Факультет
компьютерных наук

Институт
искусственного интеллекта
и цифровых наук

Создание новых методов для функциональной аннотации генов

Зайцев Сергей

Введение

Актуальность:

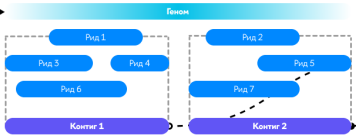
- Большое количество генов с неизвестными функциями в прокариотических геномах
- Отсутствие эффективных методов для аннотации генов de novo
- Традиционные методы не находят функции для быстро эволюционирующих белков

Цель исследования: Разработка новых ML методов для идентификации генных кластеров на основе анализа геномного контекста.
Ключевая гипотеза: Если гены часто встречаются рядом в эволюционно далеких организмах, они функционально связаны и образуют генные кластеры.

Структура геномных данных

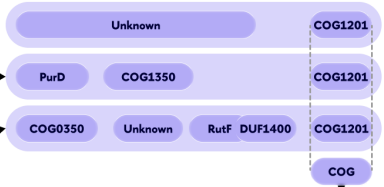
Геном

Контиги



Контиг – непрерывный участок ДНК.

Кластеры генов



Белки

Обратите внимание на разницу между кластером генов и COG!

Гены, участвующие в одних биологических процессах, имеют свойство располагаться рядом. Если мы видим одну и ту же группу генов на многих геномах, будем называть её кластером генов.

COG – по сути метка для гена, благодаря которой мы можем сказать, что один белок на первом контиге и другой на втором контиге – один и тот же ген.

Объединяем данные в датасет для исследования

Данные взяты из базы NCBI COG (обновление 2024 года).

Структура данных

- `cog-24.cog.csv`: словарь «Белок → COG». Позволяет перевести сырые ID белков в функциональные кластеры.
- `COGorg24.gene.tab.gz`: Огромный файл координат (30 млн записей). Колонки: `Gene ID`, `Coordinates` (100..500), `Strand` (+/-), `Assembly ID` (Геном), `Partition ID` (Хромосома), `Protein ID`.
- `cog-24.pathways.tab`: «Ground Truth». Список известных функциональных систем.

Объединяем данные в датасет для исследования

Итоговая структура таблицы

assembly_id	partition_id	cog_id	start	end
GCA_000008085.1	AE017199.1	COG1201	883	2691
GCA_000008085.1	AE017199.1	COG1047	2668	3189
GCA_000008085.1	AE017199.1	COG1371	3290	3748
GCA_000008085.1	AE017199.1	COG2419	3745	4665
GCA_000008085.1	AE017199.1	COG0350	4680	5090

Алгоритм

Два гена считаются соседями, если расстояние между концом одного и началом другого ≤ 1000 нуклеотидов.

$$\text{Freq(COG_A, COG_B)} = \frac{N_{shared}}{N_{total}}$$

N_{shared}: Число геномов, где COG_A и COG_B находятся рядом (≤ 1000 нт).
N_{total}: Число геномов, где COG_A (или COG_B) присутствует вообще.

Результат

COG_A	COG_B	co_occurrence	total_A	total_B	frequency	common_system
COG3649	COG5938	139	702	139	1.000000	CRISPR-Cas system
COG3407	COG3890	44	1445	44	1.000000	Isoprenoid biosynthesis
COG1681	COG6090	27	147	27	1.000000	Archaella
COG4789	COG6009	60	651	60	1.000000	Type III secretion system
COG3649	COG5941	451	702	456	0.989035	CRISPR-Cas system
COG1336	COG3337	135	262	137	0.985401	CRISPR-Cas system
COG4669	COG6028	93	552	95	0.978947	Type III secretion system
COG5720	COG5721	139	142	197	0.978873	Photosystem II
COG3736	COG6033	44	913	45	0.977778	Type IV secretion system
COG1857	COG5939	111	275	114	0.973684	CRISPR-Cas system

Метрики качества модели

Матрица ошибок:

		Ground Truth	
		+	–
Model	+	TP	FP
	–	FN	TN

Метрики:

- Precision $= \frac{TP}{TP+FP}$ — точность модели
- Recall $= \frac{TP}{TP+FN}$ — полнота модели
- F_β-score $= \frac{(1+\beta^2)\cdot\text{Precision}\cdot\text{Recall}}{\beta^2\cdot\text{Precision}+\text{Recall}}$ (β показывает, насколько Recall важнее Precision)

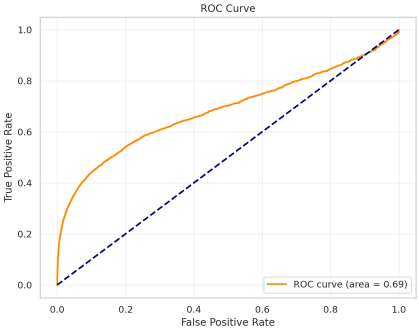
ROC AUC (Receiver Operating Characteristic)

Что это?

