УДК 547.821

БАЗА ДАННЫХ ПО СИНТЕЗАМ ПИРИДИНОВ ИЗ АЦИКЛИЧЕСКИХ РЕАГЕНТОВ: РОЛЬ НЕСТАНДАРТНЫХ ПОИСКОВЫХ ЗАПРОСОВ В РАЗРАБОТКЕ СТРАТЕГИИ КОМПЬЮТЕРНОГО СИНТЕЗА

Е. В. Бабаев, С. В. Цитовский

(кафедра органической химии)

Рассмотрена методология построения базы данных по синтезам пиридинового ядра в стандартной оболочке *ChemBase*. Предложен алгоритм внесения в базу данных нестандартной информации об электрофильно-нуклеофильной природе реагентов и структурном образе реакции (реакционном графе). Последующие поисковые запросы по введенным данным оказались эвристичными для планирования синтеза пиридинов из реагентов заданной полярности.

Настоящая работа продолжает исследования в области методологии компьютерного синтеза и дизайна новых гетероциклических структур и реакций, проводимых на химическом факультете МГУ [1—3]. Ранние публикации касались разработки формальных моделей для описания гетероциклических перегруппировок [1, 2]. Развитие этого подхода привело к созданию компьютерной программы для предсказания новых трансформаций циклов (или рециклизаций) гетероциклических систем [3].

Ключевую роль при выработке теоретической модели и прогнозировании новых реакций играло адекватное использование уже имеющейся информации о перегруппировках. Для этого в компьютерной программе была предусмотрена возможность создания элементарной базы данных по уже существующим реакциям, обращение к которой, в частности, позволяло делать выводы о степени новизны прогнозируемого структурного превращения.

В настоящей работе рассматривается дальнейшее обобщение этой методологии на реакции гетероциклизации. Во-первых, будет рассмотрено, какие именно формальные модели следует использовать для описания гетероциклизаций. Во-вторых, мы обсудим критерии, использованные нами для создания базы данных по синтезам пиридинов из ациклических структур. В-третьих, мы хотим показать, каким образом небольшая база данных по известным методам синтеза конкретного гетероцикла может помочь в разработке общей методологии компьютерного синтеза беспрецедентных примеров химических превращений.

Отдельно будут рассмотрены общие критерии, которым должны удовлетворять подобные базы данных: в частности, мы иллюстрируем, как, используя самые стандартные оболочки баз данных, например весьма распространенную среди химиков программу *ChemBase*, научиться извлекать нестандартные эвристические подсказки и правила для более общей проблемы планирования синтеза гетероциклов.

Абстрактные схемы синтеза гетероциклов

Кратко рассмотрим общие принципы формализации и классификации химических реакций [4—6]. Основная идея большинства подходов заключается в мысленной суперпозиции (отождествлении) атомов и связей исходных реагентов и конечных продуктов. Полученный «образ реакции» содержит информацию о местонахождении разрываемых и

образующихся связей. Следовательно, можно пренебречь всеми теми атомами, окружение которых не изменилось, и всеми теми связями, кратность которых осталась неизменной. Результирующий объект может быть задан в виде графа реакции, если левая и правая части выражались в виде обычных структурных формул (молекулярных графов). Другой способ — описание химического превращения в виде матрицы реакции (*BE*-матрицы [Dugundji]), если структура реагентов и продуктов задана, например, в виде матрицы связей.

Результирующий математический образ реакции — граф или матрица — показывает, как изменились ближайшее окружение атомов и порядки связей в ходе реакции (Фуджита [7] предложил удачный термин для таких объектов — «воображаемое переходное состояние реакции»). Такие объекты можно вводить в компьютер для поиска, например, аналогий между разными реакциями или для прогноза реагентов и конечных продуктов.

