# Построение вероятностного метрического пространства для моделирования зависимых от ориентации состояний

Панченко Святослав

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В. В. Стрижов

Москва,
2020 г.

# Моделирование плотности распределения ориентаций

## Задача

Восстановить плотности распределения пространственных ориентаций пары аминокислота-лиганд. Ориентация задаётся расстоянием связи $r$ и парой сферических углов $(\theta, \varphi)$.

## Требования к модели восстанавления плотности

- интерпретируемость с точки зрения эксперта;
- согласованность с ранее полученными результатами применения более простых моделей;
- описание распределения угловых величин в естественном для них пространстве.

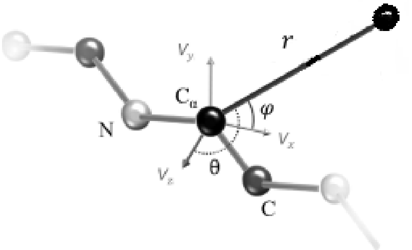
## Идея

Пары сферических углов $(\theta, \varphi)$ моделируются как реализации случайной величины, описывающейся распределением Кента, область значений которого — это сфера в трёхмерном пространстве, а не $\mathbb{R}^n$.

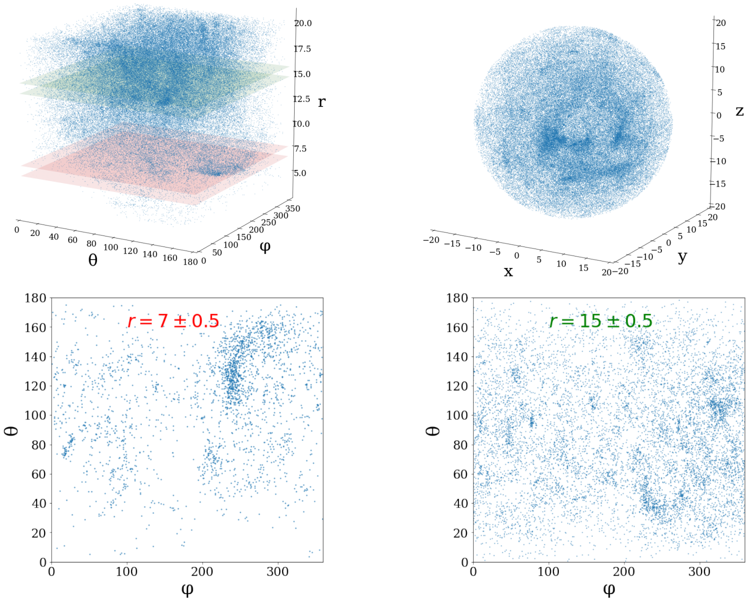
# Описание молекулярной химической связи

В данной работе исследуются взаимные пространственные ориентации различных пар молекул, образующих между собой химическую связь. Эта связь характеризуется тремя параметрами:

- $r$ — расстояние между молекулами, $r \in [3Å, 20Å]$;
- $(\theta, \varphi)$ — пара сферических углов, определяющих положение лиганда в системе координат аминокислоты, *theta* $\in [0, \pi]$, $\varphi \in [0, 2\pi]$.



# Представление выборки для пары ALA-$C_{ar}$



# Базовые публикации


- *Kent J. T.* (1982) 'The Fisher–Bingham distribution on the sphere.', J. Royal. Stat. Soc.
- *Kent J. T., Hamelryck T.* (2005) 'Using the Fisher–Bingham distribution in stochastic models for protein structure.' Leeds, Leeds University Press.
- *Whiten W.J.* (2001) 'Fitting mixtures of Kent distributions to aid in joint set identification.', Am. Stat. Ass.
- *Hamelryck, Thomas; Kent, John T.; Krogh, Anders* (2006) 'Sampling realistic protein conformations using local structural bias.' PLoS Comput. Biol.
- *Jean Diebolt, Eddie H.S.Ip* (1994) 'A Stochastic EM algorithm for approximating the maximum likelyhood estimate.' Department of Statistics, Stanford University.

# Постановка задачи восстановления плотности

## Дано

Выборка $\mathbf{X}^{a,b} = \{\boldsymbol{x}_i\}_{i=1}^n$, $\boldsymbol{x}_i = [r, \theta, \varphi]^\mathsf{T} \in \Omega$,
$\Omega = [3\text{Å}, 20\text{Å}] \times [0, \pi] \times [0, 2\pi]$.