- ROC-кривая показывает зависимость между True Positive Rate и False Positive Rate при изменении порога классификации от 0 до 1
- ROC AUC — площадь под этой кривой
- ROC AUC показывает, насколько два класса хорошо раличимы моделью

Как считается?

- $TPR = \frac{TP}{TP+FN}$ — доля правильно найденных положительных примеров
- $FPR = \frac{FP}{FP+TN}$ — доля ошибочно классифицированных отрицательных примеров
- Значение 0.5 = рандом, 1.0 = идеальная модель



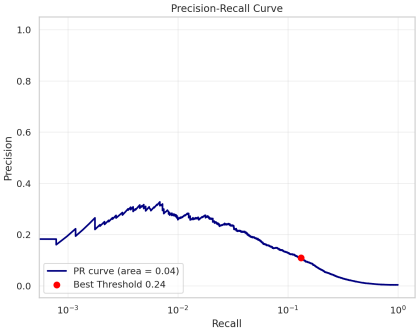
PR AUC (Precision-Recall)

Что это?

- PR-кривая показывает зависимость между Precision (точность) и Recall (полнота) при изменении порога классификации от 0 до 1
- PR AUC — площадь под этой кривой

Как считается?

- Precision $= \frac{TP}{TP+FP}$ — доля корректных предсказаний среди всех положительных предсказаний модели
- Recall $= \frac{TP}{TP+FN}$ — доля найденных положительных примеров
- PR AUC показывает, насколько хорошо модель находит положительные примеры



Если данные несбалансированы

Проблема дисбаланса классов:

- Позитивные пары (гены с общей функцией): 4%
- Негативные пары (независимые гены): 96%

Почему ROC AUC обманчив?

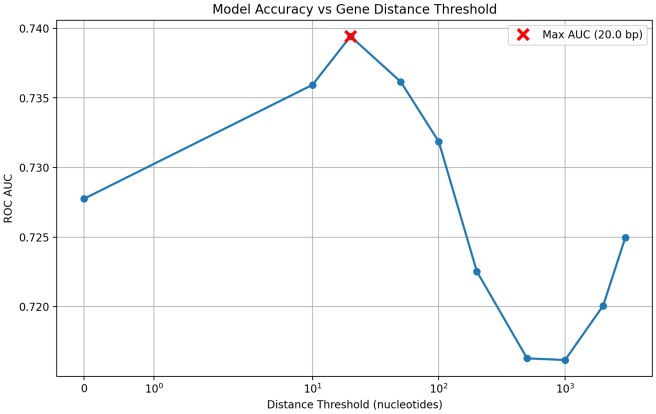
- ROC AUC использует $FPR = \frac{FP}{FP+TN}$
- Когда негативных примеров очень много (TN огромное), знаменатель остаётся огромным $\Rightarrow$ FPR остаётся низким
- Результат: ROC AUC выглядит хорошо, даже если модель плохо отличает позитивные пары

Почему PR AUC честнее?

- PR AUC использует Precision $= \frac{TP}{TP+FP}$ — нет TN в знаменателе
- Метрика смотрит только на предсказанные положительные примеры

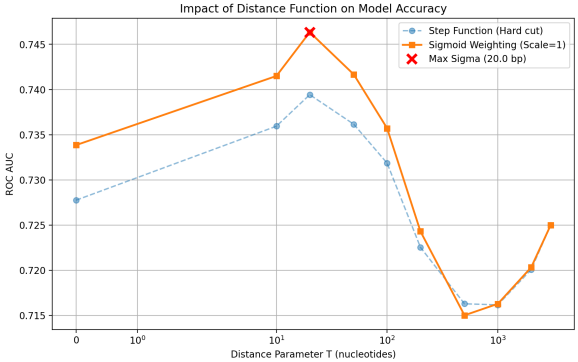
Линейные пороги

Смотрим на ROC AUC для разных порогов: 0, 10, 20, 50, 100, 200, 500, 1000, 2000, 3000.



Сигмоида

Смотрим на ROC AUC для разных порогов, но затем применяем к frequency сигмоиду: $\sigma(x) = \frac{1}{1+e^x}$.



Перспективы

Текущее ограничение: мы смотрим только на расстояние и выбор числа необоснован.
Куда двигаться:

1. Icity (Shmakov et al. 2018): Взвешивание геномов. Если у нас 100 штаммов E. coli, их надо считать как 1 событие, а не 100.
2. Spacedust (Zhang et al. 2025): Можно добавить ordering p-value: учитывать порядок генов ($A \to B$ vs $B \to A$) и направление цепи ($+/-$).
3. Разбить на таксономические группы и считать отдельно для каждой.
4. Использовать PR AUC как метрику качества модели.

Спасибо за внимание!