Практикум 6. Базы данных KEGG, GO и другие

ЧАСТЬ 1. Входные данные

У нас есть список из ID 14 генов:

CD320
AMN
CUBN
MTRR
ABCD4
MMAB
CBLIF
MTR
LMBRD1
TCN2
MMUT
MMAA
MMACHC
MMADHC

Ссылка на список:
https://kodomo.fbb.msu.ru/FBB/year_24/lectures/lists_go/list28.txt

Просто смотря на сам список, понять, что объединяет эти гены между собой, объединяет ли что-то, невозможно. Поэтому далее будем анализировать эти гены доступными нам средствами.

ЧАСТЬ 2. Групповой анализ (База данных PANTHER)

Для того, чтобы понять какие функции выполняют белки, где локализуются, в каких биологических процессах участвуют, обратимся к БД Panther.

Эта база данных является частью проекта GO.
На основе множественных выравниваний, кластеризации семейств, построенных филогенетических деревьев, различных алгоритмов белки и соответствующие им гены разделены на семейства, подсемейства. Помимо использования GO-терминов, PANTHER использует своими термины из PANTHER/X, которые тоже структурированы иерархически, но все же отличаются GO-терминов и не могут быть сопоставлены на 100%.

База данных Panther, как минимум, позволяет следующее:

1. Узнать о семействах белков, метаболических путях, клеточных процессах, молекулярных функциях.
2. Составить списки генов, связанных с определенным семейством/подсемейством белков, молекулярной функцией, биологическим процессом или метаболическим путем.

3. Всесторонне изучить список генов: визуализировать распределение биологических процессов, в которых участвуют данные белки, молекулярных функций, выполняемых ими, провести статистический анализ перепредставленности GO-категорий.

Воспользуемся функционалом Panther.
Сначала узнаем, в каких процессах в целом участвуют наши белки (рис. 1). Видно, что белки участвуют в метаболизме липидов, аминокислот, ароматических и серных соединений (возможно, те же аминокислоты). Это пока дает нам очень мало информации.

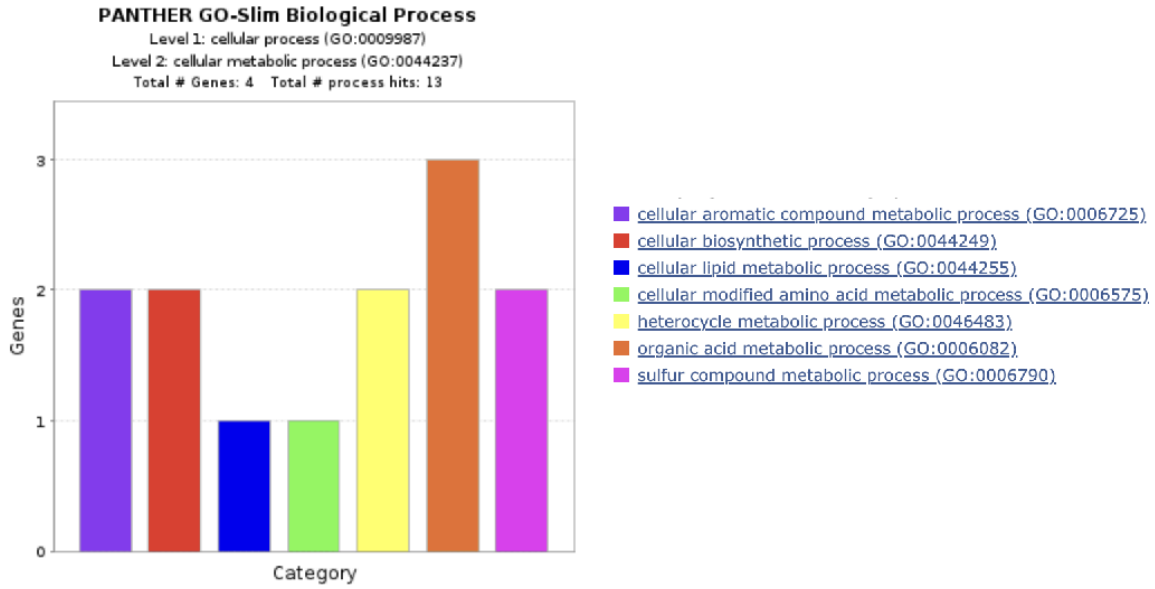

Рис. 1. Распределение клеточных метаболических процессов среди анализируемых белков.

Посмотрим где локализуются белки (рис. 2).
Множество из наших белков в цитоплазме, некоторые в мембране, но на самом деле удивляет наличие 30 хитов для обнаруженных 11 белков. Опять же, очень мало информации, нужно либо копать дальше, для каждой локализации, либо смотреть каждый белок по отдельности.

