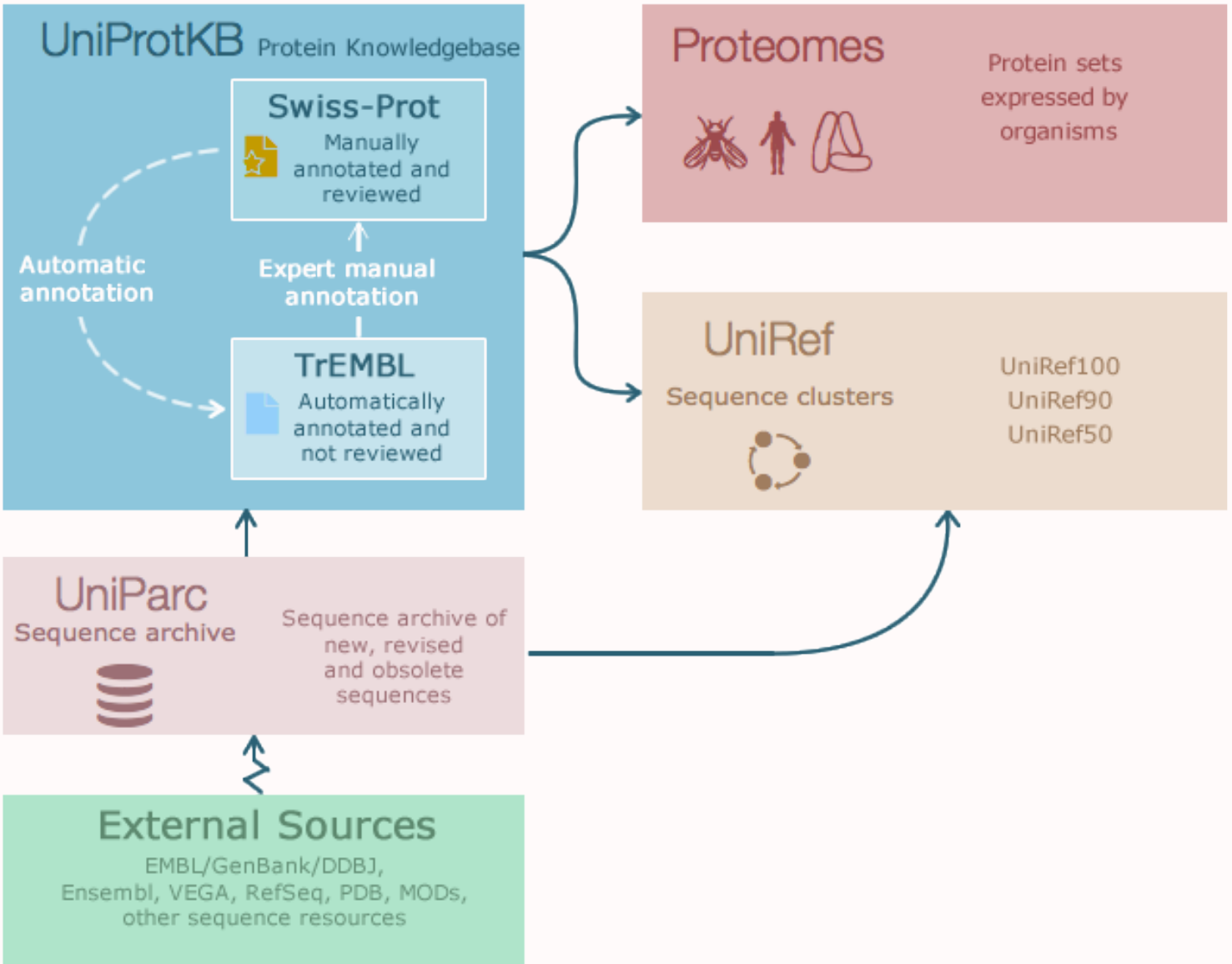
Протеомы и кластеры UniProt

UniParc, UniRef, Proteomes

Иван Русинов

Структура UniProt



UniParc

UniProt Archive

Архив уникальных последовательностей белков.

- Содержит все последовательности белков, которые когда-либо были в UniProtKB, и даже те, которые не были включены в UniProtKB по каким-либо причинам.
- Каждой уникальной последовательности присвоен идентификатор UPI, который никогда не изменяется и не удаляется.
- Запись UniParc содержит только последовательность, её хеш-сумму для проверки, ссылки на базы, в которых в какой-то момент времени хранилась такая же белковая последовательность и чуть-чуть вспомогательной информации.
- Последовательности (почти) неаннотированные.

