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

Уваров Никита Денисович

Построение вероятностных метрических пространств в задачах анализа молекулярных конфигураций

03.04.01 — Прикладные математика и физика (магистратура)

Магистерская диссертация

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

Москва
2020 г.

Содержание

Аннотация

Рассматривается задача ранжирования синтезированных молекулярных конформаций по энергетической устойчивости. Предлагается построить метрическое пространство на вероятностных распределениях взаимного расположения элементарных взаимодействующих пар аминокислота-лиганд. Вероятностные распределения восстанавливаются с использованием четырёх различных моделей: гистограммы, окна Парзена, смеси гауссиан и нейросети. Вводится критерий попарной согласованности распределений, составляется каталог согласованных экстремумов распределений для изучения экспертами. В качестве метрики, либо предметрики, используется f-дивергенция между распределениями. На основе метрики пространства строится ранжирующая функция для конформаций из данных CASF.

Ключевые слова: *вероятностное метрическое пространство, восстановление плотности, молекулярный докинг, CASF.*

1 Введение

Восстановление 3D структуры белка является ключевым шагом к пониманию его биологической функции, и в конечном счёте позволяет отбирать и синтезировать новые соединения, отвечающие заданным требованиям. Число экспериментально исследованных соединений на порядки меньше, чем число потенциальных кандидатов, получаемых лучшими вычислительными алгоритмами генерации кандидатов. Высокая стоимость и ограниченные возможности лабораторных экспериментов по синтезу соединений обуславливают необходимость разработки надёжных методов оценки устойчивости синтетических соединений.

В основе алгоритмов оценки энергетической устойчивости соединения лежит как правило либо физическая, либо статистическая [1] модель взаимодействия элементов молекулы. Физические модели аппроксимируют силовое поле с учётом всех пар элементов конформации. Статистические модели обычно используют отсечку по расстоянию между взаимодействующими элементами: таким образом достигается компромисс между сложностью модели и точностью оценки. Кроме того, для построения статистической модели вводятся дополнительные предположения о независимости расположения различных пар взаимодействующих элементов.

Использование современных моделей машинного обучения и искусственного интеллекта позволило существенно продвинуться в решении задачи молекулярного докинга с использованием статистического подхода [2, 3]. Данная работа посвящена разработке фундамента для применения метрических методов в задачах анализа молекулярных конфигураций.

2 Постановка задачи ранжирования конформаций белка CASF

Рассматривается случай взаимодействия белка и лиганда (маленькой молекулы). Белок представляется как совокупность аминокислот, каждая из которых относится к одному из 20 типов. Центральным объектом исследования является взаимное расположение аминокислоты и лиганда (рис. 1). Для описания пространственного расположения в аминокислоте выделяются три атома: углерод (C), альфа-углерод (C_α) и азот (N).

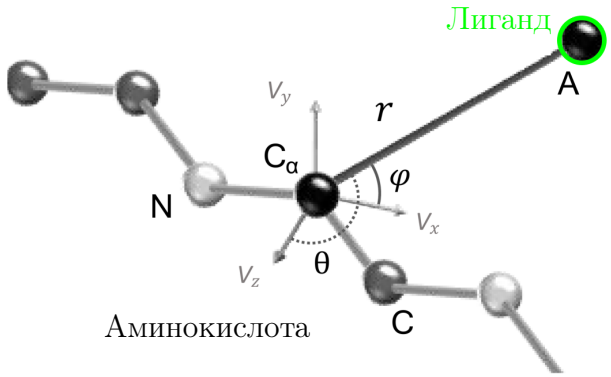


Рис. 1: Локальная система координат для пары аминокислота-лиганд

Вводится локальная система координат [1] с центром в атоме альфа-углерода C_α, осями которой являются векторы

$$\vec{V}_z = \frac{\overrightarrow{C_\alpha C} + \overrightarrow{C_\alpha N}}{|\overrightarrow{C_\alpha C} + \overrightarrow{C_\alpha N}|},\ \vec{V}_y = \frac{V_z \times \overrightarrow{C_\alpha N}}{|\vec{V}_z \times \overrightarrow{C_\alpha N}|},\ \vec{V}_x = \vec{V}_y \times \vec{V}_z.$$

В полученной системе координат взаимное расположение описывается одним вектором $\vec{r} = \overrightarrow{C_\alpha A}, \vec{r} \in \mathbb{R}^3$. Характерное расстояние $|\vec{r}|$ составляет от 4Å до 20Å.

