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

Панченко Святослав Константинович

Построение вероятностного метрического пространства для моделирования зависимых от ориентации состояний

03.03.01 — Прикладные физика и математика

Выпускная квалификационная работа бакалавра

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

Москва
2020

Содержание

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.

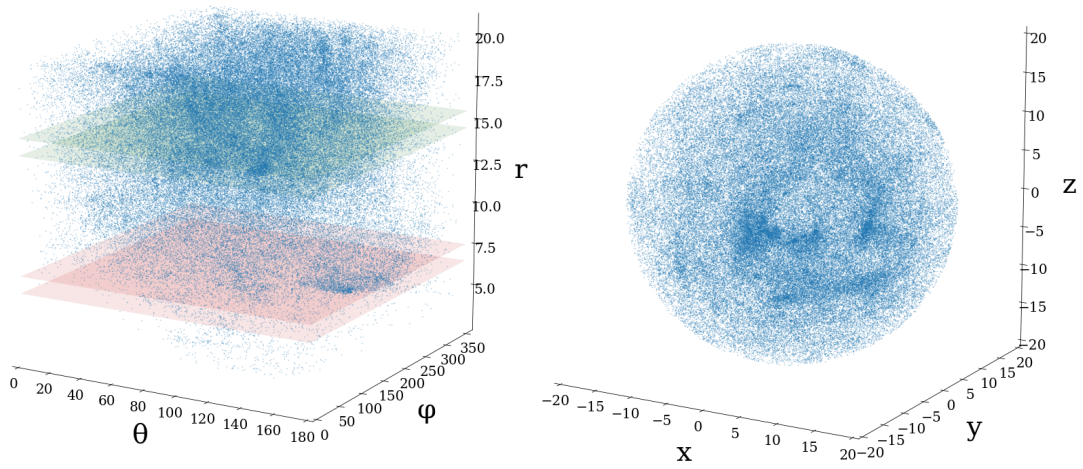

Рис. 1: Множество точек в пространстве (r, θ, φ) в сферической и декартовой системах координат. Каждая точка отвечает имеющейся в базе данных конфигурации рассматриваемой пары молекул: аминокислотного остатка ALA и лиганда C_{ar}.

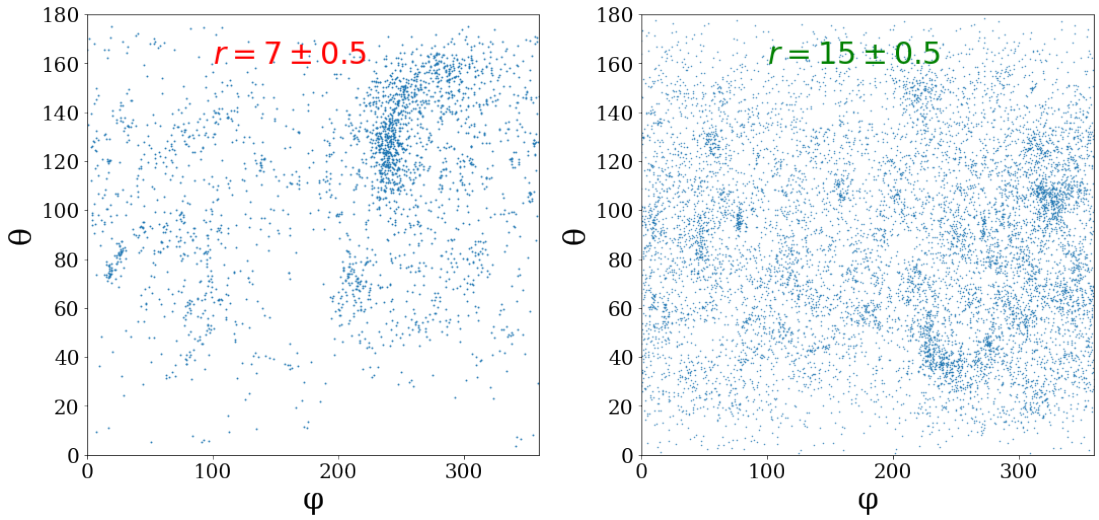

Рис. 2: Двумерная иллюстрация: множества точек между поверхностями $r = 7 \pm 0.5$ и $r = 15 \pm 0.5$, спроецированные на плоскость углов (θ, φ).

Цель работы. Целью работы является построение описанных плотностей для простого случая взаимодействия пары аминокислота-лиганд. Это взаимодействие характеризуется скромным набором из трёх параметров: расстояние между молекулами r и пара сферических углов θ, φ, определяющих положение лиганда в системе координат, связанной с аминокислотой. Плотности восстанавливаются по имеющимся данным для различных пар взаимодействующих молекул. Для задачи затруднительна постановка внешнего критерия каче-

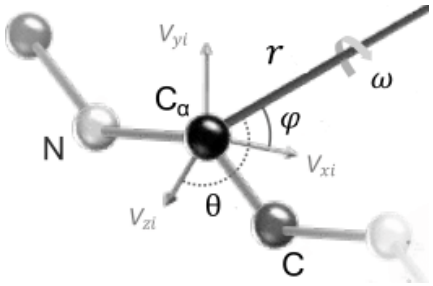

Рис. 3: Иллюстрация химической связи и её параметров (изображение взято из [5])

ства, поэтому от плотностей требуется:

- **интерпретируемость** с точки зрения эксперта, т.е. максимумы восстановленных плотностей должны соответствовать энергетически устойчивым конфигурациям молекул;

- **согласованность** с результатами восстановления, полученных независимо с применением других моделей и заведомо одобренных экспертом.

Новизна. В работах [11, 12], посвящённых созданию качественных генеративных моделей белковых структур, в качестве инструмента используется раздел теории вероятности, называемый *directional statistics*. Этот раздел изучает случайные величины, носителем которых является произвольное компактное Римановское многообразие. В частности, он исследует случайные величины, определённые на сфере, т.е. величины, с помощью которых можно естественным образом описывать совместное распределение сферических углов. Именно такие случайные величины используются в данной работе для восстановления искомых плотностей.

Основные положения, выносимые на защиту.

1. Алгоритм нахождения параметров для восстановления плотности с помощью модели смеси распределений, распределение углов в компоненте которой описывается с помощью распределения Кента с носителем на сфере.

2. Решение прикладной задачи восстановления плотности распределения взаимных пространственных ориентаций различных пар молекул.

3. Демонстрация соответствия полученного результата с результатами, полученных с применением других моделей.

1 Постановка задачи восстановления плотности

Имеется выборка объёма n:

$$\mathbf{X}^{a,b} = \{\boldsymbol{x}_i\}_{i=1}^{n}, \tag{1}$$

каждый элемент выборки $\boldsymbol{x}_i$ представляет собой тройку:

$$\boldsymbol{x}_i = [r_i, \theta_i, \varphi_i]^T \in \Omega, \tag{2}$$

где:

a — тип взаимодействующей аминокислоты;

b — тип взаимодействующего лиганда;

r_i — длина химической связи между аминокислотой a и лигандом b;

θ_i и φ_i — сферические углы, соответствующие лиганду в системе координат аминокислоты;

$\Omega = [3\text{Å},\ 20\text{Å}] \times [0, \pi] \times [0, 2\pi]$ — множество возможных значений элемента выборки.

Здесь $\boldsymbol{x}_i,\ i = \overline{1, n}$ интерпретируются как независимые в совокупности реализации трёхкомпонентного случайного вектора $X^{a,b}$ с истинной плотностью:

$$p^{a,b}(\boldsymbol{x}) = p^{a,b}(r, \theta, \varphi). \tag{3}$$

Постановка задачи восстановления плотности По имеющейся выборке требуется построить функцию:

$$\hat{p}^{a,b}(\boldsymbol{x}) = \hat{p}^{a,b}(r, \theta, \varphi), \tag{4}$$

аппроксимирующую истинную плотность $p^{a,b}(\boldsymbol{x})$. Ниже верхние индексы a, b будут опускаться.
В данном исследовании искомая функция строится на основе модели смеси распределений. Плотность случайного вектора представляется в виде суммы:

$$p(\boldsymbol{x}|\boldsymbol{w}, \mathbf{U}) = \sum_{k=1}^{K} w_j p_k(r, \theta, \varphi|\boldsymbol{u}_k),\ \ \sum_{k=1}^{K} w_k = 1,\ w_k \geq 0,\ k = \overline{1, K}, \tag{5}$$

$$(\boldsymbol{w}, \mathbf{U}) = (w_1, ..., w_K, \boldsymbol{u}_1, ..., \boldsymbol{u}_K), \tag{6}$$

где (для всех $k = \overline{1, K}$):

K — число компонент смеси;

$w_k = p(k)$ — априорная вероятность k-ой компоненты смеси;

$p_k(r, \theta, \varphi|\boldsymbol{u}_k)$ — плотность распределения k-ой компоненты смеси, характеризующаяся вектором параметров $\boldsymbol{u}_k$;

(6) — совокупность параметров модели.

