иомедицинская химия, 2004, том 50, приложение №1, с.61-67

АНАЛИЗ ВЗАИМОСВЯЗЕЙ "СТРУКТУРА-СВОЙСТВО"

УДК [681.3.06:547.1.024.].001.57.

СИСТЕМА ПРОГНОЗА КОМПЛЕКСА СВОЙСТВ ХИМИЧЕСКИХ СОЕДИНЕНИЙ

О.В Тюрина, А.А. Тюрин, В.В. Кирлан, Л.А. Тюрина

Научно-исследовательский технологический институт гербицидов и регуляторов роста растений Академии наук Республики Башкортостан
450029, Уфа, ул. Ульяновых, 65; тел.: (3472) 529384; факс(3472) 42-83-52;
эл.почта: tjurina@anrb.ru

Разработана система прогноза комплекса свойств химических соединений, включающая два модуля: модуль прогноза "взаимонезависимых свойств", при котором последовательность прогноза свойств не зависит от результатов, полученных по другим моделям, и модуль прогноза на основе многоуровневых иерархических прогнозирующих комплексов, в которых последовательность прохождения прогнозируемого объекта по комплексу зависит от результатов, полученных в каждом из предшествующих элементов комплекса. Приведены примеры применения системы для исследования связи "структура-активность-токсичность" и конструирования структур.

Ключевые слова: комплексный прогноз и дизайн, "структура-активность-токсичность".

ВВЕДЕНИЕ. При целенаправленном поиске новых биологически активных соединений существенным моментом, наряду с прогнозом, является конструирование соединений с заданным комплексом свойств. На этой стадии особенно важно учесть токсикологические характеристики. Это позволяет исключить нежелательные варианты структур на ранних стадиях планирования синтеза, предложить эффективные и безопасные соединения, тем самым снизить затраты ресурсов на синтез и испытания.

Токсикологические характеристики (такие как острая токсичность ЛД50 и др.). имеют широкий диапазон количественных значений. На практике часто ориентируются на их интервальные значения (например, классы опасности, токсичности и пр. [1,2]). При исследовании связи "структура - свойства" ориентация на определённые интервальные значения, корректно отражающие изменение этих свойств, вполне корректна и даже предпочтительна. В этом случае реализация результатов при дизайне конкретных соединений однозначна и имеет минимум неопределённости, тогда как реализация количественных результатов в процедурах дизайна однозначно невоспроизводима, тем более, что прогноз интервальных значений соответствует множеству практических задач.

МЕТОДИКА. Компьютерные расчётные эксперименты по выявлению закономерностей, связывающих строение химических соединений и их биологические свойства (выявление и исследование характера влияния признаков, формирование математических моделей распознавания и прогноза, конструирование химических соединений с заданными свойствами) проводили с использованием компьютерной системы "SARD-21" (Structure Activity Relationship & Design) [3-5]. Исходной информацией для системы "SARD-21" служат данные о строении и свойствах исследуемых химических соединений: молекулярные структурные формулы и результаты биологических испытаний.

"SARD-21" включает аналитический блок и блок конструирования. Основное назначение аналитического блока - оценка влияния разнообразных структурных параметров химических соединений на исследуемые свойства и формирование математических моделей распознавания и прогноза. Назначение блока конструирования - молекулярный дизайн структур с заданным комплексом свойств. В процессе исследования мы обращались к обоим блокам.

Математические методы, используемые в системе SARD-21, - методы теории распознавания образов, методы теории игр, теории графов.

В системе SARD-21 осуществляются следующие основные процедуры:

1. Формирование массива обучения, включающего альтернативные по активности группы соединений. При этом соблюдены некоторые требования, позволяющие повысить информативность массива: наличие структурного сходства при разнообразии анализируемых соединений; достаточная глубина альтернативы по уровню активности противопоставляемых групп, наличие минимального числа соединений в альтернативных группах (не менее 20).

2. Представление структуры соединения на принятом фрагментарном языке описания, которое включает: а) исходные дескрипторы, б) различные взаимосочетания исходных дескрипторов с учетом их связи в структурах (сложные фрагменты). Таким образом, анализируемый массив соединений кодируется с помощью фиксированных линейных дескрипторов, заданных в специальном словаре и разнообразных циклических систем, образуемых при агрегировании их составных линейных элементов.

