На правах рукописи

Рудик Анастасия Владимировна

КОМПЬЮТЕРНЫЙ ПРОГНОЗ БИОТРАНСФОРМАЦИИ КСЕНОБИОТИКОВ

03.00.28 - биоинформатика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата биологических наук

Москва - 2007

Работа выполнена в Государственном учреждении Научно-исследовательский институт биомедицинской химии им. В.Н. Ореховича Российской академии медицинских наук.

Научный руководитель:
кандидат физико-математических наук Филимонов Дмитрий Алексеевич

Официальные оппоненты:
доктор физико-математических наук, профессор Шайтан Константин Вольдемарович

доктор биологических наук, профессор Иванов Алексей Сергеевич

Ведущая организация: Институт физиологически активных веществ РАН

Защита состоится **28 июня 2007 года** в **11:00** часов на заседании Диссертационного совета Д 001.010.01 при ГУ НИИ Биомедицинской химии им. В.Н. Ореховича РАМН по адресу: 119121, Москва, ул. Погодинская, д. 10.

С диссертацией можно ознакомиться в библиотеке ГУ НИИ Биомедицинской химии им. В.Н. Ореховича РАМН.

Автореферат разослан « » мая 2007 года

Ученый секретарь Диссертационного совета,
кандидат химических наук Карпова Е.А.

1. ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

1.1. Актуальность проблемы.

Интерес к биотрансформации органических ксенобиотиков, среди которых лекарства занимают особое место по важности для человека и по контролю за их безопасностью, вызван многими причинами. Образующиеся в результате реакции биотрансформации метаболиты могут быть более токсичными, чем исходные соединения, или проявлять нежелательные побочные эффекты, а могут быть и более фармакологически активными.

Исследование процесса биотрансформации включает анализ образующихся метаболитов, их стабильности, и определение ферментов, осуществляющих реакции биотрансформации, что важно для изучения зависимости от локализации в органах и тканях действия лекарств и исследования совместимости лекарственных средств.

Существующие экспериментальные методы *in vitro* и *in vivo* – сложный, дорогостоящий, трудоемкий и длительный процесс, возможности проведения экспериментов в клинике весьма ограничены, а результаты изучения биотрансформации *in vitro* и *in vivo* не всегда совпадают с таковыми для человека.

Компьютерные методы позволяют прогнозировать биотрансформацию и метаболиты органических соединений на основе химической структуры их молекул еще до проведения химического синтеза и биологического тестирования, что, благодаря сокращению количества соединений, которые будут подвергнуты биологическому тестированию, позволяет сократить время и затраты на разработку новых лекарств.

В настоящее время известно несколько компьютерных методов прогноза возможных путей биотрансформации органических ксенобиотиков. Это, прежде всего, методы молекулярного моделирования. Они ориентированы на изучение механизмов взаимодействия определенных ферментов с субстратами. Основным их недостатком является продолжительность расчетов и необходимость использования мощных вычислительных ресурсов, что не приемлемо для высокопроизводительного скрининга. Помимо этого, необходимы данные о трехмерной структуре, как субстрата, так и фермента, информация о которых во многих случаях не доступна.

Экспертные системы для предсказания метаболизма основаны на формализации знаний экспертов и предсказывают возможные метаболические сети. Их недостатком является генерирование большого количества метаболитов, отсев которых также основан на экспертных знаниях, и, как правило, они не предсказывают ферменты, которые могут осуществлять реакции биотрансформации. [*Langowski J, 2002*].

1.2. Цель и задачи исследования.

Цель диссертационной работы – разработка нового метода компьютерного прогноза биотрансформации органических ксенобиотиков по структурным формулам.

Метод должен быть высокопроизводительным, предсказывать реакции биотрансформации и осуществляющие их ферменты, предсказывать возможные метаболиты и оценивать точность предсказания.

В ходе выполнения работы решались следующие задачи:

1. Разработать метод генерации продуктов биотрансформации.
2. Разработать метод прогноза типов биотрансформации.
3. Разработать метод прогноза сайтов биотрансформации и метаболитов.
4. Исследовать устойчивость прогноза при модификациях обучающей выборки.

1.3. Научная новизна.

В диссертационной работе предложен новый подход к прогнозу биотрансформации органических ксенобиотиков. Впервые разработан метод, позволяющий одновременно предсказывать вероятность реакций биотрансформации которым может подвергнуться субстрат, с учетом ферментной системы и конкретной изоформы фермента, ответственных за метаболизм. Для решения задачи прогноза метаболитов разработаны оригинальные дескрипторы, отражающие структурные изменения субстрата в ходе реакции биотрансформации.

1.4. Практическое значение работы.

Разработанный метод прогноза биотрансформации может быть использован при поиске и создании новых лекарственных субстанций, для прогноза биодеградации веществ в окружающей среде и идентификации структуры исходных веществ по их метаболитам.

1.5. Апробация работы.

Результаты работы были представлены на X, XI, XII, XIII и XIV Российских национальных конгрессах “Человек и лекарство” (Москва, 2003 - 2007 г.г.), 3-ей Всероссийской конференции «Молекулярное моделирование» (Москва, 2002 г.), на XII международной конференции «Новые информационные технологии в медицине, биологии, фармакологии и экологии» (Украина, Гурзуф, 2004 г.), на 4-ом Международном симпозиуме по фармацевтической химии в г. Стамбуле (Турция, 2003 г.), на 15-ом европейском симпозиуме, посвященном QSAR и молекулярному моделированию в г. Стамбуле (Турция, 2004 г.), на 2-м Международном симпозиуме «Компьютерные методы в токсикологии и фармакологии» (Греция, 2003 г.).

1.6. Публикации.

Результаты диссертации опубликованы в 15 печатных работах, включая 2 статьи в рецензируемых журналах, 11 публикаций в сборниках трудов научных конференций и один патент на программу.

1.7. Объем и структура диссертации.

Основное содержание работы изложено на 95 страницах, содержит 42 рисунка и 21 таблицу. Диссертация состоит из введения и шести глав, включая литературный обзор и 3 приложений. Список цитированной литературы содержит 116 наименований, в том числе 92 зарубежных публикации.

