техники. Сер. Токсикология. Т. 16 .— М.: ВИНИТИ, 1990 .— 280 с.

10. Путрин А. В. Дис. ... канд. техн. наук.— М., 2000.

11. Flesher J. W., Horn J., Lehner A. F. Molecular modeling of carcinogenic potential in polycyclic hydrocarbons // J. Molec Struct. (Theochem).— 1996 .— Vol. 362 .— P. 29–49.

12. Блинова В. Г., Добрынин Д. А. Язык ФКСП описания химической структуры соединения // НТИ. Сер. 2 .— 2001 .— № 6 .— С. 14–21.

Материал поступил в редакцию 05.12.03

* *
*

Работа выполнена при поддержке программы Президиума РАН "Интеллектуальные компьютерные системы" на 2001 г. проекта № 2.2 "Разработка решателей задач для интеллектуальных систем, реализующих синтез познавательных процедур".