Автоматическое выделение доменных границ в белках по пространственной структуре

Терентьев Александр

2025

Введение

- Белковые домены — структурные и функциональные единицы белка
- Автоматическое определение границ доменов важно для аннотации и анализа белков
- Цель: разработать и сравнить методы выделения доменных границ по 3D-структуре
- **Предположение:** последовательность аминокислотных остатков и их пространственное расположение содержат достаточно информации для автоматического выделения границ доменов

Формальная постановка задачи

- Дано: 3D-структура белка (PDB), последовательность остатков $R = (r_1, \ldots, r_n)$
- Требуется: построить бинарную маску $y \in \{0, 1\}^n$, где $y_i = 1$ — граница домена
- Модель: $f(\mathbf{X}, G) \to \hat{y}$, где $\mathbf{X}$ — признаки остатков, G — граф контактов

Задача: $\arg\max_f \ \mathbb{E}_{(\mathbf{X},G,y)} \ Q(f(\mathbf{X}, G), y)$

Метрики качества

- IoU (границы): $\frac{|\hat{y} \cap y|}{|\hat{y} \cup y|}$
- IoU (домены): среднее IoU по сегментам
- Boundary F1-score: F1 по найденным границам с допуском
- Mean Boundary Deviation (MBD): среднее отклонение границ

Датасет и подготовка данных

- SCOP: аннотированные домены, загрузка PDB-структур
- Формирование ground truth: маска границ по SCOP
- Train/test split, кросс-валидация

Классические методы

DOMAK:

- Границы определяются по разрывам: $d_{i,i+1} = \|\mathbf{x}_{i+1} - \mathbf{x}_i\|$
- Если $d_{i,i+1} > t$, то i — граница домена
- t — эмпирический порог (обычно 8Å)
- См. *Siddiqui, A. S., Barton, G. J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Science, 4(5), 872-884.*

SPLIT:

- Пусть k — число доменов (из разметки)
- Границы: $b_j = \left\lfloor \frac{j \cdot n}{k} \right\rfloor$, $j = 1, \ldots, k-1$
- Каждый сегмент $[b_{j-1}, b_j)$ — домен
- Простой baseline, не учитывает структуру

Ссылки:

- Siddiqui, A. S., Barton, G. J. (1995). Protein Science, 4(5), 872-884.
- Holland, T. A., Veretnik, S., Shindyalov, I. N., Bourne, P. E. (2006). Protein domain identification: a structural biology perspective. Structure, 14(7), 997-1006.

GCN-модель (DomainGCN)

- Вход: $\mathbf{X} \in \mathbb{R}^{n \times 3}$ — координаты CA, $G = (V, E)$ — граф контактов
- Граф: $A_{ij} = 1$ если $\|\mathbf{x}_i - \mathbf{x}_j\| < 8$Å
- Модель: $\mathbf{H}^{(l+1)} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}\mathbf{H}^{(l)}W^{(l)})$
- 4 слоя GCNConv, выход — вероятности классов для каждого остатка
- Функция потерь: $\mathcal{L} = -\sum_i w_{y_i} \log p_{i,y_i}$, $w_1 \gg w_0$

Ссылки:

- Kipf, T. N., Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. ICLR.
- Gainza, P. et al. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods, 17(2), 184-192.

Постановка эксперимента

- Использованы аннотированные белки из базы SCOP (Structural Classification of Proteins)
- 3D-структуры белков загружались из PDB (Protein Data Bank)
- Формирование ground truth: маска границ по SCOP-аннотациям
- Обучение на train, оценка на test (разделение 80/20)
- Сравнение с DOMAK и SPLIT по всем метрикам
- Кросс-валидация для оценки дисперсии

Результаты сравнения

Метрика	Model (mean±std)	DOMAK (mean±std)	SPLIT (mean±std)
IoU (границы)	0.1082 ± 0.1935	0.5162 ± 0.3786	0.5310 ± 0.4045
IoU (домены)	0.6122 ± 0.2596	0.8118 ± 0.2147	0.8674 ± 0.1896
Boundary F1-score	0.2544 ± 0.2732	0.6937 ± 0.3326	0.7206 ± 0.3436
Mean Boundary Deviation	6.46 ± 1.74	19.35 ± 36.36	8.28 ± 14.68

Таблица: Сравнение качества методов выделения доменных границ

- GCN превосходит классические методы по близости полученных границ, но он выдает больше фантомных границ

Примеры предсказаний

- Визуализация: истинные и предсказанные границы на белке
- Ошибки: ложные/пропущенные границы

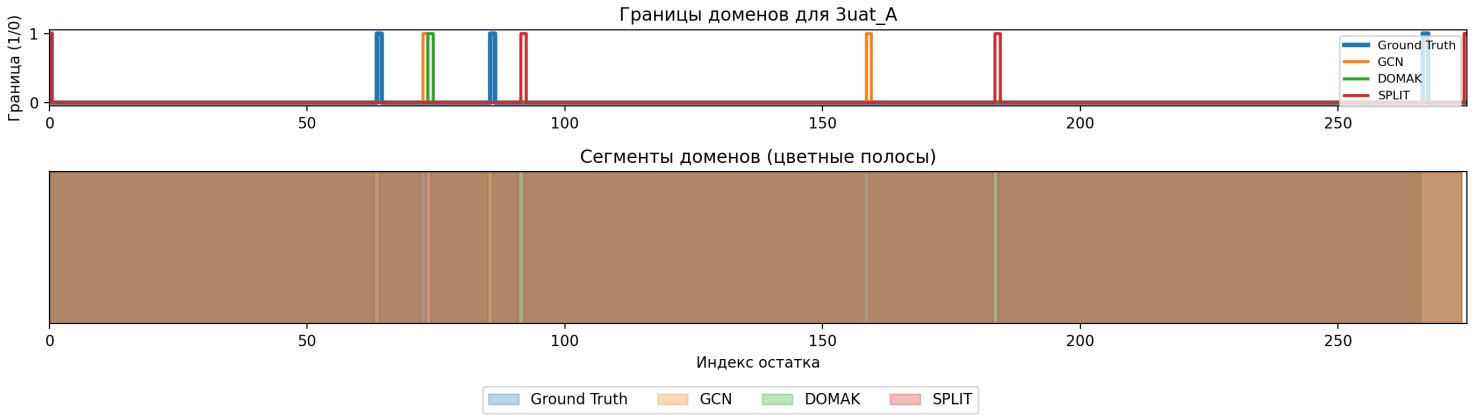

Рис.: Спрогнозированные границы доменов

Выводы

- GCN-модель успешно выделяет границы доменов по структуре
- Классические методы уступают по точности
- Возможности для улучшения: дополнительные признаки, архитектуры

Спасибо за внимание!

Вопросы?