Предположим, что взаимное расположение $\vec{r}$ не зависит от других элементов молекулярного комплекса, и таким образом является случайной величиной, рас-

пределение которой зависит от типа аминокислоты $a \in \mathcal{A}$ и типа лиганда $b \in \mathcal{B}$, $|\mathcal{A}| = 20, |\mathcal{B}| = 40$. Тогда паре $x = (a, b)$ соответствует вероятностное распределение $p_{a,b} = p_x(r, \varphi, \theta)$ на $\mathbb{R}^3$.

Задача ранжирования конформаций белка CASF [4] ставится следующим образом. Дан набор конформаций одного белка $\{z_i\}_{i=1}^{\ell}$, $z_i = \{(a, b, r, \varphi, \theta)\}_{j=1}^{\ell_i}$. Для обучающей выборки известны истинные значения энергии E_i конформаций. Необходимо построить модель, предсказания $\widetilde{E}_i$ которой ранжируют конформации по энергии. Критерием качества является точность в топ-5:

$$P@K = \frac{1}{K} \sum_{i:\text{Rank}(E_i) \leqslant K} [\text{Rank}(\widetilde{E}_i) \leqslant K], K = 5,$$

где $\text{Rank}(\widetilde{E}_i)$ - позиция $\widetilde{E}_i$ в порядке по возрастанию $\widetilde{E}_i$.

3 Построение метрического пространства

Пусть $\mu(t) : (0, \infty) \to \mathbb{R}$ - везде выпуклая функция, $f(1) = 0$, а $p_x(\vec{r})$ и $p_y(\vec{r})$ - два абсолютно непрерывных распределения на $\mathbb{R}^3$. Тогда μ-дивергенцией называется

$$D_\mu(p_x||p_y) = \int_{\mathbb{R}^3} \left(\frac{p_x(\vec{r})}{p_y(\vec{r})} \right) p_x(\vec{r}) d\vec{r}.$$

D_μ обладает следующими свойствами [5]:

1. $D_\mu(p_x||p_y) \geq 0$.
2. $D_\mu(p_x||p_y) = 0 \Leftrightarrow p_x = p_y$, если μ строго выпукла в 1 ($\alpha\mu(a) + (1 - \alpha)\mu(b) > \mu(1)$ для любых $a, b \in (0, \infty), \alpha \in (0, 1) : \alpha a + (1 - \alpha)b = 1$).
3. Отображения $p_x \to D_\mu(p_x, p_y)$ и $p_y \to D_\mu(p_x, p_y)$ являются выпуклыми в силу выпуклости $(p_x, p_y) \to D_\mu(p_x, p_y)$.
4. $D_\mu(p_x||p_y)$ не зависит от выбора меры в исходном пространстве $\mathbb{R}^3$ (в отличие от, например, квадрата L_2-расстояния между расределениями $\int |p_x(\vec{r}) - p_y(\vec{r})|^2 d\vec{r}$).

Свойства 1 и 2 позволяют называть D_μ предметрикой на пространстве вероятностных распределений. Например, KL-дивергенция получается выбором $\mu(t) = t \log t$. Чтобы дивергенция D_μ являлась метрикой, необходимо выполнение неравенства треугольника: $D_\mu(p_x, p_y) \leq D_\mu(p_x, p_z) + D_\mu(p_z, p_y)$ для любого p_z.

