МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Университет «Дубна»

Институт системного анализа и управления

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
БАКАЛАВРСКАЯ РАБОТА

Тема: Разработка и оптимизация оценочной функции для учета межмолекулярных взаимодействий в белковых комплексах

Ф.И.О. студента Никулин Даниил Андреевич

Группа 4181 **Направление подготовки** 01.03.02 Прикладная математика и информатика

Направленность (профиль) образовательной программы Математическое моделирование

Выпускающая кафедра распределенных информационно-вычислительных систем

Руководитель работы _____ /ст. преп. Полуян С.В. /

Консультант(ы) _____ /_____/

_____ /_____/

Рецензент _____ /доцент, к.б.н. Белов О.В./

Выпускная квалификационная работа
допущена к защите **«____»_____20___г.**

Заведующий кафедрой _____ /Кореньков В.В./

г. Дубна

АННОТАЦИЯ

В настоящей работе рассматриваются белковые комплексы вида белок-белок, где в качестве компонент комплекса выступают белки, представленные в полноатомном виде. При моделировании процесса образования устойчивого комплекса компонентами при их нековалентном взаимодействии друг с другом возникает необходимость в вычислении энергии такого взаимодействия. Одним из существующих методов для вычисления энергии взаимодействия является использование оценочной функции, которая для заданной пространственной конфигурации компонент позволяет приближенно оценить искомую энергию. В исследовании приведено описание разработанной оценочной функции, в которой учитываются силы межатомных взаимодействий, представленные эмпирическими потенциалами Кулона и Леннард-Джонса. Молекулы растворителя в явном виде не рассматриваются, для этого энергия сольватации вычисляется в рамках модели неявного растворителя. При помощи k-d-дерева произведена оптимизация этапа поиска взаимодействующих атомов между различными компонентами комплекса. Для тестового набора комплексов приведены результаты применения оценочной функции, которые показывают приемлемый уровень корреляции выполняемых оценок при сравнении с существующими инструментами. В различных численных экспериментах продемонстрированы результаты оптимизации оценочной функции, которые демонстрируют уменьшение времени выполнения оценки энергии взаимодействия.

In this work protein-protein complexes considered in full-atom form and consists of several components. Protein-protein complex component assembly is guided by the establishment of non-covalent interactions. To estimate the strength of such interactions at different steps of binding score functions are usually used. In this study presented score function that estimate energy for the given spatial configuration of the protein subunits in complex. Score function considers interatomic interactions through Coulomb and Lennard-Jones potentials. Solvent molecules are not considered explicitly. The implicit solvent model is used to calculate solvation energy. The k-d-tree was used to optimize the search for interacting atoms between different components of the complex. The result of using score fucntion on the test set of protein-protein complexes is presented. It demonstrates an acceptable level of correlation compared to existing tools. The work presents results of various numerical experiments for a test set of different protein-protein complexes which demonstrates an acceptable decrease in the time of score evaluation in comparision with other score instruments.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Функцию для оценки энергии взаимодействия лиганда с белком в заданной пространственной конфигурации называют «оценочной функцией» (scoring function) [1]. В настоящее время разработано множество оценочных функций, которые подразделяются на группы, исходя из принципов их построения. Например, распространено нестрогое деление на эмпирические, статистические и функции на основе силовых полей [2]. Помимо точности оценки искомой энергии взаимодействия важными критерием выбора оценочной функции является вычислительная сложность процедуры оценки, поэтому при моделировании взаимодействия на больших временных масштабах прибегают к моделям с упрощенным представлением белков [3], а также исключают конформационную подвижность, рассматривая компоненты комплекса как «твёрдые» тела, совершающие в растворителе только поступательные и вращательные движения.

Целями работы являются: разработка, реализация, оптимизация и верификация оценочной функции для учета межмолекулярных взаимодействий в белковых комплекса в рамках библиотеки для полноатомного моделирования белковых комплексов PSM [4] (protein structure modeling). Разработка библиотеки PSM проходит в рамках НИР в университете «Дубна», которая позволяет моделировать процесс образования белкового комплекса с помощью кинетического метода Монте-Карло[5]. В основе метода лежит классическая теория переходного состояния, где в процессе моделирования система движется в сторону наименьшей полной энергии по пути с наименьшими энергетическими барьерами, что позволяет модельной системе на пути к термодинамическому равновесию проходить через последовательность квазиравновесных состояний. На определенном этапе метода требуется выполнить оценку взаимодействия, как представлено на рис. 1.

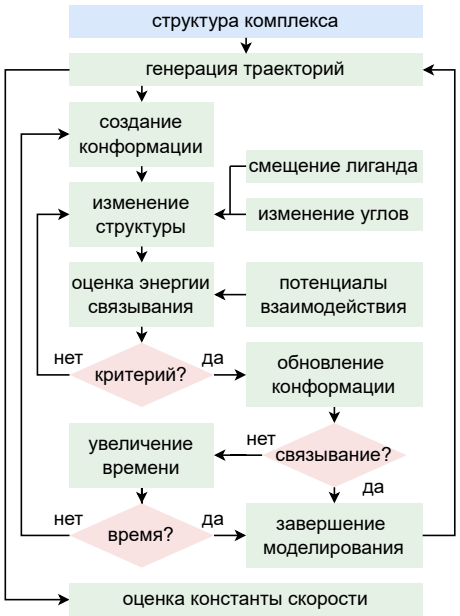

Рис. 1. Основные этапы работы кинетического метода Монте-Карло

При этом, поскольку метод является стохастическим, этап оценки взаимодействия повторяется значительное количество раз. В связи с этим актуальным становится оптимизация этого этапа без существенной потери качества выполняемых оценок. Следует отметить, что для моделирования процесса образования комплекса достаточно сформировать оценочную функцию, учитывающую только парные межатомные взаимодействия и влияние растворителя[3].

Существует множество инструментов для оценки энергии взаимодействия. Например, выполнить оценку взаимодействия возможно с помощью фреймворка Rosetta [6] или силового поля CHARMM [7]. Следует отметить, что для выполнения оценки требуется представление белкового комплекса внутренними средствами инструмента, что, как правило, довольно накладно с временной точки зрения. Например, в случае применения CHARMM перед выполнением оценки требуется перевод структуры во внутренний формат силового поля непосредственно из файла со структурой, а в случае применения Rosetta при выполнении оценки происходит обновление энергетической карты, которое также влияет на время выполнения оценки.

Разработанная в настоящей работе оценочная функция напрямую интегрирована в библиотеку PSM, что позволяет снизить вычислительную сложность выполнения оценки по сравнению с использованием внешних инструментов, что подчеркивает актуальность выполненной работы.

1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

Электростатические взаимодействия между мозаично распределенными на поверхности белка электрическими зарядами являются основным фактором, определяющим специфичность взаимодействий в белковых комплексах [8]. Поэтому при разработке оценочной функции для моделирования компонент в виде «твёрдых» тел без конформационных изменений возможно опустить оценку ковалентных и внутримолекулярных нековалетных взаимодействий.

В исследовании для описания межмолекулярных взаимодействий использовались общепринятые эмпирические парные потенциалы Леннард–Джонса и Кулона. Энергия растворителя вычислялась в рамках модели неявного растворителя EEF1 [9]. Процедура оценки энергии с использованием оценочной функции разделяется на два этапа.

На первом этапе формируется список взаимодействующих пар атомов в зависимости от заданного радиуса сферы взаимодействия между атомами указанных компонент белкового комплекса. Для того, чтобы сформировать такой список возможно использовать прямой попарный перебор всех атомов с определением евклидова расстояния. Очевидно, что такой подход очень затратен с вычислительной точки зрения, поскольку каждый компонент комплекса состоит из нескольких тысяч атомов. Указанный этап оптимизируют с использование различных структур данных. Описание выбранной структуры для оптимизации времени поиска атомов приведено в разделе 1.2.

На втором этапе для каждой пары взаимодействующих атомов выполняется оценка энергии взаимодействия, а получившиеся значения суммируются формируя общую оценку взаимодействия для компонент белкового комплекса. Существуют различные подходы для оптимизации указанного этапа, однако они выходят за рамки выполняемой работы.

Разработанная оценочная функция состоит из трёх слагаемых:

$$F_s = E_v + E_c + E_s, \tag{1.1}$$

где E_v представляет собой оценку энергии, полученную с помощью классического потенциала Леннард–Джонса «6–12», слагаемое E_c является оценкой энергии, полученной с помощью потенциала Кулона, E_s – значение оценки энергии растворителя. В следующем разделе для каждого компонента оценочной функции приведено описание с указанием теоретических основ деталей применения.

Поскольку разработка оценочной функции невозможна без учёта физических параметров атомов белка и растворителя в работе использовались данные широко распространённого силового поля CHARMM36 [7]. Принципы применения CHARMM и описание используемых файлов силового поля для выполнения оценок представлены в разделах ниже.

