2020

Human disease Network

DATA ANALYTICS

SIMONE MONTI 807994 – UNIVERSITÀ DEGLI STUDI MILANO BICOCCA

1 SOMMARIO

2 ABSTRACT

La rete che andremo ad analizzare nel corso di questo paper è caratterizzata da nodi rappresentanti le malattie dell'uomo e da archi che le mettono in relazione nel caso in cui queste posseggano dei geni condivisi, essa prende il nome di Human Disease Network. La HDN è di fondamentale importanza per risalire alla genetica comune di molte malattie e grazie a questo grafo siamo in grado di individuare schemi funzionali simili tra disturbi diversi.

Osserveremo l'esistenza di particolari malattie HUB e di come queste indichino una sorta di stato "finale" della mutazione genetica. Vedremo inoltre come la genetica delle malattie sia varia, come malattie che colpiscono lo stesso apparato possano essere completamente diverse a livello genetico e come, invece, alcuni gruppi di malattie siano geneticamente simili.

3 DATASET

3.1 PRESENTAZIONE DEL DATASET

Il dataset iniziale è dato da un grafo diretto composto da 1419 nodi divisi secondo due tipologie diverse:

- Disease: nodo corrispondente ad una particolare malattia dell'uomo,
- Gene: nodo rappresentante di un particolare gene.

Il grafo presenta 3926 archi diretti che svolgono due differenti funzioni:

- Mettere in relazione fra loro due differenti malattie se queste hanno in comune dei geni,
- Mettere in relazione malattie e geni se le prime presentano quel particolare gene all'interno del proprio patrimonio.

Questi dati sono stati reperiti dal compendio Online Mendelian Inheritance in Man[1] in cui sono contenuti i geni delle malattie e i loro fenotipi.

3.2 PRE-PROCESSING

Per andare rispondere ai quesiti posti come obiettivi è stato di nostro interesse essere a conoscenza dell'esistenza della relazione fra due malattie ma non del particolare gene in comune fra esse. Il primo step del pre-processing dunque è stato quello di andare ad eliminare tutti i nodi corrispondenti ai geni e gli archi ad essi associati trasformando la rete da bipartita a connessa.

La rete successivamente è stata trasformata in un grafo indiretto in quanto per ogni arco della rete era presente anche il suo diretto reciproco. Dopo averne dimezzato il numero la rete conta un insieme totale di 1184 archi indiretti.

Come ultimo step di pre-process sono stati rimossi quegli attributi che non portavano alcuna informazione aggiuntiva al nostro dataset.

Al termine del delle modifiche il dataset si presenta nella seguente maniera:

- Nodo:
  - Id: codice univoco di ogni nodo
  - Label: nome della malattia
  - Type: classe della malattia basata sul sistema fisiologico colpito.
- Edge:
  - Nodo source
  - Nodo target

[1] https://omim.org/

3.3 DESCRIZIONE DELLA RETE

La rete è composta da 515 nodi con un totale di 1184 archi indiretti, si tratta di una rete connessa con bassa densità e caratterizzata da una matrice di adiacenza sparsa (il diametro della rete è pari a 15). Gli archi non sono caratterizzati da particolari attributi, essi hanno il ruolo di mettere in relazione fra loro malattie che presentano una parte di patrimonio genetico comune (anche un singolo gene).

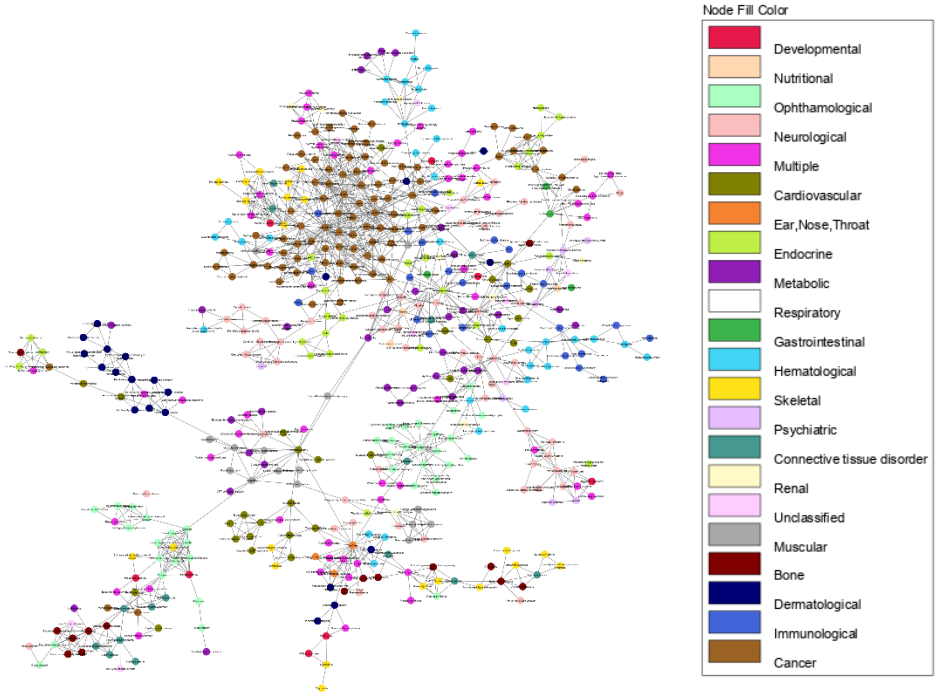

Figura 1 - Immagine della rete divisa per tipo malattia

Si può notare dalla *figura 1* come i nodi appartenenti alla parte alta dell'immagine siano molto più connessi rispetto ai nodi presenti nella metà bassa. I nodi si suddividono in 22 tipi diversi di malattie sulla base dell'apparato colpito. Inseguito è possibile osservare la distribuzione secondo il tipo (*figura 2*) e avere una proiezione della rete focalizzata sul raggruppamento dei nodi in base alla propria tipologia (*figura 3*).