2. МАТЕРИАЛЫ И МЕТОДЫ

В работе были использованы две коммерчески доступные **базы данных** (БД) – **Metabolite** (MDL, Elsevier) и **Metabolism** (Accelrys). В первой, в основном, содержится информация о биотрансформациях лекарственных соединений, для некоторых реакций есть также данные об осуществляющих реакции ферментах. Во второй содержатся данные о биотрансформации лекарственных соединений, агрохимических веществ, пищевых добавок и пестицидов. В базе данных Metabolite V 2001.1 содержатся 16922 соединений и более 55000 реакций; в базе данных Metabolism V 2002.1 содержится более 25000 реакций и 4254 соединений, 1723 из которых встречаются в БД Metabolite.

2.1. Метод прогноза типов биотрансформации

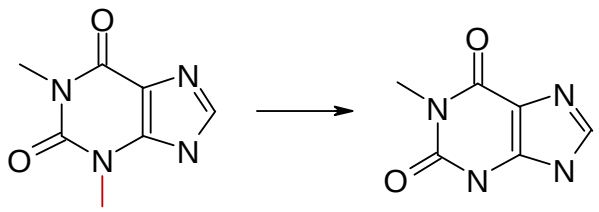
На основе информации из БД формируются обучающие выборки, состоящие из структур соединений и связанных с ними названий типов биотрансформации.

Тип биотрансформации – это, в первую очередь, название реакции биотрансформации. При наличии соответствующей информации тип биотрансформации может содержать название семейства ферментов, участвующих в биотрансформации, и данные об изоформе фермента:

Название реакции биотрансформации [(Фермент [,Изофермент])]

(данные, заключенные в квадратные скобки, не являются обязательными)

На рис. 1 приведен пример записи типов биотрансформации для реакции окислительного N-деалкилирования молекулы теофиллина цитохромом P450 (изоформа CYP1A2).



Oxidative N-Dealkylation
Oxidative N-Dealkylation (Cytochrome P450)
Oxidative N-Dealkylation (Cytochrome P450, CYP1A2)

Рис. 1. Окислительное N-деалкилирование молекулы теофилина цитохромом P450 CYP1A2.

При ***прогнозе типов биотрансформаций*** структуры органических соединений описываются ***множеством дескрипторов многоуровневых атомных окрестностей MNA*** (Multilevel Neighborhood of Atoms) (Filimonov et al, 1999*),* а ***алгоритм прогноза*** *(*в диссертационной работе проводилось исследование применимости алгоритма, реализованного в программе PASS (Д.А. Филимонов, В.В. Поройков, РХЖ, 2006, L, N2, 66-75) состоит в следующем**.**

В ходе обучения для каждого типа биотрансформации субстраты соответствующих реакций из обучающей выборки являются положительными примерами, а все остальные отрицательными примерами субстратов этого типа биотрансформации. Для каждого дескриптора D_i оценивается условная вероятность того, что органическое соединение, содержащее D_i в множестве описывающих его структуру дескрипторов, является субстратом реакции биотрансформации типа B_j :

$$p(B_j|D_i) = \sum_k f_k(B_j) g_k(D_i) / \sum_k g_k(D_i),$$

где $f_k(B_j)$ и $g_k(D_i)$ – характеристические функции принадлежности субстрата с номером k множеству положительных примеров и множеству соединений, содержащих в описании структуры дескриптор D_i, соответственно.

Априорная вероятность биотрансформации типа B_j оценивается так:

$$p(B_j) = \sum_k f_k(B_j) \sum_i g_k(D_i) / \sum_k \sum_i g_k(D_i),$$

По множеству дескрипторов MNA, описывающих структуру органического ксенобиотика, для каждого типа биотрансформации вычисляется оценка ее возможности t:

$$t = (1+(s-s_0)/(1-s*s_0))/2,$$
$$s = Sin(\sum_i ArcSin(2*p(B_j/d_i)-1)/m), \quad s_0 = 2*p(B_j)-1,$$

где суммирование ведется только по тем m дескрипторам, которые описывают данную структуру и имеются в структурах субстратов обучающей выборки. При

t=0 соединение не может быть субстратом биотрансформации данного типа, а при t=1 практически достоверно им является.

При обучении рассчитываются величины t и функции их распределения для положительных (t_t) и отрицательных (t_f) примеров.

При прогнозе сравнение вычисленного значения t с распределениями величин t_t и t_f дает оценки вероятностей быть (P_t) и не быть (P_f) данному органическому соединению субстратом реакции биотрансформации соответствующего типа.

Для ***оценки точности прогноза*** используется процедура скользящего контроля с исключением по одному. На каждом шаге процедуры одно соединение исключается из обучающей выборки, и для него выполняется прогноз. Точность прогноза каждого типа биотрансформации оценивается по инвариантному критерию точности прогноза:

$$IAP = [N(t_+ \geq t_-)/N_+N_-]*100\%, \quad (1)$$

где $N(t_+ \geq t_-)$ – число случаев, когда значение t, рассчитанное для положительного пример, выше чем t, рассчитанное для отрицательного примера, N_+ и N_- - число положительных и отрицательных примеров, соответственно.

В результате обучения создается SAR base, содержащая множества дескрипторов MNA субстратов, их типы биотрансформаций, распределения величин t_t и t_f и IAP для каждого типа биотрансформации.

2.2. Метод прогноза сайтов биотрансформации

Задача ***прогноза сайтов биотрансформации*** ставилась как задача прогноза сайтов определенной реакции биотрансформации, т.е., например, прогноз сайтов Ароматического гидроксилирования. В этом случае задача прогноза сайта определенной реакции аналогична задаче нахождения метаболита этой реакции.

Для того чтобы используемые в прогнозе дескрипторы содержали информацию о произошедших во время реакции структурных изменениях, на основе дескрипторов MNA были разработаны ***дескрипторы RMNA*** (Reacting Multilevel Neighborhood of Atom).

Главным отличием RMNA является учет сразу двух структур - структуры субстрата и продукта и дополнительные метки для атома:

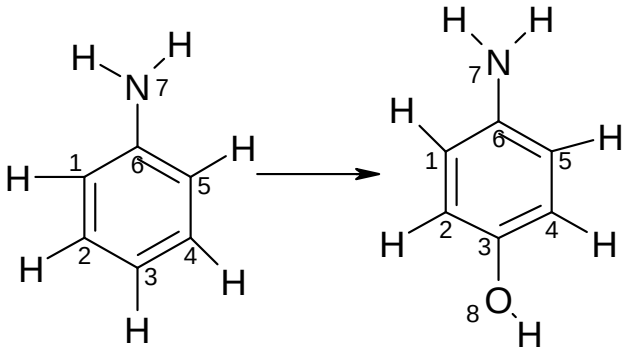
$$RNMA_1(A) = ([-]A[T]A_1[B]A_2[B]..A_i[B]\ldots),$$

где A - тип атома, *[-]*метка, обозначающая, что атом не входит в цикл, *[T]*метка для изменяющегося атома (может принимать значение «<» для присоединяемого и «>» для удаляемого атомов, соответственно), $A_1,..,A_i$…- типы ближайших соседей атома в лексикографическом порядке, *[B]* – метка для связи, изменяющейся во время биотрансформации (может принимать значение "/" в

случае, если связь между A и одним из соседних атомов разрушается, "\" - если связь образуется, "]" если изменяется тип связи в субстрате и "[" если изменяется тип связи в продукте).