1.1. Компоненты оценочной функции

1.1.1. Потенциал Кулона

Электростатический потенциал Кулона описывает взаимодействие двух постоянных точечных зарядов и определяется следующим образом

$$E_c = \sum_{i,j} \left(\frac{1}{4\pi\varepsilon_r} \frac{q_i q_j}{d_{ij}} \left[\frac{d_{ij}^2}{k^2} - \frac{2d_{ij}}{k} + 1 \right] \right), \tag{1.2}$$

где d_{ij} – евклидово расстояние между центрами атомов, q_i и q_j – фиксированные частичные атомные заряды в рассматриваемой паре атомов, ε_r – диэлектрическая константа. Атомные заряды для каждого атома в зависимости от типа получаются с помощью программы PDB2PQR [10]. Вместо первой дроби при вычислениях используется константа, применяемая в силовом поле CHARMM: 332.0716 ккал·Å·e^{-2}/моль. Последний множитель с коэффициентом $k = 14$Å определяет радиус сферы взаимодействия, для $d_{ij} > k$ вычисления не производятся.

Приведенный потенциал Кулона рассматривается не в классическом виде. Как видно в сумме 1.2 используется дополнительный полиномиальный множитель – квадратная функция, которая необходима для сглаживания значений потенциала при использовании коэффициента отсечения. Поскольку использование такого множителя применяется в силовом поле CHARMM принято решение использовать идентичный принцип вычисления потенциала. На рис. 1.1 представлены результаты вычисления потенциала Кулона без дополнительного множителя и с использованием дополнительным множителя.

1.1.2. Потенциал Леннард-Джонса

Потенциал Леннард-Джонса является моделью парного взаимодействия атомов, которая представляет собой математическую функцию описывающую силы притяжения и отталкивания между атомами на основе их расстояния. Потенциал состоит из двух компонентов, которые позволяют смоделировать эффекты притяжения и отталкивания.

Потенциал Леннард-Джонса «6–12» вычисляется по формуле

$$E_v = \sum_{i,j} \left(\varepsilon_{ij} \left[\left(\frac{R_{ij}}{d_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}}{d_{ij}} \right)^{6} \right] \right), \; R_{ij} = \frac{R_i}{2} + \frac{R_j}{2}, \; \varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}, \tag{1.3}$$

где d_{ij} – евклидово расстояние между центрами атомов, R_i и R_i – расстояния, на которых значение потенциала становится равным нулю, ε_i и ε_j – глубины потенциальных ям. Указанные параметры получены из файла топологии для соответствующих типов атомов в рассматриваемой паре атомов с индексами i и j.

Необходимо отметить, что потенциал Леннард-Джонса в классическом виде 1.3

не используется в силовом поле CHARMM. Для описания сил Ван-дер-Ваальса используется двойной экспоненциальный потенциал [11]. Он позволяет более точно оценить энергию, поскольку использует отдельные функции для оценки эффекта притяжения и отталкивания, а также отдельную процедуру для учета дальнодействующих взаимодействий. Отличия в получаемых оценках продемонстрированы в разделе 2.4.2.

На рис. 1.1 приведены значения потенциала в зависимости от расстояниям между двумя атомами: углерода ($\varepsilon = 0.11, R/2 = 2$) и водорода ($\varepsilon = 0.031, R/2 = 1.25$).

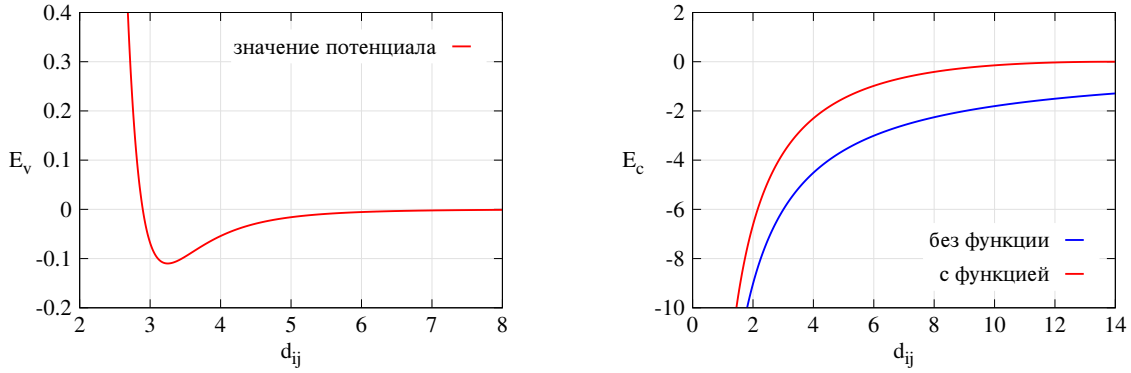

Рис. 1.1. Потенциал Леннард–Джонса. Потенциал Кулона с применением дополнительного множителя и без применения дополнительного множителя

1.1.3. Неявный растворитель

Знание эффективной энергетической гиперповерхности биологических макромолекул в растворе имеет фундаментальное значение для понимания их свойств. Эффективная энергия (потенциал средней силы) для данной конформации макромолекулы представляет собой свободную энергию системы, состоящей из макромолекулы и растворителя, усредненную по всем степеням свободы растворителя при заданной температуре. Она состоит из внутримолекулярной энергии (энергия макромолекулы в вакууме) и энергии свободного растворения (свободная энергия переноса макромолекулы из газовой фазы в раствор). Общая свободная энергия системы макромолекула-растворитель в данной области гиперповерхности является суммой средней эффективной энергии и конфигурационной энтропии. В физиологических условиях обычно предполагается, что белки стабильны в окрестности глобального минимума (родной конформации), несмотря на большой прирост конфигурационной энтропии в денатурированном состоянии. Это называется "термодинамической гипотезой"стабильности белка, и имеющиеся данные свидетельствуют о ее справедливости для большинства малых однодоменных белков. Даже там, где эргодичность, кажется, нарушается, родное состояние должно соответствовать глубокому минимуму на эффективной энергетической гиперповерхности, который отделен от глобального минимума барьерами, слишком высокими для преодоления на экспериментальных временных шкалах. Теоретические прогнозы структуры белка из последовательно-

сти аминокислот и анализ механизма складывания требуют знания эффективной энергетической гиперповерхности и метода эффективного поиска конформационного пространства для определения расположения минимумов.

Всеатомные силовые поля, используемые в молекулярной механике и динамике симуляций макромолекул, дают энгию белка в вакууме. Только введя явный растворитель и проведя симуляции, показывающие как макромолекулу, так и растворитель, можно учитывать эффекты соляции. Из результатов симуляций и теоретических соображений ясно, что эффективная гиперповерхность энергии, включающая растворитель, значительно отличается от внутримолекулярной гиперповерхности энергии. Хотя внутримолекулярная гиперповерхность энергии часто имеет глубокий минимум около нативной конформации, она может иметь другие равно глубокие или более глубокие минимумы в отдаленных областях пространства конформации. Это может возникнуть из-за ряда эффектов. Например, взаимодействия полярных-неполярных групп в газовой фазе энергетически выгодны так же, как и полярные-неполярные взаимодействия. Однако в воде полярно-неполярные взаимодействия эффективно отталкивают из-за высокого штрафа за десольватацию полярной группы. Кроме того, гидрофобная гидратация делает эффективные взаимействия между неполярными группами в воде сильнее, чем в газовой фазе. Таким образом, вода способствует стабилизации нативного состояния с зарытыми неполярными группами и помогает гарантировать его уникальность. Взаимодействия противоположных зарядов менее стабилизирующие в воде, чем в вакууме. Таким образом, проблемы, связанные с складыванием и стабильностью белков, не могут быть решены без учета соляции.

Хотя за последние 40 лет был сделан значительный прогресс в статистической теории жидкостей, все еще отсутствуют простые и точные модели водной соляции. В настоящее время наиболее надежным методом учета соляции является симуляция белка в присутствии явных молекул воды. Однако этот подход связан с большими вычислительными затратами, и большая часть времени компьютера в таких симуляциях уходит на расчет взаимодействий растворителя-растворителя. Например в симуциях денатурации барназа система состояла из 1100 атомов белка и около 9000 атомов растворителя. Включение такого большого количества атомов растворителя ставит серьезные ограничения на тип проблем, которые можно изучать. Например, хотя динамику нативных состояний можно эффективно симулировать с помощью явных моделей растворителя, симуляции раскладки (и раскладанных состояний) на временных масштабах, необходимых для выборки достаточного количества различных начальных состояний для получения значимых сходимых результатов, пока что невозможны, за исключением, возможно, маленьких пептидов.