При фиксированном числе компонент смеси K, выбранных с точностью до параметров плотностей $p_k(r, \theta, \varphi|\boldsymbol{u}_k)$ и заданной выборке (1) строятся оценки параметров смеси (6), исходя из принципа максимума правдоподобия:

$$\ln L(\mathbf{X}|\boldsymbol{w}, \mathbf{U}) = \ln \prod_{i=1}^{n} p(\boldsymbol{x}_i) = \sum_{i=1}^{n} \ln \sum_{k=1}^{K} w_k p_k(r, \theta, \varphi|\boldsymbol{u}_k) \to \max_{\boldsymbol{w}, \mathbf{U}}. \tag{7}$$

Обозначим через $(\boldsymbol{w}^*, \mathbf{U}^*)$ оптимальные параметры смеси:

$$(\boldsymbol{w}^*, \mathbf{U}^*) = \operatorname*{argmax}_{\boldsymbol{w}, \mathbf{U}} \ln L(\mathbf{X}|\boldsymbol{w}, \mathbf{U}). \tag{8}$$

Требуется выбрать плотности компонент смеси $p_k(r, \theta, \varphi|\boldsymbol{u}_k)$, подобрать число компонент смеси K и найти оптимальные параметры смеси (8) таким образом, что восстановленнная плотность

$$\hat{p}(\boldsymbol{x}) = p(r, \theta, \varphi|\boldsymbol{w}^*, \mathbf{U}^*) \tag{9}$$

удовлетворяла бы следующим требованиям:

Интерпретируемость: максимумы плотности должны быть интерпретируемы с точки зрения эксперта, т.е. должны соответствовать энергетически устойчивым конфигурациям взаимодействующих молекул.

Согласованность: максимумы плотности должны соответствовать результатам, полученным с помощью более простых моделей.

Учёт периодичности: распределение углов должно описываться с помощью случайных величин, определённых на сфере в трёхмерном пространстве.

2 Решение задачи восстановления плотности

В данном разделе предлагаются плотности компонент смеси, описывающие распределение углов в естественном для них пространстве, а также предлагается алгоритм нахождения оптимальных параметров смеси.

2.1 Распределение Кента как способ описания распределения углов

5-параметрическое распределение Фишера-Бингхама или *распределение Кента* — это распределение на единичной сфере S^2 в трёхмерном пространстве $\mathbb{R}^3$, являющееся аналогом двумерного нормального распределения на поверхности сферы. Распределение описано в одноимённой статье [13].

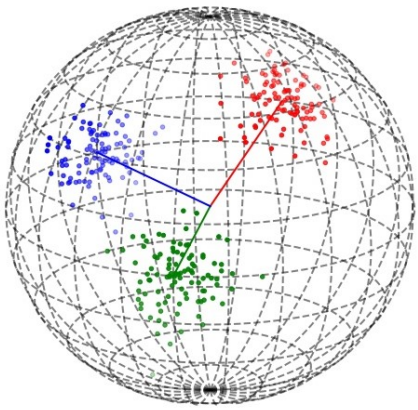

Рис. 4: Выборки из различных распределений Кента

Плотность распределения Кента задаётся выражением:

$$f(\mathbf{x}) = \frac{1}{c(\kappa,\beta)} \exp\left\{\kappa\boldsymbol{\gamma}_1^\intercal\mathbf{x} + \beta\left[(\boldsymbol{\gamma}_2^\intercal\mathbf{x})^2 - (\boldsymbol{\gamma}_3^\intercal\mathbf{x})^2\right]\right\}, \tag{10}$$

где $\mathbf{x}$ — единичный вектор:

$$\mathbf{x} = [x_1, x_2, x_3]^\intercal \in S^2 \subset \mathbb{R}^3,\ ||\mathbf{x}||_2 = 1, \tag{11}$$

а $c(\kappa, \beta)$ — нормирующая константа:

$$c(\kappa, \beta) = 2\pi \sum_{j=0}^{\infty} \frac{\Gamma(j + \frac{1}{2})}{\Gamma(j+1)} \beta^{2j} I_{2j+\frac{1}{2}}(\kappa) \left(\frac{1}{2}\kappa\right)^{-2j-\frac{1}{2}}, \tag{12}$$

где $I_v(\kappa)$ — модифицированная функция Бесселя ранга v, а $\Gamma(\cdot)$ — гамма-функция.

Параметр $\kappa > 0$ задаёт сконцентрированность распределения относительно среднего направления, а параметр β $(0 < 2\beta < \kappa)$ задаёт эллиптичность контуров равной плотности вероятности. Чем параметры κ и β больше, тем, соответственно, распределение более сконцетрировано, а контуры более эллиптичны. Вектор $\boldsymbol{\gamma}_1$ определяет среднее направление распределения, а векторы $\boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3$ определяют ориентацию контуров равной вероятности на сфере. При этом 3×3-матрица $[\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3]$ ортогональна.

От компонент единичного вектора $[x_1, x_2, x_3]^\intercal$ перейдём к сферическим углам $\theta \in [0, \pi]$, $\varphi \in [0, 2\pi]$ согласно нотации Кента, описанной в [13]:

$$x_1 = \cos\theta, \;\; x_2 = \sin\theta\cos\varphi, \;\; x_3 = \sin\theta\sin\varphi. \tag{13}$$

В выражение плотности распределения (10) подставим соотношения (13), т.е. выразим функцию плотности через сферические углы (θ, φ). В таком случае плотность распределения Кента обозначим:

$$\mathcal{K}(\theta, \varphi | \kappa, \beta, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3) := \mathcal{K}(\theta, \varphi | \boldsymbol{v}) := f([\cos\theta, \sin\theta\cos\varphi \sin\theta\sin\varphi]^T), \tag{14}$$

где f — плотность распределения как функция координат единичного вектора (10), $\mathcal{K}(\theta, \varphi | \boldsymbol{v})$ — краткое обозначение для плотности распределения Кента через вектор параметров распределения:

$$\boldsymbol{v} = [\kappa, \beta, \boldsymbol{\gamma}_1^\intercal, \boldsymbol{\gamma}_2^\intercal, \boldsymbol{\gamma}_3^\intercal,]^\intercal. \tag{15}$$

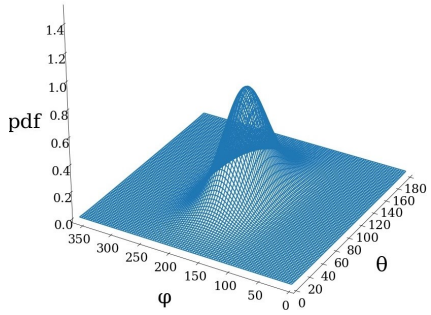

Рис. 5: Пример плотности распределения Кента, построенной в координатах (θ, φ).