Метрика получается при выборе $\mu(t)$ одним из следующих способов:

- **Total variation** расстояние, $\mu(t) = \frac{1}{2}|t-1|$, при этом D_μ является метрикой [6].
- **Расстояния Хеллингера**, $\mu(t) = (1-\sqrt{t})^2$, при этом $\widetilde{D_\mu} = \sqrt{D_\mu}$ является метрикой [6].
- **Дивергенция Йенсена-Шеннона**, $\mu(t) = t\log\frac{2t}{t+1} + \log\frac{2}{t+1}$, при этом $\widetilde{D_\mu} = \sqrt{D_\mu}$ является метрикой [6].
- **Расстояние Ле-Кама** [7], $\mu(t) = \frac{1-t}{2t+2}$, при этом $\widetilde{D_\mu} = \sqrt{D_\mu}$ является метрикой [6].

Введём метрическое пространство M_μ, элементами которого являются всевозможные вероятностные распределения $p_x(\vec{r})$ на $\mathbb{R}^3$, а в качестве метрики выступает μ-дивергенция D_μ, либо корень из неё $\widetilde{D_\mu}$ — в случаях описанных выше. Нетрудно видеть, что аксиомы 1 и 2 не теряют справедливости для $\widetilde{D_\mu}$. В дальнейшем метрику в пространстве M_μ, для которой выполнены все аксиомы, включая неравенство треугольника, обозначать будем просто D_μ. Таким образом, мы указали четыре способа выбора μ, которые позволяют получить метрическое пространство над вероятностными распределениями.

4 Применение для решения задачи CASF

Восстанавливаем для пар (a,b) взаимодействующих элементов конформации z_i наблюдаемые распределения $\widetilde{p}_{a,b}(z_i) \in M_\mu$. Пусть $p_{a,b} \in M_\mu$ - априорные распределения взаимных расположений по данным обучающей выборки PDB [8]. Тогда в качестве оценки устойчивости используется

$$\widetilde{E}_i = \sum_{a\in\mathcal{A},b\in\mathcal{B}} D_\mu(\widetilde{p}_{a,b}(z_i), p_{a,b}).$$

5 Вычислительный эксперимент

На данных конформациях из PDB [8], представленных в виде $5\cdot 10^7$ записей (a,b,r,φ,θ), проводится серия экспериментов.

1. Для каждой из 800 пар (a,b) плотность вероятности $p(\vec{r})$ восстанавливается моделями из семейства $\mathfrak{F}$: гистограммой, окном Парзена, смесью гауссиан и

нейросетью. Проверяетс согласованность полученных аппроксимаций плотности $f_i(\vec{r})$, где i — номер модели в семействе $\mathfrak{F}$. В пространстве M_μ таким образом отмечаются 800 элементов $p_{a,b} \in M_\mu$, соответствующих восстановленным по PDB плотностям.

2. Находятся согласованные экстремумы (максимумы) плотности, то есть точки r, являющиеся локальными максимумами $f_i(\vec{r})$ для всех i.

3. Положения $\vec{r}$ экстремумов агрегируются в каталог, записи в котором имеют формат

$$(a, b, (r_{\min}, r_{\max}], \varphi, \theta).$$

Экспериментально подтверждается, что такая агрегация по расстоянию имеет смысл, поскольку угловые положения каждого отдельного экстремума незначительно зависит от расстояния. Зависимость от расстояния выражена только в абсолютном значении экстремума плотности вероятности.

Критерием точности аппроксимации в пункте 1 является попарное расхождение плотностей, предсказываемых различными моделями:

$$L_1 = \sum_{1 \leqslant i < j \leqslant 4} \int_{\mathbb{R}^3} (f_i(\vec{r}, \hat{w}_i) - f_j(\vec{r}, \hat{w}_j))^2 d\vec{r},$$

$$L_2 = \sum_{1 \leqslant i < j \leqslant 4} \int\limits_{|f_i(\vec{r}, \hat{w}_i) - f_j(\vec{r}, \hat{w}_j))| > \varepsilon} d\vec{r}.$$

Структурные параметры всех моделей оптимизируются в совокупности, перебором по сетке, для достижения минимума $\frac{1}{2}(L_1 + L_2)$, который соответствует максимуму согласованности. Параметры моделей, требующие обучения (положения и ковариационные матрицы гауссиан в смеси; веса нейросети), оптимизируются при фиксированных структурных параметрах.

