UniProt Proteomes

Иван Русинов

Что такое протеом в UniProt?

В теории: совокупность белков, экспрессирующихся в одном организме.

На практике: совокупность трансляций открытых рамок считывания из полного генома.

Технически: запись в базе данных Proteomes.

- ссылки на записи UniProtKB и/или UniParc
- метаданные (статус протеома, организм, ссылка на сборку генома и т.д.)

Этапы добавления нового протеома

- Добавление новой полногеномной сборки, содержащей информацию о открытых рамках считывания, в нуклеотидный архив.
- Проверка на избыточность.
- Создание записей, оценка качества и полноты.

А дальше может происходить:

- Добавление/удаление белков.
- Перевод протеома в разряд референсных.
- Удаление протеома.

Типы протеомов в UniProt

Некоторые протеомы в UniProt имеют один из типов, перечисленных ниже. Основная часть протеомов не относится ни к одному из этих типов.

Референсные (reference) – вручную или автоматически отобранные в качестве лучшего среди доступных протеомов таксономической группы (обычно вида).

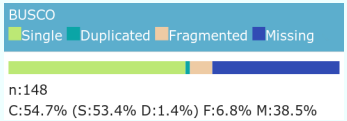
Избыточные (redundant) – слишком сильно похожие на другой протеом; для белков из таких протеомов не создаются записи в UniProtKB, только в UniParc.

Удаленные (excluded) – протеомы, удаленные вслед за геномной сборкой из RefSeq; белки из таких протеомов удаляются из TrEMBL.

Меры качества и полноты

CPD (Complete Proteome Detector) – сравнение с протеомами близких организмов на предмет отличия в размерах. По результатам присваивают один из трех статусов: Standard, Close to Standard, Outlier.

BUSCO (Benchmarking Universal Single-Copy Ortholog) – внешний алгоритм оценки качества по наличию представителей референсных ортологичных групп генов. Каждой группе ортологов присваивается один из 4 статусов: Single, Duplicated, Fragmented, Missing. Результат – процент групп из каждой категории в графическом или числовом представлении.



Пан-протеомы в UniProt

Пан-протеом (Pan proteome) – совокупность разных белков из группы близкородственных организмов.

Включает в себя:

- все белки из референсного протеома;
- по одному представителю из всех кластеров UniRef50 нереференсных протеомов, которые не содержат белков из референсного.

Пан-протеомы не выделены в отдельную базу, но и не имеют своих записей в базе Proteomes. Идентификатор пан-протеома совпадает с ID референсного протеома, входящего в его состав.