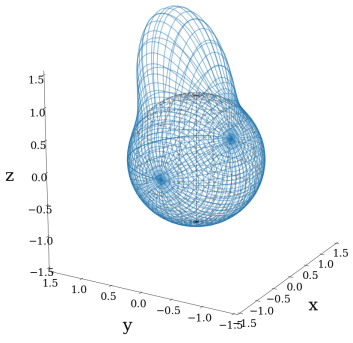

Рис. 6: Плотность распределения Кента в виде параметрической поверхности

2.2 Выбор плотности распределения компоненты смеси

Предлагается следующий способ выбора плотности распределения одной компоненты смеси $p_k(r, \theta, \varphi|\boldsymbol{u}_k)$:

1. Распределения расстояния r и пары углов (θ, φ) предполагаются независимыми, т.е. p_k представляется в виде произведения:

$$p_k(r, \theta, \varphi|\boldsymbol{u}_k) = p_k^{(1)}(r|\boldsymbol{u}_k^{(1)})p_k^{(2)}(\theta, \varphi|\boldsymbol{u}_k^{(2)}),\ \boldsymbol{u}_k^\intercal = [\boldsymbol{u}_k^{(1)\intercal}, \boldsymbol{u}_k^{(2)\intercal}].$$

   Несмотря на то, что расстояния не зависят от углов для одной компоненты, для всей смеси (5) это уже, разумеется, будет не так

2. Распределение расстояний предполагается нормальным с параметрами μ_k и σ_k^2:

$$p_k^{(1)}(r|\boldsymbol{u}_k^{(1)}) = \mathcal{N}(r|\mu_k, \sigma_k^2).$$

3. Распределение углов предполагается распределением Кента с вектором параметров $\boldsymbol{v}_k$:

$$p_k^{(2)}(\theta, \varphi|\boldsymbol{u}_k^{(2)}) = \mathcal{K}(\theta, \varphi|\boldsymbol{v}_k).$$

В итоге плотность распределения компоненты смеси имеет вид:

$$p_k(r, \theta, \varphi|\boldsymbol{u}_k) = \mathcal{N}(r|\mu_k, \sigma_k^2)\mathcal{K}(\theta, \varphi|\boldsymbol{v}_k), \tag{16}$$

$$\boldsymbol{u}_k^\intercal = [\mu_k, \sigma_k^2, \boldsymbol{v}_k^\intercal].$$

2.3 Алгоритм нахождения оптимальных параметров смеси

Воспользуемся алгоритмом Expectation-Maximization для модели смеси распределений, описанным, например, в [14] и [15]:

Algorithm 1 Итерационный алгоритм Expectation-Maximization

Input: выборка $\mathbf{X}$, число компонент смеси K, N_{iter}, начальные значения параметров $(\boldsymbol{w}^{(0)}, \mathbf{U}^{(0)})$;

1: **for** $t = 0, ...N_{\text{iter}} - 1$: **do**

2: E-шаг (expectation): оценка апостериорного распределения скрытых переменных $\mathbf{z}|\mathbf{X}, (\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})$:

$\mathbf{z} = (z_1, ..., z_n)^\intercal, z_i \in \{1, ..., K\}$ — номер компоненты, соответствующей $\boldsymbol{x}_i$,

$$x_i|(z_i = k) \sim p_k(r_i, \theta_i, \varphi_i|\boldsymbol{u}_k^{(t)}),$$
$$g_{ik}^{(t)} = \mathbb{P}\Big[z_i = k|\boldsymbol{x}_i, (\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})\Big].$$

Для всех $i = 1, ..., n$, $k = 1, ..., K$:

$$g_{ik}^{(t)} := \frac{w_k^{(t)} p_k(r_i, \theta_i, \varphi_i|\boldsymbol{u}_k^{(t)})}{\sum_{s=1}^K w_s^{(t)} p_s(r_i, \theta_i, \varphi_i|\boldsymbol{u}_s^{(t)})}.$$

3: M-шаг(maximization): максимизация матожидания правдоподобия по отношению к апостериорному распределению скрытых переменных

$$(\boldsymbol{w}^{(t+1)}, \mathbf{U}^{(t+1)}) = \underset{(\boldsymbol{w}, \mathbf{U})}{\operatorname{argmax}} \mathbb{E}_{\mathbf{z}|\mathbf{X}, (\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})} \Big[\ln L\big(\mathbf{X}, \mathbf{z}|(\boldsymbol{w}, \mathbf{U})\big)\Big]$$

.

Для всех $k = 1, ..., K$:

$$w_k^{(t+1)} := \frac{1}{n} \sum_{i=1}^n g_{ik}^{(t)},$$

$$\boldsymbol{u}_k^{(t+1)} := \underset{\boldsymbol{u}}{\operatorname{argmax}} \sum_{i=1}^n g_{ik}^{(t)} \ln p_k(r_i, \theta_i, \varphi_i|\boldsymbol{u}).$$

4: **end for**

Output: $(\boldsymbol{w}^*, \mathbf{U}^*) := (\boldsymbol{w}^{(N_{\text{iter}})}, \mathbf{U}^{(N_{\text{iter}})})$ (окончательная оценка)

Учитывая (16), формула E-шага принимает вид:

$$g_{ik}^{(t)} := \frac{w_k^{(t)} \mathcal{N}(r_i|\mu_k^{(t)}, \sigma_k^{2(t)}) \mathcal{K}(\theta_i, \varphi_i|\boldsymbol{v}_k^{(t)})}{\sum_{s=1}^K w_s^{(t)} \mathcal{N}(r_i|\mu_s^{(t)}, \sigma_s^{2(t)}) \mathcal{K}(\theta_i, \varphi_i|\boldsymbol{v}_s^{(t)})}, \ i = \overline{1, n}, k = \overline{1, K}, \tag{17}$$

а формулы M-шага для обновления значений параметров $\mathbf{u}_k$ принимают вид:

$$\mu_k^{(t+1)}, \sigma_k^{2(t+1)} = \underset{\mu, \sigma^2}{\operatorname{argmax}} \sum_{i=1}^n g_{ik}^{(t)} \ln \mathcal{N}(r_i|\mu, \sigma^2), \tag{18}$$

$$\boldsymbol{v}_k^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{v}} \sum_{i=1}^{n} g_{ik}^{(t)} \ln \mathcal{K}(\theta_i, \varphi_i|\boldsymbol{v}). \tag{19}$$

Первая из этих задач максимизации (18) имеет аналитическое решение:

$$\mu_k^{(t+1)} = \frac{1}{nw_k^{(t+1)}} \sum_{i=1}^{n} g_{ik}^{(t)} r_i;\ \sigma_k^{2(t+1)} = \frac{1}{nw_k^{(t+1)}} \sum_{i=1}^{n} g_{ik}^{(t)} (r_i - \mu_k^{(t+1)})^2. \tag{20}$$

Решению задачи (19) посвящён следующий раздел.

2.4 Обновление параметров распределения Кента

Задача максимизации взвешенного правдоподобия для распределения Кента (19), возникающая на M-шаге EM-алгоритма, не имеет аналитического решения. При этом построение численного решения неэффективно — его приходится строить на каждой итерации алгоритма. Поэтому в алгоритм вносятся модификации, призванные упростить вычисление параметров $\boldsymbol{v}_k$.

2.4.1 Стохастическая модификация EM-алгоритма

Рассмотрим следующую модификацию EM-алгоритма, описанную в работе [16] под названием SEM:

1. Для каждого объекта выборки $\boldsymbol{x}_i$ сэмплируем значение s_i из апостериорного распределения $z_i|\mathbf{X}, (\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})$:

$$s_i \sim z_i,\ \mathbb{P}\Big[z_i = k|\boldsymbol{x}_i, (\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})\Big] = g_{ik}^{(t)},\ i = \overline{1,n},\ k = \overline{1,K}. \tag{21}$$

   По сути, для каждого объекта выборки сэмплируется номер соответсвующей ему компоненты распределения.

2. Для каждого $k = \overline{1,K}$ формируем индексное множество $\mathcal{A}_k^{(t)}$:

$$\mathcal{A}_k^{(t)} = \big\{i \in \overline{1,n} \mid s_i = k\big\}. \tag{22}$$

   Множество $\mathcal{A}_k^{(t)}$ состоит из индексов тех объектов выборки, для которых был сэмплирован номер компоненты s_i, равный k.

