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 $\mathcal{M}$ - $\mathbb{R}^{3M}$.

Уменьшение размерности

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

Score Matching

Вместо $p(\mathbf{x})$ моделируется $\nabla_{\mathbf{x}} \log p(\mathbf{x})$:

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

Генерация возможна с помощью алгоритма Ланжевена:

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

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 реальной 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: прямой процесс

Здесь и далее $\mathbf{R} \in \mathrm{SO}(3)$, $\log\mathbf{R} = \mathbf{r} \in \mathfrak{so}(3)$, $*$ - композиция элементов $\mathfrak{so}(3)$.

Предложение 1 (Клыпа, 2024)

Прямой процесс на $\mathrm{SO}(3)$:

$$\mathbf{R}_t = \exp[d(t)\mathbf{r}_0 * \tilde{\mathbf{r}}(t)],\ \tilde{\mathbf{r}}(t) \sim \mathrm{IGSO}_3\left(\mathbf{Id}, \sigma^2(t)\right) \tag{3}$$

где $d(t)$ – коэффициент дрифта, $d(0) = 1$, $d(1) = 0$.

Теорема 1 (Клыпа, 2024)

Для прямого процесса 3:

$$p_t(\mathbf{R}_t|\mathbf{R}_0) = \mathrm{IGSO}_3\left(\mathbf{R}_t|\exp[d(t)\mathbf{r}_0], \sigma^2(t)\right), \tag{4}$$

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

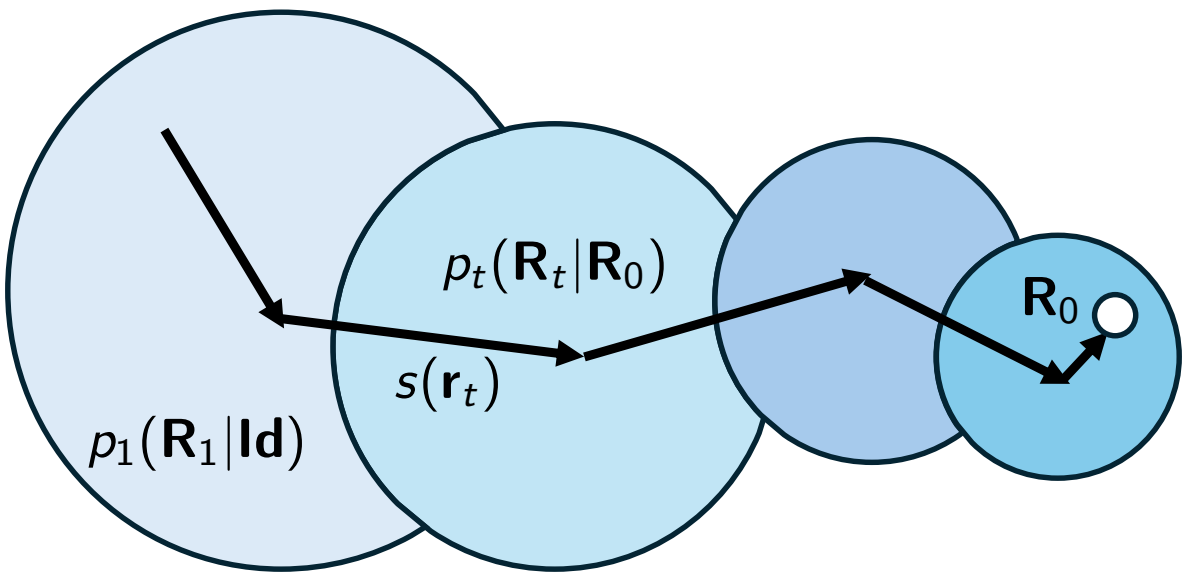
Предлагаемый подход: обратный процесс

Предложение 2 (Клыпа, 2024)

Обратный процесс на SO(3), соответствующий 3:

$$\mathbf{R}_{t-dt} = \exp[\mathbf{r}_t * g^2(t) s(\mathbf{r}_t) dt * g(t)\sqrt{dt}\tilde{\mathbf{r}}(t)], \tag{5}$$

где $s(\mathbf{r}_t) = \nabla_{\mathbf{r}} \log p_t(\mathbf{r}_t|\mathbf{r}_0)$, $g(t) = \sqrt{\frac{d}{dt}\sigma^2(t)}$.