Еще одним ограничем симуляций явной соляции является то, что разница свободной энергии не получается прямым способом. Например, точное моделирование воды в нативной и неправильно свернутой структурах не показывает, какая из них имеет наименьшую свободную энергию. Требуется особая симуляция, включающая

обратимый переход от одной конформации к другой или выборка по зонтику, чтобы можно было рассчитать необходимые различия в свободной энергии. Такие симуляции были проведены главным образом для малых молекул, таких как бутан, дипептиды и другие маленькие пептиды. Недавно потенциал средней силы для маленького белка относительно радиуса инерции в качестве параметра порядка был рассчитан с явным растворителем; для этой симуляции потребовалось 450 часов процессорного времени на 64-узловом T3E-суперкомпьютере (эквивалент 4 месяцам на стандартной рабочей станции). Таким образом, использованте этого подхода для изучения деталей сложенных и раскладанных состояний ряда белков, а также многих других проблем текущего интереса, все еще невозможно.

Чтобы преодолеть ограничения вакуумных расчетов, с одной стороны, и явных симуляций растворителя, с другой, было разработано множество невных моделей соляции для белков, которые объединяют эмпирическое силовое поле для внутримолекулярных взаимодействий в вакууме с коррекцией сольвации. Последняя получается путем рассмотрения переноса всего белка или соответствующих составных групп из газовой фазы в воду. Очень простая модель использовала потенальную энергию CHARMM и коррекцию соляции на основе площади поверхности атома; были включены пять различных типов атомов. Та же модель соляции была объединена с силовым полем AMBER Шиффера. Фратернали и ван Гунстерен предложили еще более простую модель для использования в молекурной динамике (MD); эта модель также основана на доступных поверхностях и использует только два параметра - один для неполярных и один для полярных групп.

Хотя вышеупомянутые модели оценивают эффект соляции в терминах площади поверхности, нет фундаментального теоретического обоснования для этого выбора. Модель, которая не использует площадь поверхности, - это модель гидратационной оболочки. Она предполагает, что свободная энергия гидратации группы возникает из первой гидратационной оболочки и пропорциональна объему гидратационной оболочки, доступной растворителю (т.е. не занятой другими атомами растворителя). Еще один тип модели соляции основан на контактах, которые каждая группа устанавливает с другими атомами растворителя. Чем больше контактов, тем меньше величина свобной энергии соляции группы, и контакты взвешиваются в соответствии с некоторой функцией их расстояния от группы. Эта модель похожа на подходы, использованные ранее Гибсоном, Шерагой и Левиттом. Физически, модель Колонна-Чезари-Сандера похожа на модели площади поверхности, но она намного быстрее в использовании, потому что подсчет количества контактов занимает значительно меньше времени, чем даже самые эффективные аналитические методы расчета площади поверности. Кроме того, аналитические производные свободной энергии соляции могут быть легко получены для оценки сил, необходимых для минимизации энергии и молекулярной динамики. Версия этой модели была параметризована на основе свободных энергий гидратации малых молекул. Для каждой группы назначается

параметр соляции. Модель была объединена с полем сил GROMOS и использована в стохастической динамической симуляции инибита трипсина поджелудочной железы крупного рогатого скота (BPTI). Было обнаружено, что полученные структуры были разумными, хотя отклонение от кристалличкой структуры было немного больше, чем в симуляциях с явной водой.

Другой набор моделей сольватации рассматривает весь белок сразу и основан на континуальной электростатике и линеаризованном уравнении Пуассона-Больцмана. Поскольку численные решения уравнения Пуассона-Больцмана дороги, были предложены полуаналитические или аналитические приближения. Стилл ввел простое обобщение формулы Борна на многоатомные молекулы. Позднее обобщенное уравнение Борна было объединено с методом интегрированного поля для собственных энергий, что дало полностью аналитическую трактовку электростатических энергий и сил.38 Были сделаны приложения к ряду простых систем, и вполне вероятно, что этот подход будет использоваться более широко в будущем.

Описание модели.

Эффективная энергия $W(R^M)$ макромолекулы с координатами R^M в решении может быть записана следующим образом

$$W(R^M) = H_{intra}(R^M) + \Delta G^{slv}(R^M), \tag{1.4}$$

где

- H_{intra} – внутримакромолекулярная энергия
- ΔG^{slv} – свободная энергия растворителя

Чтобы получить данное уравнение 1.4, единственное предположение состоит в том, что гамильтониан является сепарабельным; то есть это сумма терминов раствор-раствор, раствор-растворитель и растворитель-растворитель. Так обстоит дело с большинством эмпирических функций энергии, которые не включают поляризацию. Недавняя теоретическая работа по термодинамике сольватации показала, что свободная энергия сольватации ΔG^{slv} заданной конформации R^M может быть записана как интеграл по окружающему пространству; то есть,

$$\Delta G^{slv} = \int f(r)dr, \tag{1.5}$$

где $f(r)$ – плотность свободной энергии сольватации в точке r. Она состоит из энергии расствореного вещества-растворителя, энергии преобразования растворителя, энтропии раствора-растворителя и энтропии преобразования растворителя. Ожидается, что плотность свободной энергии растворителя сильно зависит от расстояния. Ее величина достигает максимума вблизи состояния раствора и стремится к нулю при отдалении от этого состояния. Когда две молекулы раствора приближаются друг к другу или изменяется конформация многоатомного раствора, сольватация каждой

группы изменяется из-за двух эффектов: исключение растворителя из пространства, которое занято другими группами раствора, и изменение плотности ориентационного распределения растворителя в пространстве, которое не занято раствором. Для электростатических взаимодействий собственная энергия зарядов относится к первой категори, а диэлектрическое экранирование ко второй. В представленной модели упускается второй эффект для неполнярных групп, поскольку ожидается, что он будет незначительно малым, но он частично учитывается для полярных грпп за счет использования диэлетричской проницаемости, котороая зависит от расстояния. Предполагается, что для многоатомного раствора свободную энергию сольватации можно записать как сумму групп, то есть,

$$\Delta G^{slv} = \sum_i \Delta G_i^{slv}, \tag{1.6}$$

где ΔG_i^{slv} – свободная энерегия сольватации группы i. Выражение 1.6 может быть формально получено путем рассмотрения энергии взаимодействия раствора и растворителя как суммы взаимодействия группы и растворителя и корреляционной функции раствора и растворителя как произведения корреляционных функций группы и растворителя. Принимая во внимание только эффект исключения растворителя, можно записать

$$\Delta G_i^{slv} = \Delta G_i^{ref} - \sum_j \int_{V_j} f_i(r) dr, \tag{1.7}$$

где ΔG_i^{ref} (эталонная свободная энергия сольватации) – свободная энергия сольватации i в молекуле, выбранной подходящим образом, в которой группа i практически полностью подвергается воздействию растворителя. Интеграл в выражении находится по объему V_j группы j, а суммирование происходит по всем группам j вокруг i. Для упрощения вычислений интеграл по $f_i(r)$ заменяется произведением $f_i(r_{ij})V_j$, то есть

$$\Delta G_i^{slv} = \Delta G_i^{ref} - \sum_{j \neq 1} f_i(r_{ij}) V_j, \tag{1.8}$$

где r_{ij} – расстояние между i и j. Выражние 1.8 сообщает, что свободная энергия сольватации группы i равна энергии в модельной системе ΔG_i^{ref} за исключением сольватации из-за присутствия окружающих групп. Предполагаемая плотность свободной энергии сольватации определяется функцией Гаусса.

$$f_i(r) 4\pi r^2 = \alpha_i exp(-x_i^2), x_i = \frac{r - R_i}{\lambda_i}, \tag{1.9}$$

где R_i – ван дер Ваальсов радиус i, равный $\frac{1}{2}$ расстояния до энергетического минимума в потенциале Леннард-Джонса, λ_i корреляционная длинна и α_i – коэффициент пропорциональности, равный

$$\alpha_i = \frac{2\Delta G_i^{free}}{\sqrt{\pi\lambda_i}}, \tag{1.10}$$

где ΔG_i^{free} – свободная энергия сольватации изолированной группы i; ΔG_i^{free} близка к ΔG_i^{ref}, но не тождественно и определеяется эмпирически, требуя, чтобы свободная энергия сольватации глубоко лежащих групп равнялась нулю.

1.2. Структура данных k-d-дерево

k-d-дерево - это структура данных, которая позволяет эффективно хранить и обрабатывать точки в многомерном пространстве. Она используется для решения задач, связанных с поиском ближайших соседей, поиском точек в заданном диапазоне и кластеризацией данных.

k-d-дерево представляет собой бинарное дерево, в котором каждый узел соответствует гиперплоскости, разбивающей пространство на две части. Каждый узел содержит точку из множества, которое нужно организовать, а также указатели на двух потомков - левого и правого.

При поиске ближайших соседей или точек в заданном диапазоне, происходит спуск по дереву, выбирая тот узел, который содержит искомую точку. Затем происходит проверка точек в этом поддереве и, если они удовлетворяют условию поиска, то происходит их добавленик в результат. Если же поддерево не содержит искомую точку, происходит переход к следующему поддереву, пока не будет найдена нужная точка или не произведен обход всех поддеревьев.

k-d-дерево имеет ряд преимуществ перед другими структурами данных, такими как массивы или хэш-таблицы. Оно позволяет эффективно хранить и обрабатывать большие объемы данных, а также быстро выполнять операции поиска и обработки данных. k-d-дерево также может быть использовано для решения задач машинного обучения, таких как классификация и кластеризация данных.