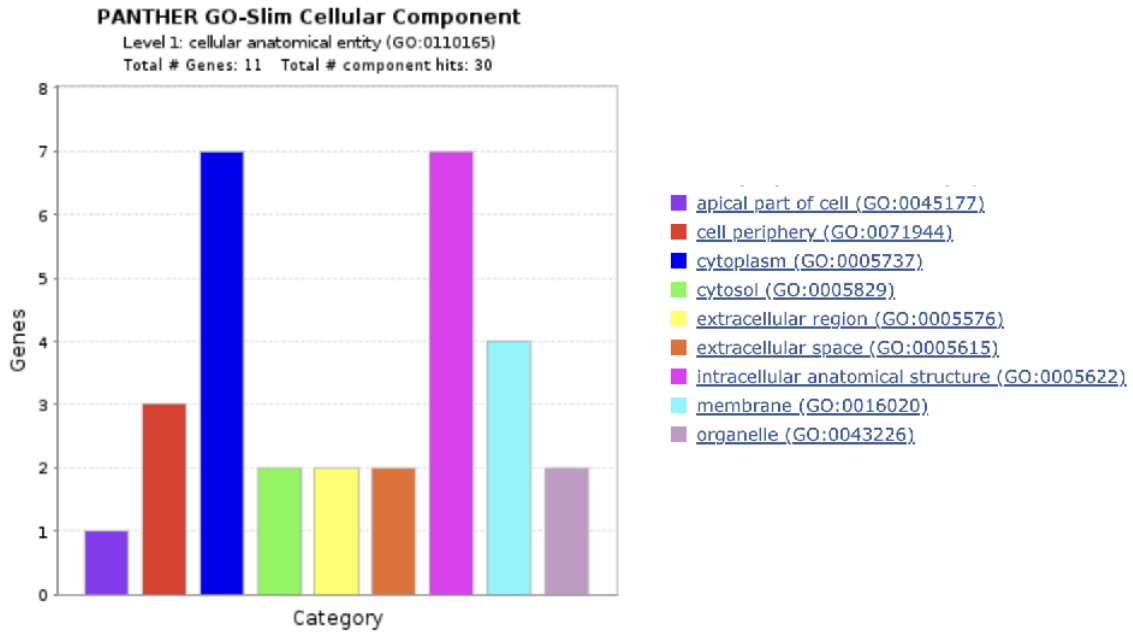

Рис. 2. Распределение локализации рассматриваемых белков.

Чтобы упростить себе жизнь, сразу (почти сразу) воспользуемся самым, на мой взгляд, полезным инструментом Panther - статистический анализ для списка ID. Возьмем для анализа аннотацию GO Biological process, в качестве референса будем использовать протеом человека.
Для самой статистики - тест Фишера, с поправкой Бонферрони на множественные сравнения (благо сам Panther предлагает это выбрать!).
И теперь обнаруживаем, что хотели, эти белки явно связаны с кобаламином (рис. 3). С огромной уверенностью (p-value $2.52 * 10^{-26}$) можно утверждать, что выборка белков участвует в метаболических процессы кобаламина, а также в его транспорте (отсюда становится понятным и такие GO-категории как метаболические процессы тетрапиррола, транспорт витаминов, метаболизм серосодержащих аминокислот (метионин необходим для реактивации кобаламина)).

	Homo sapiens (REF)	Client Text Box Input (Hierarchy) NEW!				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	▲ P value
cobalamin metabolic process	9	9	.01	> 100	+	2.52E-26
tetrapyrrole metabolic process	61	9	.04	> 100	+	4.31E-16
cobalamin transport	9	6	.01	> 100	+	3.65E-14
vitamin transport	57	6	.04	> 100	+	1.55E-08
homocysteine metabolic process	13	3	.01	> 100	+	8.18E-04
sulfur amino acid metabolic process	34	3	.02	> 100	+	1.70E-02

Рис. 3. Статистический анализ представленности GO-категорий в списке анализируемых белков.

Всё равно недостаточно информации:)
Но теперь мы знаем, что объединяет эти белки в одну группу.
По-хорошему, нужно провести еще анализ по другим аннотациям, но я перейду к следующей части

ЧАСТЬ 3. Индивидуальный анализ (Human Protein Atlas)

Выберем один ген: CBLIF.
Узнаем его функции, роль и другую информацию с помощью базы данных Human Protein Atlas (HPA).