Динамика Ланжевена с отжигом.

Эквивариантность и сходимость процессов

Лемма 1 (Клыпа, 2024)

При $p(\mathbf{r}|\mathbf{r_0}) = \mathrm{IGSO}_3\left(\mathbf{r}|\mathbf{r}_0, \sigma^2\right)$, процесс

$$\mathbf{r}_{i+1} = \mathbf{r}_i * \epsilon \nabla_{\mathbf{r}} \log p(\mathbf{r}_i|\mathbf{r_0}) * \sqrt{2\epsilon}\mathbf{z}_i, \;\; \mathbf{z}_i \sim \mathrm{IGSO}_3\left(\mathbf{Id}, \sigma^2\right) \tag{6}$$

сходится к $p(\mathbf{r}|\mathbf{r_0})$ при $i \to \infty$.

Теорема 2 (Клыпа, 2024)

Прямой процесс 3 эквивариантен относительно SO(3):

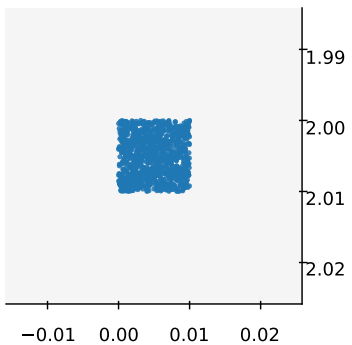
$$\mathbf{R}\mathbf{R}_t(\mathbf{R}_0) = \mathbf{R}_t(\mathbf{R}\mathbf{R}_0) \; \forall \mathbf{R} \in \mathrm{SO}(3). \tag{7}$$

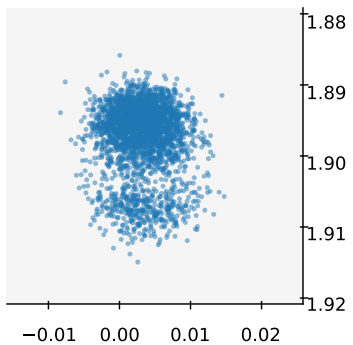
Если $s_\theta(\mathbf{r}_t)$ инвариантен относительно SO(3), то обратный процесс 5 также эквивариантен:

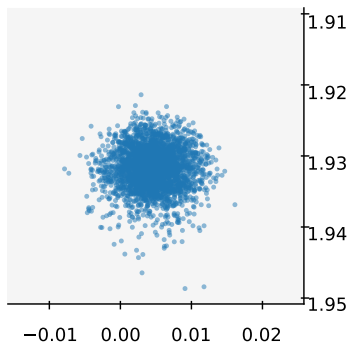
$$\mathbf{R}\mathbf{R}_t(\mathbf{R}_1) = \mathbf{R}_t(\mathbf{R}\mathbf{R}_1) \; \forall \mathbf{R} \in \mathrm{SO}(3). \tag{8}$$

Вычислительный эксперимент

Элементы группы SO(3) изображены в трехмерном пространстве при использовании представления векторов Эйлера. Для наглядности полученные распределения были спроецированы на плоскость.