RMNA дескрипторы последующих уровней строятся как конкатенация дескрипторов предыдущего уровня атома и его ближайших соседей.

Пример генерации дескрипторов разного уровня (до второго включительно) приведен на рис. 2.



Атом	RMNA0	RMNA1	RMNA2
1	C	**C-HCC**	**(C-HCC) (C-HCC)(C-NCC)(-HC)**
2	C	**C-HCC**	**(C-HCC) (C-HCC) (-HC)(C-O<CC\)**
3	C	**C-O<CC**	**(C-O<CC\ ) (C-HCC) (C-HCC) (-O<-H<C\\)**
4	C	**C-HCC**	**(C-HCC) (C-HCC) (-HC)(C-O<CC\)**
5	C	**C-HCC**	**(C-HCC) (C-HCC)(C-NCC)(-HC)**
6	C	**C-NCC**	**(C-NCC) (C-HCC) (C-HCC)(-N-H-HC)**
7	N	**-N-H-HC**	**(-N-H-HC)(C-NCC)(-H-N) (-H-N)**
8	O	**-O<-H<C**	**(-O<-H<C\\)(C-O<CC\)(-H<-O<\)**

Рис. 2. Пример RMNA дескрипторов для типа биотрансформации «Ароматическое гидроксилирование».

Алгоритм прогноза сайтов биотрансформации соответствует вышеописанному алгоритму прогноза типов биотрансформации.

2.3. Метод генерации продуктов биотрансформации

Генерация продуктов (потенциальных метаболитов) **биотрансформации** осуществляется с помощью созданных фрагментов реакции, которые представляют собой пару фрагментов структур – «фрагмент структуры субстрата» и «фрагмент структуры продукта». Во фрагменты структур включены только атомы и связи, изменяющиеся во время реакции, а также их ближайшие соседи. Чтобы отличать ароматические молекулы, добавлен тип связи «ароматическая».

Словарь фрагментов реакций был создан в ***полуавтоматическом режиме:*** вначале из БД Metabolite были автоматически выделены фрагменты

всех реакций, после чего отобраны фрагменты, соответствующие только одной реакции, и не отражающие многостадийные процессы.

При создании словаря было поставлено условие, что одному фрагменту реакции должно соответствовать одно название реакции биотрансформации; в то же время одному названию реакции биотрансформации могут соответствовать несколько фрагментов реакций.

Например, реакции биотрансформации «Алифатическое гидроксилирование» соответствуют шесть фрагментов (рис. 3):

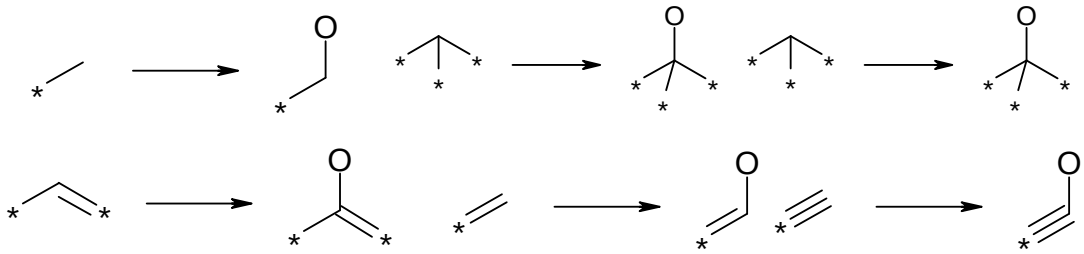


Рис 3. Пример фрагментов реакции «Алифатическое гидроксилирование».

Чтобы избежать повторов и сделать словарь компактней, было принято решение не описывать все реакции биотрансформации, а создать иерархическую систему классификации и заносить в словарь фрагментов только самые нижние уровни.

Например, реакция биотрансформации «С-гидроксилирование» включает в себя «Ароматическое гидроксилирование» и «Алифатическое гидроксилирование». В словарь фрагментов заносятся только фрагменты реакции, относящихся к нижним уровням. При генерации продуктов реакций используются фрагменты всех нижестоящих реакций. Например, при генерации продуктов «С-гидроксилирования» используются фрагменты реакций «Ароматическое гидроксилирование» и «Алифатическое гидроксилирование.

В результате работы, проделанной совместно с к.х.н. В.Г. Блиновой, и аспирантом А.В. Дмитриевым, была построена система классификации, основанная на следующих принципах:

1. Реакция биотрансформации относится к одной из 4-ех основных групп реакций: окисление, восстановление, конъюгация и гидролиз.

2. Внутри иерархии введены отношения «частное-общее», верхний уровень включает в себя все реакции нижних уровней.

Ниже, в качестве примера, приведен фрагмент классификации биотрансформации:

A. Hydroxylation
    1) C-Hydroxylation
        a. Aromatic Hydroxylation
        b. Aliphatic Hydroxylation
    2) N-Hydroxylation
    3) S-Hydroxylation
    4) P-Hydroxylation

B. Dealkylation

…

Созданный словарь фрагментов реакций используется для генерации продуктов биотрансформации. Для этого производится поиск «фрагмента структуры субстрата» в структуре субстрата и если он найден, то производится его замена на соответствующий «фрагмент структуры продукта», в результате чего генерируется структура продукта. Задача поиска фрагмента структуры в целой структуре является задачей нахождения изоморфного вложения подграфа в граф.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

3.1. Прогноз типов биотрансформации

Для прогноза типов биотрансформации было создано несколько обучающих выборок. Две выборки (одна на основе БД Metabolite, другая - Metabolism) были созданы без изменения оригинальных названий, т.е. были использованы названия реакций биотрансформации, как они указанны в базах данных. В двух других выборках использовались автоматически формализованные названия. Для формализации названий использовался созданный ***словарь фрагментов*** - если фрагмент реакции (состоящий из двух частей - фрагмента структуры до и после реакции) найден в искомой реакции, то ей приписывается соответствующее название реакции биотрансформации. Помимо этого реакции присваиваются названия реакций биотрансформации вышестоящих уровней из созданной классификации. Например, если искомой реакции соответствует название реакции биотрансформации «Aromatic Hydroxylation», то к реакции прибавляются еще три названия – «Aromatic Hydroxylation», «C-Hydroxylation» и «Hydroxylation».

