## COMPUTER SYNTHESIS

N. S. ZEFIROV

***The paper shortly discusses the methods of planning synthesis of organic compounds and building up of the "synthetic tree" with the help of computer programs. Three general problems of computer synthesis are discussed: (i) imaging and analysis of chemical structures; (ii) imaging and analysis of chemical reactions-transformations, and (iii) usage of criteria for the selection of appropriate programs. Several well-known computer programs are described and examples of their work are given.***

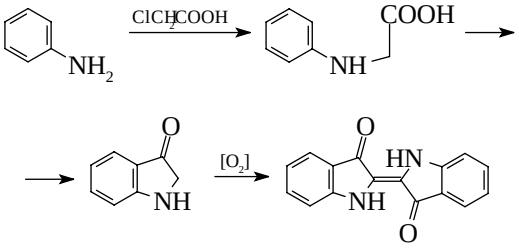
***Рассмотрены пути решения задачи планирования синтеза органических соединений и построения "дерева синтеза" с помощью компьютерных программ. Обсуждены три основных аспекта компьютерного синтеза: представление и анализ химических структур, представление и анализ химических реакций-трансформаций, критерии отбора. Дано описание некоторых известных компьютерных программ и примеры их использования.***

# КОМПЬЮТЕРНЫЙ СИНТЕЗ

Н. С. ЗЕФИРОВ

*Московский государственный университет им. М.В. Ломоносова*

Химики-органики 60–70-х годов XIX века не только выделили индивидуальное вещество (носитель красящих свойств) из природной краски, но и осуществили его синтез – сначала лабораторный, а затем и промышленный по следующей схеме:




Индиго

Синтетическая уксусная и лимонная кислоты, синтетические витамины, антибиотики и алкалоиды – список успехов органического синтеза разнообразен и огромен [1]. Объединяет подобные примеры тот факт, что в каждом случае целью синтеза являлись известные в природе соединения с известным комплексом полезных свойств. Однако, как показывает весь опыт органической химии, вещества, обладающие полезными свойствами, могут быть получены не только путем копирования природных структур. Действительно, многие (хотя далеко не все) свойства органического соединения могут быть предсказаны заранее на основании одной только структурной формулы соединения, еще не существующего ни в природе, ни в лаборатории.

Синтез целевого вещества всегда протекает по заранее составленному плану. Обычно для рационального планирования синтеза необходимо произвести как бы "разборку" молекулы, то есть представить себе, из каких ближайших предшественников эту молекулу можно синтезировать с помощью реальных реакций. Иными словами, мысленно химик при планировании синтеза идет в ретросинтетическом направлении, то есть от целевой структуры к выбору предшественников. Задача планирования синтеза в ретросинтетическом направлении формулируется следующим образом: пусть задано целевое соединение (**T** – target; рис. 1) и химик должен найти возможные пути его синтеза из других веществ-предшественников ($\mathbf{T}_1$ – $\mathbf{T}_7$ на рис. 1).

Если какие-либо из предшественников доступны, то на этом планирование заканчивается и можно приступать к экспериментальной работе по синтезу. Однако в большинстве случаев это не так:

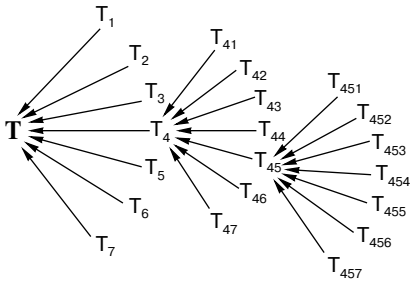



**Рис. 1.** Схематическое представление дерева синтеза для трехстадийного процесса

структуры каждого из этих предшественников необходимо проанализировать аналогичным образом и так продолжать вплоть до тех соединений, которые доступны и имеются на полке. В результате этого планирования возникает так называемое дерево синтеза (рис. 1). Дерево синтеза представляет собой схематическое изображение возможных путей синтеза данного веществ из исходных соединений через промежуточные.

Уже на этом этапе можно сделать общие рекомендации по плану синтеза. Представим себе, что мы синтезируем вещество по схеме *а* (рис. 2), последовательно получая из одной структуры последующую. Очевидно, что сбой и экспериментальная неудача на любой стадии этого "линейного" синтеза означает полную неудачу плана в целом. План *б* (рис. 2), представляющий "разветвленное дерево", оставляет возможность маневра, превращая неудачу на отдельной стадии в локальную проблему. Такой план синтеза называется конвергентным.