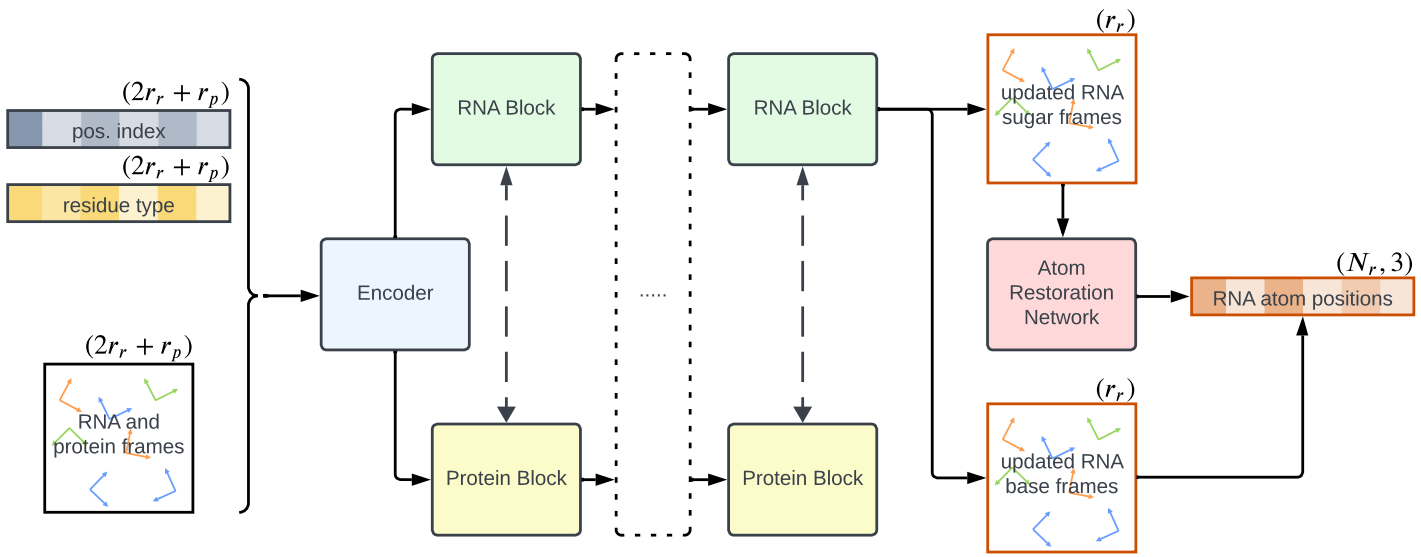


Слева изображено целевое распределение, по центру - результат генерации метода Yim et al. 2023., справа - результат предлагаемого нами метода.

Генерация трехмерных структур РНК

Постановка задачи

Генерация трехмерных структур молекул РНК, при взаимодействии с протеином: $P(geom_{RNA}|geom_{pr}, seq_{pr}, seq_{RNA})$. Предполагается, что $geom_{RNA}, geom_{pr} \in \mathrm{SE}(3)^N$.



Архитектура модели MolBindDif. Здесь r_r - количество остатков РНК, r_p - количество остатков белка, а N_r - общее количество атомов в РНК.

Результаты MolBindDif

Теорема 3 (Клыпа, 2024)

Архитектура модели MolBindDif эквивариантна относительно SO(3).

Результаты эксперимента

Процесс	$\downarrow$ rRMSD$_{rr}$, Å	$\downarrow$ rRMSD$_{rp}$, Å	$\uparrow$ lDDT$_{rr}$	$\uparrow$ lDDT$_{rp}$
Yim et al. 2023	**10.7** $\pm$ **5.5**	14.1 $\pm$ 7.8	0.17 $\pm$ 0.06	0.10 $\pm$ 0.05
Klypa, 2024	11.1 $\pm$ 5.1	**13.4** $\pm$ **8.2**	0.15 $\pm$ 0.06	**0.11** $\pm$ **0.05**
RoseTTaFoldNA	13.3 $\pm$ 6.3	17.6 $\pm$ 7.8	**0.19** $\pm$ **0.10**	0.08 $\pm$ 0.04

Результаты генерации трехмерных структур PHK. RoseTTAFoldNA не является генеративным процессом.

Выносится на защиту

1. Предложен новый порождающий процесс на SO(3).
2. Доказана эквивариантность процесса.
3. Продемонстрированы его преимущества относительно прошлых работ на синтетических данных.
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 submission