3. Рассматривая матожидание на M-шаге:

$$\mathbb{E}_{\mathbf{z}|\mathbf{X},(\boldsymbol{w}^{(t)},\mathbf{U}^{(t)})}\Big[\ln L\big(\mathbf{X}, \mathbf{z}|(\boldsymbol{w}, \mathbf{U})\big)\Big] =$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{P}\Big[z_i = k|\boldsymbol{x}_i, (\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})\Big]\big(\ln w_k + \ln \mathcal{N}(r_i|\mu_k, \sigma_k^2) + \ln \mathcal{K}(\theta_i, \varphi_i|\boldsymbol{v}_k)\big),$$

   выделим в нём слагаемое

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{P}\Big[z_i = k|\boldsymbol{x}_i, (\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})\Big] \ln \mathcal{K}(\theta_i, \varphi_i|\boldsymbol{v}_k).$$

С помощью сэмплированных на шаге 1 значений s_i, оценим это слагаемое следующим образом:

$$\sum_{k=1}^{K} \sum_{i \in \mathcal{A}_k^{(t)}} \ln \mathcal{K}(\theta_i, \varphi_i | \boldsymbol{v}_k).$$

Отсюда получаем следующий вид формулы M-шага для обновления параметров $\boldsymbol{v}_k$:

$$\boldsymbol{v}_k^{(t+1)} := \operatorname*{argmax}_{\boldsymbol{v}} \sum_{i \in \mathcal{A}_k^{(t)}} \ln \mathcal{K}(\theta_i, \varphi_i | \boldsymbol{v}). \tag{23}$$

Отметим, что при такой модификиции для обновления параметров $\boldsymbol{v}_k^{(t+1)}$ мы перешли от задачи максимизации взвешенного правдоподобия (19) к более простой задаче максимизации правдоподобия (21) по подвыборке $\{(\theta_i, \varphi_i) \mid i \in \mathcal{A}_k^{(t)}\}$.

2.4.2 Моментные оценки параметров распределения Кента

Задача обновления параметров упростилась, но всё ещё не имеет аналитического решения. Однако, в своей статье [13] автор распределения Джон Кент определяет *моментные оценки* $\hat{\boldsymbol{v}}_{\mathrm{ME}}$ (ME — Moment Estimates) и формулирует для них следующую теорему:

Теорема 1 *(Кент, 1982) Моментные оценки параметров распределения Кента* $\hat{\kappa}_{ME}, \hat{\beta}_{ME}, \hat{\boldsymbol{\gamma}}_{1,ME}, \hat{\boldsymbol{\gamma}}_{2,ME}, \hat{\boldsymbol{\gamma}}_{3,ME}$ *по выборке* $\{(\theta_i, \varphi_i)\}$ *обладают следующими свойствами:*

- *являются несмещёнными состоятельными оценками истинных значений параметров;*
- *при малых значениях отношения* $2\beta/\kappa$ *близки к оценкам максимума правдоподобия.*

Отсюда получаем окончательный вид формулы M-шага:

$$\boldsymbol{v}_k^{(t+1)} = \hat{\boldsymbol{v}}_{\mathrm{ME}}\left(\left\{(\theta_i, \varphi_i) \mid i \in \mathcal{A}_k^{(t)}\right\}\right), \tag{24}$$

где за $\hat{\boldsymbol{v}}_{\mathrm{ME}}(\{(\theta_i, \varphi_i \mid i \in \mathcal{A}_k^{(t)})\}$ обозначены оценки, построенные по выборке $\{(\theta_i, \varphi_i)\}$. В следующем подразделе описан способ построения моментных оценок.

2.4.3 Аналитические формулы для моментных оценок

Шаг 1. Имеющуюся выборку пар углов $\{(\theta_i, \varphi_i)_{i=1}^n\}$ объёма n преобразуем в выборку координат единичных векторов $\{\mathbf{x}_i\}_{i=1}^n = \{[x_{1,i}, x_{2,i}, x_{3,i}]^\intercal\}_{i=1}^n\}$ по формулам (13).

Шаг 2. Рассчитаем выборочное среднее $\overline{\mathbf{x}}$ и матрицу $\mathbf{S}$:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i, \;\; \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\intercal.$$

Шаг 4. Обозначим за θ, φ сферические углы, соответствующие $\overline{\mathbf{x}}$. Построим ортогональную матрицу поворота H:

$$\mathbf{H} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta\cos\varphi & \cos\theta\cos\varphi & -\sin\varphi \\ \sin\theta\sin\varphi & \cos\theta\sin\varphi & \cos\varphi \end{bmatrix}.$$

Шаг 5. Сформируем матрицу $\mathbf{B} = \mathbf{H}^\intercal\mathbf{S}\mathbf{H}$. Обозначим за l_1, l_2 ($l_1 > l_2$) собственные значения подматрицы

$$\mathbf{B}_L = \begin{bmatrix} b_{22} & b_{23} \\ b_{32} & b_{33} \end{bmatrix}.$$

Шаг 6. Выберем угол ψ и построим матрицу поворота $\mathbf{P}$, диагонализующую подматрицу $\mathbf{B}_L$:

$$\tan 2\psi = 2b_{23}/(b_{22} - b_{33}),\ \mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix}.$$

Шаг 7. Оценкой матрицы $\mathbf{\Gamma}$ служит матрица композиции описанных выше преобразований:

$$\hat{\mathbf{\Gamma}}_{\text{ME}} = (\hat{\boldsymbol{\gamma}}_{1,\text{ME}}, \hat{\boldsymbol{\gamma}}_{2,\text{ME}}, \hat{\boldsymbol{\gamma}}_{3,\text{ME}}) = \mathbf{HP}.$$

По сути, матрица $\mathbf{\Gamma}$ оценивается при помощи композиции двух поворотов: первый из них совмещает одну из осей декартовой системы координат со средним направлением распределения, а второй поворачивает систему координат относительно этой оси, задавая таким образом ориентацию эллиптических контуров равной вероятности.

Шаг 8. Сформируем пару вспомогательных величин:

$$r_1 = ||\overline{\mathbf{x}}||,\ r_2 = l_1 - l_2.$$

Оценка параметров κ, β (при больших значениях κ):

$$\hat{\kappa}_{\text{ME}} \approx (2 - 2r_1 - r_2)^{-1} + (2 - 2r_1 + r_2)^{-1};$$

$$\hat{\beta}_{\text{ME}} \approx (2 - 2r_1 - r_2)^{-1} - (2 - 2r_1 + r_2)^{-1}.$$

2.5 Определение числа компонент в модели смеси распределений

Число компонент смеси K выберем из эмпирически определённого интервала значений $\{1, 2, ..., 20\}$ на основе *информационного критерия Акаике*: для каждого значения из указанного диапазона находятся оптимальные параметры модели и вычисляется значение критерия:

$$\text{AIC} = 2m - 2\ln L(\mathbf{X}|\boldsymbol{w}^*, \mathbf{U}^*),$$

где m — число параметров модели, $\ln L(\mathbf{X}|\boldsymbol{w}^*, \mathbf{U}^*)$ — логарифм правдоподобия модели с оптимальными параметрами (8). Выбирается число компонент, для которого критерий принимает наименьшее значение. Аналогичный метод нахождения числа компонент предложен в [15].

2.6 Инициализация параметров смеси в алгоритме

Начальные значения весов w_k при выбранном значении K полагаются равными $1/K$, число итераций алгоритма N_{iter} полагается равным 200. Начальные значения параметров плотностей выбираются следующим образом:

1. Случайным образом выбираются K элементов $\{(\tilde{r}_k, \tilde{\theta}_k, \tilde{\varphi}_k\}_{k=1}^K$ выборки (поощряется равномерный выбор элементов по r).

2. Начальные значения параметров (для $k = \overline{1, K}$):

   $\mu_k = \tilde{r}_k$;

   $\sigma_k^2 = 1$;

   $\kappa_k = 10$;

   $\beta_k = 0$;

   $\boldsymbol{\gamma}_{1,k} = (\cos\tilde{\theta}_k, \sin\tilde{\theta}_k \cos\tilde{\varphi}_k, \sin\tilde{\theta}_k \sin\tilde{\varphi}_k)^\intercal$;

   $\boldsymbol{\gamma}_{2,k}, \boldsymbol{\gamma}_{3,k}$ подбираются произвольным образом для построения ортогональной матрицы $(\boldsymbol{\gamma}_{1,k}, \boldsymbol{\gamma}_{2,k}, \boldsymbol{\gamma}_{3,k})$.

