Συνολικός γράφος KEGG

Με βάση το αρχείο csv που φτιάξαμε προηγουμένως, κατασκευάζουμε τον συνολικό γράφο των δεδομένων από την βάση KEGG.

Μπορεί να κατασκευαστεί ως MultiDiGraph, ώστε να αποτυπώνεται η κατεύθυνση των σχέσεων:

```
MultiDiGraph with 5185 nodes and 17389 edges
```

Ή να μετατραπεί σε undirected Graph για απλοποίηση:

```
Graph with 5185 nodes and 11766 edges
```

Όσον αφορά την συνεκτικότητα του γράφου, υπάρχει ένα main component και 162 μικρά ασύνδετα components, τα περισσότερα εκ των οποίων είναι ζευγάρια.

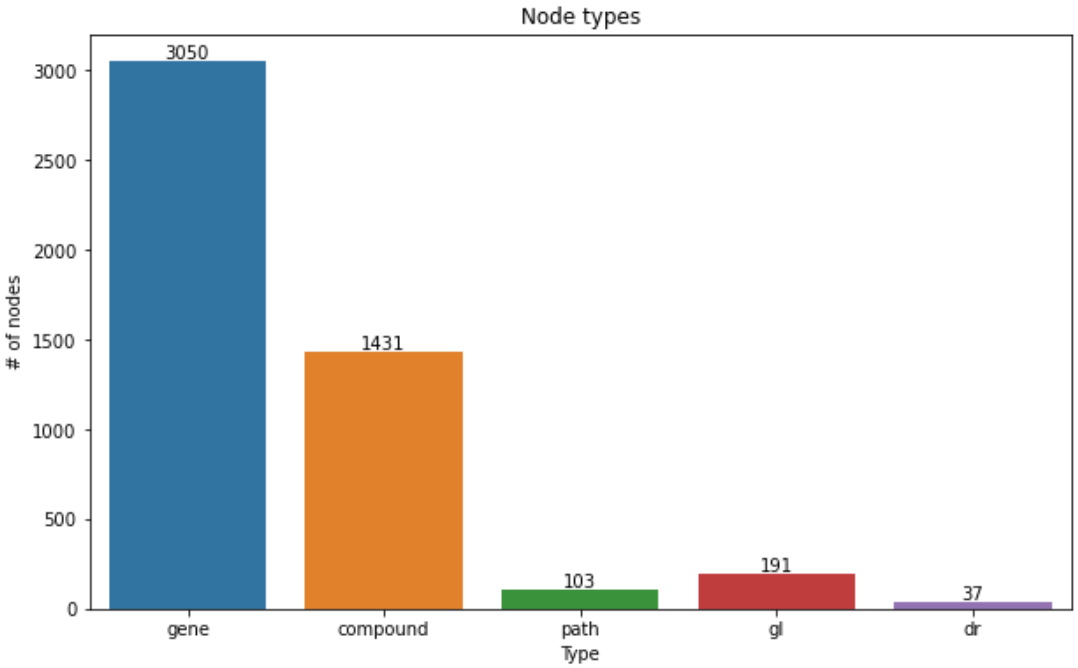
```
Number of connected components: 163
```

```
Number of nodes in each component:
[4812, 7, 7, 7, 6, 6, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3
, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

Για να μην υπάρχει bias κατά συγκεκριμένων κόμβων, είναι σκόπιμο να αφαιρεθούν οι λίγοι κόμβοι οι οποίοι είναι ασύνδετοι με το main component του γράφου. Έτσι προκύπτουν οι γράφοι:

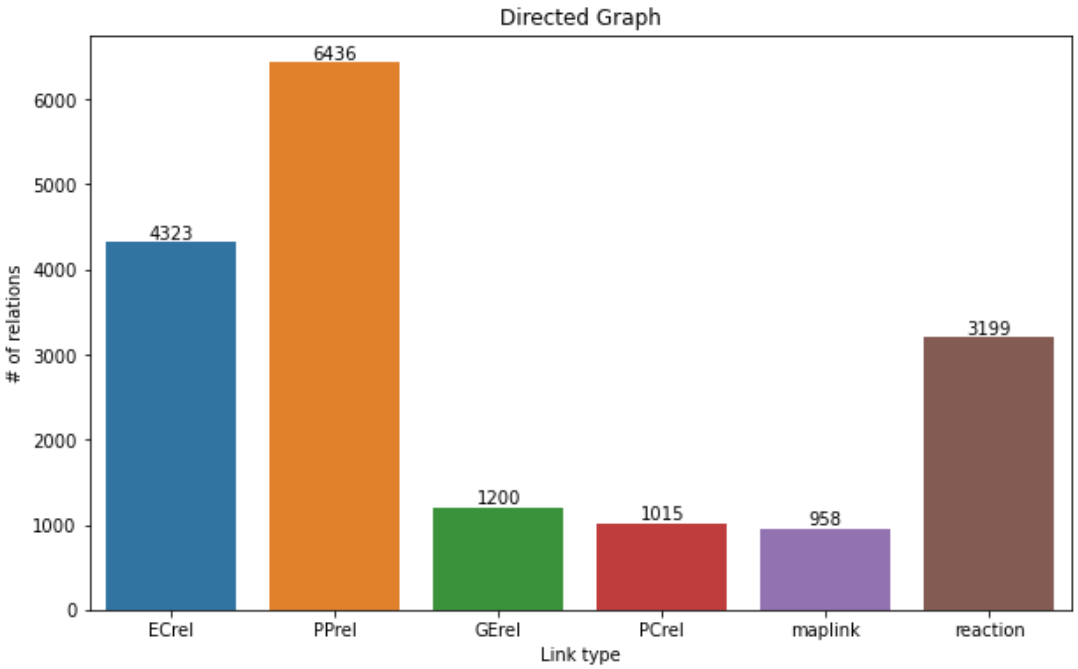
```
MultiDiGraph with 4812 nodes and 17131 edges
Graph with 4812 nodes and 11539 edges
```

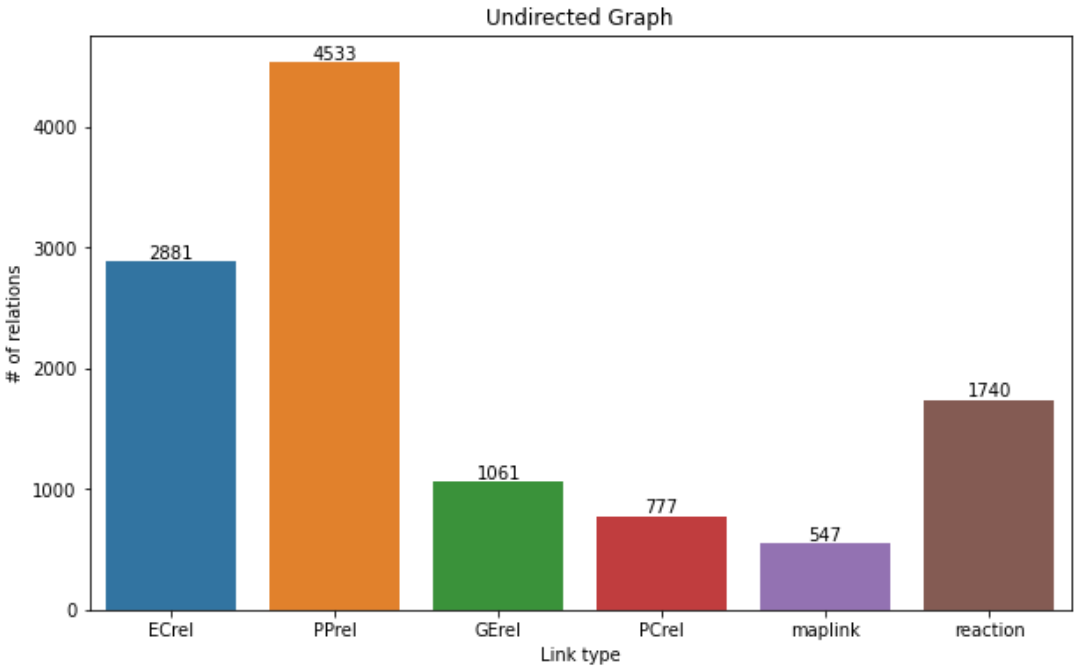
Μπορούμε να δούμε μερικά στατιστικά στοιχεία για τα δεδομένα μας:



Οι κόμοβοι gene και compound μας ενδιαφέρουν πρωτίστως, ενδεχομένως και τα glycans και drugs αργότερα. Οι κόμβοι paths μπορεί να έχουν χρήση ως συνεκτικοί κρίκοι ανάμεσα σε pathways.