Для построения k-d-дерева необходимо выбрать гиперплоскость, которая будет разбивать пространство на две части. Это можно сделать различными способами, например, выбрать гиперплоскость, которая проходит через среднюю точку множеств, или выбрать гиперплоскость, которая максимизирует расстояние между точками в разных поддеревьях.

При построении k-d-дерева также необходимо учитывать возможность перебора всех точек в заданном множестве. Для этого можно использовать различные алгоритмы, например, обход дерева в глубину или в ширину.

В целом, k-d-дерево – эффективная структура данных, которая может быть использована для решения широкого спектра задач, связанных с обработкой данных в многомерном пространстве.

1.3. Силовое поле CHARMM

1.3.1. Схема общего проекта CHARMM

Типичный исследовательский проект с использованием CHARMM можно описать в очень общих терминах на основе потока информации в программе, который схематически изображен на рисунке 1.2. Пользователь начинает проект, сначала настраивая атомную модель, представляющую систему цели исследования. Настройка состоит из импорта файла топологии "остатка"(RTF) и параметров силового поля (PRM), создания файла структуры "белка"(PSF) и сборки полной конфигурации (координат) всех атомов в системе; кавычки вокруг "остатка"и "белка"указывают на то, что используется одна и та же (историческая) нотация, когда программа применяется к молекулам в целом. Для молекул и фрагментов, которые были параметризованы, таких как белки, нуклеиновые кислоты и липиды, можно использовать стандартные файлы PRM и RTF CHARMM, и процедура настройки проста, если известны большинство координат. Если молекула не включена в стандартные библиотеки – CHARMM позволит использовать практически неограниченное разнообразие дополнительных молекулярных топологий и параметров силового поля. Для расчетов с использованием нескольких копий структуры, таких как расчеты путей реакции, в которых координаты двух конечных структур получаются из данных рентгеновской кристаллографии, требуется согласовать метки атомов на всех копиях, особенно для химически эквивалентных атомов (например, Cd1 и Cd2 Tyr). CHARMM предоставляет набор общих инструментов для облегчения настройки и манипуляции с молекулярной системой (например, преобразования координат и создание отсутствующих координат) и для наложения различных ограничений и ограничений на систему, где это уместно; ограничения позволяют изменять свойство цели исследования с энергетическим штрафом, в то время как ограничения фиксируют свойство, обычно на значения, указанные пользователем. Пользователь может указать ряд параметров для расчета несвязанных взаимодействий и выбрать любое условие из множества граничных условий для системы. Чтобы выполнить расчеты за приемлемое время, пользователь должен учитывать компромисс между точностью/сложностью и эффективностью при выборе модели, которая будет использоваться в расчетах; кроме того, ему может потребоваться параллельная компиляция кода или использование функций экономии времени, таких как таблицы поиска. В настоящее время существуют два веб-интерфейса, которые можно использовать для облегчения этапа настройки проекта CHARMM: CHARMM-GUI25 и CHARMMing.26

Проект может потребовать подготовительного этапа: например, для МД-моделирования обычная процедура заключается в минимизации
структуры системы (часто получаемой из кристаллографических или ЯМР-данных), нагреве системы до желаемой температуры и затем ее уравновешивании. После этого проект переходит на производственный этап, во время которого атомная конформация

системы может быть уточнена, исследована и отобрана с применением различных вычислительных процедур. Эти процедуры могут состоять в выполнении минимизации энергии, распространении траекторий МД или динамики Ланжевена, отборе с использованием алгоритмов Метрополиса Монте-Карло или поиска на сетке, получении различий свободной энергии термодинамики через вычисления возмущения свободной энергии, выполнении выборки путей перехода или расчете нормальных мод колебаний. С помощью таких методов можно моделировать временную эволюцию молекулярной системы, оптимизировать и генерировать конформации в соответствии с различными статистическими механическими ансамблями, характеризовать коллективные движения и исследовать энергетический ландшафт вдоль определенных путей реакции. Некоторые вычислительные методы (например, так называемые "алхимические"симуляции свободной энергии) включают рассмотрение "нереальных"промежуточных состояний для улучшения расчета физических наблюдаемых, включая свободную энергию, энтропию и изменение энтальпии из-за мутации или конформационного перехода.

Хотя во время производственного этапа проекта обычно контролируются несколько ключевых величин, дополнительные свойства системы могут потребовать определения путем постобработки данных - например, для расчета изменений свободной энергии из координат или коэффициентов диффузии из скоростей, сохраненных в течение одной или нескольких МД-траекторий. Эти производные величины могут включать временные ряды, корреляционные функции или другие свойства, связанные с экспериментальными наблюдениями. Продвинутый пользователь CHARMM в некоторых случаях может расширить функциональность программы в процессе выполнения своего проекта, создавая сценарии CHARMM написанием внешнего кода в качестве дополнения, использованием внутренних "хуков"к исходному коду CHARMM или прямое изменение одного или нескольких модулей исходного кода. После того, как такой разрабатываемый код будет приведен в соответствие стандартам кодирования CHARMM и протестирован, его следует отправить менеджеру CHARMM для рассмотрения возможности включения в будущие дистрибутивы программы.

1.3.2. Функциональная множественность CHARMM

Важной особенностью CHARMM является то, что многие конкретные вычислительные задачи (например, расчет свободной энергии или определение пути реакции) могут быть выполнены несколькими способами. Это разнообразие имеет две основные функции. Во-первых, наилучший метод часто зависит от конкретной природы изучаемой проблемы. Во-вторых, в рамках данного типа проблемы или метода, уровень аппроксимации, который обеспечивает наилучший баланс между требованиями к точности и вычислительными ресурсами, часто зависит от размера системы и ее сложности. Типичный пример возникает в классе моделей, используемых для представления влияния окружающего растворителя на макромолекулу. Наиболее

реалистичное представление обрабатывает среду растворителя, явно включая молекулы воды (а также любые контр-ионы, кристаллические соседи или мембранные липиды, если они присутствуют) и накладывая периодические граничные условия (PBC), которые имитируют бесконечную систему путем воспроизведения центральной ячейки7,8 (см. раздел IV.B.). Системы, варьирующиеся от десятков до сотен тысяч частиц, могут быть смоделированы с использованием таких моделей со всеми явными атомами на протяжении сотен наносекунд с использованием существующих вычислительных ресурсов, таких как большие кластеры узлов с распределенной памятью и параллельные архитектуры программ. Однако недостатком такого подхода к растворам является то, что большая часть времени вычислений (часто более 90%) используется для моделирования растворителя, а не тех частей системы, которые представляют основной интерес. В результате часто используется альтернативный подход, при котором влияние растворителя включается неявно с использованием эффективного среднепольного потенциала (т.е. без учета реальных молекул воды в расчете). Этот подход может значительно сократить вычислительные затраты на расчет для белка по сравнению с использованием явного растворителя, часто в 100 раз или больше, и учитывает многие равновесные свойства растворителя. Однако он вводит приближения, поэтому гидродинамика и трение растворителя, а также роль структуры воды, обычно не учитываются в подходе с неявным растворителем. В CHARMM доступно множество моделей неявного растворителя с различными профилями точности и эффективности. Промежуточный подход между моделированием всего атома с PBC и моделями неявного растворителя заключается в моделировании только небольшой области явно в присутствии уменьшенного количества молекул явного растворителя, применяя эффективный граничный потенциал растворителя (SBP) для имитации среднего влияния окружающего растворителя. Подход SBP часто выгоден в моделировании, требующем явного, атомарного представления воды в ограниченной области системы, например, при изучении реакции, происходящей в активном центре большого фермента. Выбор представления растворителя для проекта зависит от нескольких факторов, включая требования к точности расчета, тип искомых данных, размер системы и вычислительные и реальные ресурсы.

1.4. Описание входных данных

1.4.1. Формат данных PDB (Protein Data Bank)

Архив PDB представляет собой хранилище атомных координат и другой информации, описывающей белки и другие важные биологические макромолекулы. Структурные биологи используют такие методы, как рентгеновская кристаллография, ЯМР-спектроскопия и криоэлектронная микроскопия, чтобы определить положение каждого атома относительно друг друга в молекуле. Затем они размещают эту информацию, которая затем аннотируется и публикуется в архиве wwPDB.