В принципе получаемое дерево синтеза очень "густое" и содержит огромное число "ветвей". Собственно в этом и заключается одна из трудностей синтеза сложных веществ – гигантское многообразие различных вариантов и тем самым исключительная трудность выбора оптимального пути синтеза.

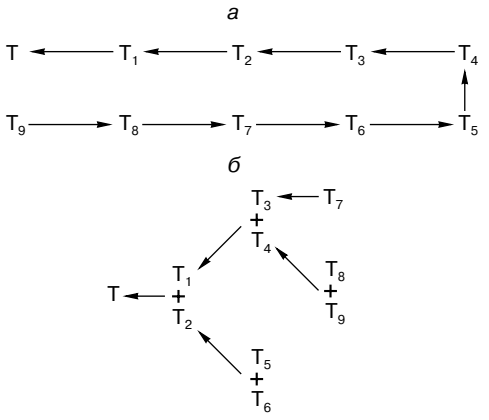
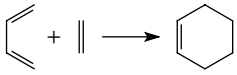



**Рис 2.** Примеры неконвергентного (*а*) и конвергентного (*б*) синтетического плана

Долгое время химики-синтетики руководствовались в своих действиях обращением к литературным данным в поиске надежных и проверенных практикой химических реакций, которые специфически пригодны для синтеза базового скелета данной целевой молекулы. Здесь нельзя не упомянуть реакцию Дильса–Альдера:



которая революционизировала синтез структур, содержащих шестичленные циклы.

Однако, как показала практика, такой поиск не всегда успешен, поскольку фактический материал по органическим реакциям настолько огромен, что можно легко пропустить и не заметить нужную реакцию, хотя она и опубликована в литературе. Более того, часто требуется открытие принципиально новых реакций и подходов, что запланировать невозможно.

Выход из этой ситуации заключается в широком применении компьютеров для планирования органического синтеза. Лишь колоссальная память и быстродействие компьютера позволяют найти и оценить огромное число возможных вариантов синтеза того или иного соединения, выбрать из них оптимальный план синтеза, который таким образом может быть осуществлен с минимальными затратами и максимальными шансами на успех.

В настоящее время компьютерный синтез уже не является совокупностью отдельных разработок, а сформировался в большое научное направление. Мы рассмотрим три основных аспекта компьютерного синтеза: 1) представление и анализ химических структур, 2) представление и анализ химических реакций-трансформаций, 3) критерии отбора.

Осуществление любой программы компьютерного синтеза начинается с ввода некоторых начальных данных, основу которых составляет информация о структуре заданной химической системы. Практически все существующие программы используют возможности ввода структуры в виде рисунка с помощью графических устройств компьютера. Обычно рисунок вводится в самом привычном для пользователя виде. Полученная графическая информация преобразуется в некоторое внутреннее представление структуры в программе. Каковы основные принципы кодирования в программе молекулярной структуры? Любая программа учитывает типы атомов в заданной системе и связи между ними, в том числе кратность связи. В некоторых программах для описания структуры химического соединения используют таблицы связности, которые указывают на ближайшее окружение каждого атома в структуре, а также могут содержать дополнительную

информацию (описывают стереохимические особенности структуры, заряды атомов и т. п.).

Приведем примеры описания химических структур в некоторых известных программах. В программе SYNGEN (SYNthesis GENeration, Дж. Хендриксон, США) описание заданной системы включает в себя только атомы скелета молекулы и связи между ними. Кроме того, для каждого атома скелета в таблице связности предусмотрены данные, касающиеся характера связей каждого конкретного атома с теми или иными функциональными группами. Программа EROS (Erzeugung von Reaktionen für die organische Synthese, Й. Гастайгер, ФРГ) хранит описание структуры в виде так называемой BE-матрицы, которая указывает порядки связей между атомами заданной системы (недиагональные элементы) и число свободных электронов на внешней валентной оболочке каждого атома (диагональ матрицы). Программа SYNCHEM (SYNthetic CHEMistry, Г. Гелернтер, США) использует обычное представление структур в виде таблиц связности. Описание заданной системы в виде таблицы связности применяется при анализе и трансформации структур, в то время как указанный код предназначен для удобства хранения и составления структурных банков данных. В ходе анализа структур, входящих в состав заданной системы, программы извлекают дополнительную информацию, необходимую для дальнейшей работы: особенности строения скелета (число и взаимосвязь циклов, цепей), число и вид функциональных групп, данные о симметрии структуры.