3. Формирование сочетаний фрагментарных дескрипторов, отражающих определённые их сочетания в исследуемых молекулах.

4. Оценка информативности всех признаков. Пределы изменения оценок от -1 до +1. Чем выше абсолютное значение информативности, тем больше вероятность влияния данного признака на проявление анализируемого свойства ("+" - положительное, "-" - отрицательное).

5. Формирование модели распознавания - решающего набора признаков (РНП). Для формирования РНП используются два основных принципа сокращения признакового пространства: принцип максимальной информативности и минимальной взаимозависимости признаков.

6. Распознавание структур. Проводится по двум методам теории распознавания образов: геометрическому подходу и методу голосования.

7. Количественная оценка относительных вкладов элементов каждой структуры в проявляемое ими целевое свойство; выявление злементов строения структур, определяюших оптимальные варианты модификации.

8. Конструирование структур с заданным свойством и распознавание их по сформированной модели.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ. Для прогноза и дизайна химических соединений с комплексом свойств разработано математическое и программное обеспечение компьютерной системы, которая включает два модуля:

1. Модуль комплексного прогноза совокупности "взаимонезависимых свойств". Он ориентирован на параллельное независимое прогнозирование по моделям, специально полученным для каждого из рассматриваемых свойств на основе системы анализа связи "структура - свойства" SARD-21, или взятым из банка моделей. Последовательность прогноза и конечный результат не зависят от результатов, полученных по другим моделям.

2. Модуль прогноза на основе многоуровневого иерархического прогнозирующего комплекса, в котором последовательность прохождения прогнозируемого объекта по комплексу зависит от результатов полученных в каждом из предшествующих элементов комплекса. Она определяется схемой прогнозирующего комплекса, и задается логическим решающим правилом.

Наличие анализируемого свойства может быть прогнозируемо для достаточно большого числа объектов с одинаковой прогностической оценкой. Хотя различие оценок чрезвычайно важно для сокращения множества новых объектов, прогнозируемых по определённому свойству и выбора из них потенциально перспективных.

Для дифференциации прогностических оценок, а, следовательно, повышения достоверности прогноза мы ввели критерии, характеризующие качество моделей, полученных на объектах обучения - **К3** и **К4**.

К3= N1/Nmax - учитывает соотношение числа признаков решающего набора (РНП) в прогнозируемых структурах ***(N1)*** и максимальное число признаков РНП, наблюдаемое на структурах обучения ***(Nmax)***. ***К4=1-R1/Rmax*** - отражает сходство с расчетным

эталоном по расстояниям до разделяющей гиперплоскости ***R1*** для прогнозируемой структуры и ***Rmax*** для структур обучения.

Данные прогноза ранжируются по каждому из четырёх критериев: ***R1*** (число голосов по алгоритму "голосование"), ***R2*** (ранг по геометрическому подходу), ***K3***, ***K4*** - новые критерии, характеризующие качество моделей на обучении, а также по суммарному рангу.

При разработке модуля прогноза "взаимозависимых" прогнозируемых свойств предложен метод на основе формируемых иерархических прогнозных комплексов.

При формировании моделей методами ТРО обычно используется дихотомическая процедура (разбиение на две альтернативные по свойствам группы). Однако, диапазон измерения некоторых свойств (например, острой токсичности) является достаточно большим и при использовании данной процедуры образуются широкие интервалы значений. Следовательно, обычная дихотомическая процедура формирования модели в подобных случаях не может быть применена.

Нами предложен метод прогноза интервальных значений на основе формируемых иерархических прогнозных комплексов. Его сущность заключается в последовательном ступенчатом сужении прогнозируемых интервалов в рамках комплекса иерархических моделей [5,6]. Прохождение структурной информации по этому комплексу определяется задаваемой логической схемой, по ходу которой обозначаются более узкие интервалы. Границы интервалов устанавливаются автоматически, путем оптимизации распознавания свойств соединений на стадии формирования моделей, а также алгоритмов прогноза.