Η κατανομή των link types είναι όπως παρακάτω:





Τα links τύπου ECrel είναι όλα compound relations, εκτός ελαχίστων εξαιρέσεων.

Τα maplinks είναι επίσης όλα compound relations.

Τα PPrel είναι activation, inhibition, binding κ.α., κυρίως γονίδιο-γονίδιο (υπάρχουν και μερικά compounds, καθώς και κάποια χωρίς όνομα σχέσης).

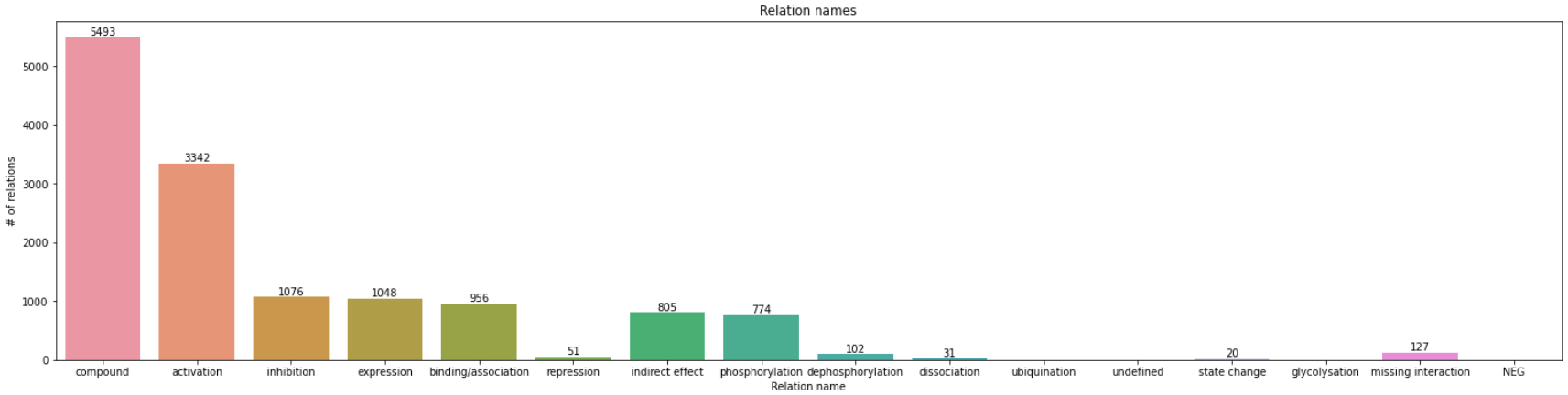
Τα PCrel είναι activation, inhibition, binding κ.α., μικτά γονίδια, compounds κλπ.

Τα GErel είναι expression (ή σπανια repression), πάντα γονίδιο-γονίδιο.

Τα reactions είναι πάντοτε μεταξύ compounds. Συνδέουν όλα τα substrates με όλα τα products.