Ниже представлены характеристики созданных выборок и рассчитанная точность прогноза при выборе четырех различных уровней дескрипторов.

Как видно из Таблицы 1, оптимальным уровнем дескрипторов для прогноза типов биотрансформаций является второй (MNA2); точность прогноза в выборке, использующей формализованные названия из БД Metabolism выше, чем в выборке, использующей оригинальные названия из БД Metabolism, а точность прогноза для выборок из БД Metabolite соотношение обратное.

Таблица 1.

Зависимость точности прогноза типов биотрансформации от уровня дескрипторов MNA.

Наименование выборки и её характеристика	Кол-во соединений	Кол-во прогнозир. типов биотрансф.	Точность прогноза по скользящему контролю (IAP-LOO)			
			MNA1	MNA2	MNA3	MNA4

Выборка I Источник - БД Metabolite (MDL&Elsevier). Оригинальные названия	14577	2317	**82.7%**	**85.0%**	**83.1%**	**81.2%**
Выборка II Источник - БД Metabolism (Accelrys). Оригинальные названия	3581	170	**81.1%**	**84.6%**	**82.8%**	**79.2%**
Выборка III Источник - БД Metabolite (MDL&Elsevier). Формализованные названия	7310	830	**79.9%**	**83.2%**	**80.8%**	**79.7%**
Выборка IV Источник - БД Metabolism (Accelrys). Формализованные названия	1695	24	**87.3%**	**89.7%**	**86.6%**	**83.8%**

Сравнение точности прогноза по выборкам, использующим формализованные названия и использующим оригинальные *названия только по пересекающимся* названиям реакции биотрансформации представлены в Таблице 2.

Практически для всех сравненных типов биотрансформаций точность прогноза при автоматическом способе формализации названий возросла, причем средняя точность прогноза увеличилась почти на 2% (а для БД Metabolism - на 3%).

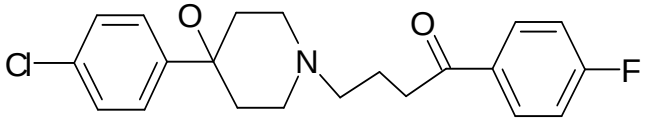
Недостатком автоматической формализации является его зависимость от словаря фрагментов реакций, что не дает возможности выполнить прогноз по некоторым присутствующим в БД реакциям биотрансформации (например, реакция «Covalent Binding» не будет описана в словаре трансформаций, а будут описаны только «Protein Binding» и «DNA Binding»).

Таблица 2. Результаты прогноза для двух выборок. В первой используются оригинальные названия из БД Metabolite, во второй используются автоматически формализованные названия

	Оригинальные названия		**Формализованные названия**	
Тип биотрансформации	**Кол-во соед.**	**IAP-LOO*,%**	**Кол-во соед.**	**IAP-LOO*, %**
Aliphatic Hydroxylation	1464	**73.3**	1977	**81.8**

Aliphatic Hydroxylation(Cytochrome P450)	427	**75.1**	516	**81.2**
Aliphatic Hydroxylation(Cytochrome	51	**80.2**	56	**81.1**
Aliphatic Hydroxylation(Cytochrome	133	**75.5**	143	**79.2**
Aromatic Hydroxylation	1564	**79.2**	1485	**81.6**
Aromatic Hydroxylation(Cytochrome	71	**85.2**	69	**87.1**
C-Hydroxylation	2570	**70.4**	3104	**76.3**
C-Hydroxylation(Cytochrome P450)	840	**75.0**	804	**77.5**
Deacylation	32	**86.6**	90	**98.0**
Hydroxylation(Cytochrome P450)	72	**75.9**	880	**76.7**
N-Dealkylation	1663	**87.6**	789	**92.1**
N-Dearylation	35	**81.5**	63	**97.2**
N-Demethylation	757	**92.2**	273	**97.1**
N-Demethylation(Cytochrome P450)	203	**88.6**	56	**94.3**
N-Hydroxylation	179	**86.4**	505	**88.7**
O-Dealkylation	821	**88.5**	787	**93.9**
O-Dealkylation(Cytochrome P450)	199	**86.9**	156	**87.2**
O-Glucuronidation	1533	**81.8**	1837	**87.0**
S-Oxidation	486	**91.7**	171	**94.8**
….	…	**..**	..	**..**
В среднем	8742	**80**	7310	**81.9**

После обучения системы на созданных выборках ***можно выполнить прогноз типов биотрансформации для органического соединения***. Ниже приведен пример прогноза для молекулы галоперидола, который выполнен на основе SAR base, построенной по обучающей выборке I (MNA 2) (см. табл. 1).

Известные типы биотрансформации для галоперидола	**Структура галоперидола и предсказанные для него типы биотрансформации**
Aromatic Hydroxylation **Aromatization** Aromatization(Cytochrome P450) Aromatization(Cytochrome P450,CYP3A) Aromatization(Reductase) C-Oxidation **Cyanidation** Decarboxylation **Dehydration** Dehydration(Cytochrome P450)	 732 Possible of 2317 Biotransformations at Pt > Pu. Pt Pf for Biotransformation: **0.996 0.002 N-Dealkylation** (Cytochrome P450,CYP2C12)