Отбор соединений в альтернативные группы из общего банка данных производится автоматически по значениям оценок их свойств, согласно начальным границам интервалов.

Для создания прогнозирующего иерархического комплекса автоматически формируется банк всех возможных моделей (M1…Mi), отвечающих заданным интервалам пороговых критериев, используемых при формировании РНП для отбора "признаков-претендентов", и заданной нижней границей распознавания соединений в каждой из альтернативных групп (>70%).

Далее отбираются оптимальные по числу признаков и уровню распознавания *РНП*, общие или индивидуальные для двух алгоритмов (геометрического подхода и метода голосования). Эти РНП представляют собой рабочие модели и являются элементами прогнозирующего комплекса (*ЭПК*).

Структура прогнозирующего комплекса определяется логической схемой, которой задаются пути прохождения прогнозируемого объекта в зависимости от того, к какой из альтернативных групп отнесен исследуемый объект.

Алгоритмическая запись приведенного ниже фрагмента комплекса по маршруту отнесения объекта к альтернативным группам (в соответствии с границей разделения) по маршрутам A, B, AA, AB выглядит следующим образом: {/M1,M2,M3}, {A/M1,M2,M8,M9,M11}, {B/M7,M8,M9},{AA/M28,M29,M30}, {AB/M7,M8,M9}.

Поскольку каждая совокупность моделей (M1-M3; M28-M30 и т.д.) отражает определённый интервал со своими границами, а общее количество моделей может быть достаточно большим, то их целесообразно сгруппировать в элемент прогнозирующего комплекса (*ЭПК*). В этом случае алгоритмическая запись становится более компактной, а использование схемы более наглядным (рис.1).

Для корректного распознавания соединений в каждом *ЭПК* предусмотрена процедура голосования решений, принимаемых по каждой модели *ЭПК*. Выбор решения на каждом шаге производится процедурой голосования решений моделей (нечётное количество) *ЭПК* с определёнными характеристиками: алгоритм распознавания ("геометрический подход" или "голосование"), качественный и количественный состав *РНП*, уровень распознавания (% правильного распознавания соединений обучения в альтернативных группах).

В соответствии с предложенным методом для прогноза острой токсичности $ЛД_{50}$ арил-гетерил производных оксикарбоновых кислот были разработаны 108 рабочих моделей. Исследовано более 500 химических соединений. При отборе химических структур использованы справочные и литературные данные по токсичности соединений [7-11] и экспериментальные данные, полученные в НИТИГ АН РБ и БГМУ (кафедра гигиены и санитарии).

Моделируемые интервалы охватывают значения токсичности от 0,1 до 28000 мг/кг. Достоверность (уровень распознавания) созданных моделей составляет 72-97%.

На основе 108 моделей организован 21 ЭПК со следующими интервалами и границами разделения альтернативных групп (нижняя-средняя-верхняя, мг/кг): 0,1-(150,