2.7 Окончательный вид алгоритма поиска оптимальных параметров

Приведем окончательный вид предложенного алгоритма, согласно выражениям $(17), (20) - (22), (24)$:

Algorithm 2 Модификация алгоритма Expectation-Maximization

Input: выборка $\mathbf{X}$, число компонент смеси K, N_{iter}, начальные значения параметров $(\boldsymbol{w}^{(0)}, \mathbf{U}^{(0)})$;

1: **for** $t = 0, ...N_{\text{iter}} - 1$: **do**

2: E-шаг (expectation): оценка апостериорного распределения скрытых переменных.
Для всех $i = 1, ..., n$, $k = 1, ..., K$:

$$g_{ik}^{(t)} := \frac{w_k^{(t)} \mathcal{N}(r_i|\mu_k^{(t)}, \sigma_k^{2(t)}) \mathcal{K}(\theta_i, \varphi_i|\boldsymbol{v}_k^{(t)})}{\sum_{s=1}^{K} w_s^{(t)} \mathcal{N}(r_i|\mu_s^{(t)}, \sigma_s^{2(t)}) \mathcal{K}(\theta_i, \varphi_i|\boldsymbol{v}_s^{(t)})}.$$

3: S-шаг(sampling): сэмплирование из апостериорного распределения скрытых переменных.
Для всех $i = 1, ..., n$, $k = 1, ..., K$:

$$s_i \sim z_i, \ \mathbb{P}(z_i = k|\boldsymbol{x}_i, (\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})) = g_{ik}^{(t)}, \ i = \overline{1, n}.$$

$$\mathcal{A}_k^{(t)} = \left\{ i \in \overline{1, n} \mid s_i = k \right\}.$$

4: M-шаг(maximization): максимизация матожидания правдоподобия для весов w_k и параметров μ_k, σ_k^2; максимизация правдоподобия для параметров $\boldsymbol{v}_k$.
Для всех $k = 1, ..., K$:

$$w_k^{(t+1)} := \frac{1}{n} \sum_{i=1}^{n} g_{ik}^{(t)},$$

$$\mu_k^{(t+1)} = \frac{1}{n w_k^{(t+1)}} \sum_{i=1}^{n} g_{ik}^{(t)} r_i,$$

$$\sigma_k^2 = \frac{1}{n w_k^{(t+1)}} \sum_{i=1}^{n} g_{ik}^{(t)} (r_i - \mu_k^{(t+1)})^2.$$

$$\boldsymbol{v}_k^{(t+1)} = \hat{\boldsymbol{v}}_{\text{ME}}(\{(\theta_i, \varphi_i \mid i \in \mathcal{A}_k^{(t)})\},$$

5: **end for**

Output: $(\boldsymbol{w}^*, \mathbf{U}^*) := (\boldsymbol{w}^{(N_{\text{iter}})}, \mathbf{U}^{(N_{\text{iter}})})$ (окончательная оценка)

3 Вычислительный эксперимент

В экспериментальной части демонстрируется работа алгоритма нахождения оптимальных параметров смеси для задачи восстановления плотности распределения пространственных ориентаций различных пар вида аминокислота-лиганд.

3.1 Экспериментальные данные

Данные представляют собой 47916041 пятерку значений, элементы каждой пятерки: a — индекс аминокислоты, b — индекс лиганда и тройка r, θ, φ. Индексы аминокислоты принимают 21 различное значение, индексы лиганда — 40, соответственно они образуют 840 пар. Сформированные пары используются для разделения данных на 840 соответствующих выборок (r, θ, φ). Для каждой из 840 выборок должна быть построена восстановленная плотность $\hat{p}^{a,b}(r, \theta, \varphi)$.

3.2 Восстановление плотности распределения пространственных конфигураций пары ALA-C_{ar}

Ниже приведены результаты восстановления плотности на примере пары аминокислоты-лиганда ALA-C_{ar} с индексами 0-2. Программная реализация и полученные изображения расположены в репозитории [17], реализация распределения Кента основана на результатах [18].

По имеющейся выборке из 111801 тройки (r, θ, φ) ищутся оптимальные параметры смеси с помощью модификации алгоритма Expectation-Maximization, описанной в предыдущем разделе. Параметры подбираются для числа компонент из диапазона $1, 2, ..., 20$. Оптимальное с точки зрения критерия AIC число компонент оказывается равно 18.
Для проверки сходимости алгоритма и контроля переобучения организуется следующая процедура:

1. Данные разбиваются на 5 фолдов в соответствии со схемой кросс-валидации.

2. Оптимальные параметры для смеси из 18 компонент ищутся для 5 обучающих выборок, полученных отбрасыванием одного из фолдов. Отброшенная часть выборки объявляется тестовой. Размеры обучающей и тестовой выборок n_{train} и n_{test} соответственно.

3. На каждой итерации алгоритма вычисляются средние по объёмам выборок логарифмы правдоподобия:

$$\frac{\ln L(\mathbf{X}_{\text{train}}|\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})}{n_{\text{train}}} \text{ и } \frac{\ln L(\mathbf{X}_{\text{test}}|\boldsymbol{w}^{(t)}, \mathbf{U}^{(t)})}{n_{\text{test}}}.$$

Средние по кросс-валидации значения этих отношений, взятые с противоположным знаком, приведены на следующем графике.

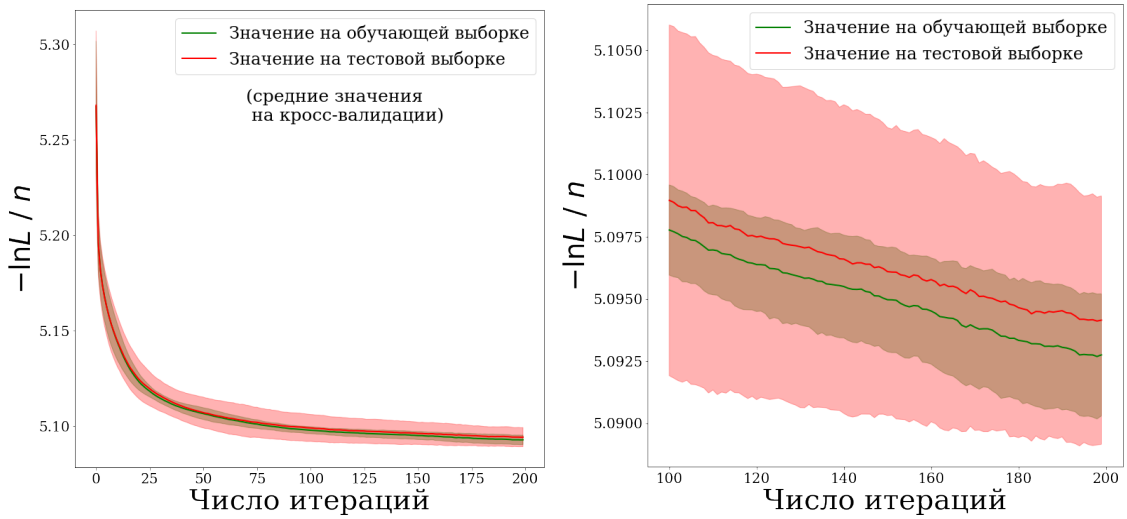

Рис. 7: Среднее на кросс-валидации значение отношения логарифма правдоподобия к объёму, соответственно, обучающей и тестовой выборок. График иллюстрирует сходимость алгоритма и отсутствие переобучения.

Поскольку график восстановленной функции плотности, зависящей от трёх переменных, является поверхностью в четырёхмерном пространстве, иллюстрировать результаты восстановления предлагается с помощью проекций графика на трёхмерное пространство $r = \text{const}$ для разных значений r.

Ниже приведены иллюстрации для значений $r = 7$Å, $r = 11$Å и $r = 15$Å. Каждое значение r проиллюстрировано 4 изображениями:

1. Множество точек выборки в небольшом слое $r \pm \delta$, где $\delta = 0.5$. Данное изображение позволяет примерно обозначить скопления точек в небольшой окрестности заданного значения r.