Dehydration(Cytochrome P450,CYP3A)	0.996 0.002 N-Demethylation (Cytochrome
Dehydration(Reductase)	P450,CYP2A2)
Dehydrogenation	**0.997 0.005 Dehydration** (Cytochrome
Dehydrogenation(Cytochrome P450)	P450,CYP2E1)
Dehydrogenation(Cytochrome	**0.996 0.004 Aromatization** (Peroxidase)
CYP3A)	0.991 0.003 N-Demethylation (Cytochrome
Dehydrogenation(Reductase)	P450,CYP2C6)
Dehydroxylation	**0.994 0.005 Reduction** (Peroxidase)
Dehydroxylation(Cytochrome P450)	**0.994 0.008 Aromatic Hydroxylation** (Monoamine
Dehydroxylation(Cytochrome P450,CYP3A)	Oxidase)
Dehydroxylation(Reductase)	**0.989 0.004 N-Dealkylation** (Cytochrome
Hydrogenation	P450,CYP2C6)
Hydrogenation(Aldoketoreductase)	**0.966 0.003 Dehydrogenation** (Cytochrome
Hydrogenation(Carbonyl Reductase)	P450,CYP2A6)
Hydrogenation(Cytochrome P450)	**0.980 0.019 Oxidation (**Peroxidase)
Hydrogenation(Cytochrome P450,CYP2D6)	**0.991 0.036 Oxidation** (Lipoxygenase)
Hydrogenation(Cytochrome P450,CYP3A4)	**0.925 0.615 Reduction** (Flavoprotein-Linked
Hydrogenation(Ketoreductase)	Monooxygenase)
Hydrogenation(Reductase)	**0.910 0.186 Dehydrogenation** (Aldehyde Oxidase)
Hydrolysis	
N-Dealkylation	…
N-Dealkylation(Cytochrome P450)	***0.783 0.003 Quaternization (Cytochrome P450)***
N-Dealkylation(Cytochrome P450,CYP3A4)	***0.727 0.198 Reduction (Cytochrome P450,CYP2D6)***
N-Dealkylation(Peroxidase)	…
N-Oxidation	**0.719 0.235 Reduction** (N-Acetyltransferase)
Oxidation	**0.672 0.029 Dehydration**
Quaternization	…
Quaternization(Cytochrome P450)	***0.622 0.079 Aromatic Hydroxylation***
Quaternization(Cytochrome P450,CYP3A)	***0.621 0.012 Dehydration (Cytochrome P450)***
Quaternization(Reductase)	***0.585 0.041 Reduction (Cytochrome P450)***
Reduction	0.697 0.192 Epoxidation (N-Acetyltransferase)
Reduction(Aldoketoreductase)	0.608 0.105 Optical Resolution (Cytochrome P450)
Reduction(Carbonyl Reductase)	**0.594 0.111 Cyanidation** (Cytochrome P450)
Reduction(Cytochrome P450)	0.590 0.098 Hydroxylation (Cytochrome
Reduction(Cytochrome P450,CYP2D6)	P450,CYP1A1)
Reduction(Cytochrome P450,CYP3A4)	**0.568 0.068 N-Oxidation** (Cytochrome
Reduction(Ketoreductase)	P450,CYP3A4)
Reduction(Reductase)	**0.549 0.055 Oxidation**(Cytochrome P450,CYP2B1)
	0.528 0.036 N-Dealkylation(Cytochrome
	P450,CYP3A4)

Утолщенным шрифтом выделены спрогнозированные типы биотрансформации, у которых совпадает название реакции биотрансформации с известным для этого соединения (левый столбец в Таблице). Курсивом и утолщенным шрифтом выделены спрогнозированные типы биотрансформации, у которых совпадает название реакции биотрансформации и метаболизирующие ферменты с известными для этого соединения.

Как видно, на верхних строчках результатов прогноза, в основном, находятся типы биотрансформации, найденные экспериментально для галоперидола.

3.2. Прогноз сайтов биотрансформации

Задача прогноза сайтов биотрансформации имеет существенное отличие от задачи прогноза типов биотрансформации. В обоих случаях для прогноза необходимо иметь отрицательные примеры. При прогнозе типов биотрансформаций отрицательными примерами являются все субстраты в обучающей выборке, которые не являются субстратами реакций данного типа биотрансформаций. А при прогнозе сайтов биотрансформации есть проблема отсутствия отрицательных примеров. Чтобы определить, в каком положении пойдет реакция, желательно иметь примеры положений, в которых эта реакция не произойдет. Эти примеры и будут являться отрицательными. В базах данных таких примеров нет. Есть только реакции рассматриваемого типа биотрансформации (например, реакции ароматического гидроксилирования), которые рассматриваются алгоритмом прогноза как положительные, и реакции совершенно других типов биотрансформаций, которые не целесообразно брать как отрицательные, так как в них не отражено, в каком положении интересующая нас реакция биотрансформации не произойдет.

Поэтому **отрицательные примеры** для задачи прогноза сайтов биотрансформации **генерируются автоматически**. Например, при генерации отрицательных примеров ароматического гидроксилирования гидроксил будет добавлен ко всем свободным ароматическим углеродам. После чего из полученной выборки удаляются положительные примеры, тем самым получается выборка, содержащая только отрицательные примеры.

Обучающие выборки для прогноза сайтов биотрансформации создавались двумя способами. В первом случае в качестве положительных примеров были взяты все реакции определенной биотрансформации (например, реакции Алифатического гидроксилирования) из базы данных. Во втором случае происходил автоматический отбор реакций. А именно - выделялись фрагменты структур субстрата и метаболита, изменяющиеся во время реакции. После этого выделенный фрагмент сравнивался с существующими в словаре фрагментов реакций. Если искомый фрагмент был найден, то данная реакция рассматривалась в качестве положительного примера. Эта процедура проводилась для отсечения многостадийных реакций.

В Таблице 3 представлены характеристики выборок из БД Metabolite, использованные для прогноза на основе оригинальных названий биотрансформаций. Как и в случае прогноза типов биотрансформации, был проведен эксперимент по выбору оптимального уровня дескрипторов для прогноза сайтов биотрансформации.

Из Таблицы 3 видно, что для прогноза метаболитов биотрансформации «Ароматическое гидроксилирование» лучше всего использовать четвертый уровень дескрипторов, а для остальных, напротив, RMNA4 дает наихудший результат. Это связано с тем, что при прогнозе ароматического гидроксилирования для учета ароматического кольца целиком необходимо учитывать большее количество соседей, а для других реакций информации о ближайших соседях реакционного центра более чем достаточно для прогноза.

Таблица 3. Характеристика выборок из БД Metabolite (положительные примеры не отбираются) для прогноза метаболитов.

	Количество примеров		IAP, %		
Химическая трансформация	положител.	отрицател.	RMNA2	RMNA3	RMNA4
Aliphatic Hydroxylation	2329	10849	71.41	76.12	76.41
Aromatic Hydroxylation	2773	5104	99.42	95.81	88.91
N-Demethylation	1100	352	82.74	78.21	73.91
O-Glucuronidation	1765	2257	98.47	97.81	97.25

При отборе положительных примеров произошло ухудшение точности почти для всех реакций. Это можно объяснить тем, что, получая более «чистую» выборку, мы можем потерять часть положений, по которым пойдет реакция (например, если этот сайт указан совместно с сайтом другой реакции). Если не производится отбор положительных реакций, то сайты других реакций вносят свой вклад в дескрипторы, что может повысить точность прогноза (отличать положительные примеры от отрицательных становится проще).