Эта идея была адаптирована в ранних работах [1, 2] для нужд гетероциклической химии, а именно для описания рециклизаций гетероциклов. Особенность подхода заключалась в том, что предельно упрощенный образ реакции (рециклизационный граф) тем не менее содержал скелеты исходного и конечного циклов. Такое «неполное» абстрагирование резко упростило весьма сложную проблему описания рециклизаций и привело к созданию иерархической (и периодической) классификации гетероциклических перегруппировок, использованной позднее [3] для компьютерного прогнозирования новых трансформаций гетероциклов.

Естественным обобщением этого подхода могло бы явиться аналогичное моделирование гетероциклизаций, т. е. способов сборки гетероциклического ядра из ациклических предшественников. К счастью, химики весьма давно разработали для гетероциклизаций адекватный формальный образ. Достаточно раскрыть любой обзор, обобщающий синтезы произвольного гетероцикла (например, [8]), чтобы увидеть, как именно это делается. Обычно изображается структура целевого гетероцикла, на которой пунктиром помечаются связи, образующиеся в ходе реакции.

Такие объекты (называемые в англоязычной литературе *disconnection schemes*), очевидно, не являются ни исходными реагентами, ни конечным циклическим продуктом. Это отображение между ними, т. е. граф реакции в упомянутом выше смысле. Нетрудно видеть естественную комбинаторную природу этого отображения: число способов расставить пунктиры по циклу в такой схеме полностью соответствует числу мыслимых методов, которыми можно собрать целевой гетероцикл. (Будем называть далее такие отображения схемами синтеза.)

Комбинаторная природа схем синтеза неоднократно привлекала внимание химиков. Имеется ряд работ (см., например, обзор [6]), где детальный анализ «придуманной» схемы синтеза приводил к экспериментальному обнаружению неизвестного ранее подхода к конструированию гетероциклического ядра. Именно так, «на кончике пера» были спрогнозированы и открыты неизвестные ранее синтезы ядра индазола и тиазола [6]. Между тем сам ход рассуждений и логика перехода (от весьма абстрактной схемы к совершенно конкретным реагентам для гетероциклического синтеза) практически никогда не оказывались предметом самостоятельного анализа со стороны специалистов по компьютерной химии.

Хотя в методологии компьютерного синтеза существует ряд подходов к дизайну конкретных реагентов для синтеза целевых молекул

[6], их использование нередко затруднительно в связи с эффектом «комбинаторного взрыва» и появлением множества бесперспективных предшественников — шумовых структур. Решение такого рода проблем могло бы заключаться в разработке эффективных критериев отбора наиболее приемлемых реагентов. Это в свою очередь требует тщательного отбора исходной информации о том, какие синтезы данного гетероцикла осуществляются легко и селективно, а какие — нет.

Заметим, что именно этот аспект проблемы — анализ «хороших» и «плохих» синтезов — наиболее детально разрабатывался химиками-экспериментаторами, например, в обзорах по синтезам того или иного гетероцикла. Обычно для одной и той же схемы синтеза приводится обстоятельный анализ, какие именно реагенты годятся для данной схемы, а какие нет (см., например, [8]). Таким образом, существует разрыв между имеющимся у экспериментаторов реальным знанием о закономерностях той или иной схемы синтеза и неумением теоретиков компьютерного синтеза этим знанием воспользоваться.

Три класса реагентов для гетероциклизаций

Подавляющее большинство гетероциклизаций является гетеролитическими процессами, где связи C—C, C—X или X—Y (X и Y — гетероатомы) формируются в процессах электрофильного или нуклеофильного присоединения или замещения. Одной из важнейших идей, учитывающей доминирование гетеролитических процессов в синтезах, а также упрощающей и систематизирующей типы ациклических структур для синтеза гетероциклов, является разделение реагентов на три группы [8, 9]: биэлектрофилы, бинуклеофилы и бифункциональные соединения (т. е. содержащие на концах цепи электрофильный и нуклеофильный центры).