Следующим необходимым этапом составления программы компьютерного синтеза являются представление и анализ химических реакций. На этом этапе можно выделить два принципиально различных подхода: эмпирический (трансформации заданной системы осуществляются на основе сведений об известных органических реакциях) и неэмпирический (трансформации генерируются без привлечения фактических сведений). В первом случае химические реакции должны быть заранее систематизированы и закодированы, во втором случае для поиска трансформаций применяется комбинаторный алгоритм (набор некоторых логико-комбинаторных инструкций). Эмпирическое направление имеет то преимущество, что в этом случае программа обычно предсказывает правдоподобные пути синтеза, а большинство критериев отбора в неявном виде содержится в описании каждого конкретного превращения. К сожалению, такие программы неспособны предложить принципиально новый синтетический путь или найти новую реакцию, так как ограничены конкретной библиотекой трансформаций. Программы неэмпирического направления лишены этого недостатка, однако требуют включения в программу строгих критериев отбора, чтобы избежать получения нереальных или малоинтересных результатов.

Большинство программ относится к эмпирическому направлению и использует в своей работе библиотеки трансформаций, для которых характерны следующие общие черты. Во-первых, описание трансформации включает перечисление структурных фрагментов, которые должны присутствовать в заданной системе, для того чтобы структурная трансформация могла осуществиться. Во-вторых, описание должно содержать тесты, которые определяют принципиальную возможность осуществления данного превращения и/или определяют приоритетность описываемой трансформации. Кроме того, описание включает также перечень структурных изменений, которые необходимо произвести, чтобы получить структуры, соответствующие результату применения данной трансформации. Наконец, необходим идентификатор каждой трансформации. Указанные аспекты присущи всем описанным в литературе библиотекам трансформаций, однако некоторые из них могут содержать дополнительную информацию. Например, в программах LHASA (Logic and Heuristic Applied to Synthetic Analysis, Ю. Кори, США) и SECS (Simulation and Evaluation of Chemical Synthesis, У. Уипке, США) каждой трансформации приписывается краткое определение структурных изменений в заданной системе, которые вызывает данная трансформация (образование – разрыв связи, замыкание – раскрытие цикла, введение – удаление функциональных групп). Кроме того, в этих программах содержатся сведения об условиях реакций (например, о температуре и вспомогательных реагентах). В системе REACT (REACTion path synthesis program for the petrochemical industry, Р. Говинд, Г. Пауэрс, США), предназначенной специально для изучения химико-технологических процессов, большое внимание уделено описанию технологических условий для успешного осуществления конкретных процессов.

В отличие от эмпирических в программах неэмпирического направления трансформации осуществляются не на основе данных библиотеки, а в результате применения некоторых логических конструкций. Например, программы EROS и TOSCA (Topological Synthesis design by Computer Application, Ю. Зандер, ФРГ) используют ряды так называемых генераторов реакций, то есть инструкций, которые в самом общем виде описывают перераспределения связей в ходе химических реакций. В одной из последних версий программы EROS используют пять генераторов, описывающих большинство органических реакций:

1) $X: + I - J \longrightarrow I - X - J$

2) $I - X - J \longrightarrow X: + I - J$

3) $I - J + K - L \longrightarrow I - K + J - L$

4) $I - J + K - L + M - N \longrightarrow N - I + J - K + L - M$

5) $I - L + K - M + X: \longrightarrow M: + I - K + L - X,$

где I, J, K, L, M, N – реакционные центры, то есть атомы, связи между которыми изменяют свой порядок на единицу. Центр X соответствует атому, изменяющему в ходе реакции свою валентность на две единицы (например, карбенному центру в реакции присоединения карбена). Разрыв связи между центрами I и J может означать как разрыв, так и уменьшение порядка связи. Аналогично образование связи может соответствовать как реально образованной связи, так и увеличению кратности уже имеющейся. Очевидно, что для различных центров и связей один и тот же генератор реакции будет порождать различные химические превращения. Например, третий генератор может соответствовать как присоединению по двойной связи, так и реакции замещения. Таким образом, генераторы реакций в применении к конкретной системе могут порождать наряду с хорошо известными превращениями также и совершенно новые трансформации.