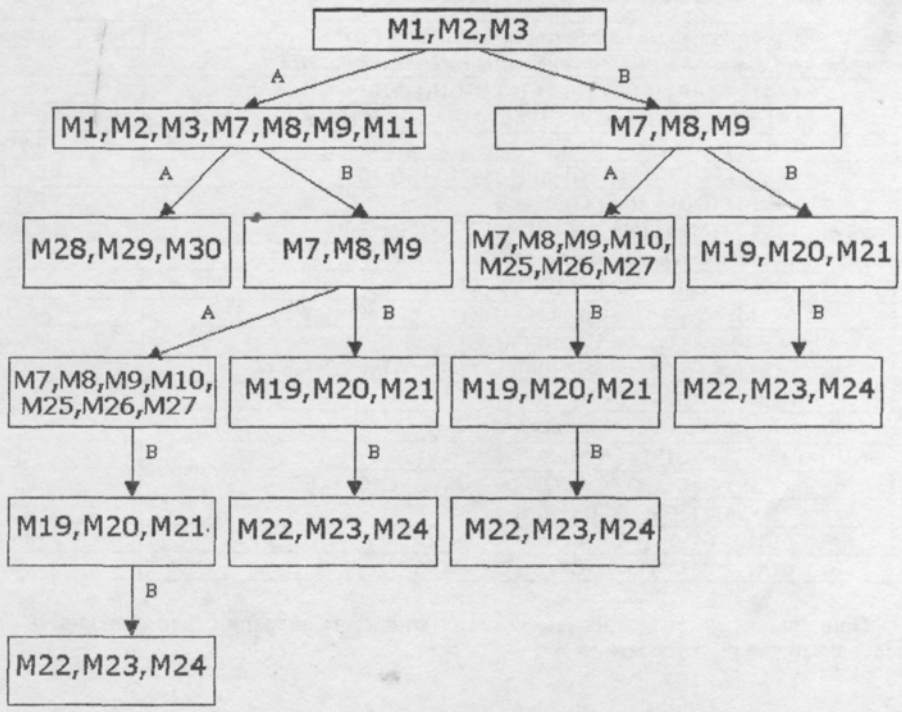

Рисунок. 1.
Логическая схема организации моделей для иерархического прогноза. интервалов токсичности

500, 800, 1000, 1500, 1800, 2000, 2300, 2500, 3000, 3500, 4500, 5000)-28000; 0,1-150-5000; 0,1-150 / 5000-28000; 150-(500,1000,2000,5000)-5000; 500-1000-5000; 1000-2000-5000. Число соединений в альтернативных группах (A/B) приведены соответственно для каждого ЭПК: 1 - 27/173; 2 - 58/142; 3 -75/125; 4 - 90/110; 5 -103/97 ; 6 -112/88; 7 - 116/84; 8 - 124/76; 9 - 130/70; 10 - 139/61; 11 -146/54 ; 12 -151/49; 13 -161/39 ; 14 - 74/39; 15 - 27/134; 16 - 27/39; 17 - 135/39; 18 - 32/103; 19 - 34/71; 20 - 29/45; 21 - 25/22. На их основе этих ЭПК создан прогнозирующий комплекс (рис.2).

Апробация комплекса на 30 экзаменационных соединениях (среди них эфиры фенокси-γ-масляной, уксусной, пропионовой кислот, синтезированных в НИТИГ АН РБ, а также амидосодержащих производных феноксиуксусной и β-(оксисульфинил)изомасляной кислот) показала уровень распознавания более 70 %, который с позиций теории распознавания образов рассматривается как удовлетворительный [3]. Высокотоксичные соединения распознаются полностью правильно. Для небольшого числа соединений произошло завышение расчётных интервалов токсичности по отношению к опытным данным. Подобная ошибка классифицируется как "ложная тревога", она может быть рассмотрена как более безопасная при прогнозе токсичности.

РНП моделей для следующих ***ЭПК***: 1 (границы 0,1-150-28000), 15 (0,1-150-5000) и 13 (0,1-5000-28000), - содержат логические сочетания фрагментов, участвующие в распознавании соответственно высоко-, умеренно- и малотоксичных соединений (табл.1).

По мере прохождения по комплексу 1 (рис.2) прогнозируются следующие интервалы токсичности (мг/кг): 0.1-150, 150-500, 500- 800, 800-1000, 1000-1500, 1500-2500, 2500-3500, 3500-4500, 3500-5000, более 5000. Апробация комплекса на экзаменационных соединениях показала удовлетворительный уровень распознавания - более 70 %.

Аналогично разработан комплекс для широкого круга гетероциклических соединений. Комплексы применены для прогноза токсичности более 150 структур потенциально активных соединений. Выявлены структуры малодозных пестицидов и антигельминтиков умеренно- (в интервале 1000-5000 мг/кг) и малотоксичных.

Модели, не отобранные при экзамене, занесены в компьютерный банк и могут быть применены для формирования других прогностических комплексов.