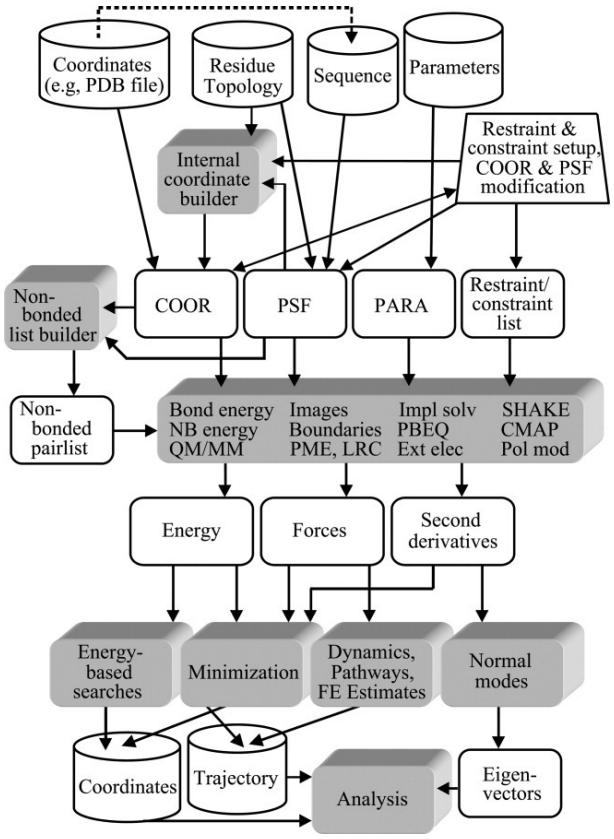

Рис. 1.2. Общая схема проекта CHARMM

Данные PDB Первичная информация, хранящаяся в архиве PDB, состоит из файлов координат биологических молекул. В этих файлах перечислены атомы в каждом белке и их трехмерное расположение в пространстве. Эти файлы доступны в нескольких форматах (PDB, mmCIF, XML). Типичный файл в формате PDB включает в себя большой раздел «заголовка», текста, в котором резюмируется белок, информация о цитировании и детали структурного решения, за которым следует последовательность и длинный список атомов и их координат. Архив также содержит экспериментальные наблюдения, которые используются для определения этих атомных координат.

Визуализация структур Возможно просматривать файлы PDB напрямую с помощью текстового редактора, часто наиболее полезно использовать программу просмотра или визуализации для их просмотра. Онлайн-инструменты, такие как на веб-сайте RCSB PDB, позволяют искать и изучать информацию под заголовком PDB, включая информацию об экспериментальных методах, химии и биологии белка.

Потенциальные проблемы При изучении архива PDB может возникуть ряд проблем. Например, многие структуры, особенно определяемые кристаллографией, включают информацию только о части функциональной биологической сборки. Кроме того, во многих записях PDB отсутствуют части молекулы, которые не наблюдались в эксперименте. К ним относятся структуры, включающие только положения альфа-углерода, структуры с отсутствующими петлями, структуры отдельных доме-

нов или субъединиц более крупной молекулы. Кроме того, в большинстве записей кристаллографической структуры отсутствует информация об атомах водорода.

Центральным хранилищем данных о структуре биологических макромолекул является Protein Data Bank, доступный по адресу rcsb.org.

Уникальным идентификатором каждой структуры в Protein Data Bank является PDB ID. Данный идентификатор состоит из четырех символов, первый из которых, как правило, цифра; буквы в PDB ID принято писать заглавными, хотя большинство программ в данном случае не чувствительны к регистру символов. Примеры PDB ID: 4L9K, 1BTI, 2R33.

В файле данного формата каждая колонка обладает строго определенную длину и содержит определенную информацию: 1-6 – имя записи; 7-11 – серийный номер атома; 13-16 – имя атома; 17 – альтернативное имя атома; 18-20 – имя остатка; 22 – идентификатор цепи; 23-26 – номер остатка; 27 – код для вставки остатков; 31-38 – координата X атома; 39-46 – координата Y атома; 47-54 – координата Z атома; 55-60 – вместимость; 61-66 – температура; 77-78 – символ элемента; 79-80 – заряд атома.

1.4.2. Формат данных pqr

Надежные модели электростатических взаимодействий важны для понимания событий раннего молекулярного распознавания, где доминируют дальнодействующие межмолекулярные взаимодействия и эффекты сольватации на биомолекулярные процессы. В то время как явные электростатические модели, которые рассматривают растворенное вещество и растворитель в атомарных деталях, являются общими, эти подходы обычно требуют тщательного уравновешивания и отбора проб для сведения интересующих свойств в интересующий статистический ансамбль. Континуальные подходы, которые интегрируют важные, но в значительной степени неинтересные степени свободы, жертвуют численной точностью в пользу надежной, но качественной точности и эффективности, устраняя необходимость отбора проб и уравновешивания, связанную с явными моделями растворов и растворителей.

Хотя существует выбор между несколькими моделями неявной сольватации, одна из самых популярных моделей неявной растворимости для биомолекул основана на уравнении Пуассона–Больцмана (ПБ). Уравнение ПБ обеспечивает глобальное решение для электростатического потенциала ϕ внутри и вокруг биомолекулы путем решения уравнения в частных производных

$$\Delta\epsilon\Delta\phi - \sum_{i}^{M} c_i q_i e^{-\beta(q_i\phi+V_i)} = \rho \tag{1.11}$$

Растворитель описывается объемной диэлектрической проницаемостью растворителя ε_s и свойствами подвижных ионов $i = 1, ..., M$ описываемыми их зарядами q_i, концентрациями c_i и стерическим потенциалом взаимодействия ионов с раствором V_i. Биомолекулярная структура включена в уравнение через V_i, функцию диэлек-

трического коэффициента и функцию распределения заряда ρ. Диэлектрическая проницаемость ϵ часто устанавливается равной постоянному значению ε_{min} внутри молекулы и резко меняется на молекулярной границе до значения ε_s, которое описывает объем растворителя. Форма границы определяется размером и расположением атомов раствора, а также специфическими для модели параметрами, такими как характерный размер молекулы растворителя. Распределение заряда ρ обычно представляет собой сумму дельта-распределений Дирака, расположенных в центрах атомов. Наконец, $\beta = (5\kappa T)^{-1}$ — обратная тепловая энергия, где κ – постоянная Больцмана, а T — температура. Потенциал ϕ может использоваться в различных приложениях, включая визуализацию, другие структурные анализы, моделирование диффузии и ряд других расчетов, требующих глобальных электростатических свойств. Теория ПБ является приближенной и, как следствие, имеет несколько хорошо известных ограничений, которые могут повлиять на ее точность, особенно для сильно заряженных систем или высоких концентраций солей. Несмотря на эти ограничения, методы ПБ по-прежнему очень важны и популярны для биомолекулярного структурного анализа, моделирования и симуляции.

Было разработано несколько пакетов программного обеспечения, которые решают уравнения Пуассона-Больцмана для оценки энергий, потенциалов и других свойств сольватации. Наиболее значимые (на основе пользовательской базы и цитирований) из них: CHARMM, AMBER, DelPhi, Jaguar, Zap, MIBPB и APBS. Тем не менее, APBS и связанный программный пакет PDB2PQR служат большому сообществу, насчитывающему около 27 000 пользователей, путем создания веб-серверов, связанных с веб-сайтом APBS, которые поддерживают подготовку биомолекулярных структур и быстрое решение методом конечных разностей. уравнения Пуассона–Больцмана, которые дополнены набором инструментов анализа. Еще более широкий набор функций и более гибкая конфигурация доступны, когда APBS и PDB2PQR запускаются из командной строки на платформах Linux, Mac и Windows, и которые можно запускать локально или через веб-службы, предоставляемую разработанным NBCR набором инструментов Opal. Этот инструментарий позволяет переносить вычислительную нагрузку для научных приложений с интенсивным использованием процессора на удаленные вычислительные ресурсы, такие как ресурсы, предоставляемые Национальным ресурсом биомедицинских вычислений (NCBR). Наконец, APBS может работать с другими программами молекулярного моделирования, такими как AMBER, CHARMM, NAMD, Rosetta и TINKER. Общая поддержка интеграции APBS со сторонними программами обеспечивается библиотекой iAPBS.

Электростатические расчеты начинаются с определения структуры молекулы и параметров заряда и размера составляющих ее атомов. Составляющие атомы обычно группируются по типам со значениями заряда и размера, определяемыми типом атома в различных файлах силового поля, разработанных для неявных расчетов растворителя. APBS включает эту информацию в расчеты в формате «PQR». PQR

— это формат файла неизвестного происхождения, используемый несколькими программными пакетами, такими как MEAD и AutoDock. Файл PQR просто заменяет столбцы температуры и заполнения плоского файла PDB на заряд ператома (Q) и радиус (R). Существуют гораздо более элегантные способы реализации функциональности PQR с помощью более современных расширяемых форматов файлов, таких как mmCIF или PDBML; тем не менее, простой формат PDB по-прежнему является одним из наиболее широко используемых форматов, и поэтому дальнейшее использование формата PQR поддерживает широкую совместимость инструментов и рабочих процессов биомолекулярного моделирования. Программное обеспечение PDB2PQR является частью пакета APBS, который был разработан для помощи в преобразование файлов PDB в формат PQR. В частности, PDB2PQR автоматически настраивает, выполняет и оптимизирует структуру для расчетов электростатики Пуассона-Больцмана, выводя файл PQR, который можно использовать с APBS или другим программным обеспечением для моделирования.

2. ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1. Реализация оценочной функции

Оценочная функция была реализована на языке программирования C++ и добавлена в проект PSM в виде набора модулей с соответствующем разбиением на модуль интерфейса и модуль реализации. Для нахождения взаимодействующих атомов был реализован алгоритм прямого перебора всех возможных пар атомов.

Кратко процедура прямого перебора выглядит следующим следующим образом. Обрабатываются списки атомов хранящиеся в ассоциативных контейнерах, где в качестве ключа выступает идентификатор белковой цепи. При рассмотрении каждой пары атомов в контейнере происходит сравнение расстояний между этими атомами и формирование нового списка взаимодействующих атомов.

Очевидно, что в случае, например, фиксированного количества атомов в каждой цепи временная сложность такого алгоритма перебора возрастает со следующей асимптотикой: $O(n^2 \cdot k \cdot (k-1)/2)$, где k – количество цепей, n – количество атомов в одной цепи. В случае различного числа атомов в цепях асимптотика прямого перебора также останется квадратичной.

Для ускорения процесса нахождения взаимодействующих пар атомов использовалась структура данных k-d-дерево. Оптимизированный алгоритм с использованием k-d-дерева работает следующим образом. происходит перебор цепей ассоциативного контейнера и строится соответствующее k-d-дерево. При этом требуется построение $k-1$ деревьев, где k – количество цепей в комплексе. Сложность алгоритма $O(k^2 \cdot m \cdot build)$, где k – количество цепей, n – количество атомов в цепи, по которой не построено k-d-дерево, $build = O(n \log_2 n)$ – сложность постройки дерева. В общем случае, сложность поиска всех соседей в k-d-дереве составляет $O(\log_2 n + k)$, где n – количество точек в дереве, а k – количество найденных соседей в заданном радиусе сферы взаимодействия атома. Также следует отметить, построенное k-d-дерево обладает фиксированной пространственной сложностью.

В случае отсутствия изменения в позициях атомов некоторых компонентов комплекса нет необходимости строить дерево для каждой оценки. Например, в случае если в комплексе только два компонента и где только один меняет пространственную позицию достаточно построить дерево один раз для неподвижного компонента. В рамках исследования такой подход был реализован и обозначен как k-d-opt.

2.2. Оптимизация с использованием k-d-дерева

В данной работе для ускорения расчетов оценочной функции была реализована структура данных k-d-дерево на языке программирования C++ с поддержкой модулей. Поскольку библиотека PSM для полноатомного моделирования белка реализована на

языке программирования C++ язык реализации оценочной функции и структуры данных k-d-дерева идентичен.

Структура данных представляет собой отдельный подключаемый модуль в рамках библиотеки и представляет собой C++ структуру KDTree, которая хранит в себе указатель на корневой узел дерева и содержит соответствующие методы для построения и поиска взаимодействующих атомов. Листинг кода представлен в Приложении А.

2.3. Результаты численных экспериментов

Выполнение оценок разработанной оценочной функцией возможно только для полноатомных структур, представленных в формате PDB. Поскольку оценочная функция разрабатывалась для моделирования процесса образования комплекса кинетическим методом Монте-Карло, для проверки качества выполняемых оценок был сформирован набор тестовых комплексов, взятый из хранилища SKEMPI [12], представленный в работах [3; 13]. При этом из общего списка комплексов был выделен ограниченный набор структур. Исключены комплексы, содержащие разрывы в главных цепях, а также комплексы, состоящие из трёх и более цепей.

Перед проведением численного эксперимента для всех комплексов проведена предварительная подготовка. С помощью пакета CHARMM выполнено восстановление структур в полноатомный вид, поскольку представленные в базе данных PDB структуры могут не включать определенные атомы. Затем средствами пакета CHARMM выполнена минимизация энергии. В результате для каждого комплекса формируются три стартовых PDB файла, для которых генерируются PQR файлы, содержащие частичные заряды для каждого атома. На последнем этапе производится оценка энергии средствами CHARMM и разработанной оценочной функцией. Сформированный список комплексов и начальных файлов (PDB и PQR) представлен в репозитории [14].

Для проведения верификации оценочной функции использовался облачный сервис университета «Дубна», в рамках которого использовался следующий процессор: Intel Xeon CPU E5-2650. Численные эксперименты, демонстрирующие применение k-d-дерева и сравнение времени поиска взаимодействующих атомов другими инструментами, выполнено на процессоре AMD Ryzen 7 5700X на базовой частоте процессора равной 4.2 ГГц.

2.3.1. Поиск взаимодействующих атомов

На рис. 2.1 представлено сравнение времени поиска взаимодействующих атомов для двух белков различными инструментами.

Тестовый белковые комплексы 1TM1 и 1KXQ выбраны из библиотеки SKEMPI. Их различает разное число атомов. Второй комплекс содержит почти в два раза больше атомов. Комплексы выбраны специально для объективности сравнения оценок при изменении числа атомов. Комплекс 1TM1 имеет два компонента (цепи), которые

содержит 3939 и 1059 атомов соответственно. Комплекс 1KXQ также двухкомпонентный, каждая цепь содержит 7613 и 1786 атомов соответственно. Построение k-d-дерева во всех случаях происходило для первой цепи белкового комплекса.

Как видно из результатов, алгоритм с применением прямого перебора всех атомов работает в несколько раз медленнее других подходов. Однако, при сравнении с Rosetta, несмотря на то, что коэффициент ускорения присутствует и больше единицы, получаемое ускорение незначительно. Это связано с тем, что в Rosetta реализован собственный механизм построения карты взаимодействия атомов, который, в том числе, использует структуру данных k-d-дерево.

Обозначения на рисунках:

1. nn – алгоритм прямого перебора всех атомов;
2. rosetta – программный комплекс Rosetta;
3. k-d – алгоритм с построением k-d-дерева при каждом поиске;
4. k-d-opt – алгоритм с предварительным построением k-d-дерева.

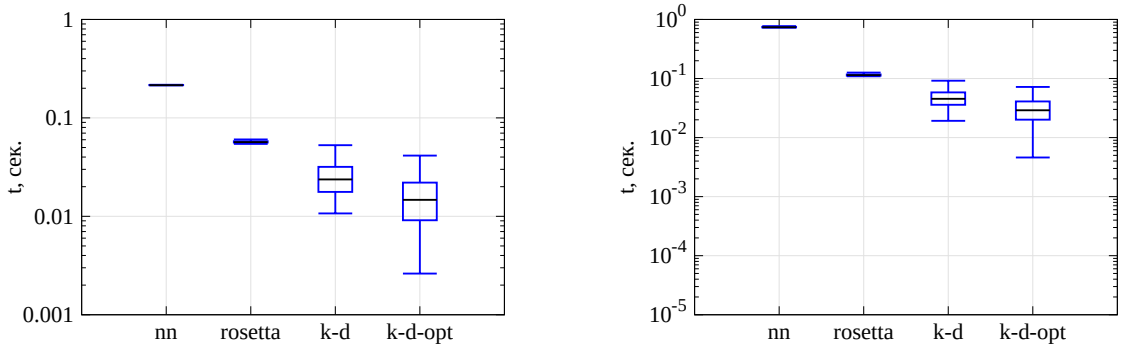

Рис. 2.1. Время поиска взаимодействующих пар атомов для комплексов 1TM1 и 1KXQ

На рис. 2.2 продемонстрированы коэффициенты ускорения соответствующие времени поиска взаимодействующих атомов разработанной функции в отношении с другими алгоритмами. Всего выполнено 9154 оценки для белка 1TM1 и 9479 для комплекса 1KXQ. Следует отметить, что на представленных рисунках число контактов уникально и не повторяется.

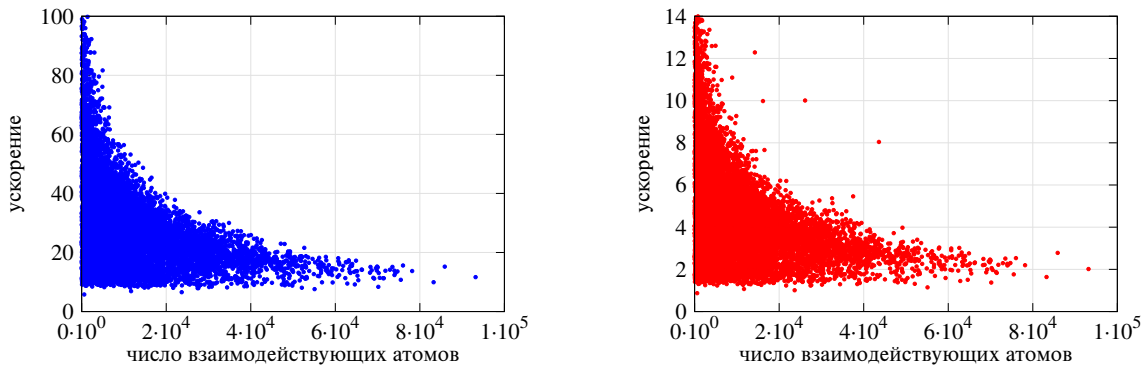

Рис. 2.2. Получаемое ускорение при использовании алгоритма k-d-opt для комплекса 1KXQ по сравнению с алгоритмом nn (слева) и Rosetta (справа)

2.3.2. Верификация выполняемых оценок

На рисунках 2.3 и 2.4 приведены результаты численных оценок для 84 комплексов. Средствами пакета CHARMM выполнена оценка энергии для представленных в разработанной оценочной функции компонент. Для полученных значений рассчитан линейный коэффициент корреляции Пирсона.