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

Η κατανομή των relation names είναι όπως παρακάτω (δεν περιλαμβάνονται τα reactions):



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		*		

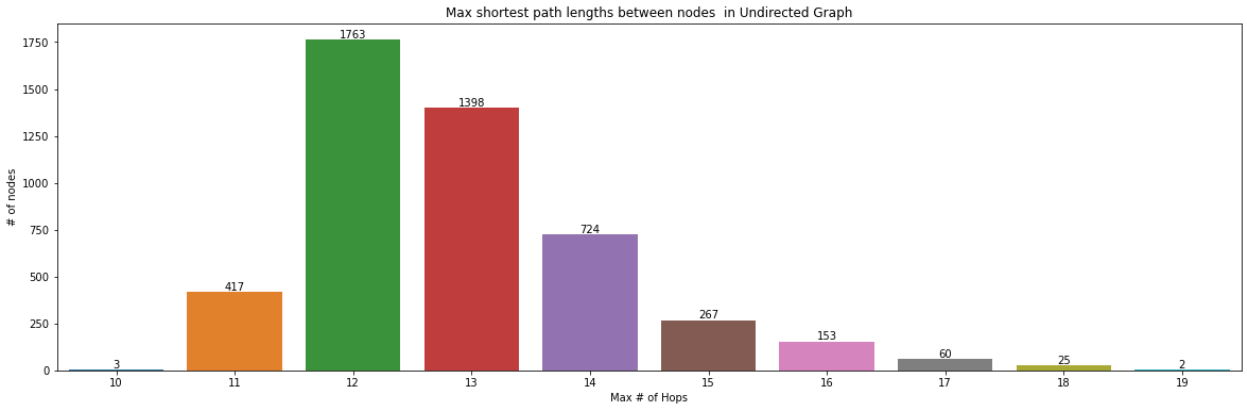
Μέγιστες αποστάσεις μεταξύ κόμβων

Undirected Graph

Στα παρακάτω γραφήματα φαίνονται οι μέγιστες αποστάσεις που έχει κάθε κόμβος με οποιονδήποτε άλλο κόμβο του γράφου. Εξετάζεται μόνο το main component, επομένως όλοι οι κόμβοι συνδέονται με όλους με αρκετά βήματα (Hops).

Δηλαδή αν για τον κόμβο «Α» ο πιο απομακρυσμένος κόμβος είναι ο «Β» με απόσταση 13 Hops, τότε ο κόμβος «Α» είναι ένας από τους 1398 κόμβους της κόκκινης στήλης στο επόμενο γράφημα.

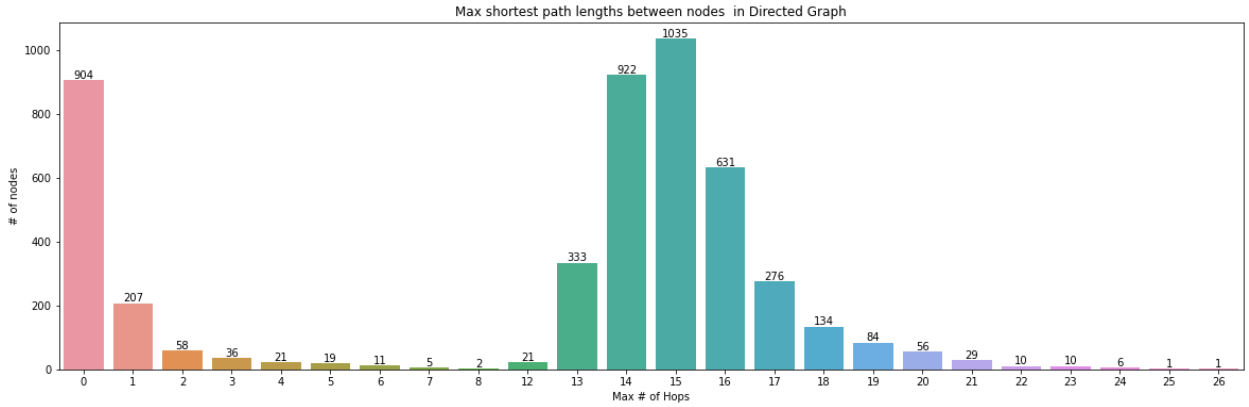
Αυτό το γράφημα αφορά τον undirected γράφο, επομένως οι σχέσεις θεωρούνται αμφίδρομες. Δεν υπάρχει επομένως έννοια αιτίου-αιτιατού, απλώς συσχέτιση.



