Курс “Введение в Хемоинформатику”

Дисциплина относится к вариативной части учебного цикла магистерских программ на физическом факультете МГУ и является спецкурсом по выбору кафедры физики полимеров и кристаллов.

Аннотация. Предпосылкой становления и быстрого роста нового научного направления – Хемоинформатики явился беспрецедентный рост объемов информации в сфере естественных наук –физики, химии, биологии и смежных с ними - биохимии, материаловедении, медицины и др., сопровождающийся быстрым развитием производительности компьютеров и компьютерных технологий. Реализация научно-исследовательских проектов в этих областях требует привлечения новейших методов информатики, разработанных для работы с научной информацией. Одной из задач хемоинформатики является создание компьютерных способов организации, хранения и систематизации данных, содержащих информацию о структурах и свойствах химических объектов в Базах данных, а также развитию методов поиска и анализа этих данных. Важнейшим направлением хемоинформатики является прогнозирование свойств новых химических объектов и конструирования соединений и материалов с заранее заданными свойствами. Стратегия основана на поиске закономерностей, связывающих микроскопические свойства молекулярных компонентов (молекулярную структуру) этих объектов с их макроскопическими (физико-химическими и биологическими) свойствами путем построения статистических моделей, количественно или качественно связывающие параметры структуры и свойств. Модели “структура-свойство” (SAR/QSPR/QSAR) строят с помощью методов машинного обучения. Такой подход применим к различным типам молекул и свойств и позволяет найти сложные непрямые взаимосвязи между микро- и макроскопическими параметрами соединений и материалов. Методы хемоинформатики востребованы во всех областях современных исследований– физике, химии, биологии, медицине, материаловедении, нанотехнологиях, электронике и др. Специалисты, владеющие методами хемоинформатики, широко востребованы в академической науке и в прикладных областях.

Цели курса. Курс направлен на подготовку специалистов, способных проводить научно-исследовательскую работу с применением методов хемоинформатики.

Задачи курса. Приобретение знаний о базовых понятиях научного направления хемоинформатика, знакомство с набором практических методов хемоинформатики возможностями их применения в научно-практических целях. В том числе: (1) освоение практических навыков работы с базами данных, содержащих информацию о структурах и свойствах химических объектов; (2) знакомство с методами машинного обучения, применяемыми для исследования взаимосвязей между микро- и макроскопическими

свойствами химических объектов и построения моделей QSPR; (3) приобретение практического опыта работы с компьютерными программами для построения моделей “структура-свойство” и прогнозирования свойств соединений и материалов.

Базовые курсы, необходимые для слушателей: общий курс математики, математическая статистика, линейная алгебра, программирование и информатика, английский язык.

Структура и содержание: 36 учебных часа (14 лекций+ 4 семинарских занятий)

План курса «Введение в Хемоинформатику»

Лекция 1. Вводная. Роль Хемоинформатики в современных научных исследованиях.. Основные задачи и методы Хемоинформатики: организация хранения, анализа и поиска информации о структуре и свойствах химических объектов в базах данных; компьютерное представление структуры химических объектов; моделирование взаимосвязи между микро- и макроскопических свойствами химических объектов (SAR/QSAR/QSPR), прогнозирование характеристик новых объектов и конструирование материалов с заранее заданными свойствами на основе построения моделей “структура -свойство” с помощью методов машинного обучения. Общий протокол Хемоинформатики.

Тема 1. Представление химических объектов в Хемоинформатике.

Лекция 2. Основные типы химических объектов и способы описания их строения. Строение молекул. Методы определения молекулярной структуры. Типы химических связей и их характеристики. Геометрия молекул. Конфигурация и конформация. Типы изомерии. Понятие о динамической стереохимии. Кислотные и основные свойства молекул. Супрамолекулярные системы.
Методы описания электронной структуры молекулярных систем. Общие представления о современных методах квантовой химии. Молекулярные орбитали. Распределение электронной плотности. Парциальные заряды на атомах. Молекулярный электростатический потенциал. Поверхность потенциальной энергии (ППЭ) молекулярной системы.

Лекция 3. Виды и особенности представлений химических структур в Хемоинформатике. Кодирующие представления. Структурная диаграмма. Понятие молекулярных графов. Базовые элементы теории графов. Линейные нотации как представления графов (SMILES, их правила и форматы; нотации SMARTS; SLN). Коды InChI. Векторные представления графов, битовая строка. Структурные ключи, молекулярные отпечатки, хэшированные молекулярные отпечатки.

Лекция 4. Виды и особенности представлений химических структур в Хемоинформатике (продолжение). Матричные представления графов, виды матриц. Таблицы связности. Структуры Маркуша. Трехмерные представления молекул. Координатные представления. Виды трехмерных представлений. Стандартные форматы файлов в Хемоинформатике. Основные форматы файлов химических структур (mol, sdf, mol2,). Конвертация между представлениями различного уровня 1D-2D-3D. Основные программы конвертации.

