УДК 579.25

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ СИГМА-ФАКТОРОВ ИЗ СБОРОК БАКТЕРИАЛЬНЫХ ГЕНОМОВ БАЗЫ ДАННЫХ NCBI

Кушнарева А.Д. (**Университет ИТМО**)

Научный руководитель – в.н.с, к.б.н А.С. Комиссаров (Университет ИТМО)

Аннотация. Сигма-факторы представляют собой бактериальные факторы транскрипции, обеспечивающие связывание РНК-полимеразы с промоторами, избирательно увеличивая транскрипцию конкретных генов в зависимости от условий окружающей среды. В ходе работы был проведен анализ всех существующих белков, аннотированных как сигма-факторы, из референсных сборок прокариотических геномов базы данных NCBI. В результате получены данные о разнообразии и распространенности различных сигма-факторов у бактерий.

Введение. Регуляция экспрессии генов позволяет клетке контролировать синтез белков, необходимых для ее жизненного цикла или для адаптации к изменениям окружающей среды. Самым важным шагом в регуляции генов является инициация транскрипции, для которой ДНК-зависимая РНК-полимераза является ключевым ферментом. Для инициации транскрипции требуется дополнительный полипептид, известный как сигма-фактор (σ). Этот белок связывается с РНК-полимеразой и отвечает за инициацию транскрипции на конкретном промоторе. Сигма-факторы принято разделять на два семейства - σ70 и σ54, в свою очередь внутри σ70 выделяют 4 группы. На сегодняшний день знания о структуре, функциональной роли и регуляции сигма-факторов ограничены, поскольку большинство исследований затрагивало первичные сигма-субъединицы и проводилось на модельных организмах, таких как *E. coli* и *B. subtilus*. Текущее количество геномных и протеомных данных позволяет изучить разнообразие и распределение σ-факторов в других микроорганизмах.

Основная часть.

Данные. В качестве объекта исследования были выбраны сигма-факторы из 200 000 сборок бактериальных геномов.

Методы: Бактериальные сборки были проаннотированы при помощи утилиты Prokka v1.14.6 и PGAP, затем из полученных данных отбирались белки, аннотированные как сигма-факторы. Белки были скластеризованы на основании сходства последовательностей с помощью программы MMseqs2, после чего полученные кластеры обрабатывались при помощи собственных скриптов на языке Python, с помощью которых была получена информация о распределении сигма-факторов по таксонам, о том, какими генами они кодируются, а также информация о длине и частоте встречаемости в бактериальных геномах. Для визуального представления взаимосвязей между анализируемыми сигма-факторами при помощи инструмента PhyML v3.0 было построено филогенетическое дерево.

Результаты. В результате анализа 3 271 193 белков были скластеризованы в 443 098 кластеров на основе сходства аминокислотных последовательностей. В качестве первичного анализа были отобраны первые 300 кластеров с наибольшим количеством входящих в них последовательностей сигма-факторов.

В результате были выявлены представители RpoH, RpoS, RpoE, RpoD, RpoF сигма-факторов и сигма-факторы семейства σ54.
Наиболее крупный кластер содержит 46 660 белков, аннотированных как RpoE. Данный тип сигма-субъединиц относится к 4-й группе семейства сигма70, также известной как группа сигма-факторов экстрацитоплазматической функции (ECF).
Второй по размеру кластер включает в себя RpoS сигма-факторы, представителей второй группы семейства сигма-70, играющих роль в стационарной фазе роста и при ответе на стресс. К третьему кластеру относились RpoH сигма-факторы теплового шока, которые активируются при воздействии высоких температур на бактерию, позволяя ей выживать в экстремальных условиях.
Анализ таксономического распределения белков выявил следующий процент содержания сигма-факторов среди бактерий рода *Escherichia (40.60%), Salmonella (24.04%), Klebsiella (21.04%) Shigella (4.69%), Enterobacter (3.78%), Yersinia (2.21%),Serratia (1.47%) , Cronobacter (1.05%)* для первого кластера; *Escherichia (58.91%), Klebsiella (32.74%), Shigella (6.03%), Pantoea (0.71%), Pectobacterium (0.48%), Raoultella (0.37%), Dickeya (0.26%), Erwinia 0.25%* для второго кластера; *Salmonella (52.05%), Klebsiella (45.43%), Citrobacter (1.31%), Raoultella (0.55%), Escherichia (0.05%), Yokenella (0.04%), Enterobacter (0.02%)* для третьего кластера.
Были также выявлены таксон-специфичные сигма-факторы (то есть наблюдавшиеся только у одного рода бактерий) для таких родов как *Salmonella, Staphylococcus, Pseudomonas, Mycobacterium, Streptococcus, Listeria* и др. Так, например, самым крупным таксон-специфичным кластером оказался кластер, содержащий RpoS сигмы, принадлежащие к роду *Salmonella*. RpoS играет важную роль для патогенности этих бактерий, поскольку гены, регулируемые этим сигма-фактором, защищают бактерии от различных стрессовых состояний внутри хозяина, включая осмотический и окислительный стресс.
Был также осуществлен филогенетический анализ путем построения дерева на основании множественного выравнивания репрезентативных последовательностей из кластеров. В результате были получены клады для всех вышеупомянутых сигма-факторов. Построенное дерево позволило идентифицировать те сигма-факторы, которые не имели аннотаций или были проаннотированы неверно, а также проследить эволюционные связи между группами.

Выводы. Полученные нами результаты позволяют расширить область знаний о существующих сигма-факторах и их распространении среди бактерий. Поскольку на сегодняшний день предсказание сигма-факторов является нетривиальной задачей за счет недостатка информации о белковых доменах, эти данные также позволят решить эту задачу и проводить функциональную аннотацию существующих сигма-факторов путем построения более точных по сравнению с существующими HMM профилей для идентификации сигма-субъединиц у немодельных бактериальных видов и, как следствие, определения их роли в патогенности и адаптации микроорганизмов.