Следует отметить, что представленная на рисунках оценка включает в себя внутримолекулярные взаимодействия. Для этого в оценочной функции и в пакете CHARMM использовалась так называемая схема 1-3, где при формировании списка взаимодействующих пар атомов исключаются пары, которые связаны ковалентно (схема 1-2), а также пары, «соединенные» одним общим атомом. При рассмотрении компонент комплекса в виде «твёрдых» тел внутримолекулярные взаимодействия изменяться не будут, поэтому при моделировании процесса образования комплекса достаточно выполнить их оценку только один раз и затем рассматривать взаимодействия только между атомами компонент комплекса.

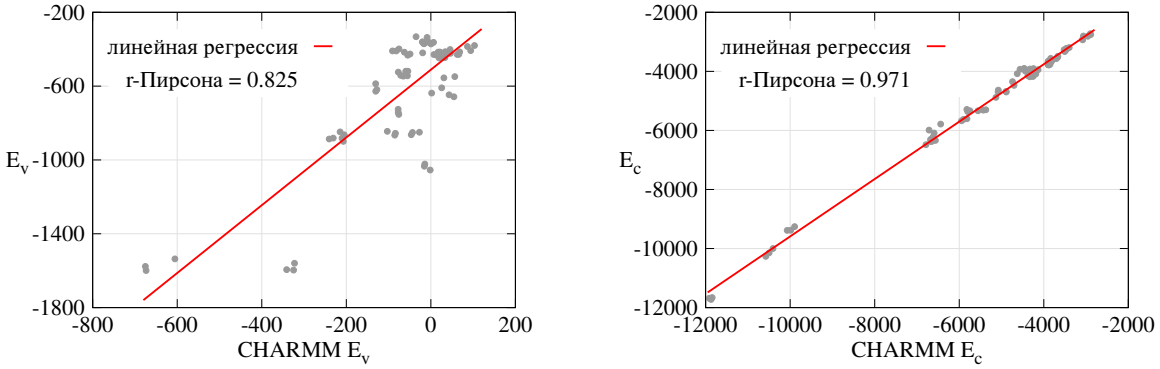

Рис. 2.3. Результаты численного эксперимента для потенциала Леннард-Джонса и для потенциала Кулона

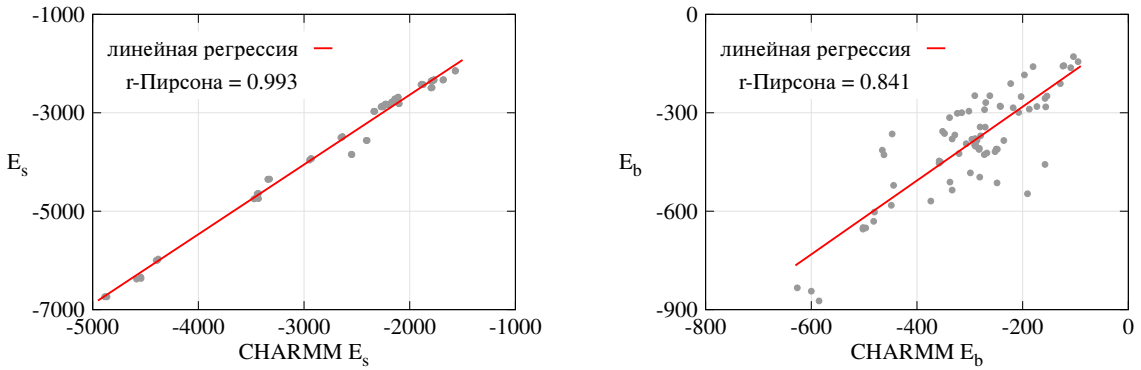

Рис. 2.4. Результаты численного эксперимента для неявного растворителя и для оценки энергии связывания

В простейшем случае процесс связывания описывается моделью вида ключ-замок, которая представляется в виде $A + B \leftrightarrow AB$, где A и B являются компонентами комплекса. С помощью разработанной целевой функции возможно оценить энергию этого взаимодействия. На рисунке 2.4 продемонстрирована оценка энергии связывания

без учета растворителя, которая определяется по следующей формуле:

$$E_b = \left[E_v^{AB} + E_c^{AB}\right] - \left[E_v^A + E_c^A + E_v^B + E_c^B\right]. \tag{2.1}$$

В численном эксперименте в начальном PDB файле представлен образованный комплекс AB. При определении энергии связывания (2.1) вычисляется разница между оценкой энергии комплекса в связанном состоянии и оценками энергий в свободном состоянии для каждого компонента в отдельности. В данном случае оценка энергии растворителя исключена для сравнения результатов оценки взаимодействия только на основе двух слагаемых оценочной функции.

На рис. 2.5 для тестового набора белков продемонстрировано сравнение оценок энергии связывания, которые были вычислены с помощью разработанной оценочной функции (обозначено F_s). Для наглядности посчитана линейная регрессия и коэффициент корреляции Пирсона.

В отличии от оценки E_b, слагаемые которой представлена в выражении 2.1, полная энергия связывания включает в себя все три компонента (включая растворитель) и вычисляется как разница между оценкой энергии комплекса в связанном состоянии и оценками энергий в свободном состоянии, т.е. когда компоненты комплекса друг с другом не взаимодействуют. Представленная на рис. 2.5 оценка энергии связывания $\mathrm{F_s}$ определяется следующим образом:

$$\mathrm{F_s} = F_s^{AB} - \left[F_s^A + F_s^B\right], \tag{2.2}$$

где F_s^{AB}, F_s^A и F_s^B найденные с помощью разработанной оценочной функции 1.1 значения.

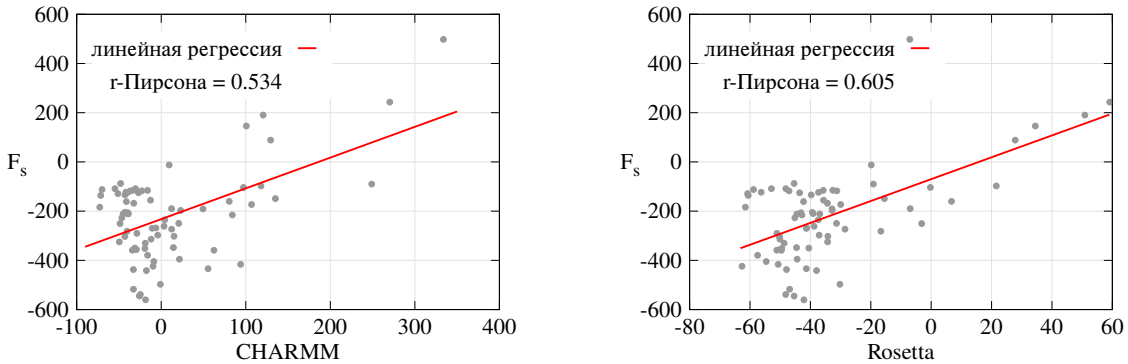

Рис. 2.5. Сравнение оценок энергии связывания разработанной оценочной функции и оценок вычисленных с помощью других инструментов

ЗАКЛЮЧЕНИЕ

В результате выполненной работы достигнуты следующие результаты.

1. Разработана и реализована оценочная функция с использованием набора параметров силового поля CHARMM в рамках библиотеки PSM.
2. Выполнена оптимизация процедуры поиска взаимодействующих пар атомов с помощью применения структуры данных k-d-дерево.
3. Проведены различные численные эксперименты, демонстрирующие приемлемую высокую корреляцию оценок с результатами силовых полей CHARMM и Rosetta.
4. На примере двух белков продемонстрировано преимущество применения структуры данных k-d-дерево для поиска взаимодействующих пар атомов.

Результаты работы представлялись на всероссийской конференции «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем 2023», которая проходила 17-21 апреля 2023 году в Москве. Тезисы доклада представлены в работе [15].

Задача разработки и реализации оценочной функции поставлена моим научным руководителем, который также является руководителем научно-исследовательской работы проводимой в университете. Разработанная оценочная функция используется в рамках программного пакета PSM, позволяет гибко настраивать все параметры потенциалов и быстро выполнять оценку энергии взаимодействия в комплексах.

Следует отменить, что несмотря на то, что применение структуры данных k-d-дерево позволило в несколько раз уменьшить временную сложность поиска атомов даже в случае большого количества взаимодействующих атомов, время поиска попадает в довольной широкий диапазон значений. Указанное отклонение можно значительно уменьшить с помощью использования других структур данных. Применение других структур данных для поиска взаимодействующих атомов может стать темой дальнейших исследований.