Семинар 1. Ввод и редактирование структур молекул с использованием интерактивных графических редакторов. Создание файлов в стандартных форматах, содержащих целевое представление молекул. Работа с программой MarvinSketch из комплекса ChemAxon. Перекодировка представлений молекул с использованием свободно доступного программного обеспечения (программа OpenBabel). Визуализация файлов, содержащих структуры: малых молекул (с помощью программы MarvinView из комплекса ChemAxon), кристаллов

низкомолекулярных соединений и неорганических материалов (с помощью программы Mercury), а также макромолекул (с помощью программы MarvinSpace из программного комплекса ChemAxon, а также программного комплекса Chimera).

Тема 2. Моделирование взаимосвязи “структура-свойство” (SAR/QSAR/QSPR, structure-activity relationships/quantitative structure-activity/property relationships)

Лекция 5. Методология моделирования взаимосвязи “структура-свойство”. Концепция молекулярных дескрипторов. Классификация и характеристики. Топологические(2D) дескрипторы: фрагментные дескрипторы, топологические индексы. Трехмерные (3D) дескрипторы: геометрические, дескрипторы поверхности. Фармакофорные дескрипторы. Физико-химические дескрипторы. Квантово-химические дескрипторы. Дескрипторы молекулярных полей. Дескрипторы молекулярного подобия. Компьютерные программы и ресурсы для расчета дескрипторов.

Лекция 6. Построение и валидация моделей “структура-свойство”. Предобработка данных. Общие принципы построения моделей “структура-свойство”. Метод наименьших квадратов. Понятие о переобучении и принцип оптимальной сложности моделей. Принципы отбора дескрипторов. Общие принципы валидации моделей. Понятие о внутреннем и внешнем, перекрестном и скользящем контроле. Количественные показатели качества регрессионных моделей. Количественные показатели качества классификационных моделей. Оценка качества моделей для виртуального скрининга: ROC-кривые. Понятие об области применимости моделей.

Лекция 7. Регрессионные методы машинного обучения, используемые для построения моделей “структура-свойство”. Множественная линейная регрессия. Метод частичных наименьших квадратов (PLS). Регрессия на опорных векторах. Многослойные нейронные сети.

Лекция 8. Классификационные методы машинного обучения, используемые для построения моделей “структура-свойство”. Метод ближайших соседей. Машина опорных векторов. Деревья решений. Случайный лес. Метод «наивного» Байеса.

Семинар 2. Программы моделирования “структура-свойство”. Построение регрессионных и классификационных моделей “структура-свойство” с помощью программных комплексов OChem, WEKA и NASAWIN.

Лекция 9. Программы моделирования “структура-свойство”. 1. Построение регрессионных моделей “структура-свойство” с помощью программного комплекса ISIDA-QSPR

Семинар 3. Программы моделирования “структура-свойство”. 2. Построение регрессионных моделей “структура-свойство” с помощью программного комплекса ISIDA-QSPR .

Лекция 10. Информатика материалов. Особенности построения моделей «структура-свойство» для разных типов материалов. Моделирование свойств наноматериалов, кристаллов, керамики, сплавов металлов, гетерогенных катализаторов, поверхностно-активных веществ и др.

Тема 3. Базы данных в Хемоинформатике

Лекция 11 . Общие сведения о химических базах данных и их особенностях. Классификация баз данных. Характеристика важнейших баз данных, содержащих информацию о структурах и свойствах соединений, а также информацию о спектрах и кристаллах (CAS/SciFinder, Cambridge Structural Database , PubChem , ZINK, Protein Data Bank, ChemSpider, GDB-13, Polinfor и др. Виды поиска в базах данных. Структурный поиск. Подструктурный поиск. Поиск по молекулярному сходству. Поиск по структурам Маркуша. Поиск в базах данных трехмерных структур. Понятие о фармакофорах, поиск по фармакофорам.

Семинар 4. Работа on-line с общедоступными химическими базами (PubChem , ZINK, ChemSpider). Создание базы данных по структурам и свойствам химических соединений с использованием программного комплекса ChemAxon.

Тема 4. Избранные методы Хемоинформатики

Лекция 12. Общее представление о методах машинного обучения «без учителя». Понятие о методах понижения размерности данных. Метод главных компонент. Самоорганизующиеся карты Кохонена. Понятие о картографии химического пространства и пространства материалов. Понятие о методах кластерного анализа.

Лекции 13 и 14. Виртуальный скрининг и принципы его использования для дизайна новых химических структур и материалов с заданными свойствами. «Воронка» виртуального скрининга. Типы фильтров для виртуального скрининга. Понятие о молекулярном докинге. Методы формирования виртуальных библиотек химических соединений и материалов.

Программное обеспечение:

ChemAxon (MarvinSketch, MarvinView, MarvinSpace) - www.chemaxon.com
OpenBabel - openbabel.org
Mercury - http://www.ccdc.cam.ac.uk/Solutions/CSDSystem/Pages/Mercury.aspx
Chimera - http://www.cgl.ucsf.edu/chimera/download.html
IDISA_QSPR vpsolovev.ru/programs/isidaqspr
Weka http://www.cs.waikato.ac.nz/ml/weka/downloading.html
OChem https://ochem.eu