2. Двумерное изображение плотности в виде *colormap* на плоскости (φ, θ) с указанием максимумов плотности $\hat{p}(r, \theta, \varphi)$, попавших в заданный диапазон значений r. Размер точки, соответствующей максимуму, зависит от высоты и сконцентрированности пика.

3. Трёхмерное изображение графика функции плотности смеси распределений с найденными алгоритмом параметрами, значение r зафиксировано.

4. Трёхмерное изображение восстановленной плотности в виде параметрической поверхности
$$R(\theta, \varphi) = 1 + \hat{p}(r = r_i, \theta, \varphi)$$
в декартовой системе координат (x, y, z). Служит иллюстрацией периодичности восстановленной плотности по углам (θ, φ).

3.2.1 Иллюстрации для $r = 7$Å

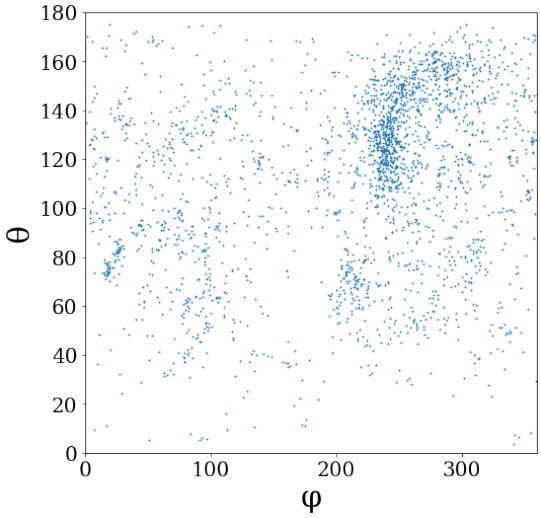

Рис. 8: Множество элементов выборки в диапазоне расстояний $r = 7 \pm 0.5$, спроецированное на плоскость (φ, θ).

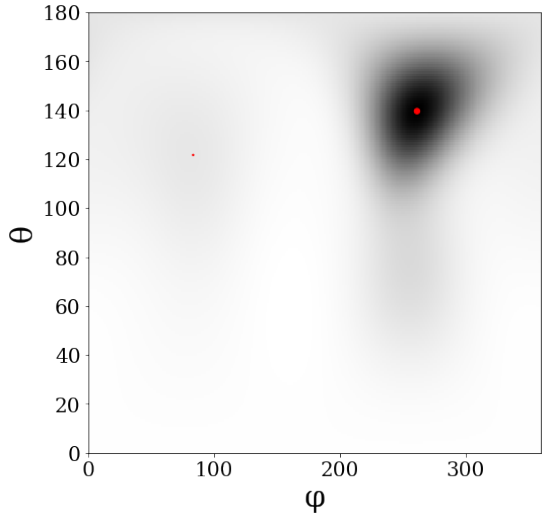

Рис. 9: Двумерное полутоновое изображение восстановленной плотности $\hat{p}(r = 7\text{Å}, \theta, \varphi)$; красная точка соответствует максимуму плотности, попавшему в диапазон $r = 7 \pm 0.5$.

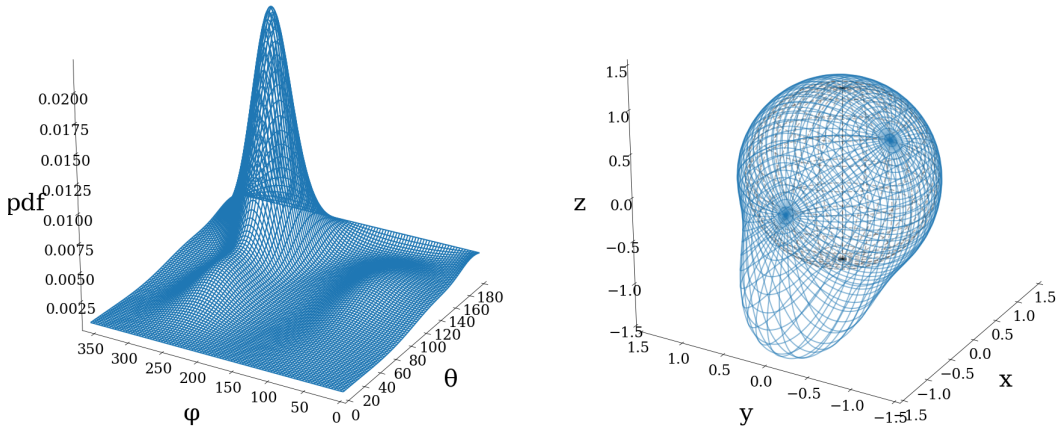

Рис. 10: Трёхмерное изображение восстановленной плотности $\hat{p}(r = 7\text{Å}, \theta, \varphi)$: в виде графика функции переменных (θ, φ) (слева) и в виде поверхности (справа).

3.2.2 Иллюстрации для $r = 11$Å

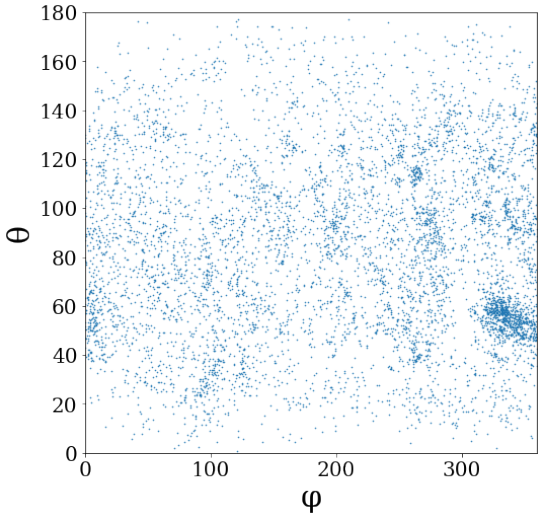

Рис. 11: Множество элементов выборки в диапазоне расстояний $r = 11 \pm 0.5$, спроецированное на плоскость (φ, θ).

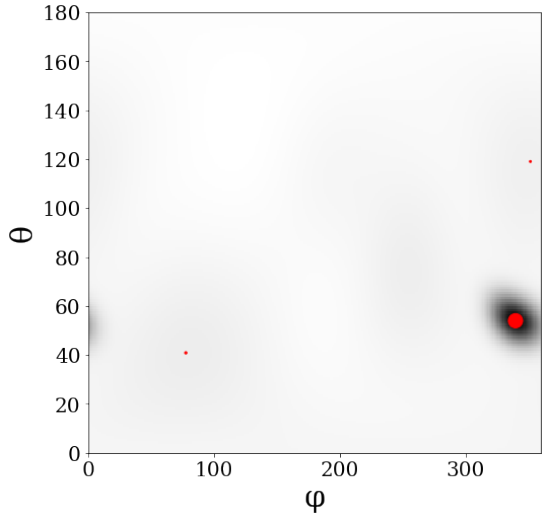

Рис. 12: Двумерное полутоновое изображение восстановленной плотности $\hat{p}(r = 11\text{Å}, \theta, \varphi)$; красная точка соответствует максимуму плотности, попавшему в диапазон $r = 11 \pm 0.5$.

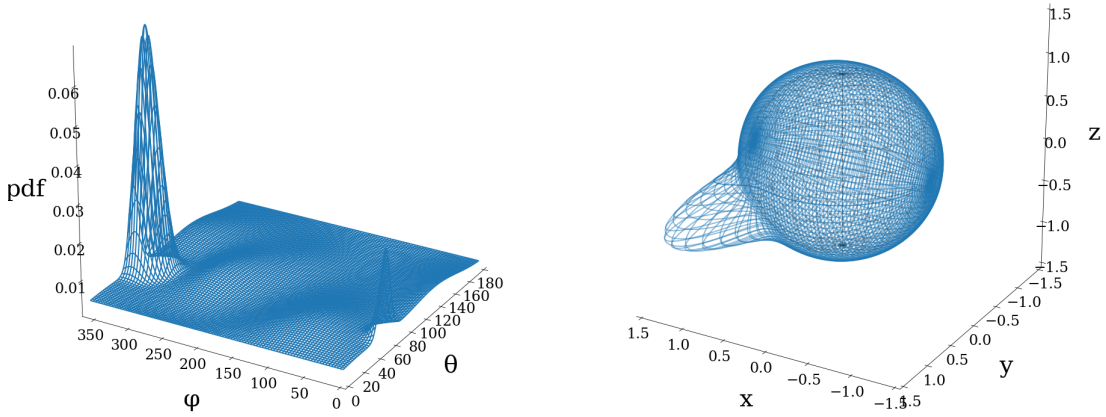

Рис. 13: Трёхмерное изображение восстановленной плотности $\hat{p}(r = 11\text{Å}, \theta, \varphi)$: в виде графика функции переменных (θ, φ) (слева) и в виде поверхности (справа).