Эта база данных позволяет:

1. Посмотреть распределение белка по тканям.
2. Посмотреть экспрессию этого белка как в разных клеточных линиях (раковых в том числе), так и мозгу, крови, других органах (при заболевании тоже).
3. Посмотреть на его взаимодействие межбелковое и метаболическое.

Посмотрим где экспрессируется выбранный ген (рис. 4). Больше всего этот ген экспрессируется в желудке и мужских половых органах. Как это связано с кобаламином? Скорее всего всасывание в желудке этого витамина. Убедимся в этом

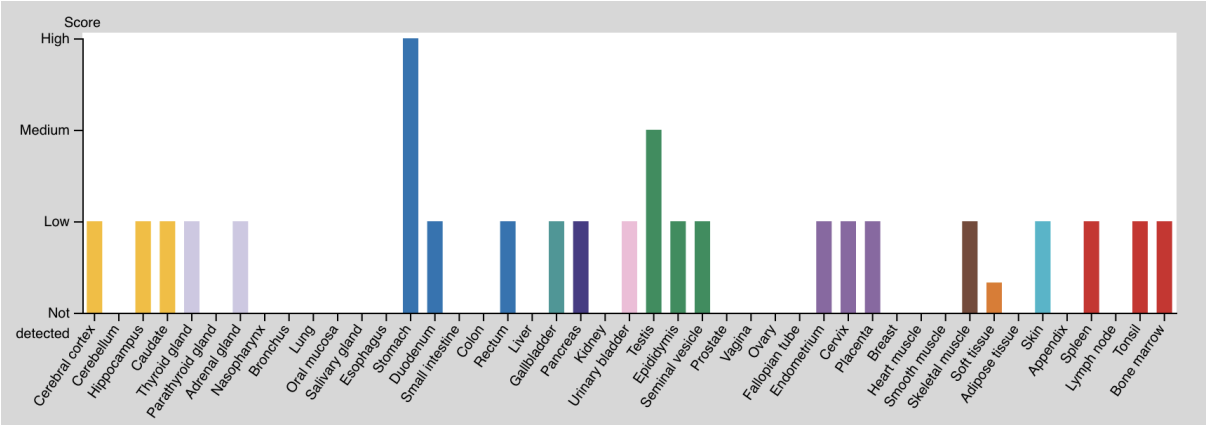

Рис. 4. Экспрессия CBLIF по органам.

Для того, чтобы понять, что именно делает этот белок, посмотрим с кем он взаимодействует (рис. 5). Этот белок взаимодействует с другим белком из списка (CUBN), а тот в свою очередь с третьим (AMN), а в их локализации наблюдается интересная закономерность - секреторный(CBLIF)-внутриклеточный(CUBN)-секреторный(AMN).

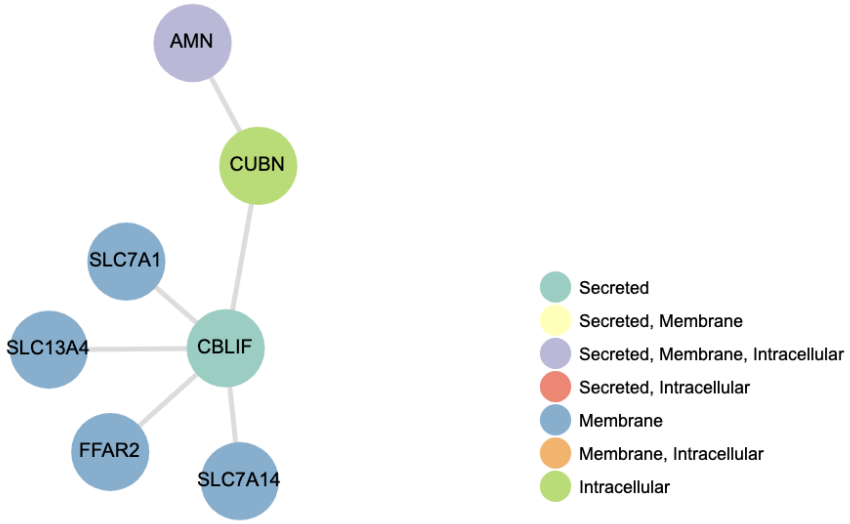

Рис. 5. Карта взаимодействия CBLIF с другими белками. В легенде цветом указана локализация белка.

Теперь меня полностью не смущает такое разнообразие в GO-категориях клеточной локализации (рис. 2), а вдобавок посмотрев, что CBLIF выполняет транспортную функцию, функцию всасывания (в разделе Interaction$Metabolic), можно с уверенностью заявить, этот белок (CBLIF), ответсвенен за всасывание кобаламина в желудке, что напрямую влияет на метаболизм кобаламина. Круг замкнулся.