

Порождающие модели белковых структур

Роман Сергеевич Клыпа

Научный руководитель: к.ф.-м.н. С. В. Грудинин

Кафедра интеллектуальных систем ФПМИ МФТИ

Специализация: Интеллектуальный анализ данных

Направление: 03.04.01 Прикладные математика и физика

2024

Постановка задачи

Задача генерации трехмерных объектов

Необходимо построить модель G , порождающую объекты согласно $p(\mathcal{M})$, где $\mathcal{M} - \mathbb{R}^{3M \times N}$.

Уменьшение размерности

Наличие твердых связей в $\mathbb{R}^{3M \times N}$ позволяет уменьшить пространство до $SE(3)^N$, что можно отождествить с $SO(3)^N \times \mathbb{R}^{3N}$.

Score Matching

Вместо $p(\mathbf{x})$ моделируется $s_\theta(\mathbf{x})$:

$$\mathcal{L}(\theta) = \mathbb{E}_{p(\mathbf{x})} \left[\|s_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x})\|^2 \right]. \quad (1)$$

Генерация возможна с помощью алгоритма Ланжевена:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon} \mathbf{z}_i, \quad i = 0, 1, \dots, K, \quad \mathbf{z}_i \sim \mathcal{N}(0, I). \quad (2)$$

Цель исследования

Поставленные цели

- ▶ Предложить прямой и обратный диффузионные процессы для $SO(3)$.
- ▶ Предложить способ обеспечить эквивариантность процессов ($g_1 f(g_2) = f(g_1 g_2)$).
- ▶ Показать преимущества предложенных методов на конкретной задаче.

Существующие подходы и их недостатки

- ▶ Предложенные ранее методы не гарантируют сходимость прямого процесса к шуму (Yim et al. 2023).
- ▶ Предложенные ранее методы являются эмпирическими.

Предлагаемый подход

Предложение 1 (Клыпа, 2024)

$$\mathbf{R} \overset{\text{exp}}{\underset{\text{log}}{\rightleftharpoons}} \mathbf{r}$$

Прямой процесс на $SO(3)$:

$$\mathbf{R}_t = \exp[d(t)\mathbf{r}_0\tilde{\mathbf{r}}(t)], \quad (3)$$

где $d(t)$ - 'коэффициент дрейфа', $d(0) = 1$, $d(1) = 0$.

Обратный процесс:

$$\mathbf{R}_{t-dt} = \exp[\mathbf{r}_t * g^2(t)s(\mathbf{r}_t)dt * g(t)\sqrt{dt}\tilde{\mathbf{r}}(t)], \quad (4)$$

где $s(\mathbf{r}_t) = \nabla_{\mathbf{r}} \log p_t(\mathbf{r}_t|\mathbf{r}_0)$.

Здесь $\mathbf{R} \in SO(3)$, $\mathbf{r} \in \mathfrak{so}(3)$, $\tilde{\mathbf{r}}(t) \sim \text{IGSO}_3(\text{Id}, \sigma^2(t))$.

Функция $g(t) = \sqrt{\frac{d}{dt}\sigma^2(t)}$ отвечает за масштабирование времени.

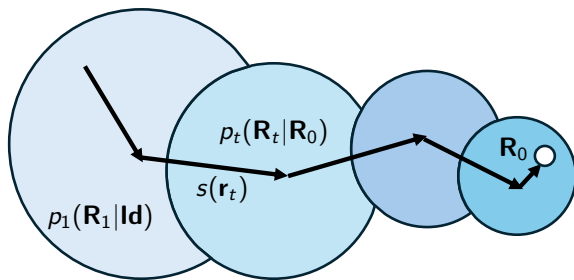
Предлагаемый подход

Теорема 1 (Клыпа, 2024)

Для прямого процесса (3):

$$p_t(\mathbf{R}_t|\mathbf{R}_0) = \text{IGSO}_3(\mathbf{R}_t|\exp[d(t)\mathbf{r}_0], \sigma^2(t)), \quad (5)$$

в частности $p_1(\mathbf{R}_1|\mathbf{R}_0) = \text{IGSO}_3(\mathbf{R}_1|\text{Id}, \sigma^2(1))$.



Обратный процесс.

