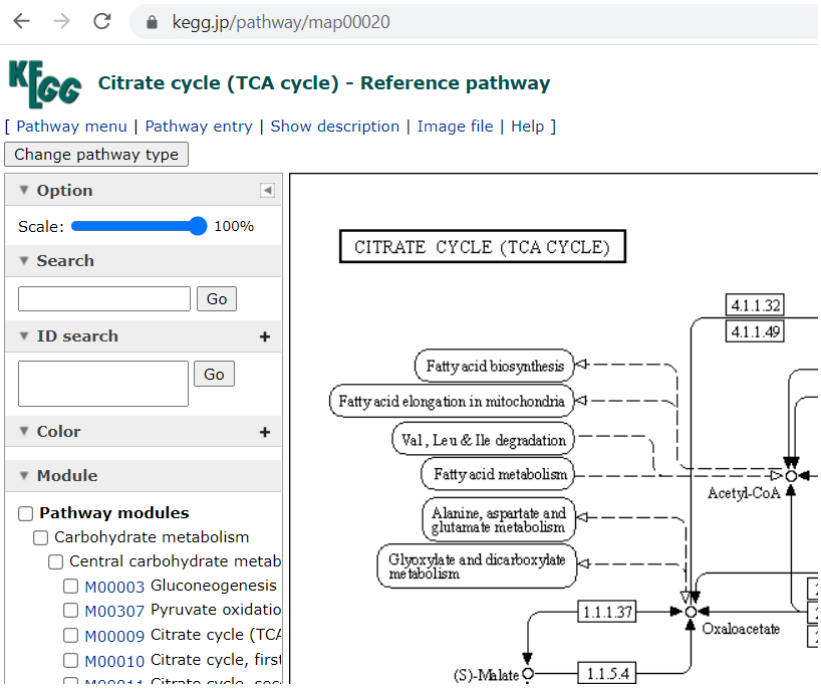
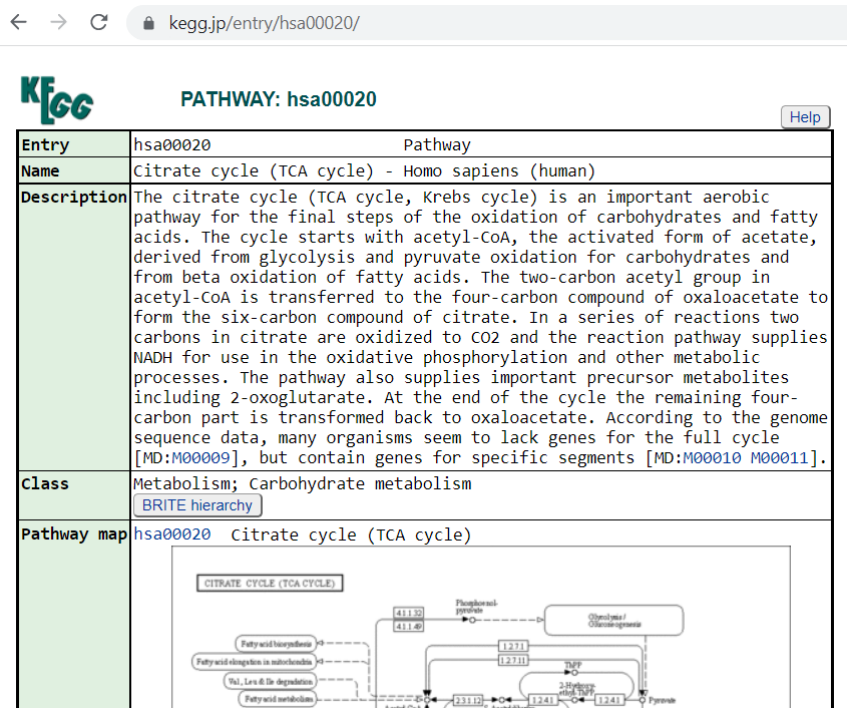
KEGG Pathways Dataset

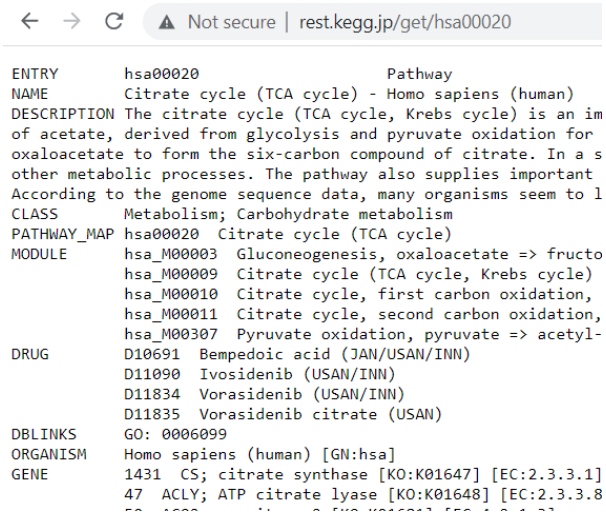
Συλλέγουμε δεδομένα από τη βάση δεδομένων Pathway του KEGG.
Κάθε entry είναι ένα map, είναι γενικό και μπορεί να αφορά πολλούς οργανισμούς. Το url για την προβολή του είναι όπως παρακάτω:



Αν θέλουμε να ασχοληθούμε συγκεκριμένα με τι ισχύει στον ανθρώπινο οργανισμό, μπορούμε να κρατήσουμε τον κωδικό του map (εδώ 00020) και να αλλάξουμε το url ως:



Η ίδια πληροφορία από το rest api KEGG:



Από το api, μας δίνεται πρόσβαση στο αρχείο kgml (αν υπάρχει) του κάθε map. Είναι ένα αρχείο xml το οποίο περιλαμβάνει entries και relations μεταξύ αυτών.

Με ένα script (prep.py), κατεβάζουμε όλα τα αρχεία kgml που υπάρχουν από την βάση δεδομένων. Βρίσκουμε ότι 207/558 maps δεν αφορούν τον ανθρώπινο οργανισμό, άλλοι 25/558 maps έχουν καταγεγραμμένα entries αλλά δεν έχουν πίνακα relations ή reactions μεταξύ τους. Έτσι, τα χρήσιμα maps είναι 326/558, δηλαδή έχουν και entries και relations ή reactions μεταξύ τους.
Για κάθε map, σώζουμε σε ξεχωριστό αρχείο τα entries, τα relations και τα reactions που το αφορούν.

Το αρχείο entries έχει τη μορφή:

id	name	type	link	gene_names
18	hsa:226 hsa:229 hsa:230	gene	https://www.kegg.jp/dbget-bin/	ALDOA, ALDA, GSD12, HEL-S-87p
42	hsa:217 hsa:219 hsa:223 h	gene	https://www.kegg.jp/dbget-bin/	ALDH2, ALDH-E2, ALDHI, ALDM..
45	cpd:C00033	compound	https://www.kegg.jp/dbget-bin/	C00033
46	path:hsa00030	map	https://www.kegg.jp/dbget-bin/	Pentose phosphate pathway

Το αρχείο relations έχει τη μορφή:

entry1	entry2	link	value	name	pathway
73	75	ECrel	90	compound	hsa00010
73	74	ECrel	90	compound	hsa00010
67	47	maplink	88	compound	hsa00010
67	46	maplink	94	compound	hsa00010

Το αρχείο reactions έχει τη μορφή:

head id	head name	tail id	tail name	link type	relation name	relation value	entry1	entry2	pathway
cpd:C15973	cpd:C15973	cpd:C15972	cpd:C15972	reaction	rn:R07618	reversible	74	71	hsa00020
cpd:C00026	cpd:C00026	cpd:C05381	cpd:C05381	reaction	rn:R00621	irreversible	76	77	hsa00020
cpd:C00068	cpd:C00068	cpd:C05381	cpd:C05381	reaction	rn:R00621	irreversible	72	77	hsa00020
cpd:C05381	cpd:C05381	cpd:C16254	cpd:C16254	reaction	rn:R03316	irreversible	77	73	hsa00020

Στις αντιδράσεις έχουμε substrates τα οποία παράγουν products. Μπορεί ένα ή παραπάνω substrates να χρειάζονται για την παραγωγή ενός ή περισσότερων products.
Τα substrates και products που αναφέρονται είναι πάντα compounds.
Για την καταγραφή των συσχετίσεων καταγράφονται όλοι οι συνδυασμοί. Δηλαδή όλα τα substrates μιας αντίδρασης έχουν σχέση με όλα τα products της αντίδρασης.