По результатам исследования связи "структура - активность - токсичность" получены оценки влияния признаков на гербицидную активность и токсичность (рис.3),

Таблица. Некоторые элементы РНП основных моделей

№	Субструктурные дескрипторы, входящие в ***РНП***	Информативность
Распознают чрезвычайно- и высокотоксичные соединения на ЭПК 1		
1	(-CH_3)-(-NH-) !* (-CH_2-)-(>CH-) ! (-CH_3)-(>N-)	0,615
2	(-O-1,2,5-зам. тиазолил)!(-CH_2het-)**-(>C<)!(>CH-CH_2-)	0,563
3	(>C=O) - (-O-) - (>C=C<)	0,511
4	(-CH_2-) - (>CH-) ! (-CH_3) - (>N-) ! (>C<) - (-O-)	0,478
5	(-CH_3) - (-NH-) - (>C=O)	0,443
6	(-CH_3) - (>N-) ! (>C<) - (-O-) ! (>CH-) - (>C=C<)	0,405
Распознают умереннотоксичные соединения на ЭПК 15		
7	(>C=C<)-(Cl) ! (>C=C<)-(CF_3) ! (-CH_3)-(-O-)	-0,491
8	(-CH_2het-)-(-O-) ! (>CH-)-(>C=O) ! (>C=O)-(>C=C<)	-0,462
9	(>CH-)-(>C=O) ! (-CH_2het-)-(>C=O) ! (-NH-)-(>C=C<)	-0,460
10	(-CH_2het-)-(>C=O) ! (>CH-)-(-O-) ! (>C=O)-(>C=C<)	-0,398
11	(-OH) ! (-N=C<) ! (-CH_2het-)	-0,348
Распознают малотоксичные соединения на ЭПК 13		
12	(-NH-) - (-N=C<) ! (-NH-) - (>C=C<) ! (>C=C<) - (F)	-0,295
13	(-NH-) - (>C=C<) ! (>C=O) - (>C=C<) ! (>C=C<) - (F)	-0,245
14	(-N=C<) - (>C=C<) - (>C=C<)	-0,192
15	(-CH_3) - (-O-) - (>C=O)	-0,147
16	(-O-) - (-N=C<) - (>C=C<)	-0,120

Примечание: *- знак логической математической операции "дизъюнкция" ("или"); **- метиленовая група при гетероатоме.

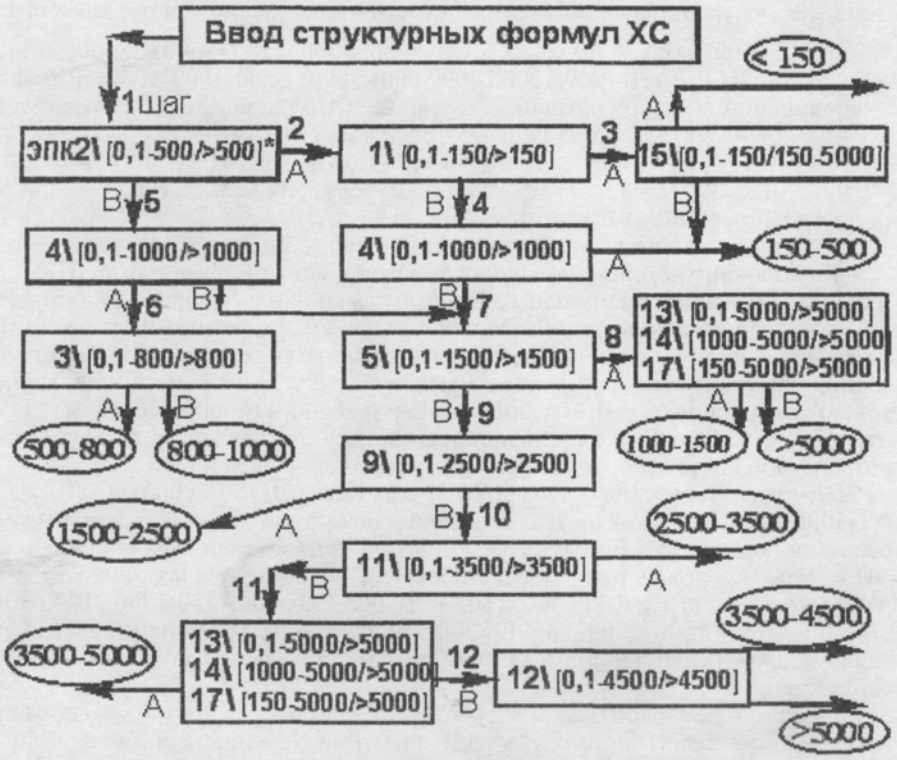

Рисунок. 2.
Комплекс 1 прогноза интервалов токсичности (арил)гетерилпроизводных оксикарбоновых кислот
*- в квадратных скобках указаны границы разделения альтернативных групп по ЛД50

