Применение диффузионных моделей для генерации последовательностей аминокислот

Мещанинов Вячеслав Павлович

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия
E-mail: meshchaninov.viacheslav@gmail.com, vetrovd@yandex.ru

Научный руководитель *— Ветров Дмитрий Петрович*

Задача генерации белков становится ключевой областью академических исследований, потенциально влияющей на биоиоинформатику, синтетическую биологию и терапию на основе белков. Даже с учетом растущей популярности задачи условной генерации [1], генерация последовательностей аминокислот – это основополагающий и жизненно важный шаг. Причина проста: глубокое понимание и способность генерировать в безусловном режиме закладывают прочную основу для более специализированной и детализированной условной генерации и последующего дообучения.

В данной работе мы предлагаем DiMA – диффузионную модель генерации последовательностей аминокислот с использованием языковой модели белков. Как показано на Рис. 1 мы применяем ESM-2 [2] для получения непрерывного представления последовательности аминокислот, на котором мы обучаем диффузионную модель. Во время генерации модель принимает на вход чистый гауссовский шум и итеративно расшумляет его. Мы тщательно оцениваем качество и разнообразие генерируемых белков, а также способность модели выучивать распределение обущающих данных. Мы используем большое количество метрик для оценки производительности модели, используя разные представления белков: последовательности аминокислот, 3-D структуры, эмбеддинги языковых моделей белков.

Обучение модели Предлагаемый метод состоит из трех частей. Первая часть представляет собой предварительно обученный энкодер белков ($\mathcal{E}$), который выучивает осмысленное пространство векторов, соответствующее исходному белковому пространству. Вторая часть представляет собой диффузионную модель ($\mathcal{F}$), которая генерирует векторы скрытого пространства энкодера белков из гауссовского шума. Третья часть представляет собой декодер ($\mathcal{D}$), который отображает сгенерированные вектора в последовательности аминокислот.

Энкодер отображает последовательность аминокислот $y = [y_1, ..., y_s]$ длины s в вектор $x = [x_1, ..., x_s] \in R^{s \times d}$, $x = \mathcal{E}(y)$, $d = 320$. Затем мы нормализуем вектор так, чтобы каждая компонента отдельного вектора в последовательности x имела нулевое среднее значение и единичную дисперсию $z_0 = Normalize(x)$. Это преобразование позволяет адаптировать дискретное представление белка для стандартной гауссовской диффузии.

Затем мы обучаем диффузионную модель, $\hat{z}_\theta()$, восстанавливать z_0 из $z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t}\varepsilon$ используя следующую функцию потерь:

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{\varepsilon \sim \mathcal{N}(0, \mathbf{I}), t \sim U[0;1]} ||z_0 - \hat{z}_\theta(z_t, t)||^2 \tag{1}$$

где α_t — функция, задаваемая расписанием диффузионной модели.

Генерация белка Важнейшим аспектом фазы генерации является определение длины сгенерированной последовательности. В то время как в процессе обучения длина последовательности напрямую определяется векторным представлением белка, во время генерации мы выбираем сэмплируем длину из эмпирического распределения, наблюдаемого в обучающем наборе данных. Сэмплирование длины является важным аспектом нашей модели, поскольку ее отсутствие приводит к неадекватному распределению длины в сгенерированной выборке. Мы используем маску внимания для ввода информации о длине последовательности в сеть. Процесс генерации начинается с сэмплирования чистого гауссовского шума и длины. Используя фиксированное количество шагов T, мы итеративно генерируем векторное представление белка $\hat{z}_0$. Затем мы денормализуем каждый вектор и используем декодер для отображения последовательности векторов в последовательность аминокислот.

Результаты Для оценивания качества модели мы проводим ее сравнение с аналогами на датасете SwissProt. Мы сравниваем предлагаемый метод DiMA с авторегрессионными подходами (NanoGPT), сверточными сетями (SeqDesign), генеративно состязательными сетями (ProteinGAN) и диффузионным подходом (EvoDiff). Мы обучаем данные методы на том же наборе данных до сходимости, чтобы гарантировать честное сравнение. Для оценки качества метода мы используем большое количество метрик. Для оценки правдоподобия сгенерированных белков мы используем pLDDT, ESM-2 pppl, scPerplexity, TM-score, BLAST. Для оценки схожести

распредления реальных и сгенерированных данных используются FPD, MMD, OT.

Оценивание качества сгенерированных последовательностей на наборе данных SwissProt показывает, что предлагаемый метод DiMA превосходит все аналоги и выдает значения метрик, точно согласованные с набором обучающих данных. Результаты представлены в Таблице [1].

Иллюстрации

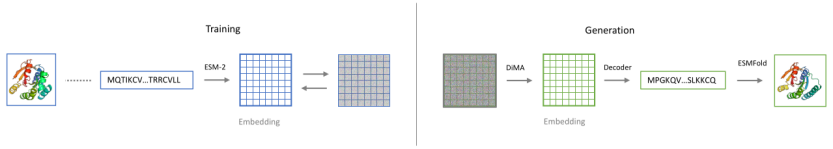

Рис. 1. Схема предлагаемой диффузионной модели для генерации белков.

Model	pLDDT (↑)	ESM-2 pppl (↓)	scPerplexity (↓)	TM-score (↑)	BLAST (↑)	FPD (↓)	MMD (↓)	OT (↓)
Dataset	80.7	5.35	1.88	0.80	100	0.13	0.00	1.08
Random sequences	25.0	21.54	2.77	0.33	0	3.97	0.20	3.88
nanoGPT	61.0	8.18	2.04	0.63	43	1.24	0.03	2.53
EvoDiff-OADM	37.1	15.77	2.44	0.42	12	1.49	0.11	2.63
SeqDesign	43.1	11.89	2.35	0.41	17	3.53	0.19	5.12
proteinGAN	30.4	16.48	2.57	0.33	0	2.94	0.17	3.98

Таблица. 1. Сравнение качества генерируемых белков предлагаемой модели DiMA и аналогов на наборе данных SwissProt.

В данной работе мы предлагаем диффузионный метод DiMA для генерации последовательностей аминокислот, работающий поверх языковой модели белков. Используя большое колчиество метрик мы оцениваем качество, разнообразие, сходство распределения и биологическую значимость сгенерированных последовательностей. Результаты демонстрируют, что DiMA превосходит другие методы генерации белков.

Литература

1. Madani A. Large language models generate functional protein sequences across diverse families // Nature Biotechnology, 2023, P. 1099–1106.
2. Zeming L. Evolutionary-scale prediction of atomic-level protein structure with a language model // Science, 2023, P. 1123–1130.