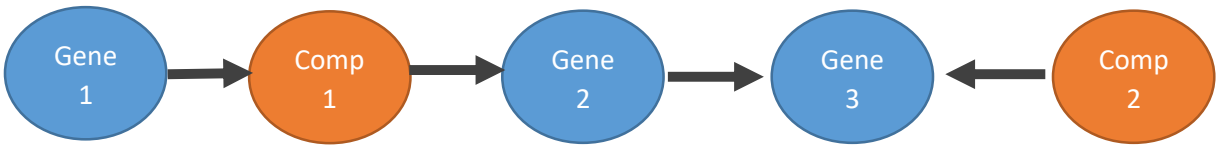
Φαίνεται ότι για τους περισσότερους κόμβους, το πιο απομακρυσμένο ζευγάρι βρίσκεται εντός 11-15 Hops. Δηλαδή πάνω από 11 Hops, θα ήταν λογικό να θεωρηθεί ότι δεν υπάρχει συσχέτιση μεταξύ των 2 κόμβων.

Directed Graph

Στον Directed γράφο, παρ' όλο που εξετάζεται μόνο το main component το οποίο είναι ίδιο με του undirected γράφου, υπάρχει η έννοια της κατεύθυνσης και του αιτίου – αποτελέσματος. Έτσι, για όλους τους κόμβους του γράφου υπάρχει κάποιος κόμβος που είναι απρόσιτος.
Το παρακάτω γράφημα δείχνει τις μέγιστες αποστάσεις μόνο εφ' όσον υπάρχει διαδρομή από τον ένα κόμβο προς τον άλλο.



Για παράδειγμα, έστω ότι το παρακάτω σχήμα είναι ένας κατευθυνόμενος γράφος.



Αν να χρησιμοποιηθούν οι κατευθύνσεις ως δεδομένο στην εκπαίδευση ενός μοντέλου, θα πρέπει να προβλεφθούν ως έξοδος, τουλάχιστον τα παρακάτω:

Head	Tail	Output
Gene 1	Comp 1	True / High Probability
Comp 1	Gene 1	False / Low Probability
Comp 1	Gene 2	True / High Probability
Gene 2	Comp 1	False / Low Probability
Gene 2	Gene 3	True / High Probability
Gene 3	Gene 2	False / Low Probability
Comp 2	Gene 3	True / High Probability
Gene 3	Comp 2	False / Low Probability

Ενδεχομένως το γεγονός ότι το ίδιο ζευγάρι μπορεί να συναντηθεί και ως True και ως False, αναλόγως ποιος κόμβος είναι Head ή Tail, να δυσκολέυσει την εκπαίδευση του μοντέλου.