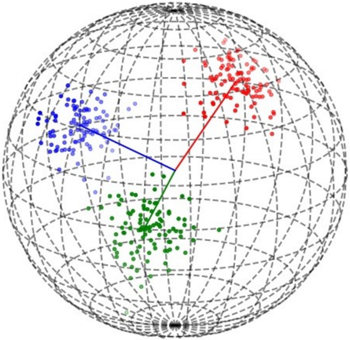
## Модель восстанавливаемой плотности $p(\boldsymbol{x}|\boldsymbol{w}, \mathbf{U})$

$$p(\boldsymbol{x}|\boldsymbol{w}, \mathbf{U}) = \sum_{k=1}^{K} w_k p_k(r) f_k(\theta, \varphi), \ \sum_{k=1}^{K} w_k = 1, \ w_k \geq 0,$$

- $K$ — число компонент смеси;
- $(\boldsymbol{w}, \mathbf{U})$ — совокупность параметров модели;
- $w_k = p(k)$ — априорная вероятность $k$-ой компоненты;
- $p_k(r) = \mathcal{N}(r|\mu_k, \sigma_k)$ — нормальное распределение;
- $f_k(\theta, \varphi) = \mathcal{K}(\theta, \varphi|\boldsymbol{v}_k)$ — распределение Кента.

# Распределение Кента

*5-параметрическое распределение Фишера-Бингхама* или *распределение Кента* — это аналог двумерного нормального распределения на сфере в трёхмерном пространстве.



Примеры выборок из различных распределений Кента

# Функция плотности распределения Кента

$$f(\mathbf{x}) = \frac{1}{c(\kappa,\beta)} exp\left\{ \kappa\gamma_1^T\mathbf{x} + \beta\left[(\gamma_2^T\mathbf{x})^2 - (\gamma_3^T\mathbf{x})^2\right]\right\},$$

где $\mathbf{x}$ — единичный вектор, $3\times 3$-матрица $[\gamma_1, \gamma_2, \gamma_3]$ ортогональна, а $c(\kappa,\beta)$ — нормирующая константа:

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

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

# Переход к угловым переменным в распределении Кента

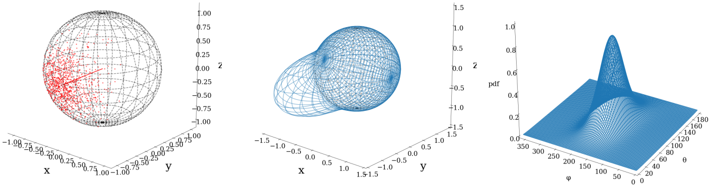
От компонент единичного вектора $[x_1, x_2, x_3]^\mathsf{T}$ перейдём к сферическим углам $\theta \in [0, \pi]$, $\varphi \in [0, 2\pi]$:

$$x_1 = \cos\theta,\ \ x_2 = \sin\theta\cos\varphi,\ \ x_3 = \sin\theta\sin\varphi,$$

В таком случае плотность распределения обозначим

$$\mathcal{K}(\theta, \varphi | \kappa, \beta, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3) \text{ или кратко } \mathcal{K}(\theta, \varphi | \boldsymbol{v})$$

.




Иллюстрации плотности распределения типичного представителя семейства распределений Кента

# Нахождение параметров смеси

## Модификация алгоритма Expectation-Maximization

1. E-шаг (expectation): оценка скрытых переменных.

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

2. S-шаг (sampling): сэмплирование из апостериорного распределения скрытых переменных.

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

Составим индексные множества $\mathcal{A}_k^{(t)} = \left\{ i \in \overline{1, n} \mid s_i = k \right\}$, соответстующие тем элементам выборки, для которых сэмплирован номер компоненты, равный $k$.

## Модификация алгоритма Expectation-Maximization

3. M-шаг(maximization): максимизация взвешенного правдоподобия для весов $w_k$ и параметров $\mu_k, \sigma_k^2$; максимизация правдоподобия для параметров $\boldsymbol{v}_k$.

Для всех $k = 1, ..., K$:

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

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

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

# Моментные оценки

При такой модификиции нахождение оптимальных параметров $\mathbf{v}_k$ на М-шаге — это нахождение оценок максимального правдоподобия на подвыборке $\{(\theta_i, \varphi_i) \mid i \in \mathcal{A}_k^{(t)}\}$. Эти оценки приблизим *моментными оценками* $\mathbf{v}_{\mathrm{ME}}$, для которых справедлива следующая теорема:

**Theorem**

*(Кент, 1982) Моментные оценки параметров распределения Кента $\kappa_{ME}, \beta_{ME}, \gamma_{1,ME}, \gamma_{2,ME}, \gamma_{3,ME}$ 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.*

Оценки моментов находятся по аналитическим формулам, предложенным Кентом.

# Вычислительный эксперимент

## Цель

Восстановить плотности распределения пространственных ориентаций различных пар вида аминокислота-лиганд.