После обучения систем на созданных выборках ***можно выполнить прогноз сайтов биотрансформации для органического соединения***.

Для нового субстрата генерируются все возможные продукты определенной химической реакции, для каждой пары «субстрат-продукт» формируется набор RMNA дескрипторов, и рассчитываются величины P_t и P_f - вероятности того, что сгенерированный продукт будет и не будет являться метаболитом, соответственно.

В Таблице 4 приведен пример прогноза положения ароматического гидроксилирования для субстрата, приведенного на рис. 4, при использовании обучающей выборки без отбора положительных примеров.

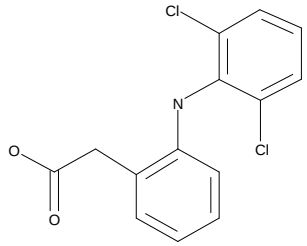
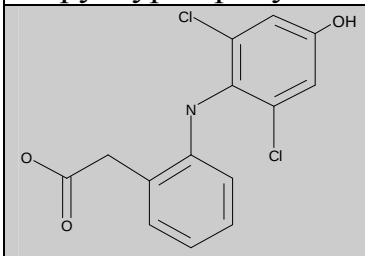
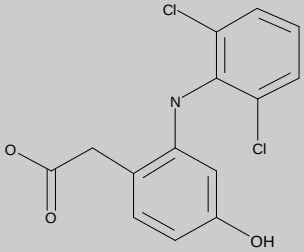
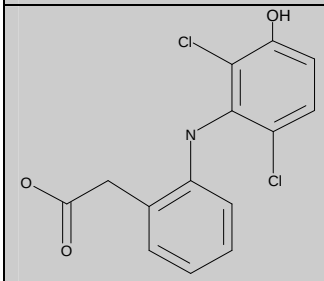
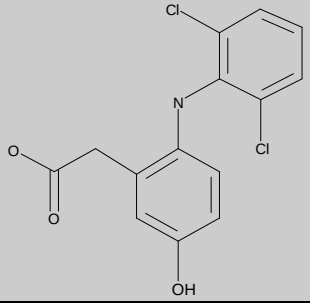
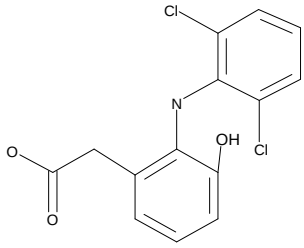
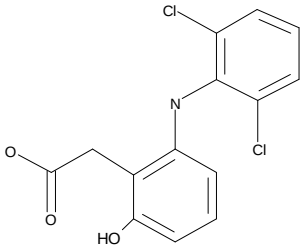


Рис. 4. Структура субстрата {2-[(2,6 - dichlorophenyl) amino]phenyl} acetic acid.

Все продукты были сгенерированы автоматически, после чего для них был выполнен прогноз. Серым цветом выделены ячейки, в которых находятся экспериментально найденные метаболиты этого субстрата (есть информация в БД).

В Таблице 4 приведен «идеальный случай», когда для экспериментально найденных метаболитов разность $P_t - P_u$ положительная, для остальных продуктов реакции ароматического гидроксилирования - отрицательная. Средняя точность прогноза по всей выборке (см. табл. 3) составляет 76%.

Таблица 4. Прогноз сайта ароматического гидроксилирования для субстрата (рис. 8). Представлены продукты и разности $P_t - P_f$.

Структура продукта	P_t-P_f	Структура продукта	P_t-P_f
	0.828		0.503
	0.072		0.758
	-0.039		-0,032

3.3. Исследования устойчивости прогноза при модификациях обучающей выборки.

Состав и объем обучающей выборки существенным образом влияют на точность прогноза. Анализ полученных ранее данных по робастности PASS (V.V. Poroikov et al., 2000) позволил предположить следующую полуэмпирическую формулу:

$$IA = IA_{inf}/(1 + L_-/N_- + L_+/N_+) = A(1 - m_-/N_- - m_+/N_+)/(1 + L_-/N_- + L_+/N_+),$$

где IA – инвариантная точность, связанная с инвариантной точностью прогноза (1) следующей формулой: IA=2IAP-1, IA_{inf} – предельно достижимая точность, N_+ и N_- – число положительных и отрицательных примеров, соответственно, m_- – количество примеров, ошибочно классифицированных как отрицательные (не найденные, еще не исследованные положительные примеры), m_+ – количество примеров, ошибочно классифицированных как положительные, например, из-за ошибок в использованных для формирования выборки данных, или из-за ошибок исследователей, и т.п. Коэффициенты $L_- \geq 1$ и $L_+ \geq 1$ отражают разнообразие примеров в выборке, и чем оно больше, тем больше величина L_+ и/или L_-.

Если предположить, что L_+ и L_- не зависят или слабо зависят от N_+ и N_- то можно попытаться оценить числитель $A(1 - m_-/N_- - m_+/N_+)$ при изменении объема выборки с сохранением пропорций m_-/N_- и m_+/N_+ (включение/удаление положительных и/или отрицательных примеров независимо от результатов прогноза) через оценку асимптотического поведения знаменателя $(1 + L_-/N_- + L_+/N_+)$.

Изменения объема выборки с сохранением пропорций m_-/N_- и m_+/N_+ можно добиться, удаляя случайным образом субстраты, т.е. и положительные и отрицательные примеры с их участием. Это эксперимент №2 () по модификации обучающей выборки. Другие исследованные в диссертационной работе способы модификации обучающей выборки:

Эксперимент №1 () - из обучающей выборки случайным образом удалялась часть положительных примеров (они автоматически попадали в отрицательные примеры);

Эксперимент №3 () - из обучающей выборки случайным образом удалялась часть положительных примеров.

Эксперимент №4 () - из обучающей выборки случайным образом удалялась часть отрицательных примеров.

Оценка точности прогноза была выполнена с усреднением 100 реализаций каждой модификации обучающей выборки на основе инвариантного критерия точности по методу скользящего контроля с исключением по одному.