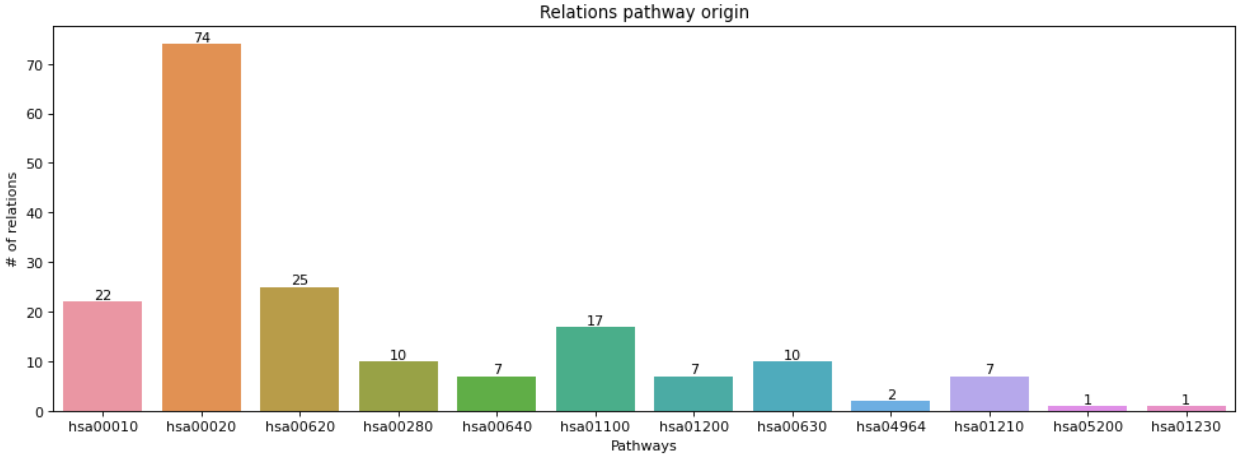
Ανακατασκευή του κύκλου του Κρεμπς

Από τον συνολικό γράφο αφαιρούνται όλοι οι κομβοι εκτός αυτών που περιλαμβάνονται στον κύκλο του Κρεμπς. Προκύπτουν οι γράφοι:

```
Directed:      MultiDiGraph with 36 nodes and 183 edges
Undirected:    Graph with 36 nodes and 71 edges
```

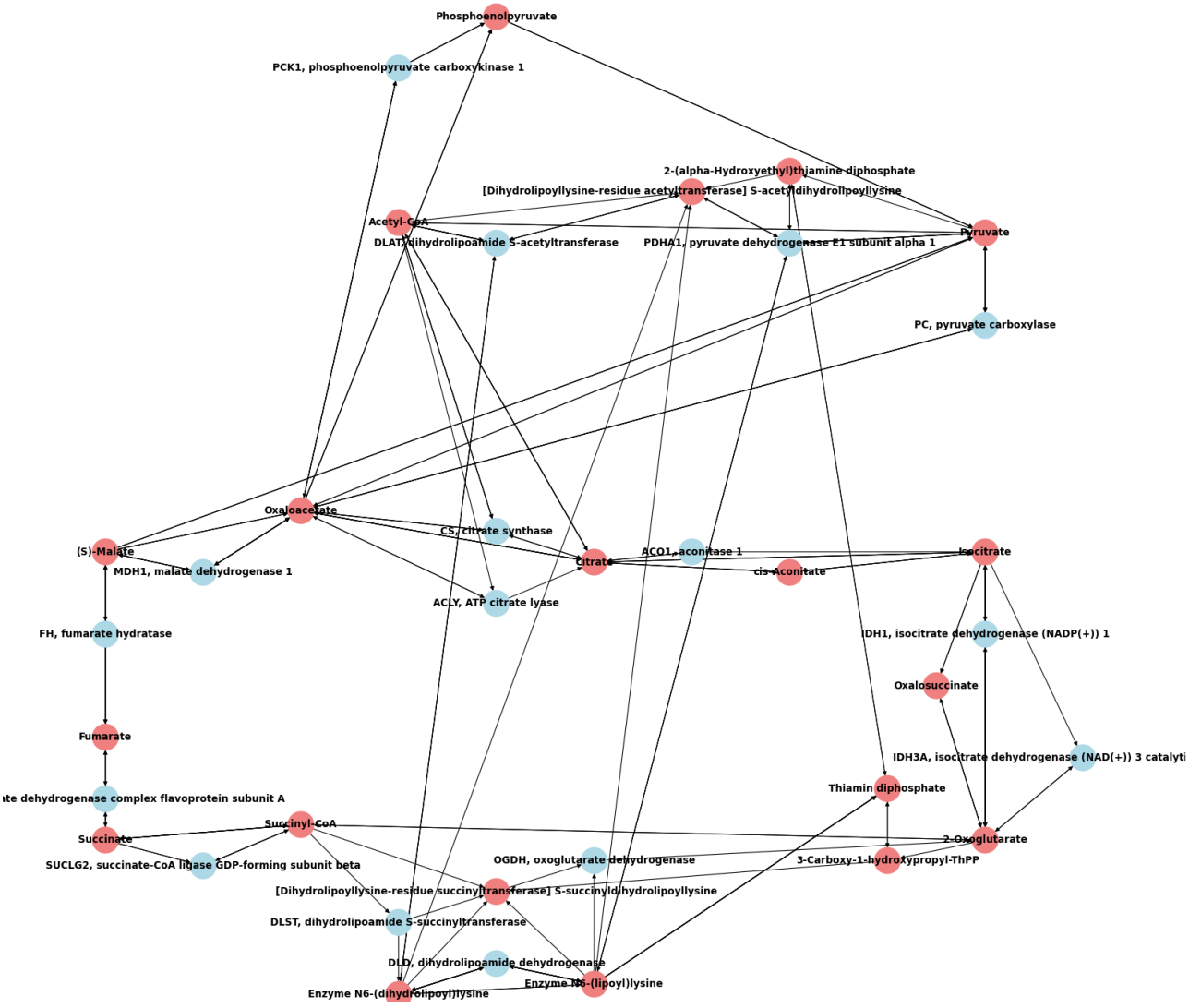
Οι ακμές μειώνονται σε λιγότερες από τις μισές στον undirected γράφο. Επομένως υπάρχουν σίγουρα διπλότυπες ακμές με την ίδια κατεύθυνση από διαφορετικά maps.

Οι ακμές του directed γράφου προκύπτουν από τα παρακάτω pathways:

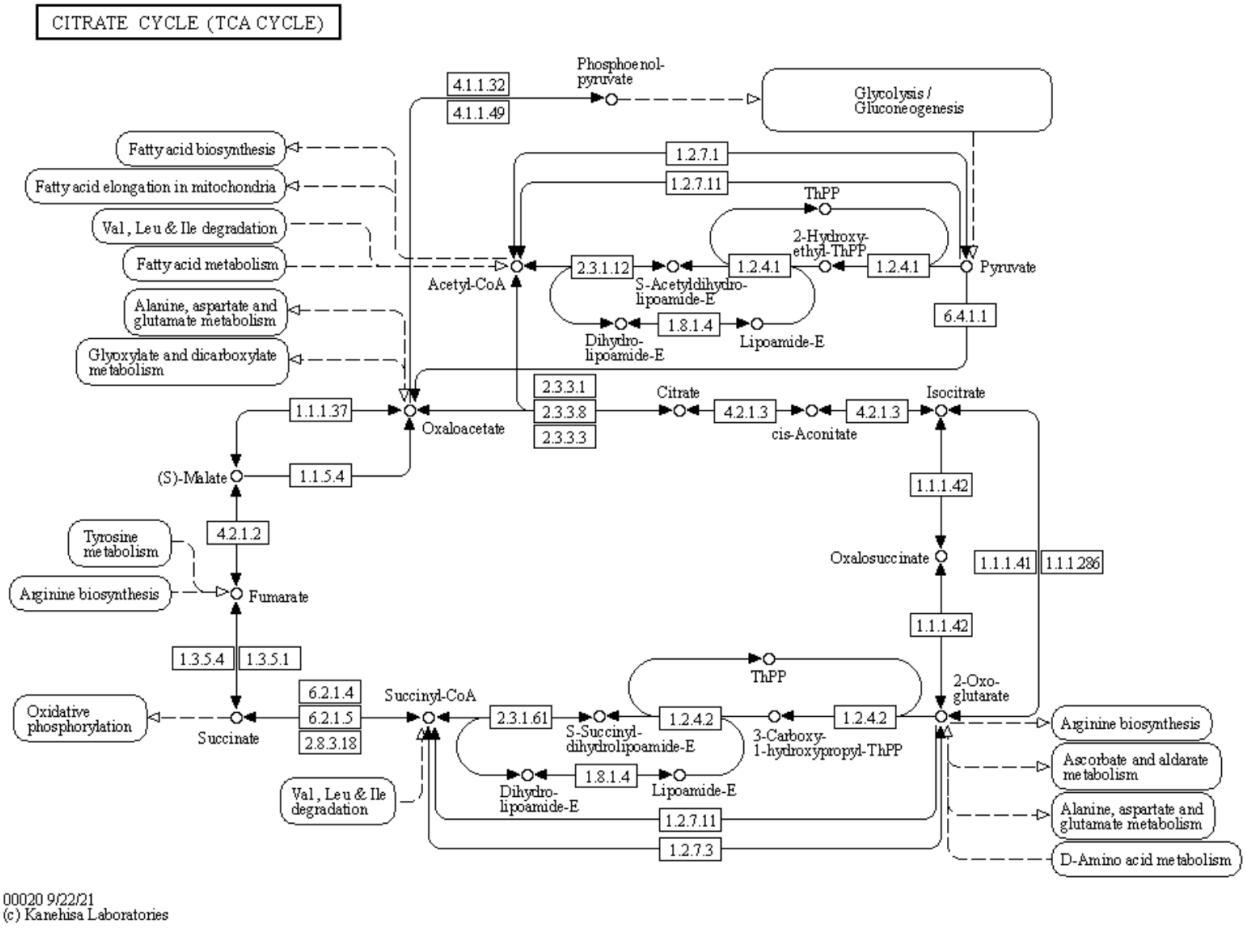


Map id	*Ονομασία*
hsa00010	Glycolysis / Gluconeogenesis