СПИСОК ЛИТЕРАТУРЫ

1. Лизунов А. Ю., Зайцева Н. И., Зосимов В. В. Учет взаимодействий между атомами лиганда в задаче докинга с помощью потенциала усредненных энергий // Труды МФТИ. — 2014. — Т. 6, № 1.
2. Liu J., Wang R. Classification of Current Scoring Functions // Journal of Chemical Information and Modeling. — 2015. — Т. 55, № 3. — DOI: 10.1021/ci500731a.
3. Dhusia K., Su Z., Wu Y. Using Coarse-Grained Simulations to Characterize the Mechanisms of Protein–Protein Association // Biomolecules. — 2020. — Т. 10. — DOI: 10.3390/biom10071056.
4. Репозиторий библиотеки PSM (protein structural modeling). — URL: https://vcs.uni-dubna.ru/PoSV.th/psm (дата обр. 30.05.2023).
5. Voter A. Introduction to the kinetic Monte Carlo method // NATO Science Series. — 2007. — Т. 235. — DOI: 10.1007/978-1-4020-5295-8_1.
6. Alford R. F. [и др.]. The Rosetta all-atom energy function for macromolecular modeling and design // Journal of Chemical Theory and Computation. — 2017. — Т. 13, № 6. — С. 3031—3048. — DOI: 10.1021/acs.jctc.7b00125.
7. Brooks B. [и др.]. CHARMM: The biomolecular simulation program // Journal of Computational Chemistry. — 2009. — Т. 30. — DOI: 10.1002/jcc.21287.
8. Хрущев С. С. [и др.]. Моделирование белок-белковых взаимодействий с применением программного комплекса многочастичной броуновской динамики ProKSim // Компьютерные исследования и моделирование. — 2013. — Т. 1. — DOI: 10.20537/2076-7633-2013-5-1-47-64.
9. Lazaridis T., Karplus M. Effective energy function for proteins in solution. — 1999. — DOI: 10.1002/(SICI)1097-0134(19990501)35:2<133::AID-PROT1>3.0.CO;2-N.
10. Jurrus E. [и др.]. Improvements to the APBS biomolecular solvation software suite // Protein Science. — 2018. — Т. 27. — DOI: 10.1002/pro.3280.
11. Wu X., Brooks B. A double exponential potential for van der Waals interaction // AIP Advances. — 2019. — Т. 9. — DOI: 10.1063/1.5107505.
12. Jankauskaite J. [и др.]. SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation // Bioinformatics. — 2019. — Т. 35. — DOI: 10.1093/bioinformatics/bty635.
13. Qin S., Pang X., Zhou H. Automated Prediction of Protein Association Rate Constants // Structure. — 2011. — Т. 19. — DOI: 10.1016/j.str.2011.10.015.
14. Репозиторий со списком комлексов и результатми эксперимента. — URL: https://vcs.uni-dubna.ru/psm/data (дата обр. 30.05.2023).

15. Полуян С. В., Никулин Д. А., Ершов Н. М. Разработка и верификация оценочной функции для учета межмолекулярных взаимодействий в белковых комплексах // Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems 2023 (ITTMM 2023) Book of Abstracts. — 2023.

ПРИЛОЖЕНИЕ А

```
export module kdtree;
/// system
import <vector>;
import <tuple>;
/// internal
import atom;
namespace psm
{
namespace kd
{
 struct KdNode
 {
  psm::Atom* point;
  KdNode* left;
  KdNode* right;
  KdNode() {}
  KdNode (psm::Atom* p)
   : point(p)
   , left(nullptr)
   , right(nullptr) {}
  ~KdNode()
  {
   delete left;
   delete right;
  }
 };
 export struct KDTree
 {
  typedef std::vector
    <std::tuple<psm::Atom*,
                psm::Atom*,
                double,
                double>> atom_pairs_type;
  KdNode* root_;
  KDTree(std::vector<psm::Atom*>& atoms);
  void search(psm::Atom* point,
              const double distance,
                           atom_pairs_type& pairs_data) const;
  KdNode* build(std::vector<psm::Atom*>& points,
                const int depth,
                const int start,
                const int end);
  void search(const KdNode* node,
              psm::Atom* point,
              const double distance,
              const int depth,
              atom_pairs_type& pairs_data) const;
  ~KDTree();
 };
} // kd
} // psm
```

Листинг 5.1. Интерфейс модуля k-d дерево

```
module kdtree;
/// system
import <vector>;
import <tuple>;
import <cmath>;
/// internal
import atom;
namespace psm
{
namespace kd
{
 KDTree::KDTree(std::vector<psm::Atom*>& points)
 {
  root_ = build(points, 0, 0, points.size() - 1);
 }
 KDTree::~KDTree()
 {
  delete root_;
 }
 KdNode* KDTree::build(std::vector<psm::Atom*>& points,
                       const int depth,
                       const int start,
                       const int end)
 {
  if(start > end)
   return nullptr;
  int mid = (start + end) / 2;
  if(depth % 3 == 0)
  {
   std::nth_element(points.begin() + start,
                    points.begin() + mid,
                    points.begin() + end + 1,
   [](psm::Atom* a, psm::Atom* b)
   {
    return a->get_coordinates().x < b->get_coordinates().x;
   });
  }
  else if(depth % 3 == 1)
  {
   std::nth_element(points.begin() + start,
                    points.begin() + mid,
                    points.begin() + end + 1,
   [](psm::Atom* a, psm::Atom* b)
   {
    return a->get_coordinates().y < b->get_coordinates().y;
   });
  }
  else
  {
   std::nth_element(points.begin() + start,
                    points.begin() + mid,
                    points.begin() + end + 1,
   [](psm::Atom* a, psm::Atom* b)
   {
    return a->get_coordinates().z < b->get_coordinates().z;
   });
  }
  KdNode* node = new KdNode();
  node->point = points[mid];
  node->left = build(points, depth + 1, start, mid - 1);
  node->right = build(points, depth + 1, mid + 1, end);
  return node;
 }
 void KDTree::search(const KdNode* node,
                     psm::Atom* point,
                     const double distance,
                     const int depth,
                     atom_pairs_type& pairs_data) const
 {
  if(node == nullptr)
        return;
```

```
  double dd = node->point->get_coordinates()
                   .squared_distance(point->get_coordinates());
  double d = std::sqrt(dd);
  if(distance >= d)
        pairs_data.push_back(
                        std::make_tuple(node->point, point, d, dd));
  if(depth % 3 == 0)
  {
   if(point->get_coordinates().x - distance
     < node->point->get_coordinates().x)
   {
        search(node->left, point, distance, depth + 1, pairs_data);
   }
   if(point->get_coordinates().x + distance
     > node->point->get_coordinates().x)
   {
        search(node->right, point, distance, depth + 1, pairs_data);
   }
  }
  else if(depth % 3 == 1)
  {
   if(point->get_coordinates().y - distance
     < node->point->get_coordinates().y)
   {
    search(node->left, point, distance, depth + 1, pairs_data);
   }
   if(point->get_coordinates().y + distance
     > node->point->get_coordinates().y)
   {
    search(node->right, point, distance, depth + 1, pairs_data);
   }
  }
  else
  {
   if(point->get_coordinates().z - distance
     < node->point->get_coordinates().z)
   {
    search(node->left, point, distance, depth + 1, pairs_data);
   }
   if(point->get_coordinates().z + distance
     > node->point->get_coordinates().z)
   {
    search(node->right, point, distance, depth + 1, pairs_data);
   }
  }
 }
 void KDTree::search(psm::Atom* point,
                     const double distance,
                     atom_pairs_type& pairs_data) const
 {
  search(root_, point, distance, 0, pairs_data);
 }
} // kd
} // psm
```

Листинг 5.2. Реализация модуля k-d дерево

```
export module energy;
/// system
import <iostream>;
import <utility>;
import <string>;
import <vector>;
import <tuple>;
import <map>;
import <unordered_map>;
import <fstream>;
import <chrono>;
/// internal
import atom;
import protein;
import mover;
import residue;
import ienergy;
import kdtree;
namespace psm
{
 // ...
 export class KdScoreFunction : public psm::ScoreFunction
 {
 public:
  using ScoreFunction::init;
  using ScoreFunction::get_atom_pairs_size;
  void exclusion_nbond() override
  {
   pairs_data.clear();
   for (auto iter = atoms_ptrs.begin();
        iter != atoms_ptrs.end(); ++iter)
   {
    kd::KDTree tree1(iter->second);
    for (auto it = std::next(iter); it != atoms_ptrs.end(); ++it)
    {
     for (auto& atom : it->second)
     {
      tree1.search(atom, chdata.CUTNB, pairs_data);
     }
    }
   }
  }
 };
} // psm
```

Листинг 5.3. Реализация оценочной функции с использованием k-d дерева