Реакции «Алифатическое гидроксилирование» и «О-глюкуронидация» прогнозируются достаточно хорошо даже при значительных модификациях выборки (см. Табл. 5), поэтому эксперименты по расчету предельно достижимой точности для этих реакций не проводились.

Реакции «Ароматическое гидроксилирование» и «N-деметилирование» прогнозируются хуже. Результаты экспериментов представлены в Таблице 6.

Таблица 5. Зависимость от объема выборки точности прогноза метаболитов для двух реакций биотрансформации.

	20%	40%	60%	80%	100%
Эксперимент №1					
Aliphatic Hydroxylation	81.6	87.5	89.3	94.67	99.4
O-Glucuronidation	76.6	81.3	85.8	92.8	98.47
Эксперимент №2					
Aliphatic Hydroxylation	97.7	98.0	98.2	98.7	99.4
O-Glucuronidation	95.7	96.9	97.8	98.1	98.47

Таблица 6. Сводная таблица зависимости точности прогноза метаболитов от состава выборки для двух реакций биотрансформации.

	50%	60%	70%	80%	90%	100%
Aromatic Hydroxylation						
Эксперимент №1	63.07	65.2	67.76	70.0	73.2	76.4
Эксперимент №2	70.62	72.0	73.5	74.2	75.4	76.4
Эксперимент №3	74.6	74.97	75.53	76.05	76.27	76.4
Эксперимент №4	72.0	73.13	74.12	75.06	75.9	76.4
N-Demethylation						
Эксперимент №1	58.66	60.70	63.12	67.66	72.03	82.70
Эксперимент №2	76.3	78.7	79.9	80.7	81.8	82.70
Эксперимент №3	82.4	82.65	82.74	82.65	82.69	82.70
Эксперимент №4	81.1	81.4	81.9	82.37	82.58	82.70

По данным Таблицы 6 были рассчитаны предельно достижимые точности прогноза. Для этого мы построили графики зависимостей 50/(IAP-50) от 1/(1-p), где p - доля отбрасываемых данных.

Для реакции «Ароматическое гидроксилирование» получены оценки: предельно достижимая точность прогноза = 86.07% , L-/N- = 0.09 (L_0=459), L_+/N_+ =0,26 (L_+=720).

Для реакции «N-деметилирование» получены оценки: предельно достижимая точность прогноза = 92.67%, L_-/N_- = 0.23 (L_-=81), L_+/N_+ =0,29 (L_+= 319).

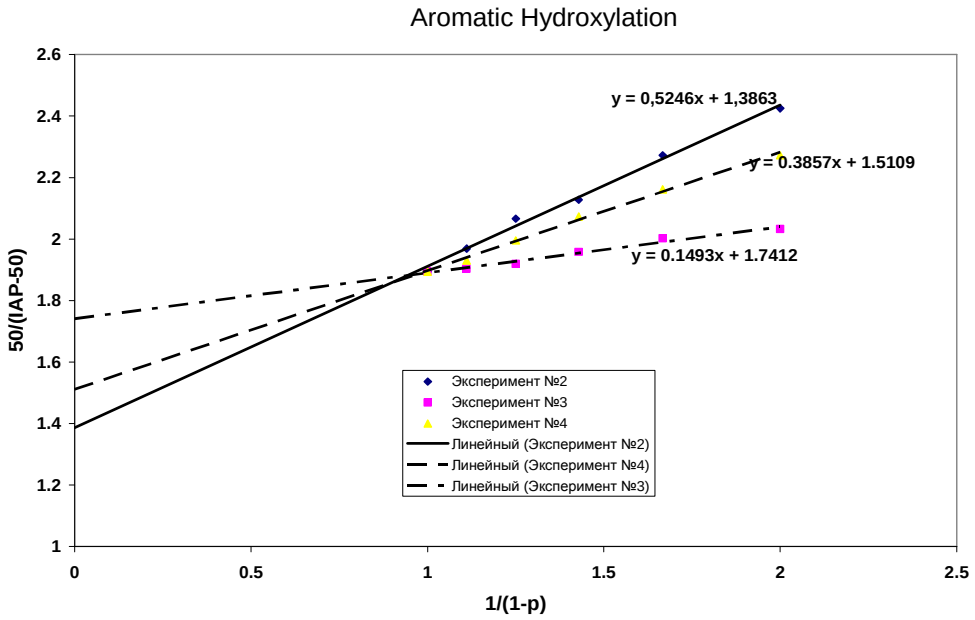


Рис. 5. Аппроксимация данных прогноза по трем экспериментам на бесконечный объем выборки для реакции «Ароматическое гидроксилирование».

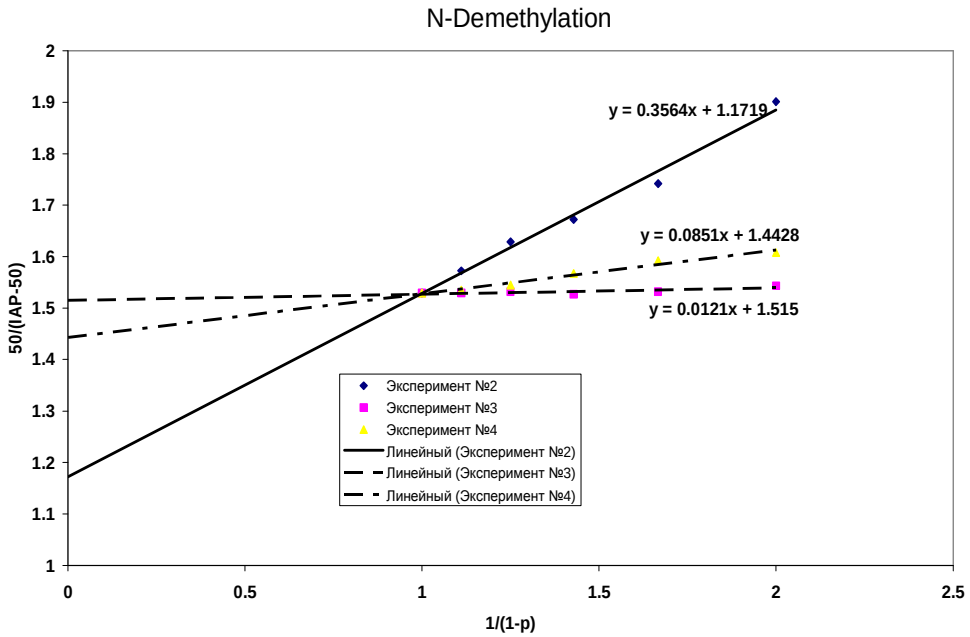


Рис. 6. Аппроксимация данных прогноза по трем экспериментам на бесконечный объем выборки для реакции «N-деметилирование».