hsa00020	Citrate cycle (TCA cycle)
hsa00620	Pyruvate metabolism
hsa00280	Valine, leucine and isoleucine degradation
hsa00640	Propanoate metabolism
hsa01100	Metabolic pathways (Γενικό και περιέχει μόνο reactions)
hsa01200	Carbon metabolism
hsa00630	Glyoxylate and dicarboxylate metabolism
hsa04964	Proximal tubule bicarbonate reclamation
hsa01210	2-Oxocarboxylic acid metabolism
hsa05200	Pathways in cancer
hsa01230	Biosynthesis of amino acids

Παρακάτω βρίσκεται μια αναπαράσταση του κύκλου του Κρεμπς, βάση των δεδομένων του γράφου. Οι θέσεις είναι χειροκίνητες ώστε να θυμίζουν την εικόνα από το KEGG. Στην αναπαράσταση τυπώνονται μόνο compounds και genes. Τα compounds είναι σε κόκκινο χρώμα ενώ τα γονίδια σε γαλάζιο.



Η εικόνα του κύκλου από το KEGG



Από την αναπαράσταση του γράφου λείπουν όσα γονίδια δεν υπάρχουν στον ανθρώπινο οργανισμό.

Το μόνο ανθρώπινο γονίδιο που λείπει είναι το

hsa:8802 “SUCLG1, succinate-CoA ligase GDP/ADP-forming subunit alpha”

το οποίο θεωρήθηκε σε μεγάλο βαθμό συγγενές με το

hsa:8801 “SUCLG2, succinate-CoA ligase GDP-forming subunit beta”

και έτσι παραλήφθηκε στην διαδικασία συλλογής και συγκεκριμένα λόγω επιλογής του πρώτου μόνο ονόματος από κάθε entry (αρχείο “Curate.py”)

Ενδεχομένως αντίστοιχες περιορισμένες παραλείψεις να υπάρχουν και στα υπόλοιπα pathways.