Μια εξήγηση των παραπάνω υπάρχει στο link: https://www.kegg.jp/kegg/xml/docs/

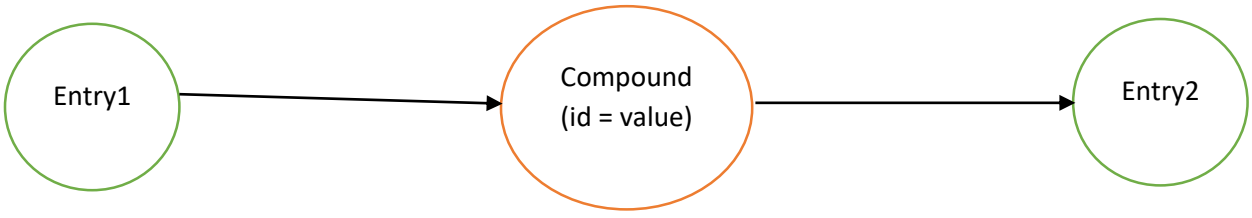
Μερικές απο τις πληροφορίες που μας αφορούν:

attribute name	data type	explanation
entry1	idref.type	the first (from) entry that defines this relation
entry2	idref.type	the second (to) entry that defines this relation
type	relation-type.type	the type of this relation

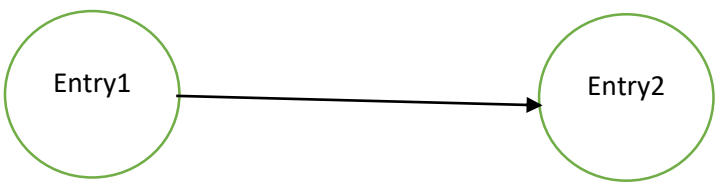
attribute value	explanation
ECrel	enzyme-enzyme relation, indicating two enzymes catalyzing successive reaction steps
PPrel	protein-protein interaction, such as binding and modification
GErel	gene expression interaction, indicating relation of transcription factor and target gene product
PCrel	protein-compound interaction
maplink	link to another map

name	value	ECrel	PPrel	GErel	Explanation
compound	Entry element id attribute value for compound.	*	*		shared with two successive reactions (ECrel) or intermediate of two interacting proteins (PPrel)
hidden compound	Entry element id attribute value for hidden compound.	*			shared with two successive reactions but not displayed in the pathway map
activation	-->		*		positive and negative effects which may be associated with molecular information below
inhibition	--|		*		
expression	-->			*	interactions via DNA binding
repression	--|			*	
indirect effect	..>		*	*	indirect effect without molecular details
state change	...		*		state transition
binding/association	---		*		association and dissociation
dissociation	-+-		*		
missing interaction	-/-		*	*	missing interaction due to mutation, etc.
phosphorylation	+p		*		molecular events
dephosphorylation	-p		*		
glycosylation	+g		*		
ubiquitination	+u		*		
methylation	+m		*		

Από τα παραπάνω και με επαλήθευση με παραδείγματα, συμπαιράναμε ότι όταν relation name == compound τότε ισχύουν οι εξής σχέσεις



Επομένως όταν το ονομα του relation είναι compound, τότε καταγράφουμε τις 2 σχέσεις που φαινονται παραπάνω, αλλιώς καταγράφουμε μία σχέση:



Τα παραπάνω γίνονται με το script curate.py και αποθηκεύονται σε ένα αρχείο (All _relations-Curated.csv) με μορφή:

	head id	head name	tail id	tail name	pathway	link type	relation name	relation value	entry1	entry2
0	hsa:130589	GALM	cpd:C00267	C00267	hsa00010	ECrel	compound	90.0	73.0	75.0
1	cpd:C00267	C00267	hsa:3098	HK1	hsa00010	ECrel	compound	90.0	73.0	75.0
3	cpd:C00267	C00267	hsa:2645	GCK	hsa00010	ECrel	compound	90.0	73.0	74.0
5	cpd:C00267	C00267	hsa:2538	G6PC1	hsa00010	ECrel	compound	90.0	73.0	76.0

Τέλος με το GrabNames.py, βρίσκουμε τα πλήρη ονόματα των ουσιών, βάση των ids και τα συμπληρώνουμε.

Προκύπτει το τελικό αρχείο (All _relations-Curated-full names.csv) με μορφή:

head id	head name	tail id	tail name	pathway	link type	relation name	relation value	entry1	entry2	head full name	tail full name
hsa:130589	GALM	cpd:C00267	C00267	hsa00010	ECrel	compound	90.0	73.0	75.0	GALM, galactose mutarotase	alpha-D-Glucose
cpd:C00267	C00267	hsa:3098	HK1	hsa00010	ECrel	compound	90.0	73.0	75.0	alpha-D-Glucose	HK1, hexokinase 1
cpd:C00267	C00267	hsa:2645	GCK	hsa00010	ECrel	compound	90.0	73.0	74.0	alpha-D-Glucose	GCK, glucokinase
cpd:C00267	C00267	hsa:2538	G6PC1	hsa00010	ECrel	compound	90.0	73.0	76.0	alpha-D-Glucose	G6PC1, glucose-6-phosphatase catalytic subunit

Σε αυτό το αρχείο είναι αποθηκευμένες 17385 συσχετίσεις.