Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Панченко Святослав Константинович

Геометрическая алгебра для решения задачи декодирования сигналов

03.03.01 — Прикладные физика и математика

Выпускная квалификационная работа магистра

Научный руководитель:
д.ф.-м.н. Стрижов Вадим Викторович

Москва
2020

Содержание

[ДАЛЕЕ СЛЕДУЕТ ШАБЛОН, АКТУАЛЬНЫЙ ТЕКСТ РАБОТЫ БУДЕТ ПРЕДОСТАВЛЕН 31-ОГО МАЯ]

Аннотация

Рассмотрена задача восстановления плотности распределения трёхмерного случайного вектора, две компоненты которого представляют собой пару сферических углов. Требуется, чтобы полученные плотности были интерпретируемы с точки зрения эксперта, согласовывались с ранее полученными результатами, а модель восстановления учитывала периодичность углов. Предлагается рассматривать значения пары углов как реализации случайного вектора, областью значений которого является сфера в трёхмерном пространстве. Искомая плотность в таком подходе моделируется в виде смеси, в каждой компоненте которой углы распределены в соответствии с распределением Кента. Параметры смеси находятся с помощью модификации алгоритма Stochastic Expectation-Maximization. Проведено восстановление плотностей распределения пространственных ориентаций различных пар молекул. Демонстрируется, что результаты восстановления согласуются с мнением эксперта и результатами других моделей.

Ключевые слова: *трёхмерная структура белка, восстановление плотности распределения, модель смеси распределений, распределение Кента, алгоритм Stochastic Expectation-Maximization.*

Введение

Актуальность темы. Трёхмерная структура белковой молекулы — это ключ к пониманию её биологических функций и свойств. Однако, определение строения белковой цепи, обычно с помощью рентгеновской кристаллографии или спектроскопии ядерного магнитного резонанса, весьма дорого и трудоёмко. Поэтому число экспериментально определённых белковых структур существенно меньше числа идентифицированных белковых цепочек. Отсюда возникает задача предсказания по последовательности образующих молекулу компонент её трёхмерной структуры. С проблемой можно ознакомиться в работах [1, 2].

Решение данной задачи — одна из самых важных целей биоинформатики и теоретической химии. Оно позволит значительно улучшить существующие генеративные и предсказательные модели в области исследования молекулярных последовательностей. Полученные при помощи этих моделей данные активно используются в медицине и биотехнологии.

Существующие решения. Один из фундаментальных подходов к решению данной задачи — построение потенциала, функции, минимумы которой соответствуют энергетически устойчивым конфигурациям молекул, образующих химическую связь. Имеются два основных подхода к построению таких потенциалов:

Физический подход: в этом подходе потенциал строится на основе законов молекулярной химии, описывающих взаимомодействия молекул. Такие потенциалы, как правило, точны, но их вычисление весьма трудоёмко. Они лучше подходят для описания одной конкретной цепочки и мало применимы для анализа произвольных последовательностей. Примеры использования подобных подходах рассмотрены в исследованиях [3, 4].

Статистический подход: в этом подходе предполагается стохастическая модель порождения данных: для каждой пары потециально взаимодействующих молекул на основе известных и изученных молекулярных структур строятся функции плотности совместной вероятности параметров химической связи, определяющих взаимную пространственную ориентацию этих молекул. С помощью полученных плотностей формируются статистические потенциалы, описывающие структуру неизученной молекулярной цепи на основе вероятностного обобщения известных структур. Многочисленные исследования в этой области представлены в статьях [5–10].

Второй, статистический, подход существенно опирается на оценку совместных плотностей распределения величин, характеризующих молекулярную связь. Для определённого набора хорошо изученных молекул существуют базы данных, содержащие информацию о параметрах химических связей, которую эти молекули формировали между собой в исследованных структурах. Одна пара таких молекул характеризуется десятками тысяч зарегистрированных конфигураций с различными параметрами. Такое множесто конфигураций объясняется тем, что рассматриваемая пара молекул входила в состав громадного количества различных молекулярных последовательностей, каждый элемент которых мог повлиять на значения этих параметров.