5.1 Семейство моделей для восстановления плотности

5.1.1 Гистограмма

Алгоритм построения гистограммы приведён в [9].

$$f(\vec{r}) = \frac{1}{\ell} \sum_{t=1}^{\ell} [\vec{r_t} \in I_{ijk}],\ i_{\vec{r}} = \left\lfloor \frac{r}{dr} \right\rfloor,\ j_{\vec{r}} = \left\lfloor \frac{\varphi}{d\varphi} \right\rfloor,\ k_{\vec{r}} = \left\lfloor \frac{\theta}{d\theta} \right\rfloor,$$

где $I_{ijk} \subset \mathbb{R}^3$ - интервал гистограммы, $\sqcup I_{ijk} = \mathbb{R}^3$. Структурными параметрами являются размеры интервалов по радиусу dr и по углу $d\varphi, d\theta$.

5.1.2 Окно Парзена

$$f(\vec{r}) = \frac{1}{\ell} \sum_{i=1}^{\ell} K\left(\frac{|\vec{r} - \vec{r}_i|}{h}\right).$$

Структурными параметрами являются ширина окна h и тип ядра $K(t)$. Перебираются гауссовское ядро $K(t) \propto \exp(-\frac{t^2}{2})$, треугольное ядро $K(t) \propto max(1 - |t|, 0)$ и ядро Епанечникова $K(t) \propto max(1 - t^2, 0)$.

5.1.3 Смесь гауссиан

$$f(\vec{r}) = \sum_{i=1}^{N} w_i \frac{1}{\sqrt{(2\pi)^N |\Sigma_i|}} \exp\left(-\frac{1}{2}(\vec{r} - \vec{\mu_i})^T \Sigma_i^{-1} (\vec{r} - \vec{\mu}_i)\right), \qquad \sum_{i=1}^{N} w_i = 1.$$

Структурным параметром является число компонентов смеси N.

5.1.4 Нейросеть

Используется простая полносвязная сеть с двумя скрытыми слоями:

$$f(\vec{r}) = W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 \cdot \vec{r})), \quad \sigma(\vec{x})_i = \tanh(x_i),$$

$$W_1 \in \mathbf{M}_{50x3}(\mathbb{R}), W_2 \in \mathbf{M}_{50x50}(\mathbb{R}), W_3 \in \mathbf{M}_{1x50}(\mathbb{R}).$$

На вход подаются координаты в декартовой системе координат, поскольку такая параметризация не содержит особых точек. При обучении используется выход окна Парзена в качестве целевой функции; восстановление плотности без привлечения вспомогательной модели [10, 11] было реализовано, но требовало значительно большего объёма вычислений для сходимости.

