# РАСПОЗНАВАНИЕ ХАРАКТЕРНЫХ 3D-ПОДСТРУКТУР В СЛОЖНЫХ МОЛЕКУЛАХ

**Аникин Н.А., Бобриков В.В., Кузьминский М.Б., Мускатин А.Ю.**
Институт органической химии им. Н.Д.Зелинского РАН, г.Москва

**Аникин Н.А.**
**Бобриков В.В.**
**Кузьминский М.Б.**
**Мускатин А.Ю.**
Институт органической химии им. Н.Д.Зелинского РАН, г.Москва

Задача распознавания химически характеристичных подсистем - фрагментов больших молекул - актуальна для ряда химических, в т. ч. биохимических задач, особенно при поиске в больших молекулярных базах данных (БД) [1]. Наиболее общим и сложным является трехмерное (3D) распознавание заданного фрагмента молекулы по геометрии (декартовым координатам всех атомов) большого числа молекул из БД.

Это более общо, чем 2D-представления связанности атомов типа SMILES [2, 3]. При этом химически разумно использование межатомных расстояний для выявления как валентных межатомных контактов, так и химически значимых характеристичных невалентных контактов, в первую очередь - характерных водородных связей (например, в парах нуклеотидов ДНК). Соответствие всего набора межатомных расстояний фрагментов органических молекул и протеинов в [4, 5] позволило идентифицировать отдельные конформеры фрагментов, но не молекулы как таковые.

Для обеспечения возможности нахождения в БД всех молекул-конформеров, имеющих произвольно заданную трехмерную группу атомов (фрагмент молекулы), в т. ч. с системой межмолекулярных водородных связей, авторами создана программа оригинальной системы поиска соответствующих молекулярных структур. Это особенно актуально при формирование сложной структуры многих биомолекул (вторичная, третичная и четвертичная структура протеинов, образование двойной спирали ДНК) и специфических супрамолекулярных структур, что связано со строго целенаправленной системой межмолекулярных связей. Наша система поиска с использованием химических 3D-индексов для молекул и их фрагментов (индексы молекул и их фрагментов являются естественным объединением индексов их атомов) может считаться расширением известного в литературе языка ATDL [6].

В начале многоэтапного поиска искомого фрагмента, задаваемого пользователем, сначала наша программа выделяет определенную связанную группу атомов во фрагменте молекулы, по которому будет проводиться поиск. Это - пока еще часть фрагмента, назовем ее группировкой атомов. Для каждого атома из этой группировки формируется "индекс" обобщенного молекулярного окружения (валентного плюс межмолекулярного)

а) вначале находятся все атомы - "валентные соседи" данного, удаленные от него ближе расстояния валентного контакта, равного сумме определенным образом масштабированных атомных радиусов;

б) дополнительно отыскиваются все атомы - "межмолекулярные соседи" данного, находящиеся на расстоянии больше расстояния валентного контакта, но меньше расстояния межмолекулярного контакта, превышающего расстояния валентного контакта на специально подобранную величину.

Далее эти атомные индексы еще могут дополняться. Для этого в группировке находится атом с наиболее "редким", по возможности - уникальным индексом, и индексы атомов его ближайшего, более отдаленного окружения, и т.д. После этого для каждой из имеющихся в БД молекул делается попытка найти точно такую группировку.

Для каждого атома молекул из БД формируется индекс обобщенного молекулярного окружения. Проверяется возможность найти атом с индексом, совпадающим с наиболее редким индексом заданной группировки. Если такого атома в данной молекуле нет, то заведомо нет и всей заданной группировки. Если такой атом есть, то продолжается поиск среди его валентных соседей: совпадают ли их индексы атомов для молекул из БД с индексами соответствующих атомов в заданной пользователем группировке. Если найдено хотя бы одно несовпадение таких индексов, то такого атома в данной молекуле нет, и заведомо нет и всей заданной группировки.

В тех отобранных к данному моменту выполнения программы молекулах из БД, в которых найдено полное совпадение индексов атомов группировки (т.е. совпала данная динамически расширяемая нашей программой группировка), снова ищется наиболее редкий индекс, и по нему производится попытка продолжения процесса расширения текущей группировки (найденной к этому моменту части заданного фрагмента), "встроенной" в отобранные молекулы. При наличии альтернатив расширения с идентичным редким индексом вводятся дополнительные эвристические критерии выбора направления расширения, более перспективного для скорейшего завершения процесса поиска. При

успешности такого наращивания процесс продолжается, а при "неудаче" - происходит переход к предыдущей стадии, и возобновляется наращивание по другому направлению (с другим соседним атомом). При успешном окончании этого процесса группировка заведомо содержится в каждой из отбираемых далее молекул.

В тех молекулах, где невозможно наращивание по всем направлениям, которые могли бы привести к успешному полному наращиванию, не содержится заданной группировки атомов. В результате пользователь получает список всех молекул, полностью содержащих заданный фрагмент (включая систему межмолекулярных связей).

Данный тип поиска применим к произвольным органическим и неорганическим молекулам, включая протеины, ДНК, супрамолекулярные структуры, с заданной трехмерной структурой, что принципиально шире возможностей часто применяемого 2D-описания молекулярных структур, особенно при поиске специфических крупных фрагментов со сложной пространственной структурой (с заданной вторичной/третичной структурой протеинов, с двойной спиралью ДНК) и специфических супрамолекулярных структур.

Литература


1. Huang Z., Shen H.T., Zhou X.F., Huang Zi, Shen Heng Tao, Zhou Xiaofang Localized co-occurrence model for fast approximate search in 3D structure databases //IEEE Transactions on Knowledge and Engineering.2008. 20, №4. 519-531.
2. Karthikeyan M., Bender A. Encoding and decoding graphical chemical structures as two-dimensional (PDF417) barcodes // J. Chem. Inf. and Modeling 2005 .45, №3. 572-580.
3. Васильев П.М., Спасов А.А. Языки фрагментарного кодирования структуры соединений для компьютерного прогноза биологической активности // Российский химический журнал. 2006. L, № 2. с. 108-127.
4. Kato H., Takahashi Y. Development of a three-dimensional substructure search program for organic molecules // Bull. Chem. Soc. Japan. 1997. 70, №1. 123-127.
5. Kato H., Takahashi Y. Three-dimensional structural feature search of proteins // Bull. Chem. Soc. Japan. 1997. 70, №7. 1523-1529.
6. Pedretti A., Vistoli G., Atom-Type Description Language: a universal language to recognize atom types implemented in VEGA program // Theor. Chem. Accounts, 109. 2003. №4. 229-232.