Единственным принципиальным различием реагентов внутри каждого из трех классов оказывается размер цепи N между полярными функциями. (Подразумевается цепь, вносимая реагентом в цикл.) Так, среди реагентов с двумя электрофильными функциями на концах цепи можно выделить биэлектрофилы следующих подклассов:

(1,1) — например, геминальные дихлориды, альдегиды, кетоны, производные карбоновых кислот;

(1,2) — например, дихлорэтан, фенацилбромид, глиоксаль;

(1,3) — 1,3-дикарбонильные соединения и т. д.;

(1,N) — биэлектрофилы с большей длиной цепи.

Аналогично можно классифицировать бинуклеофилы или бифункциональные реагенты [8, 9].

Представляло оправданный интерес, какие именно полярные типы реагентов доминируют в синтезе того или иного гетероцикла. (Общеизвестно, например, что α-галогенкетоны весьма часто используются в синтезе азолов, но редко — при получении азинов [8].) Ответ на поставленный вопрос мог бы играть ключевую роль в компьютерном ретросинтезе гетероциклов, а именно в переходе от абстрактных схем синтеза к конкретным классам реагентов. Такой ответ следовало искать, например, в сопоставлении полярной природы реагентов, используемых в синтезе тех или иных гетероциклических структур.

Компьютерная база данных по синтезам пиридинов

Мы полагали, что разработка каких-либо «хороших» правил синтеза гетероциклических структур может проводиться лишь на основе анализа конкретного обзорного материала по синтезам гетероциклов.

Такие правила можно в дальнейшем совершенствовать, сопоставляя их для разных гетероциклов.

С этой целью представлялось разумным вводить исходную информацию о синтезах конкретных гетероциклов в компьютерную базу данных (далее — БД) для дальнейшего статистического анализа.

Выбор объектов. В качестве простейшей гетероциклической системы было выбрано ядро пиридина; в дальнейшем предполагалось сопоставить методы синтеза этого ядра с методами синтеза ближайших родственных структур (например, бензо-, аза- и гетероаналогов). В первую очередь отбирались максимально ненасыщенные структуры (собственно пиридины, пиридоны и т. д.). В дальнейшем БД пополнялась частично или полностью гидрированными структурами.

Источники формирования БД. С целью добиться максимальной репрезентативности и охвата литературы разных лет за первичные источники были выбраны основные обзоры и монографии по химии пиридина [8, 9—15]. По мере накопления информации анализировались реферативные журналы последних лет.

Выбор оболочки для БД. В качестве стандартной оболочки по многим причинам была выбрана система *ChemBase*. Во-первых, эта программа получила едва ли не самое широкое распространение среди химиков. Во-вторых, она обладает мощными возможностями графического ввода и обработки информации не только о структурах *но и* о реакциях. Наконец, представлялось любопытным, можно ли «заставить» стандартную базу данных отвечать на нестандартные запросы, например, о полярной природе реагентов, используемых в синтезе. Помимо стандартного (внутреннего) формата хранения информации о реакциях, программа позволяет трансформировать ее в форматы, читаемые другими компьютерными программами, что предполагалось использовать в дальнейшем.

Отбор реакций. Реакции отбирались по следующим критериям:

Ограничение гетероциклизациями. Методы синтеза, являющиеся перегруппировками или рециклизациями, не рассматривались.

Ограничения на гетеролитический механизм. Требование однозначности описания гетероциклизаций в терминах взаимодействия электрофильных и нуклеофильных центров привело к тому, что ряд перициклических реакций не рассматривался.

Селективность. Реакции должны приводить преимущественно к одному продукту.

Лабораторные условия протекания. Ряд парофазных промышленных синтезов пиридинов не анализировался.

Произвольное число компонент для синтеза. В этом случае база данных покрывала большинство примеров именных реакций [16].

Форма представления информации. Для каждого синтеза в базе данных отводилось окно (entry) со следующими полями:

Графическая часть содержала структурные формулы исходных гетероциклических систем.

Текстовая часть состояла из 11 полей, включая, в частности, идентификационный номер, авторов работы, обзорный источник и выходные данные оригинальной статьи, название синтеза (если реакция являлась именной) и поле комментария (условия синтеза, реагенты, выход).