5.2 Результат восстановления плотности

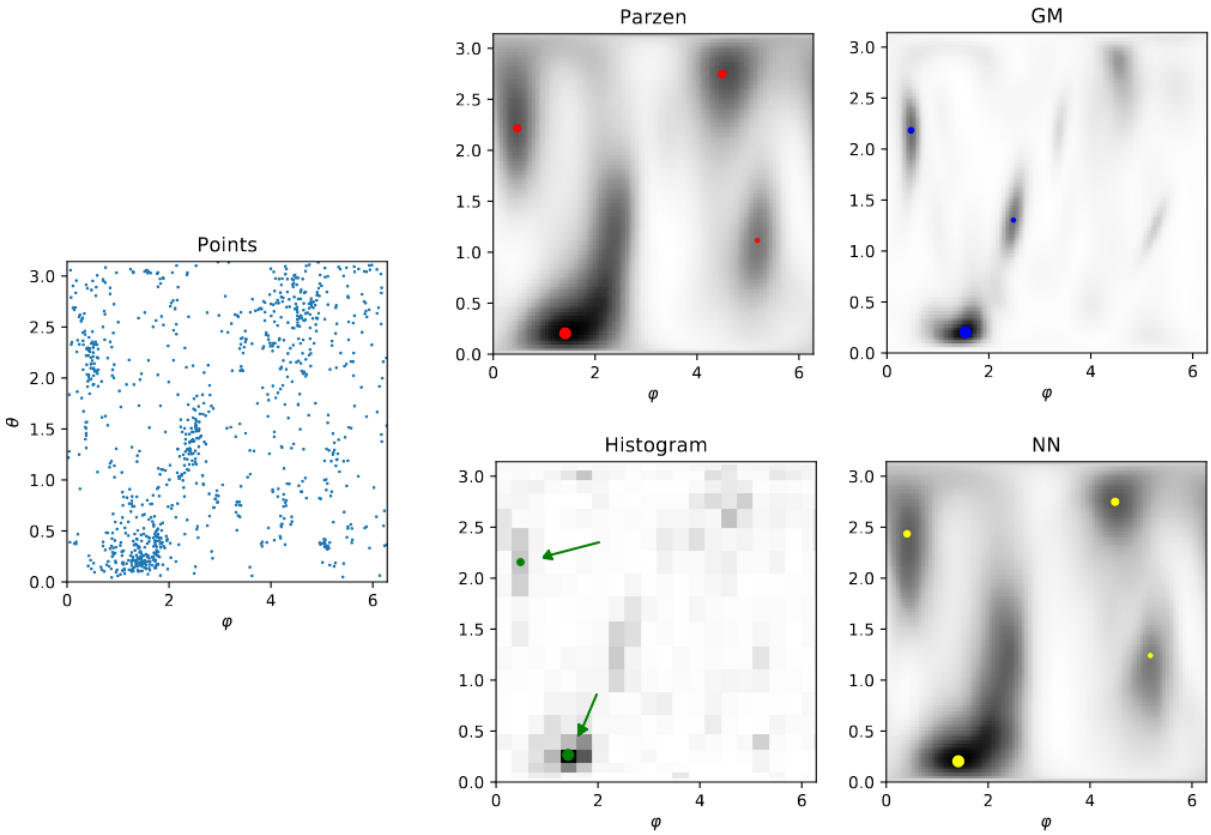


Рис. 2: Плотность, восстановленная различными моделями для $x = (4, 0)$, $r = 5$Å

На рис. 2 приведён пример исходных данных для $x = (4, 0)$ и восстановленой различными моделями плотности. Стрелками отмечены согласованные экстремумы, попадающие в каталог.