Для описания химической информации в программах неэмпирического направления может применяться и так называемый формально-логический подход (программа FLAMINCOES – Formal-Logical Approach to Molecular Interconversions, Н. С. Зефиров, С. С. Трач, Россия), в основе которого лежит представление любого процесса в виде суммарного результата, а именно в виде совокупности структурных изменений, происходящих при переходе от исходной химической системы (ХС) к конечной. Система может состоять из одного или нескольких веществ, структуры которых представлены в виде химических графов. Важнейшие типы органических реакций формально описываются как результат циклического перераспределения связей (ЦПС) при переходе от исходной к конечной системе. В настоящее время существует также программа COMPASS (Н. С. Зефиров, Д. Л. Лушников, Е. В. Гордеева), базирующаяся на сочетании чисто комбинаторных методов с эмпирическими правилами ретросинтетического анализа и, таким образом, как бы объединяющая эмпирический и неэмпирический подходы. Для специального круга реакций – карбокатионных перегруппировок – существует программа ICAR (Н. С. Зефиров, В. В. Щербухин, Е. В. Гордеева), в которой формально-логический подход используется для описания этих многостадийных процессов.

Итак, мы показали, каким образом химическая информация кодируется и формализуется в некоторых известных компьютерных системах. Однако главная проблема компьютерного синтеза – это создание и формализация критериев отбора, которые позволяют значительно сократить количество операций, с тем чтобы программа генерировала в первую очередь самые вероятные пути синтеза. Оперирование критериями отбора как раз и придает программам черты искусственного интеллекта.

В программах эмпирического направления критерии отбора могут использоваться на трех стадиях поиска: 1) выбор определенной стратегии синтеза, которая ограничивает число и вид превращений; 2) оценка и отбраковка конкретных трансформаций до их применения в заданной системе; 3) оценка и отбраковка конкретных предшественников, полученных в результате трансформации. Выбор определенной стратегии может осуществляться автоматически или с участием пользователя. Применение нескольких стратегий (например, структурно-ориентированной, стереохимической, топологической и т. п.) обычно позволяет найти эффективные и элегантные пути синтеза. Вопрос об априорной оценке вероятности протекания реакции решается в компьютерных программах с помощью формальных и эмпирических критериев отбора. Трансформация считается формально возможной, если в системе присутствует структурный фрагмент, формально необходимый для осуществления данной трансформации. Эмпирические критерии отбора реализуются на основе более глубокого анализа структуры системы, условий реакции, состава реагентов. В целях более детального анализа превращений может привлекаться и дополнительная информация. Наряду с оценкой трансформаций в программах эмпирического направления проводится и оценка самих предшественников. Отбраковка вариантов может производиться, например, по следующим критериям отбора: 1) неправильное значение валентности атома, 2) два одинаковых по знаку заряда на разных атомах, 3) нестабильная комбинация функциональных групп, 4) наличие антиароматической системы, 5) тройная связь в малом цикле и т. д.

Основная проблема, с которой сталкиваются программы неэмпирического направления, заключается в том, чтобы отобрать из всего множества формально возможных путей синтеза наиболее вероятные и интересные с химической точки зрения. В таких программах можно выделить следующие задачи, для решения которых применяются критерии отбора: 1) ограничение типов трансформаций, 2) ограничение применимости трансформаций, 3) оценка и отбраковка генерированных предшественников. Критерии отбора предшественников во многом сходны с аналогичными критериями в программах эмпирического направления, поэтому остановимся на критериях отборов первых двух типов. Возможность выбора типов трансформаций в явном виде присутствует в программе EROS. Во-первых, здесь предусмотрена ситуация, когда пользователь исключает из списка генераторов реакций, хранящихся в машинной памяти, генераторы, отвечающие тем процессам, которые, по его мнению, не могут протекать в заданной системе. Принимается, что все пять генераторов реакций могут служить для поиска предшественников. Во-вторых, имеется возможность наложить определенные ограничения на вид и размер дерева синтеза, в частности указать максимальное число его уровней и число структур на каждой стадии. Как было показано ранее, описание