Стандартные запросы по синтезам пиридинов, предусмотренные в БД

Созданная база данных позволяет осуществить поиск исходного реагента или продукта по выбору пользователя, вынося результаты в отдельный список. Аналогично возможен поиск по текстовым полям (по авторам, годам публикаций работ или, например, именных реакций). Признаки, по которым осуществляется поиск, можно объединять в логические выражения. Например, не представляет труда за несколько секунд получить информацию об использовании какого-либо соединения в синтезах какого-либо производного пиридина за последние 20 лет.

Дополнительная возможность БД — поиск по подструктуре (для чего необходимо указать произвольный фрагмент молекулы, например функциональную группу или набор атомов углерода, соединенных связями заданной кратности). Кроме того, несколько таких фрагментов можно объединять в логические выражения. Заметим, что уже в такой стандартной форме количество примеров, занесенных в БД, порождает новое качество: если структурная схема синтеза какой-либо целевой структуры неизвестна, то можно запрашивать информацию об аналогичных синтезах. В результате созданная БД имеет определенную методическую ценность: с ее помощью нетрудно проводить первоначальное ознакомление (например, студентов) с многообразием синтетических подходов в химии гетероциклов.

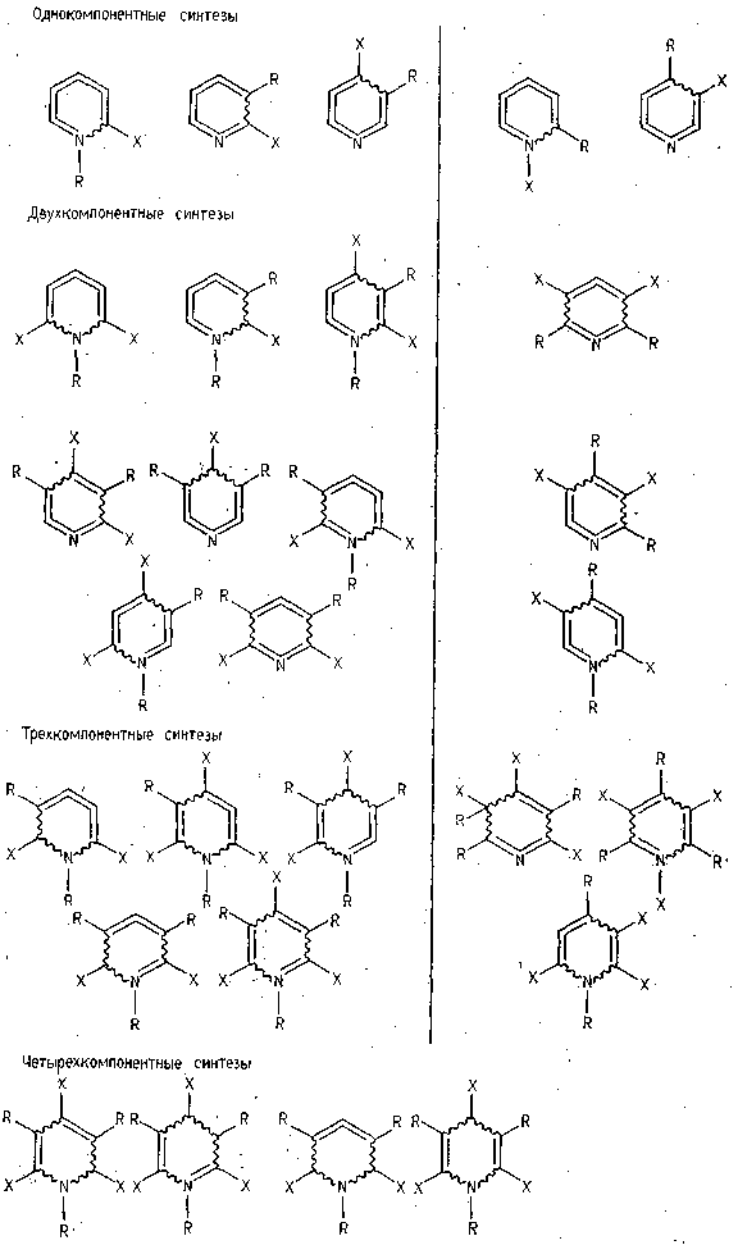
Способы адаптации БД для ответов на нестандартные запросы о полярной структуре реагентов