Как видно, предельно достижимая точность для двух исследуемых реакций примерно на 10% больше, чем достигнутая в настоящее время. Пополняя обучающую выборку новыми данными по этим реакциям биотрансформации, можно увеличить точность прогноза метаболитов этих реакций.

4. ВЫВОДЫ

1. Разработан метод генерации продуктов биотрансформации.
2. Разработанный метод прогноза типов биотрансформации позволяет по структурной формуле органического соединения выполнить прогноз типов биотрансформации, включая ферментные системы и изоформы ферментов, участвующие в биотрансформации, со средней точностью более 85%.
3. Разработанный метод прогноза метаболитов (прогноз сайтов биотрансформации) позволяет по структуре субстрата с учетом реакции биотрансформации прогнозировать возможные метаболиты со средней точностью более 80%.
   Созданные RMNA дескрипторы отражают изменения в субстрате в ходе биотрансформации.
4. Исследования устойчивости прогноза при модификации обучающей выборки показали, что даже при удалении 50% объема выборки точность прогноза для некоторых реакций остается выше 85%. Рассчитанная предельно достижимая точность прогноза сайтов ароматического гидроксилирования и N-деметилирования составила 86% и 92% соответственно.

Публикации по теме диссертации

Статьи

1. Borodina Yu., Sadym A. (Rudik A.V.), Filimonov D., Blinova V., Dmitriev A., Poroikov V. Predicting Biotransformation Potential from Molecular Structure. // J. Chem. Inf. Comput. Sci.- 2003. - Vol. 43(5) - P. 1636-1646.
2. Бородина Ю.В., Садым А.В. (Рудик А.В.), Филимонов Д.А., Поройков В.В. Компьютерный прогноз биотрансформации ксенобиотиков. //Аллергия, астма и клиническая иммунология.- 2003- Т. 7 (8).- С.85-89.
3. Sadym A.V. (Rudik A.V.), Borodina Yu.V., Filimonov D.A., Poroikov V.V., Dmitriev A., Blinova V. Computer predicting of drig-like substances' metabolism: from biotransformation reactions to metabolic network.// Abstr. 4th International Symposium on Pharmaceutical Chemistry – Istanbul (Turkey). – 2003. – P. 259-260.
4. Y. Borodina, A. Sadym (A. Rudik), D. Filimonov & V. Poroikov. Computer predicting of biotransformation potential from molecular structure. //Abstr. Computational Methods in Toxicology and Pharmacology. Integrating Internet Resources. (CMTPI-2003) – Greece. – 2003. – P. 39.
5. A.V. Sadym (Rudik A.V.), Y.V. Borodina, A. Dmitriev, V.G. Blinova, D.A. Filimonov, V.V. Poroikov. Computer generation of metabolites for drug-like molecules.//Abstr. Computational Methods in Toxicology and Pharmacology. Integrating Internet Resources. (CMTPI-2003). – Greece. – 2003 – P.67.
6. Бородина Ю.В., Садым А.В. (Рудик А.В.), Филимонов Д.А., Поройков В.В. Компьютерный прогноз биотрансформации лекарственных веществ.// Сборник трудов X Российского национального конгресса «Человек и лекарство» . – Москва. – 2003. – С. 697.
7. Бородина Ю.В., Рудик А.В., Филимонов Д.А., Блинова В.Г., Дмитриев А.В., Харчевникова Н.В., Поройков В.В. Компьютерный прогноз биотрансформации ксенобиотиков.//Сборник материалов XII Международной конференции и дискуссионного научного клуба «Новые информационные технологии в медицине, биологии, фармакологии и экологии» . – Гурзуф (Украина). – 2004. – С. 77-79.
8. Рудик А.В., Бородина Ю.В., Дмитриев А.В., Филимонов Д.А., Поройков В.В., Блинова В.Г., Харчевникова Н.В. Компьютерный прогноз биотрансформации по структуре химических соединений.// Сборник материалов XI Российского национального конгресса «Человек и лекарство. – Москва. – 2004. – С. 891.
9. A.V. Dmitriev, A.V.Rudik, Y.V. Borodina, D.A. Filimonov, V.V.Poroikov, V.G. Blinova, N.V. Kharchevnikova. Computer predicting of aromatic hydroxylation site for drug-like compounds. //Abstr. The. 15th European Symposium on

Quantitative Strukture-Activity Relationships & Molecular Modelling (Euro QSAR 2004). – Turkey. – 2004. – P. 104.

10. Borodina Yu., Rudik A., Filimonov D., Kharchevnikova N., Dmitriev A., Blinova V., Poroikov, V. A New Statistical Approach to Predicting Aromatic Hydroxylation Sites. Comparison with Model-Based Approaches.// J. Chem. Inf. Comput. Sci. –2004. – Vol. 44(6). – P. 1998-2009.
11. Поройков В.В., Филимонов Д.А., Бородина Ю.В., Рудик А.В. Свидетельство об официальной регистрации программы для ЭВМ METAPREDICT № 2004610666 от 12 марта 2004 г., Москва: Российское агенство по патентам и товарным знакам.
12. Дмитриев А.В., Харчевникова Н.В., Рудик А.В., Поройков В.В. Разработка метода для предсказания региоселективности изоформы 3А4 цитохрома Р450. // Сборник материалов XII Российского национального конгресса «Человек и лекарство» . – Москва. – 2005. – С. 750.
13. Дмитриев А.В., Рудик А.В., Блинова В.Г. ,Филимонов Д.А., Поройков В.В. Предсказание классов биотрансформации ксенобиотиков на основе базы фрагментов. // Сборник материалов XIII Российского национального конгресса «Человек и лекарство» . – Москва. – 2006. – С. 14.
14. Рудик А.В., Филимонов Д.А., Поройков В.В. Анализ устойчивости алгоритма прогноза сайтов биотрансформации органических ксенобиотиков. // Сборник материалов XIV Российского национального конгресса «Человек и лекарство» . – Москва. – 2007. – С. 315.
15. Дмитриев А.В., Рудик А.В., Блинова В.Г., Филимонов Д.А., Поройков В.В. Предсказание сайтов биотрансформации окисления с использованием программы Metapredict и набора фрагментов. // Сборник материалов XIV Российского национального конгресса «Человек и лекарство». – Москва. – 2007. – С. 283.