Pr 7

Герои сегодняшней сцены

Наши гена интереса на остаток дня такие:

```
ENPP1
RFK
ACP5
ENPP3
FLAD1
ACP1
ACP2
BLVRB
MALAT1
```

Мотивация

Работать будем с базами данных GO и KEGG. Стандартно - скажете вы. Но я готов предоставить свою мотивацию. GO интересна тем, что является огромным графом биологических терминов, соединённых рёбрами разного "цвета", такая попытка обобщения накопившейся суммы биологических знаний мне по нраву. KEGG же мне любопытен тем, что это БД по метаболическим путям, что мне кажется невероятно полезной историей - раньше много думал про то, а есть ли где-то обобщение всей БХи разных организмов, и, оказывается, есть. Так что всё это кажется мне куда интереснее, чем красивые сайты с кучей интерактива.

Возможности GO

Gene Onthology (GO) позволяет черпать информацию из глобального графа биологических знаний для произвольного гена. На пользовательском уровне это означает, что GO позволяет нам для заданного гена найти слова, которые обычно вместе с ним произносят и причины, почему это так. У причин бывают разные источники аннотаций, каждому из которых в зависимости от задачи можно доверять в большей или в меньшей степени. Кроме того, слова фильтруются по "специфичности" к нашему гену - нам, как правило (но не всегда!) интересны те слова, вероятность встретить которые по случайным причинам мала. Такой отбор проводится с помощью точного теста Фишера (его в GO делает Panther'ная часть), который является обычным расчётом вероятности того, что гипергеометрическая случайная величина примет конкретное значение. Описанная процедура носит название *GO enrichment analysis*.

GO enrichment analysis

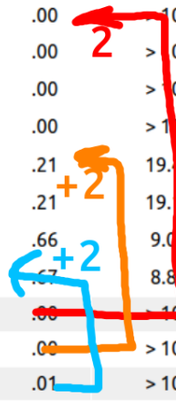
Итак, поглядим на выдачу GOea для нашего замечательного набора генов (поиск осуществлялся среди генов человека).

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
flavin adenine dinucleotide biosynthetic process	4	2	.00	> 100	+	7.92E-07	1.21E-02
↳flavin-containing compound biosynthetic process	4	2	.00	> 100	+	7.92E-07	6.04E-03
↳flavin-containing compound metabolic process	6	2	.00	> 100	+	1.98E-06	6.03E-03
↳flavin adenine dinucleotide metabolic process	5	2	.00	> 100	+	1.32E-06	6.71E-03
↳nucleotide metabolic process	529	4	.21	19.46	+	2.78E-05	4.23E-02
↳nucleoside phosphate metabolic process	537	4	.21	19.17	+	2.95E-05	4.08E-02
↳phosphate-containing compound metabolic process	1711	6	.66	9.03	+	7.89E-06	2.00E-02
↳phosphorus metabolic process	1737	6	.67	8.89	+	8.62E-06	1.88E-02
riboflavin metabolic process	5	2	.00	> 100	+	1.32E-06	5.03E-03
nucleoside triphosphate catabolic process	12	2	.00	> 100	+	8.70E-06	1.66E-02
phosphate ion homeostasis	20	2	.01	> 100	+	2.50E-05	4.24E-02

Сразу в глаза бросается большой блок про FAD и нуклеотиды. Видно, что есть 2 гена, продукты которых участвуют в синтезе непосредственно FAD, ещё 2 - в метаболизме нуклеотидов, и ещё 2 (то есть теперь уже 6) генов имеют отношение к метаболизму фосфорсодержащих соединений. По мимо FAD-блока мы видим, что есть гены, ответственные за гомеостаз PO_4^{3-}, метаболизм рибофлавина и катаболизм нуклеозидтрифосфатов.