3.2.3 Иллюстрации для $r = 15$Å

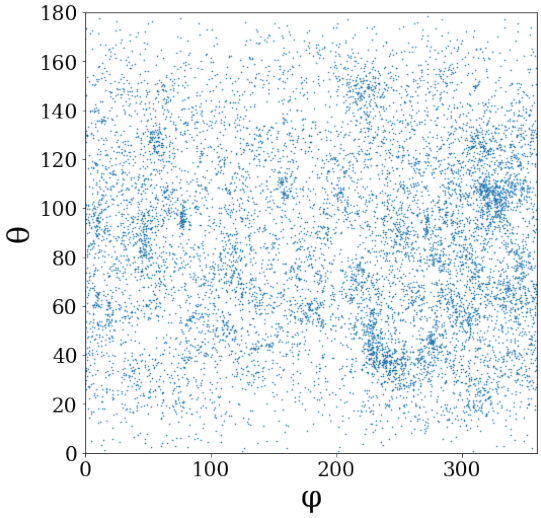

Рис. 14: Множество элементов выборки в диапазоне расстояний $r = 15 \pm 0.5$, спроецированное на плоскость (φ, θ).

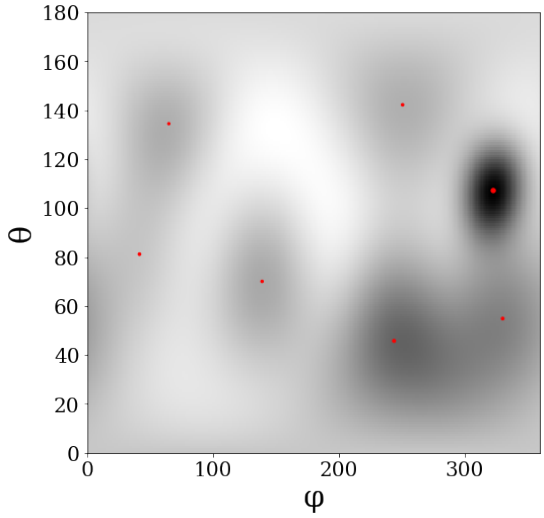

Рис. 15: Двумерное полутоновое изображение восстановленной плотности $\hat{p}(r = 15\text{Å}, \theta, \varphi)$; красная точка соответствует максимуму плотности, попавшему в диапазон $r = 15 \pm 2$.

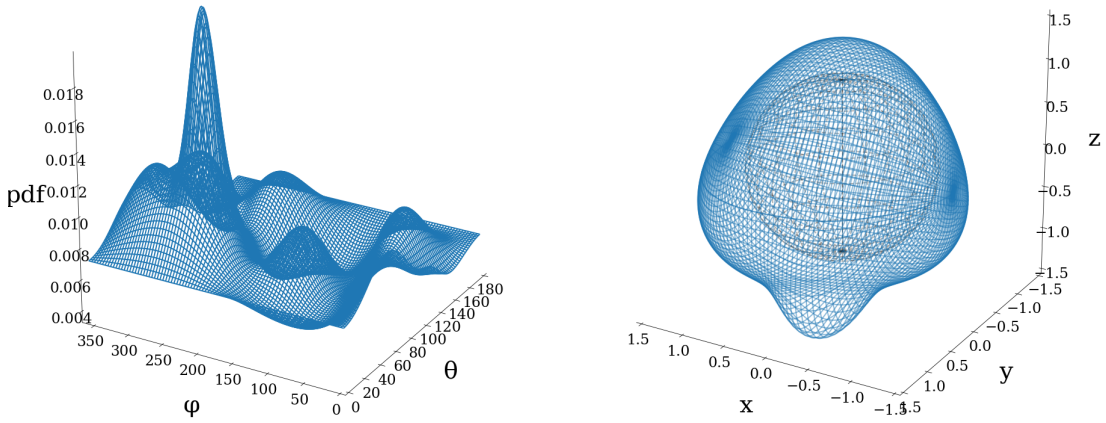

Рис. 16: Трёхмерное изображение восстановленной плотности $\hat{p}(r = 15\text{Å}, \theta, \varphi)$: в виде графика функции переменных (θ, φ) (слева) и в виде поверхности (справа).

3.3 Установление соответствия с другими моделями

Ниже демонстрируется соответствие проекций восстановленных плотностей для значений $r = 7$Å и $r = 11$Å с результатами восстановления плотности на основе 4-ёх моделей: модель окна Парзена-Розенблатта (Parzen), модель смеси гауссиан (GM), гистограмма плотности (Hist) и простая нейросетевая модель — однослойный перцептрон (NN). Иллюстрации для вспомогательных моделей взяты из работы [19].

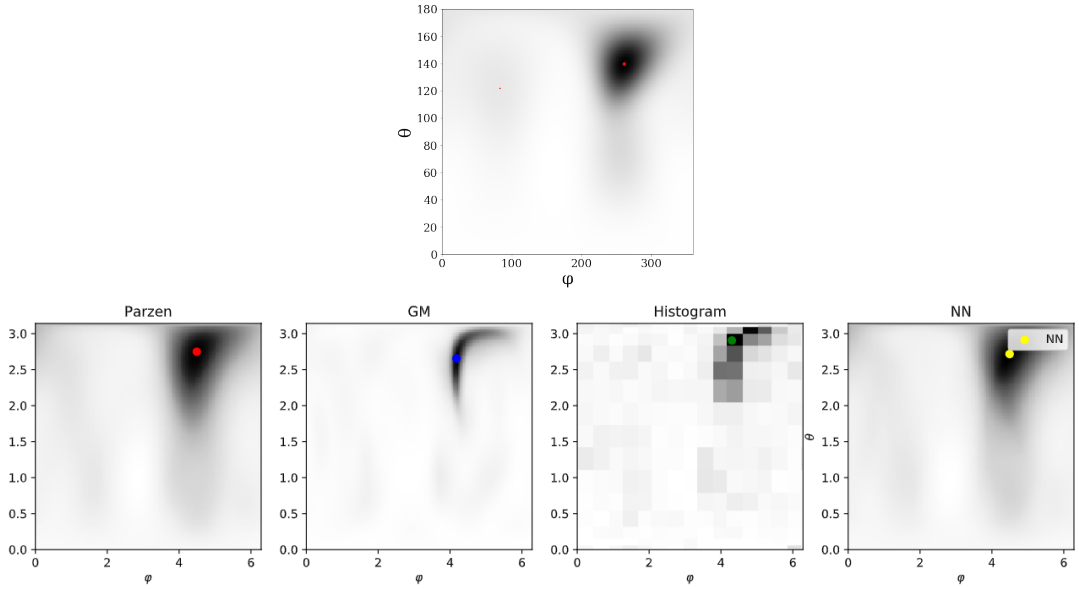

Рис. 17: Соответствие восстановленной плотности (сверху) результатам, полученным с помощью других моделей восстановления (снизу) для $r = 7$Å.