Эквивариантность и сходимость процессов

Теорема 2 (Клыпа, 2024)

Прямой процесс (3) эквивариантен относительно $SO(3)$:

$$\mathbf{R}\mathbf{R}_t(\mathbf{R}_0) = \mathbf{R}_t(\mathbf{R}\mathbf{R}_0) \quad \forall \mathbf{R} \in SO(3). \quad (6)$$

Если $s_\theta(\mathbf{r}_t)$ эквивариантен относительно $SO(3)$, то обратный процесс (4) также эквивариантен:

$$\mathbf{R}\mathbf{R}_t(\mathbf{R}_1) = \mathbf{R}_t(\mathbf{R}\mathbf{R}_1) \quad \forall \mathbf{R} \in SO(3). \quad (7)$$

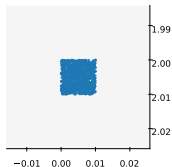
Лемма 1 (Клыпа, 2024)

При $p(\mathbf{r}|\mathbf{r}_0) = \text{IGSO}_3(\mathbf{r}|\mathbf{r}_0, \sigma^2)$, процесс

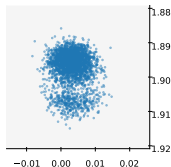
$$\mathbf{r}_{i+1} = \mathbf{r}_i * \epsilon \nabla_{\mathbf{r}} \log p(\mathbf{r}|\mathbf{r}_0) * \sqrt{2\epsilon} \mathbf{z}_i, \quad \mathbf{z}_i \sim \text{IGSO}_3(\mathbf{Id}, \sigma^2) \quad (8)$$

сходится к $p(\mathbf{r}|\mathbf{r}_0)$ при $i \rightarrow \infty$.

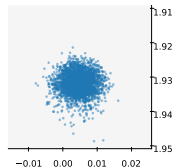
Вычислительный эксперимент



Целевое
распределение.



Генерация метода
Yim et al. 2023.



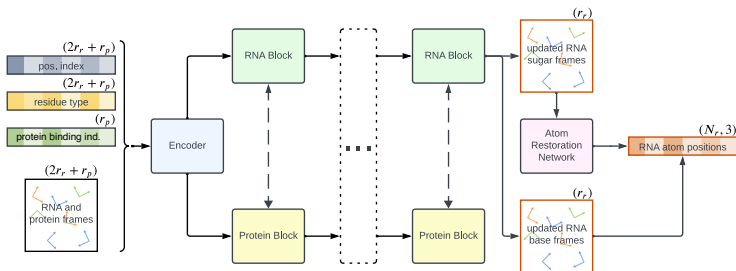
Генерация метода
Klyukha, 2024.

Элементы группы $SO(3)$ изображены в трехмерном пространстве при использовании представления axis-angle. Для наглядности полученные распределения были спроецированы на плоскость.

Применение к реальной задаче

Постановка задачи

Генерация 3D структур молекул РНК, при взаимодействии с протеином: $P(\text{geom}_{RNA} | \text{geom}_{pr}, \text{seq}_{pr}, \text{seq}_{RNA})$. Предполагается, что $\text{geom}_{RNA}, \text{geom}_{pr} \in \text{SE}(3)^N$.



Архитектура модели MolBinDif.

Применение к реальной задаче

Теорема 3 (Клыпа, 2024)

Архитектура модели MolBinDif эквивариантна относительно $SO(3)$.

Результаты эксперимента

Процесс	$\downarrow \text{rRMSD}_{rr}, \text{\AA}$	$\downarrow \text{rRMSD}_{rp}, \text{\AA}$	$\uparrow \text{IDDT}_{rr}$	$\uparrow \text{IDDT}_{rp}$
Yim et al. 2023	10.7 ± 5.5	14.1 ± 7.8	0.17 ± 0.06	0.10 ± 0.05
Клыпа, 2024	11.1 ± 5.1	13.4 ± 8.2	0.15 ± 0.06	0.11 ± 0.05

Таблица: Результаты эксперимента.

Выносятся на защиту

1. Предложен новый порождающий процесс на $SO(3)$.
2. Доказана эквивариантность процесса.
3. Продемонстрированы его преимущества относительно прошлых работ на демонстрационном примере.
4. Продемонстрирована его сходимость на реальных данных.

Публикации

1. Roman Klypa, Kliment Olechnovič, Ben Shor, Dina Schneidman-Duhovny, Sergei Grudinin. *MolBindDif: Protein-conditioned RNA structure diffusion*
OpenReview preprint, ICML 2024