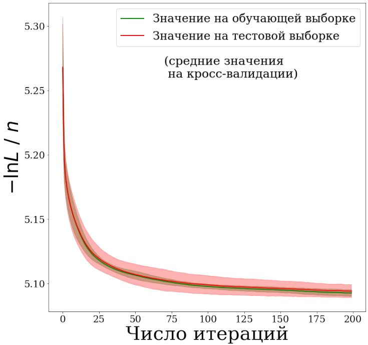
## Данные

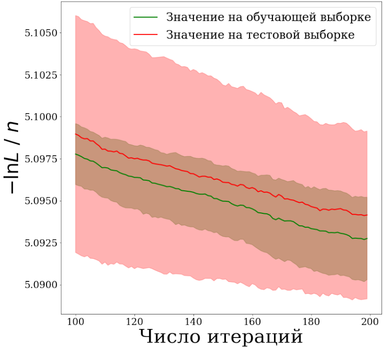
Данные представляют собой 47916041 пятерку значений, элементы каждой пятерки: $a$ — индекс аминокислоты, $b$ — индекс лиганда и тройка $r, \theta, \varphi$. Индексы аминокислоты и лиганды образуют 840 пар и используются для разделения данных на 840 выборок $(r, \theta, \varphi)$, каждая из которых соответсвует своей взаимодействующей паре.

## Описание эксперимента

Для каждой из 840 выборок строится восстановленная плотность $\hat{p}^{a,b}(r, \theta, \varphi) = p(r, \theta, \varphi | \boldsymbol{w}^*, \mathbf{U}^*)$.

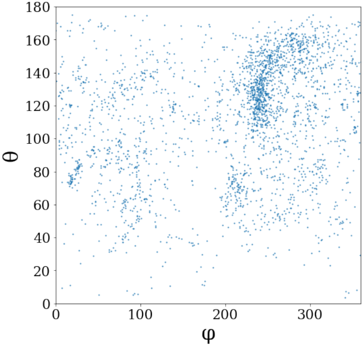
# Иллюстрация свойств алгоритма



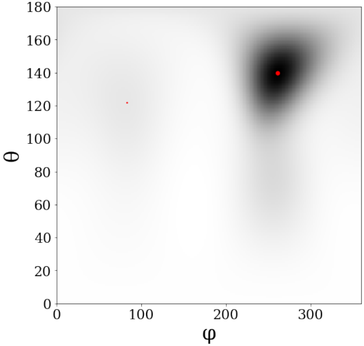




Среднее на кросс-валидации значение отношения логарифма правдоподобия к объёму, соответственно, обучающей и тестовой выборок. График иллюстрирует сходимость алгоритма и отсутствие переобучения.

# Результаты восстановления, пара $0-2$, $r=7$Å

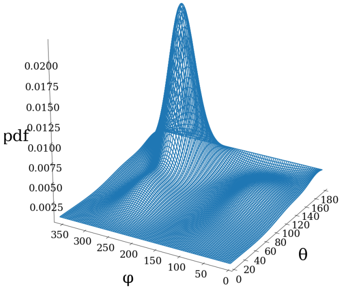


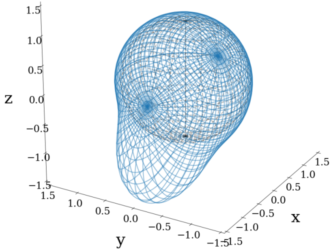

Множество элементов выборки в диапазоне расстояний $r = 7 \pm 0.5$, спроецированное на плоскость $(\varphi, \theta)$.




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

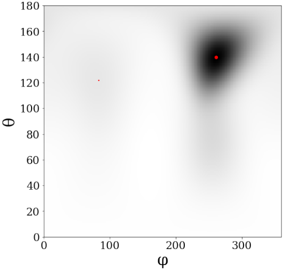
# Результаты восстановления, пара $0-2$, $r = 7Å$

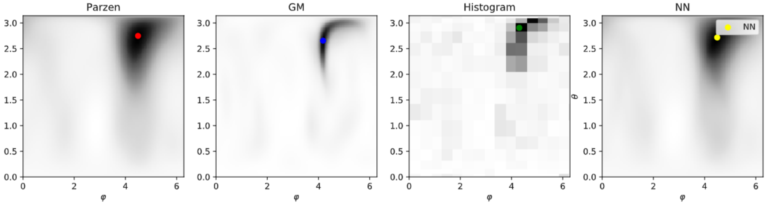



Трёхмерное изображение восстановленной плотности $\hat{p}(r = 7Å, \theta, \varphi)$: в виде графика функции переменных $(\theta, \varphi)$ (слева) и в виде поверхности (справа).

# Соответствие результатам простых моделей, $r = 7Å$




Соответствие восстановленной плотности (сверху) результатам, полученным с помощью других моделей восстановления (снизу).

# Выносится на защиту

1. Предложен алгоритм нахождения параметров смеси распределений Кента для моделирования параметров химической связи пары аминокислота-лиганд
2. Проведен анализ восстановленных плотностей, установлено соответствие найденных максимумов с результатами, полученными с помощью более простых моделей.