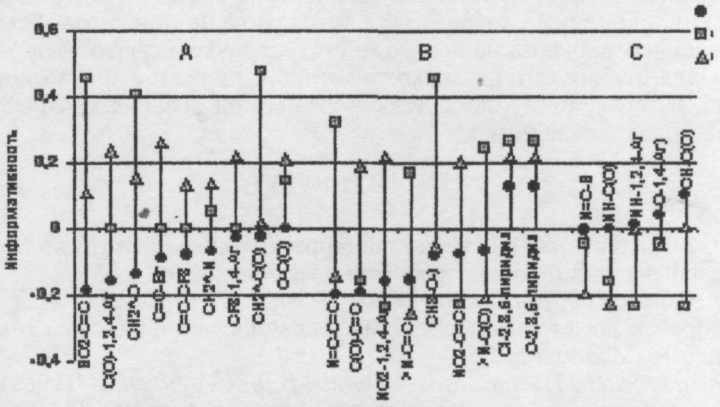

Рисунок 3.
Влияние фрагментов на гербицидную активность по нескольким моделям (М-6, М-9) и токсичность (Т).

проведено конструирование потенциальных гербицидов с учётом токсичности. Очерёдность замены фрагментов базовой структуры рассчитана на основе критериев теории игр (рис.4).

Для проведения оптимального конструирования с учётом токсичности использована модель с интервалами прогноза 150-5000 и 5000-28000 мг/кг. По влиянию на токсичность первыми к замене определены "токсофорные" фрагменты фенильной части структуры: хлор, оксигруппа и этиленовые группы при атомах хлора. На последних местах замены стоят "антитоксофорные" фрагменты триазольного цикла: -N=C< и >N-. В рамках гербицидной активности первыми к замене стоят циклические фрагменты, а центром активности определена метильная группа мостика. В результате модификации базовой структуры А (рис.4) сконструированы 24 структуры, переданные для синтеза.

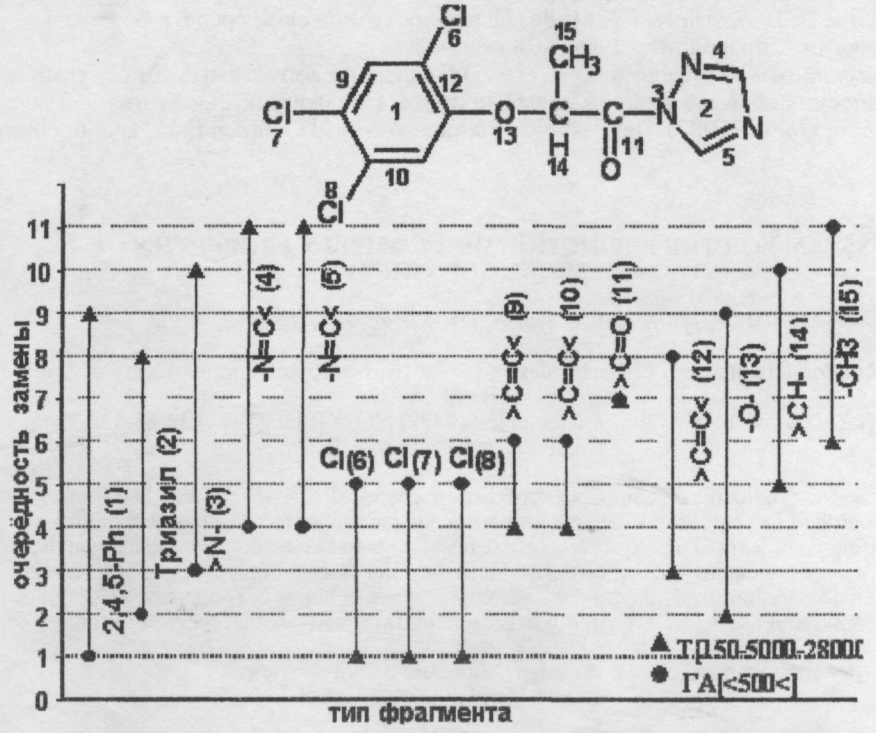

Рисунок 4.
Очерёдность замены фрагментов базовой структуры А по гербицидной активности (ГА) и токсичности (Т).