5.3 Экстремумы восстановленной плотности вероятности

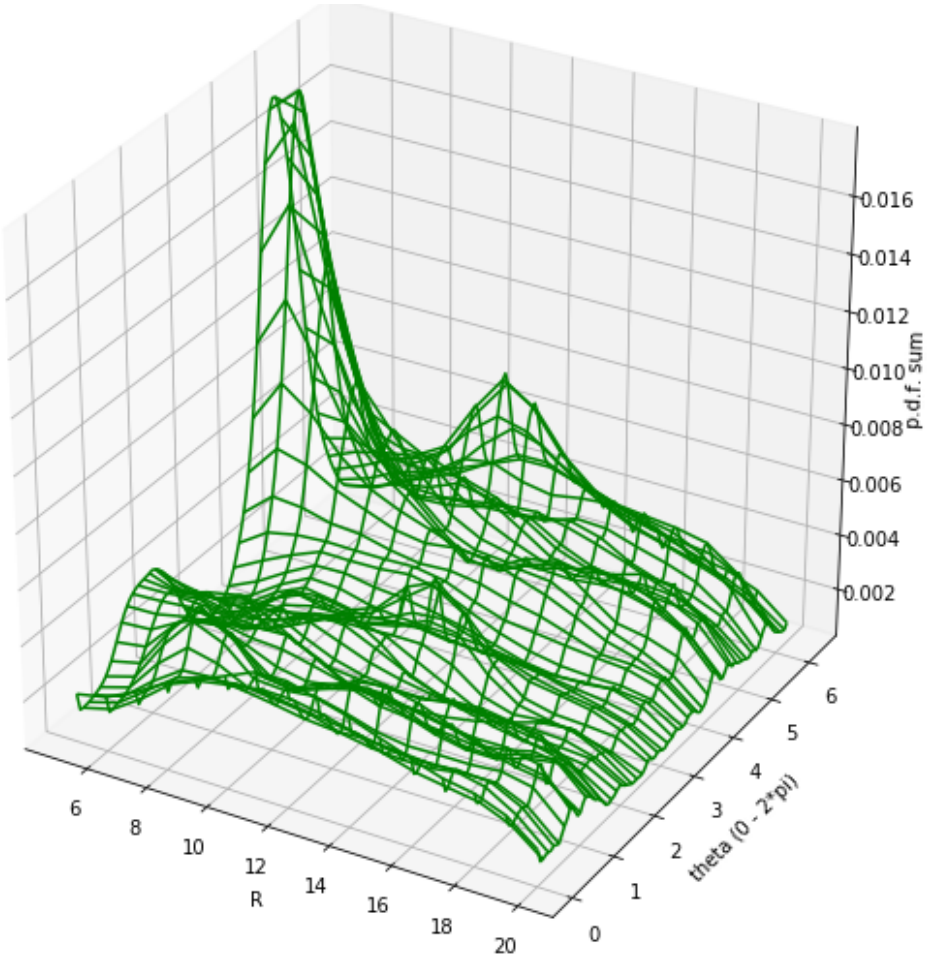


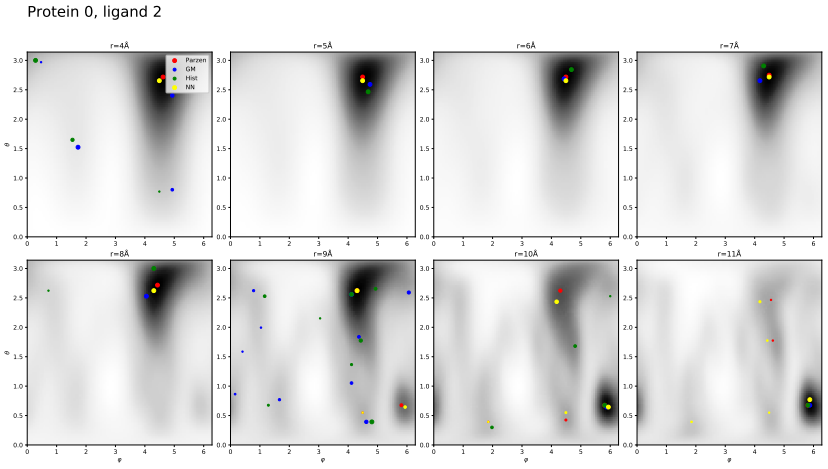
Рис. 3: Восстановленная плотность $f(r, \theta)$ (проинтегрированная по углу φ)

В начале исследования выдвигалась гипотеза о равномерности распределения при $r \to \infty$, то есть практически — на расстояниях порядка $r \sim 20$Å. Эта гипотеза была отвергнута простым статистическим тестом χ^2, что послужило началом эксперимента по нахождению экстремумов плотности в распределениях.

На рис. 3 представлен характер зависимости восстановленной плотности от расстояния и одного из углов. Видно, что положение экстремумов на сфере меняется в зависимости от расстояния незначительно, что обуславливает наличие диапазонов расстояния $[r_{\min}, r_{\max}]$ в структуре каталога.