Какие выводы можно по этому всему провести?
Видимо, мы имеем дело с группой генов, продукты которых связаны с синтезом FAD и работой с нуклеозидтрифосфатами. Можно предположить, что те 2 гена, что не входят в группу `nucleoside metabolic process`, но включены в `phosphate-containing compound metabolic precoesses` - это и есть два гена, ниже проануотированные как `phosphate ion homeostasis`, и, вероятно, их функция состоит в отщеплении/присоединении фосфата. Если в отщеплении, то эти же два гена могут оказаться генами с аннотацией `nucleoside triphosphare catabolic process`. Ещё, по-видимому, 2 гены, связанных с `riboflavin matabolic process` - первые 2 гена в FAD-блоке. В общем, считаю правильным подкрепить идеи, которые озвучил, картиночкой:

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
flavin adenine dinucleotide biosynthetic process	4	2	.00	100	+	7.92E-07	1.21E-02
↳flavin-containing compound biosynthetic process	4	2	.00	> 100	+	7.92E-07	6.04E-03
↳flavin-containing compound metabolic process	6	2	.00	> 100	+	1.98E-06	6.03E-03
↳flavin adenine dinucleotide metabolic process	5	2	.00	> 100	+	1.32E-06	6.71E-03
↳nucleotide metabolic process	529	4	.21	19.46	+	2.78E-05	4.23E-02
↳nucleoside phosphate metabolic process	537	4	.21	19.17	+	2.95E-05	4.08E-02
↳phosphate-containing compound metabolic process	1711	6	.66	9.03	+	7.89E-06	2.00E-02
↳phosphorus metabolic process	1737	6	.67	8.89	+	8.62E-06	1.88E-02
riboflavin metabolic process	5	2	.00	> 100	+	1.32E-06	5.03E-03
nucleoside triphosphate catabolic process	12	2	.00	> 100	+	8.70E-06	1.66E-02
phosphate ion homeostasis	20	2	.01	> 100	+	2.50E-05	4.24E-02



Под +2 понималось, что именно из-за конкретных двух генов у нас увеличивается каунт на 2. При этом приведённая картинка точно верна не полностью - при таком положении дел у нас бы было 2 некартированных гена, а мы видим, что он только 1:

Results

	Reference list	upload_1
Uniquely Mapped IDS:	20592 out of 20592	8 out of 8
Unmapped IDs:	0	1
Multiple mapping information:	0	0

А теперь проверим, не соврал ли нам веб-сервис по поводу вероятностей, которые он нам предоставил (на картинке `raw p-value`, но, поскольку мы использовали опцию `Fisher exact`, это, вообще-то, не p-value).
Например, проведём вычисление для колонки `phosphorus metabolic process`. Мы подали на вход выборку из 8 генов, и 6 из них были проаннотированы как гены, участвующие в метаболизме PO_4^{3-}, вероятность того, что это произошло по случайным причинам, равна $\frac{C_{20}^6 C_{20586}^2}{C_{20592}^8}$. Считаем и получаем:

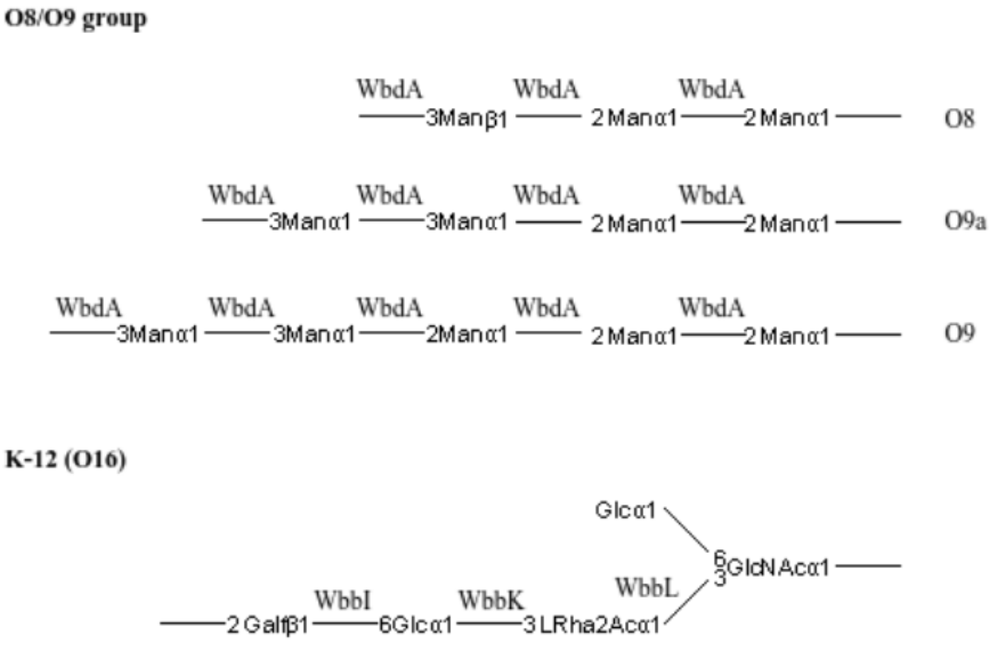
```
from math import comb
comb(20, 6)*comb(20586, 2)/comb(20592, 8)
>>> 1.0256507540768115e-17
```