ЗАКЛЮЧЕНИЕ И ВЫВОДЫ. Разработана система прогноза комплекса свойств химических соединений, которая включает модуль прогноза "взаимонезависимых свойств", и модуль прогноза на основе многоуровневых иерархических прогнозирующих комплексов для оценки интервальных значений с широким диапазоном оценок, а также базу знаний, которая может быть использована при дизайне химических соединений с заданным комплексом свойств.

ЛИТЕРАТУРА.

1. ГОСТ 12.1.007-76. "Система стандартов безопасности труда. Вредные вещества. Классификация и общие требования безопасности."
2. *Заугольников С. Д., Кочанов М. М., Лойт А. О., Ставчанский И. И.* (1978) Экспрессные методы определения токсичности и опасности химических веществ, Медицина, Л.
3. *Кадыров Ч.Ш., Тюрина Л.А., Симонов В.Д., Семенов В.А.* (1989) Машинный поиск химических препаратов с заданными свойствами, ФАН, Ташкент
4. *Тюрина Л.А., Кадыров Ч.Ш., Симонов В.Д.* (1989) Машинный поиск закономерностей строение - биологическое действие химических соединений, Итоги науки и техн. Сер. Органическая химия, ВИНИТИ, М.
5. *Кирлан А.В., Тюрина О.В., Кирлан В.В., Кирлан С.А., Тюрина Л.А., Лукманова А.Л.* (2000) в кн. Труды межд. н-т.конф. Современные информационные технологии (Contemporary Information Technologies), Информационные технологии в научном эксперименте, Пенза, Россия.
6. *Кирлан В.В.* (2003) Прогноз и молекулярный дизайн гетероорганических соединений с комплексом заданных свойств (разработка методов, программная и практическая реализация), Автореф. дисс. канд. наук, Башкирский государственный университет, Уфа.
7. *Мельников Н. Н., Козлов В.А.* (1996) Перспективы создания и производства отечественных гербицидов. Агрохимия., №6, 74-80.
8. *Мельников Н. Н.* (1993) Современные направления создания новых пестицидов. Защита растений., № 10, С. 80-118.
9. *Симонов В.Д., Базунова Г.Г.* (1988) Перечень химических средств борьбы с сорняками. Справочник.- Уфа., 138 с.
10. *Беспамятнов Г.П., Кротов Ю.А.* (1985) Предельно допустимые концентрации химических веществ в окружающей природе. Справочник. Л.- Химия., 528 с.
11. *Worthing Ch. R.* (1987) The Pesticide Manual. A World Compendium. Eighth edition.

SYSTEM FOR THE PREDICTION OF A COMPLEX PROPERTIES OF CHEMICAL COMPOUNDS

O.V. Tjurina, A.A. Tjurin, V.V. Kirlan, L.A. Tjurina

Research Institute of Technology of Herbicides and Plant Growth Regulators, Academy of Sciences, Republic of Bashkortostan
Ufa, Uljanovich, 65, 450029, Russia; tel: (3472) 529384; fax (3472) 42-83-52;
e-mail: tjurina@anrb.ru

A system for prediction of complex properties of chemical compounds has been developed. It includes the module of forecast of independent properties and module of forecast based on multilevel hierarchical predicting complexes. The sequence of passages of object via the complex depends on the results obtained at each previous element of this complex. More than 100 models of the forecast of intervals LD50 have been obtained. On their basis the predicting complexes are generated. The examples of application of system for analysis of "structure-activity-toxicity" relationships and design of structures are given.

Keywords: complex prediction and design, «structure-activity-toxicity»