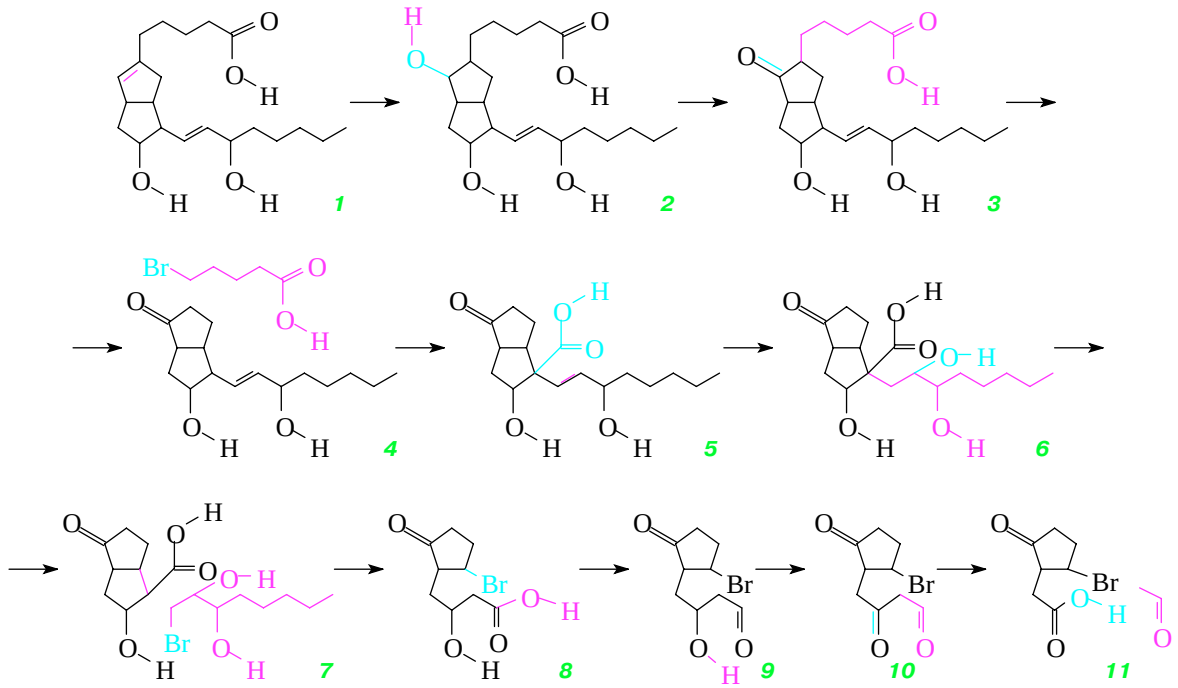


**Рис. 3.** Воспроизведение программой COMPASS синтеза изокарбоциклина

любой трансформации в программах неэмпирического направления в общем виде можно представить как набор реакционных центров, указав, каким образом перераспределяются связи между ними. Следовательно, критерий отбора, который может ограничить применимость тех или иных трансформаций, заключается в выборе потенциальных реакционных центров или реакционноспособных связей из всего множества атомов и связей данной системы. Так, например, определяются связи, которые в принципе могут разрываться в ходе дальнейших превращений. В автоматическом режиме реакционноспособными объявляются кратные связи, связи C–X, H–X, X–X (X-гетероатом), а также смежные с ними связи. Ароматические связи не считаются реакционноспособными. Пользователь может произвольным образом корректировать список реакционноспособных связей.

В заключение приведем пример работы программы COMPASS для компьютерного синтеза изокарбоциклина. На рис. 3 показаны ключевые стадии образования скелета изокарбоциклина по ретросинтетическому пути: из конечного соединения до начальных структур. Подчеркнем, что именно этим выданным компьютером способом и осуществляется практический синтез изокарбоциклина.

Приведенный пример, так же как и все изложенное выше, убедительно доказывает тот факт, что компьютеры необходимы для решения задач планирования синтеза, прогнозирования направления реакции, изучения перегруппировочных процессов. Более того, формализованное представление химической информации позволяет осуществить стратегическое планирование химического эксперимента с участием новых или малоизученных процессов.

Можно считать, что компьютерный синтез является чрезвычайно перспективным направлением органической химии и в ближайшем будущем компьютер высокого класса станет (и уже становится) непременным оборудованием лаборатории органического синтеза.

## ЛИТЕРАТУРА

1. *Бочков А.Ф.*, *Смит В.А.* Органический синтез. М.: Наука, 1987. 304 с.

* * *

Николай Серафимович Зефиров, действительный член Российской Академии наук, профессор, президент Российского общества медицинской химии, директор Института физиологически активных веществ РАН, зав. кафедрой органической химии химического факультета Московского государственного университета им. М.В. Ломоносова, лауреат премии им. М.В. Ломоносова (МГУ), премии им. А.М. Бутлерова, премий ВХО им. Д.И. Менделеева, лауреат Государственной премии СССР. Автор около 900 работ и четырех монографий.