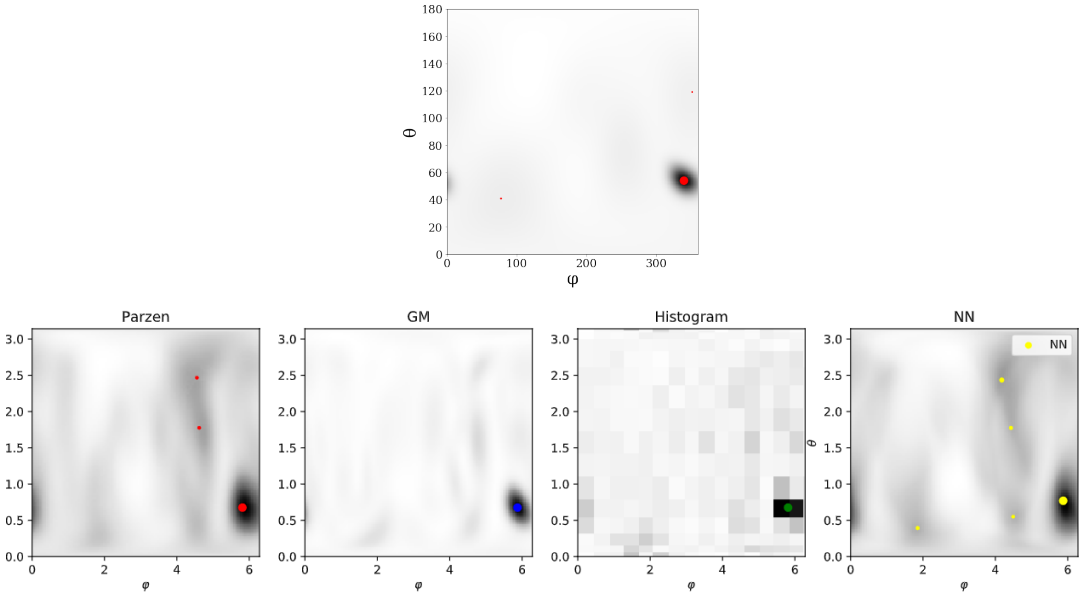

Рис. 18: Соответствие восстановленной плотности (сверху) результатам, полученным с помощью других моделей восстановления (снизу) для $r = 11$Å.

Заключение

Предложен алгоритм нахождения параметров смеси распределений для построения аппроксимирующей плотности, в которой распределение угловых переменных описывается с помощью случайных величин с носителем на сфере в трёхмерном пространстве. Алгоритм учитывает особенности распределения при выполнении итерационного обновления параметров. Применение алгоритма позволило восстановить плотности распределения взаимных пространственных ориентаций взаимодействующих молекул. Полученные плотности согласуются с результатами применения более простых моделей и, как следствие, с мнением эксперта.

Обозначения

$\mathbf{X}^{a,b} = \{\boldsymbol{x}_i\}_{i=1}^n$ — выборка объёма n, краткое обозначение $\mathbf{X}$;

$\boldsymbol{x}_i = [r_i, \theta_i, \varphi_i]^\intercal$ — элемент выборки;

a — тип взаимодействующей аминокислоты;

b — тип взаимодействующего лиганда;

r_i — длина химической связи между аминокислотой a и лигандом b;

θ_i и φ_i — сферические углы лиганда в системе координат аминокислоты;

$\Omega = [3\text{Å}, 20\text{Å}] \times [0, \pi] \times [0, 2\pi]$ — множество возможных значений $\boldsymbol{x}_i$;

$\hat{p}^{a,b}(r, \theta, \phi)$ — искомая плотность, аппроксимирующая истинную $p^{a,b}(r, \theta, \phi)$, краткое обозначение $\hat{p}(r, \theta, \phi)$;

$p(\boldsymbol{x}|\boldsymbol{w}, \mathbf{U}) = \sum_{k=1}^K w_j p_k(r, \theta, \varphi|\boldsymbol{u}_k)$ — модель смеси распределений;

K — число компонент смеси;

$w_k = p(k)$ — априорная вероятность k-ой компоненты смеси;

$p_k(r, \theta, \varphi|\boldsymbol{u}_k)$ — плотность распределения k-ой компоненты смеси;

$\boldsymbol{u}_k$ — вектор параметров распределения компоненты смеси;

$(\boldsymbol{w}, \mathbf{U}) = (w_1, ..., w_K, \boldsymbol{u}_1, ..., \boldsymbol{u}_K)$ — совокупность параметров модели;

$(\boldsymbol{w}^*, \mathbf{U}^*) = \underset{\boldsymbol{w}, \mathbf{U}}{\operatorname{argmax}} \ln L(\mathbf{X}|\boldsymbol{w}, \mathbf{U})$ — оптимальные параметры модели;

$\mathcal{K}(\theta, \varphi|\kappa, \beta, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3)$ — функция плотности распределения Кента, краткое обозначение $\mathcal{K}(\theta, \varphi|\boldsymbol{v})$;

$\boldsymbol{v} = [\kappa, \beta, \boldsymbol{\gamma}_1^\intercal, \boldsymbol{\gamma}_2^\intercal, \boldsymbol{\gamma}_3^\intercal,]^\intercal$ — вектор параметров распределения Кента;

$\mathcal{N}(r|\mu_k, \sigma_k^2)$ — функция плотности нормального распределения с параметрами μ_k, σ_k^2;

$g_{ik}^{(t)}$ — апостериорные вероятности значений скрытых переменных $\mathbf{z}$ на t-ой итерации ЕМ-алгоритма;

$s_i \sim z_i$ — значения, сэмплированные из апостериорного распределения;

$\mathcal{A}_k^{(t)}$ — индексное множество объектов выборки, для которых был сэмплирован номер компоненты, равный k;

$\hat{\boldsymbol{v}}_{\text{ME}}$ — моментные оценки.

Список литературы

[1] Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip K. Paliwal, and Yaoqi Zhou. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in Bioinformatics*, 19(3):482–494, 2018.

[2] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[3] M. Scott Shell, S. Banu Ozkan, Vincent Voelz, Guohong Albert Wu, and Ken A. Dill. Blind test of physics-based prediction of protein structures.

[4] Cezary Czaplewski, Agnieszka Karczynska, Adam K. Sieradzan, and Adam Liwo. Unres server for physics-based coarse-grained simulations and prediction of protein structure, dynamics and thermodynamics. *Nucleic Acids Research*, 46(Webserver-Issue):W304–W309, 2018.

[5] José Ramón López-Blanco and Pablo Chacón. Korp: knowledge-based 6d potential for fast protein and loop modeling. *Bioinformatics*, 2019.

[6] Petr Popov and Sergei Grudinin. Knowledge of native protein-protein interfaces is sufficient to construct predictive models for the selection of binding candidates. *Journal of chemical information and modeling*, 55 10:2242–55, 2015.

[7] Maria Kadukova and Sergei Grudinin. Convex-pl: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. *Journal of computer-aided molecular design*, 31(10):943—958, October 2017.

[8] Dennis M. Krüger, José Ignacio Garzón, Pablo Chacón, and Holger Gohlke. Drugscoreppi knowledge-based potentials used as scoring and objective function in protein-protein docking. *PLOS ONE*, 9(2):1–12, 02 2014.

[9] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp) - round xiii. *Proteins*, 2019.

[10] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020, 2019.

[11] Thomas Hamelryck, John T. Kent, and Anders Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Computational Biology*, 2(9), 2006.

[12] Parthan Kasarapu. Minimum message length based mixture modelling using bivariate von mises distributions with applications to bioinformatics, July 05 2016. Comment: arXiv admin note: text overlap with arXiv:1506.08105.

[13] J. T. Kent. The fisher-bingham distribution on the sphere. *J. Royal Stat. Soc. Series B (Methodological)*, 44(1):71–80, 1982.

[14] M. Aitkin and G. Tunnicliffe Wilson. Mixture models, outliers, and the EM algorithm. *Technometrics*, 22:325–331, 1980.

[15] David Peel, William J. Whiten, and Geoffrey J. Mclachlan. Fitting mixtures of kent distributions to aid in joint set identification, March 15 1999.

[16] Gilles Celeux, Didier Chauveau, and Jean Diebolt. On stochastic versions of the em algorithm. info:eu-repo/semantics/report; reports, HAL CCSD, 1995.

[17] Panchenko Sviatoslav. Modification of a Stochastic EM algorithm. https://github.com/PanchenkoSviatoslav/Panchenko2020Thesis, 2020.

[18] Daniel Fraenkel. Kent Distribution. https://github.com/edfraenkel/kent_distribution, 2017.

[19] Uvarov Nikita. Probabilistic Metric Spaces. https://github.com/Intelligent-Systems-Phystech/ProbabilisticMetricSpaces, 2020.