Введение в Хемоинформатику
Т.И. Маджидов и др. 2013-2016 г. Казань

Ч.1. Компьютерное представление химических структур
Ч.2. Химические базы данных
Ч.3. Моделирование «Структура-Свойство»
Ч.5. Методы машинного обучения

- **Хемоинформатика-** это мультидисциплинарное научное направление, возникшее на стыке химии, биологии, фармакологии, математики и информатики. Оно занимается обработкой накопленных экспериментальных данных о существующих химических элементах, а также развивает подходы, позволяющие заранее предсказывать химические, физические и биологические свойства новых, в том числе еще не синтезированных соединений.

Направления хемоинформатики

- Разработка компьютерных методов работы со структурной химической информацией, включая создание и оперирование химическими базами данных;
- Моделирование связи между структурами химических соединений и их свойствами;
- Компьютерное планирование синтеза химических соедиений и предсказание путей химических превращений;
- Автоматическая расшифровка структур химических соединений при помощи спектральных методов физико-химического анализа;
- Молекулярный дизайн с использованием данных по структурам биологических мишеней

Основные понятия хемоинформати

- Химическое пространство – набор химических объектов, для которых определено отношение, описывающее их сходство друг с другом
- Дескриптор – это числовой результат некоторого стандартного эксперимента, либо финальный результат математической процедуры, которая однозначно трансформирует структурную информацию о химическом объекте в число

Ч. 1. Представление молекул

- Легкость обработки при помощи компьютера. (Графическое изображение структурной формулы понятно химику, но крайне сложно при использовании компьютеров и поэтому не является кодирующим
- Высокая емкость. Хранимая информация должна занимать наименьший объем при максимальной полезности
- Эффективность. Желательно, чтобы для работы с кодирующими представлениями могли применяться высокоэффективные алгоритмы обработки информации
- Уникальность. Желательно, чтобы одной молекуле соответствовало одно представление. Процесс выбора уникально представления из множества возможных вариантов называется канонизацией.
- Однозначность. Каждому представлению в идеальном случае должна соответствовать только одно молекула. (Не удовлетворяет брутто-формула).

Ч.2. Химические базы данных

- Классификация баз данных. (1 Библиографические, полнотекстовые, фактографические.
- Структурный поиск в химических базах данных: поиск по структуре, поиск по подструктуре, поиск по подобию
- Важнейшие базы данных

Ч.3. Моделирование «структура-свойство»

- Задачей моделирования «структура-свойство» является создание статистических моделей, которые на основании структуры могут предсказать их свойства. Исторически, эти методы ассоциируются с исследованием биологической активности молекул, поэтому за отраслью закрепилось название QSAR- (Quantittative Structure-Activity Relationships). Вместе с тем, моделирование «структура-свойство» используется также в создании полимеров, материалов, катализаторов, композитов, реагентов, экстрагентов, ПАВ, ионных жидкостей и в целом для предсказания полезных для практ. целей свойств: спектров, растворимости, температур плавления, кипения и т. д.

Ч. 4. Методы машинного обучения

- Машинное обучение – это раздел искусственного интеллекта, рассматривающий методы построения алгоритмов и на их основе программ, способных обучаться. Обучение обычно ведется путем предъявления эмпирических данных (называемых прецедентами или наблюдениями), в которых выявляются закономерности, и на их основе строятся модели, позволяющие в дальнейшем прогнозировать определенные характеристики (называемые ответами) для новых объектов.