Нашей целью являлось выяснение полярной природы реагентов, наиболее часто используемых для синтезов пиридинового ядра. Следует подчеркнуть, что этот тип информации не содержится в явном виде в структурных формулах реагентов: необходимо определенное мысленное усилие, чтобы приписать данному реагенту в данной реакции определенный полярный тип. Естественно, что и программа *ChemBase* с самого начала, казалось бы, не могла быть адаптирована для решения таких задач. Таким образом, мы стояли перед выбором: либо создавать новую оболочку графического ввода структур, либо адаптировать уже существующую программу с мощным потенциалом молекулярной графики для наших целей.

В этой связи мы обратились к предусмотренной в программе *ChemBase* возможности ввода в БД структуры интермедиатов, по которым в дальнейшем можно проводить поиск, аналогичный поиску по реагентам или продуктам. Мы полагали, что именно в виде воображаемого интермедиата можно отобразить полярный тип реагента. Другими словами, представлялось целесообразным «замаскировать» абстрактную схему синтеза под интермедиат, указав на ней каким-либо образом электрофильно-нуклеофильную природу центров, формирующих связи цикла.

Решение такой задачи оказалось далеко не тривиальным. Во-первых, программа автоматически присваивает любому атому символ совершенно конкретного элемента. Во-вторых, абстрактная схема синтеза есть, вообще говоря, связный помеченный граф, имеющий два «цвета» ребер — обычные (скелетные связи реагента, перешедшие в цикл) и пунктирные (циклические скелетные связи, возникшие в ходе реакции). Следовательно, требовалось различение между пунктирной

Известные структурные типы синтезов пиридинового ядра, содержащиеся в базе данных



Двойные линии в приведенных диаграммах отвечают скелетным связям реагентов, переходящим в скелет пиридинового цикла. Волнистой линией показаны формируемые связи цикла. Символы R и X отвечают природе реакционных центров в реагенте (R — нуклеофильность соседнего атома, X — его электрофильность). Линия отделяет наиболее типичные синтезы (слева) от чрезвычайно редких (справа)

и обычной связью, которые в системе *ChemBase* отождествляются. Наконец, программа была нечувствительна к различению каких-либо меток на атомах (пусть даже предусмотренных при графическом вводе

символов, например, зарядов, изотопных меток и т. д.), отождествляя их при поиске.

Единственно допустимым решением, как «обмануть» программу, оказался следующий алгоритм «know-how»:

Стандартные схемы синтеза изображаются в виде воображаемых интермедиатов, которые программа способна сопоставлять между собой независимо от реагентов или продуктов.

Скелетные связи реагента (переходящие без изменений в цикл) на схеме синтеза выражаются двойными связями, а формируемые связи — простыми. В этом случае программа, различая простые и двойные связи, ведет поиск по структуре интермедиата в целом, т. е. фактически по схемам синтеза. Для удобства пользователя простые связи маскируются волнистой линией, воспринимаемой компьютером как обычная простая связь.

Электрофильный и нуклеофильный концы реагента выражаются введением в схему синтеза «фантом-групп» двух типов, соответственно X и R. (Символы X и R — единственные зарезервированные программой обозначения функциональных групп, которые различимы при поиске.)

Типичные примеры схем синтеза, вводимые пользователем таким путем, приведены на схеме.

Некоторые результаты анализа ответов БД на нестандартные запросы

В настоящее время БД находится в стадии формирования. Тем не менее анализ уже нескольких сотен введенных реакций привел к заключениям, в какой-то мере превосходящим наши первоначальные ожидания. В первую очередь мы стремились ввести в БД максимальное число неэквивалентных схем синтеза. Для этого вводимая в БД структура «интермедиата» сопоставляется с другими, уже имеющимися в памяти компьютера. Лишь в случае отсутствия такой схемы в БД принимается решение о первоочередности ввода самого химического уравнения. Таким образом, на схеме приведены все (известные авторам) неэквивалентные схемы синтеза пиридинов с указанием полярной природы реагентов.