Например, UPI0000000004

Database	Identifier	Version	Organism	First seen	Last seen	Active
UniProtKB/Swiss-Prot	P04195	1	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	1988-11-01	2021-04-07	Yes
UniProtKB/TrEMBL	A0A2I2MC48	1	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	2018-02-28	2021-04-07	Yes
UniProtKB/TrEMBL	Q6LDV9	1	Vaccinia virus	2006-04-18	2021-04-07	Yes
UniProtKB/TrEMBL	V5R1H0	1	Vaccinia virus WAU86/88-1	2015-07-22	2021-04-07	Yes
UniProtKB/TrEMBL	Q76ZR9	1		2004-07-05	2011-10-19	No
RefSeq	YP_232995	1	Vaccinia virus	2005-10-06	2021-01-04	Yes
EMBL CDS	AAA48264	1	Vaccinia virus	2003-03-12	2021-01-25	Yes
EMBL CDS	AAO89392	1	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	2003-06-16	2021-01-25	Yes
EMBL CDS	ABD52586	1	Vaccinia virus	2007-03-31	2021-01-25	Yes
EMBL CDS	AHB23552	1	Vaccinia virus WAU86/88-1	2014-01-06	2021-01-25	Yes
EMBL CDS	AQY54886	1	Vaccinia virus	2017-04-08	2021-01-25	Yes
EMBL CDS	SOU90125	1	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR))	2018-01-26	2021-01-25	Yes
USPTO	ADS58156	1		2011-03-07	2020-11-27	Yes
PRF	3315290DX		Vaccinia virus	2007-12-07	2009-09-01	Yes
TREMBLNEW	AAA48264			2003-03-29	2004-06-11	No
TREMBLNEW	AAO89392			2003-04-18	2004-06-11	No
PIRARC	A01146			2003-03-31	2003-04-04	No
PIRARC	A35014			2003-03-31	2003-04-04	No
PIR	CRVZW			2003-04-11	2005-01-04	No

UniRef

UniProt Reference Clusters

Кластеры записей по сходству последовательностей.
UniProtKB + UniParc без ссылок на UniProtKB

UniRef100	идентичные на 100% последовательности и их фрагменты.
UniRef90	кластеры самых длинных представителей из кластеров UniRef100, идентичных на 90% и похожих по длине (не короче 80% самой длинной последовательности в кластере). Принадлежность кластеру UniRef90 распространяется на все остальные записи из кластера UniRef100 без проверок.
UniRef50	аналогично UniRef90.

Последовательности длины 10 и более короткие включены только в UniRef100, и кластеризуются только при совпадении длины.

Сид и репрезентативная последовательность

Seed – самая длинная последовательность в кластере, с которой сравниваются остальные последовательности для проверки принадлежности кластеру.

Representative – наиболее хорошо аннотированная последовательность, используется для аннотации кластера (название и длина последовательности).

Случаются приколы 😕

Например, кластер UniRef90_P81108

Cluster members	Entry name		Protein names	Organisms	Organism IDs	Related clusters	Length	Role
P81108	**THIO_CLOSG**		**Thioredoxin (Fragment)**	**Clostridium sporogenes**	1509	UniRef100_P81108	**40**	**Representative**
A0A1V9IK41	A0A1V9IK41_CLOSG		Thioredoxin	Clostridium sporogenes	1509	UniRef100_P81108	106	
A0A0B4W3E0	A0A0B4W3E0_CLOBO		Thioredoxin	Clostridium botulinum Prevot_594	1408284	UniRef100_P81108	106	
A0A1J1CWE3	A0A1J1CWE3_CLOBO		Thioredoxin	Clostridium botulinum	1491	UniRef100_A0A1J1CWE3	106	Seed
J7T6P1	J7T6P1_CLOS1		Thioredoxin	Clostridium sporogenes (strain ATCC 15579)	471871	UniRef100_A0A1J1CWE3	106	
A0A0D0ZXA6	A0A0D0ZXA6_CLOBO		Thioredoxin	Clostridium botulinum B2 450	1379739	UniRef100_A0A1J1CWE3	106	
A0A6M0YC23	A0A6M0YC23_CLOSG		Thioredoxin	Clostridium sporogenes	1509	UniRef100_A0A6M0YC23	106	
A0A1S9I145	A0A1S9I145_9CLOT		Thioredoxin	Clostridium tepidum	1962263	UniRef100_A0A1S9I145	106	
A0A6M0T4F3	A0A6M0T4F3_CLOBO		Thioredoxin	Clostridium botulinum	1491	UniRef100_A0A6M0T4F3	106	
A0A6M0XX80	A0A6M0XX80_CLOSG		Thioredoxin	Clostridium sporogenes	1509	UniRef100_A0A6M0T4F3	106	
A0A0M1IUU5	A0A0M1IUU5_9CLOT		Thioredoxin	Clostridium sp. L74	1560217	UniRef100_A0A0M1IUU5	106	
UPI000666DA61			thiol reductase thioredoxin	Clostridium sporogenes	1509	UniRef100_UPI000666DA61	106	
UPI000D0CC3E6			thiol reductase thioredoxin	Clostridium botulinum	1491	UniRef100_UPI000D0CC3E6	106	
UPI001748E097			thioredoxin	Clostridium botulinum	1491	UniRef100_UPI001748E097	106	
UPI0005F06029			thiol reductase thioredoxin	Clostridium sporogenes	1509	UniRef100_UPI0005F06029	106	

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).

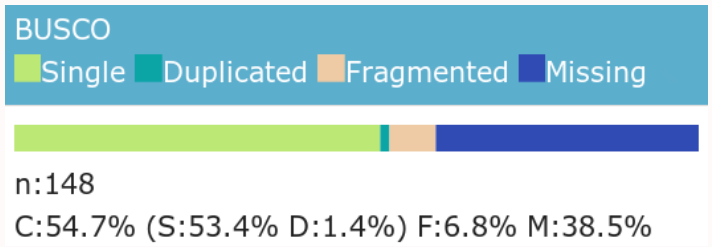
Избыточные (redundant) – слишком сильно похожие на другой протеом; для белков из таких протеомов не создаются записи в UniProtKB, только в UniParc.

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.