И что-то это совсем не `8.62e-6`. Видимо, есть ещё какая-то обёртка над Фишером, призванная решать проблему с множественным тестированием.

Хорошо, мы теперь представляем, куда копать. В метаболизм, вперёд на KEGG!

KEGG (А точнее, PATHWAY)

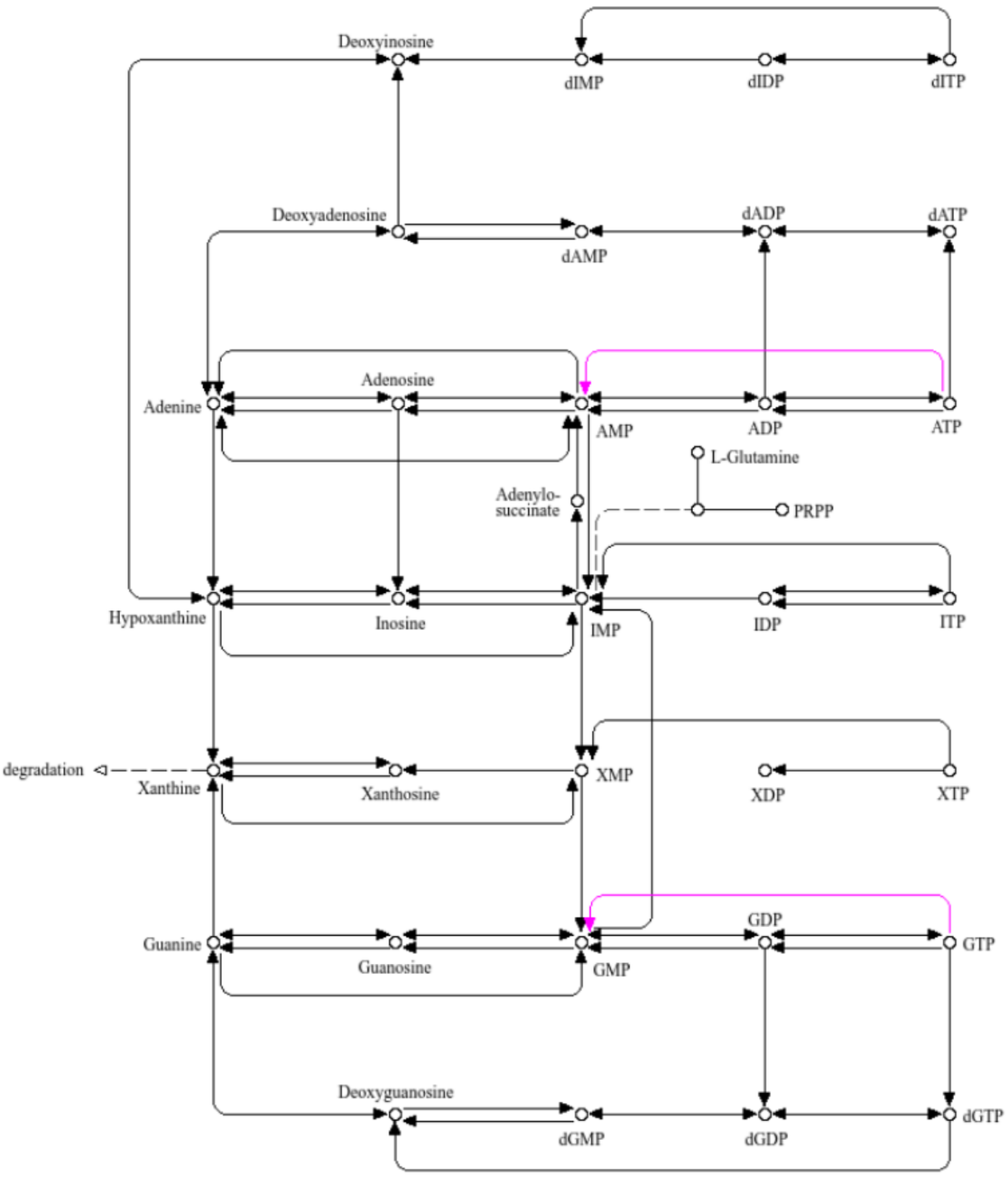
Kyoto Enciclopedia of Genes and Genes (KEGG) - японская коллекция коллекций, в осовном, биохимического толка. Так, например, коллекция GLYCANS включает в себя картинки с известными к моменту последнего обновления паттернами в структурах полисахаридов в живых организмах:

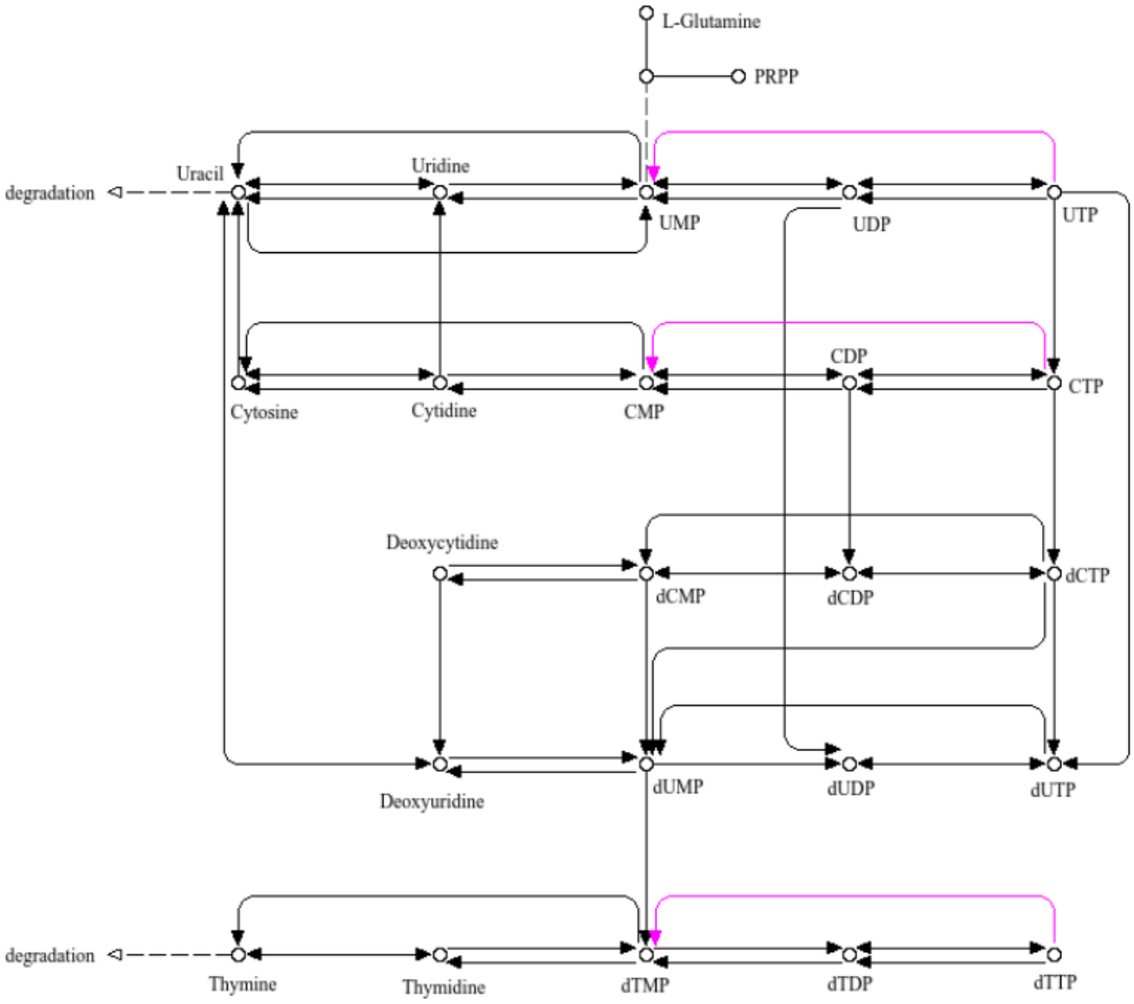


Pathway - коллекция метаболических графов, в каждом из которых рёбра могут быть ассоциированы с конкретными генами. Структурной единицей Pathway является карта (к слову, структура базы по гликанам устроена так же, то есть как набор постеров про разные семейства полисахаридов).

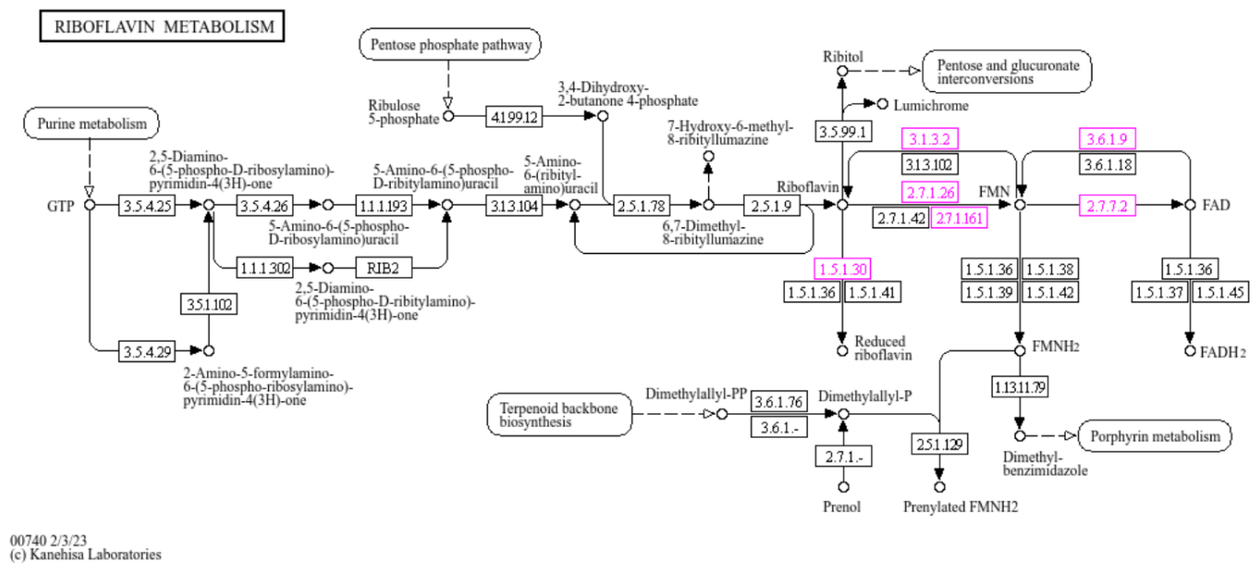
Отправляемся на метаболическую карту синтеза нуклеотидов и ищем там наши белки из списка. Из всех генов тут, оказывается, можно найти только ENPP3, здесь он катализирует отщепление отщелпение пирофосфата от ATP, CTP, UTP, GTP и dTTP.
На карте это выглядит так (описанные реакции отмечены фиолетовым цветом):

NUCLEOTIDE METABOLISM





Карты про метаболизм фосфора не приносят находок совсем, а вот [карта про метаболизм рибофлавина](#) очень радует - на ней обнаружились все гены, кроме ENPP3 (что логично, ведь он другими задачами занимается).



00740 2/3/23
(c) Kanehisa Laboratories

Таким образом, мы получили довольно ясную картину: мы имеем дело с генами, белковые продукты которых обеспечивают синтез FAD из рибофлавина и распад FAD в обратную сторону, а также восстановление рибофлавина. Правда, не очень понятно, как в эту картину вписывается реакция, катализируемая ENPP3: разве что можно предположить, что ENPP3 входит в состав комплекса, синтезирующего FAD из FMN, где не просто гидролизует ATP до AMP, а участвует в его присоединении к FMN.

Ну вот такие дела пока что)