Вертикальная черта на схеме отвечает предварительной статистике распределения синтезов пиридинов по классам: слева располагаются те синтезы, по которым имеются сотни или десятки примеров. Справа — синтезы, примеры которых единичны.

Анализируя реагенты левой части схемы, нетрудно видеть, что обнаруживается явный детерминизм предпочтительного расположения полярных функций в реагентах. Так, например, требуется, чтобы нуклеофильные центры в цепи реагента находились именно в тех местах, которые отвечают либо β-положениям формируемого пиридинового цикла, либо самому пиридиновому атому азота. Наоборот, электрофильные центры фиксированы в тех мастах цепи реагентов, которые перейдут в α- или γ-положения пиридинового цикла. Заметим, что в 99% случаев это правило строго соблюдается для одно- и двухкомпонентных синтезов максимально ненасыщенных пиридинов. Отклонения же наблюдаются в единичных примерах многокомпонентных синтезов, а также при получении гидрированных структур. Таким образом, проявляется отчетливая, неизвестная ранее взаимосвязь типа «структура—синтез», которую можно использовать в компьютерном дизайне новых гетероциклизаций. Детальный анализ этой взаимосвязи явился предме-

том отдельного сообщения [17]; здесь мы продемонстрировали путь, который к ней приводит.

Возможность использования программы *ChemBase* (Версия 1.50, Molecular Design Ltd., San Leandro) была любезно предоставлена А. В. Чепраковым, МГУ (Disk Serial BAI 2150271). Настоящая работа была включена в учебные планы кафедры органической химии (дипломная работа С. В. Цитовского, июнь 1993 г.).

СПИСОК ЛИТЕРАТУРЫ

[1] Бабаев Е. В., Зефиров Н. С.//ХГС. 1992. № 6. С. 808.
[2] Babaev E. V., Zefirov N. S.//Bull. Soc. Chim. Belg. 1992. **101**. N 1. P. 67.
[3] Babaev E. V., Lushnikov D. E., Zefirov N. S.//J. Am. Chem. Soc. 1993. **115**. P. 2416.
[4] Chemical Information Systems: Beyong the Structure Diagram. Chichester, 1990.
[5] Herges R.//Tetrahedron Comput. Methodol. 1988. N 1. P. 15.
[6] Barone P., Chanon M.//Computer Aids to Chemistry. Chichester, 1986. 411 p.
[7] Fujita J.//J. Chem. Inf. Comput. Sci. 1988. 28. P. 128.
[8] Comprehensive Heterocyclic Chemistry. Oxford, 1984. V. 1—8.
[9] Katritzky A. R. Handbook of Heterocyclic Chemistry. Oxford, 1985. P. 382.
[10] Maier-Bode H., Altpeter J./Das Pyridine und seine Derivate in Wissenschaft und Technik. Halle, 1934. 351 S.
[11] Чумаков Ю. И. Пиридиновые основания. Киев, 1965.
[12] Schofield K. Heteroaromatic Nitrogen Compounds. Pyrroles and Pyridines. L., 1967.
[13] Brody F., Ruby R. Pyridine and Its Derivatives. N. Y., 1960. Part 1. P. 99.
[14] Boodman N. S., Hawthorne J. O., Maskiantonio P. X., Simon A. W.//Pyridine and Its Derivatives. Suppl. 1. N. Y., 1974. P. 185.
[15] Pyridine and Its Derivatives. Suppl. 5. N. Y., 1984.
[16] Вацуро К. В., Мищенко Г. А. Именные реакции в органической химии. М., 1976.
[17] Бабаев Е. В.//ХГС. 1993. № 7. С. 937.

Поступила в редакцию
14.09.93