5.4 Каталог экстремумов

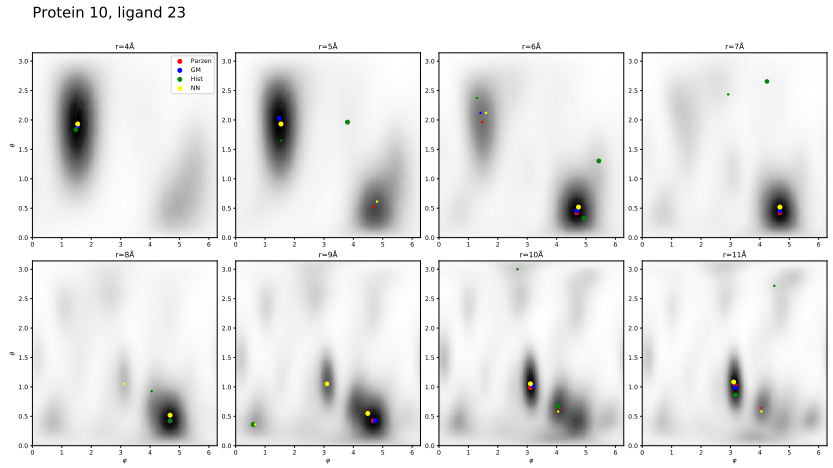
На рис. 4 и рис. 5 приведены примеры страниц каталога. Трёхмерная плотность вероятности визуализирована в сечениях по расстоянию в диапазоне от 4Åдо 11Å. Точками различных цветов отмечены экстремумы восстановленной разными моделями плотности. Согласованным экстремумам отвечают скопления точек, содержащие все цвета.



	Parzen	GM	Hist	NN
Parzen				
GM	28.98			
Hist	34.78	23.33		
NN	33.72	54.28	58.24	

MAPE, %

Рис. 4: Запись каталога экстремумов для $x = (0, 2)$



	Parzen	GM	Hist	NN
Parzen				
GM	30.89			
Hist	60.94	48.88		
NN	39.61	59.71	86.75	

MAPE, %

Рис. 5: Запись каталога экстремумов для $x = (10, 23)$

6 Заключение

Исследована проблема молекулярного докинга белка и лиганда. Решена задача восстановления плотности вероятности взаимодействия белок-лиганд различными моделями: окном Парзена, гистограммой, смесью гауссиан и нейросетью. Построен каталог экстремумов плотности и введён критерий согласованности экстремумов. Отвергнута гипотеза о равномерности плотности на больших расстояниях. На основе полученных распределений построено метрическое пространство для использования в задачах анализа молекулярных конфигураций. Предложен способ решения задачи ранжирования конформаций CASF с использованием построенного пространства. Исходный код эксперимента размещён в [12].

Список литературы

[1] **López-Blanco, J. R., & Chacón, P. 2019.** KORP: knowledge-based 6D potential for fast protein and loop modeling. Bioinformatics.

[2] **Kadukova, M., & Grudinin, S. 2017.** Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization. Journal of computer-aided molecular design.

[3] **Kadukova, M., & Grudinin, S. 2018.** Docking of small molecules to farnesoid X receptors using AutoDock Vina with the Convex-PL potential: lessons learned from D3R Grand Challenge 2. Journal of computer-aided molecular design.

[4] **Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., & Wang, R. 2018.** Comparative assessment of scoring functions: The CASF-2016 update. Journal of chemical information and modeling.

[5] **Csiszár, I. 1967.** Information-type measures of difference of probability distributions and indirect observation. Studia scientiarum Mathematicarum Hungarica.

[6] **Endres, D. M., & Schindelin, J. E. 2003.** A new metric for probability distributions. IEEE Transactions on Information theory.

[7] **Le Cam, L. 2012.** Asymptotic methods in statistical decision theory. Springer Science & Business Media.

[8] **Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., & Bourne, P. E. 2000.** The protein data bank. Nucleic acids research.

[9] **Мотренко, А.П., Рудаков, К.В., & Стрижов, В.В. 2016.** Учет влияния экзогенных факторов при непараметрическом прогнозировании временных рядов. Вестник Московского университета. Серия 15. Вычислительная математика и кибернетика.

[10] **Zhang, S. 2018.** From CDF to PDF — A Density Estimation Method for High Dimensional Data. arXiv preprint arXiv:1804.05316.

[11] **Magdon-Ismail, M., & Atiya, A. F. 1999.** Neural networks for density estimation. In Advances in Neural Information Processing Systems.

[12] Исходный код эксперимента к работе “Построение вероятностных метрических пространств в задачах анализа молекулярных конфигураций”. URL: `https://github.com/Intelligent-Systems-Phystech/ProbabilisticMetricSpaces` (дата обращения: 14.06.2020).