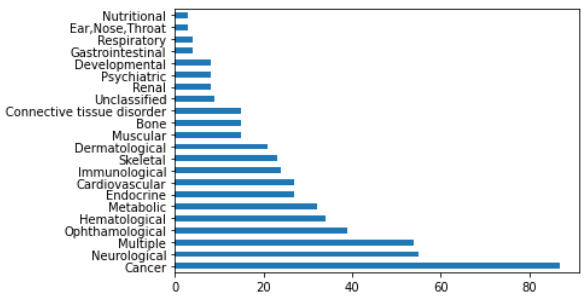

Figura 2 – distribuzione del tipo dei nodi della rete

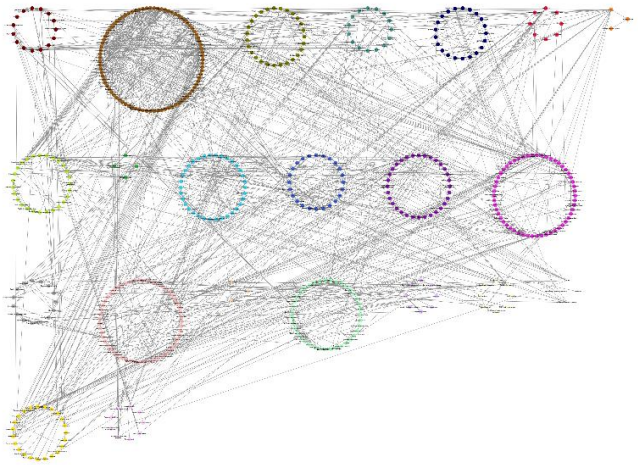

Figura 3 - rete con nodi raggruppati secondo la tipologia

4 ANALISI DELLA RETE

4.1 COME SI COMPORTANO I NODI CON GRADO DIVERSO?

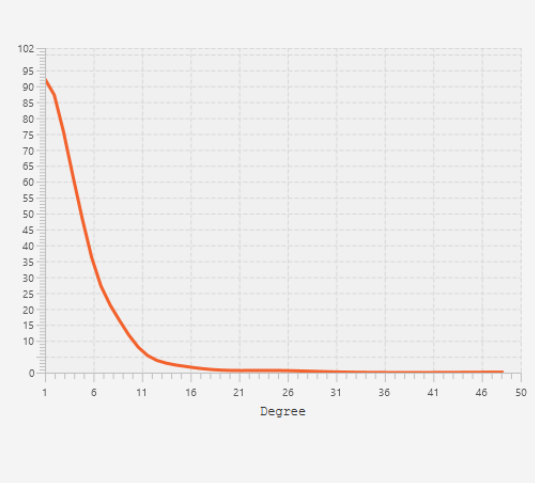

Figura 4 – distribuzione del grado dei nodi

Il grafico soprastante ci mostra come la rete sia caratterizzata dalla presenta di un elevato numero di nodi con basso grado, da pochi nodi con grado di medie dimensioni (6%) e un numero ancora inferiore di nodi HUB (1,75%). Abbiamo considerato i nodi come piccoli fino al grado 9, medi dal grado 10 al 19 e veri e propri HUB quelli con un grado pari o superiore a 20.

Colon cancer	Cancer	50
Breast cancer	Cancer	30
Gastric cancer	Cancer	27
Thyroid carcinoma	Cancer	26
Leukemia	Cancer	25
Deafness	Ear,Nose,Thr...	25
Diabetes mellitus	Endocrine	24
Pancreatic cancer	Cancer	23
Prostate cancer	Cancer	20

Figura 5 – elenco dei nodi HUB

Molti dei nodi HUB appartengono alla categoria Cancer (Neoplasie), questo comportamento, che verrà analizzato in seguito, è dovuto al fatto che questa tipologia di nodi è numerosa e con un elevata connettività fra i propri membri.

Abbiamo cercato di capire come il grado di un nodo potesse determinare il suo comportamento all'interno della Human Disease Network. Nella *figura 6* possiamo osservare i nodi colorati in maniera differente in base al proprio degree. Il nodo color verde (Colon cancer) corrisponde al nodo con grado massimo (50), i nodi di colore più o meno rosso indicano i nodi con un degree medio/alto e infine la maggioranza di nodi color giallo/arancione chiaro corrispondono ai nodi più piccoli. La rete è data da un grafo disassortativo, strutturato in modo tale che nodi avente un grado simile siano scarsamente collegati fra loro mentre è favorito il numero di archi tra nodi di diverse dimensioni.

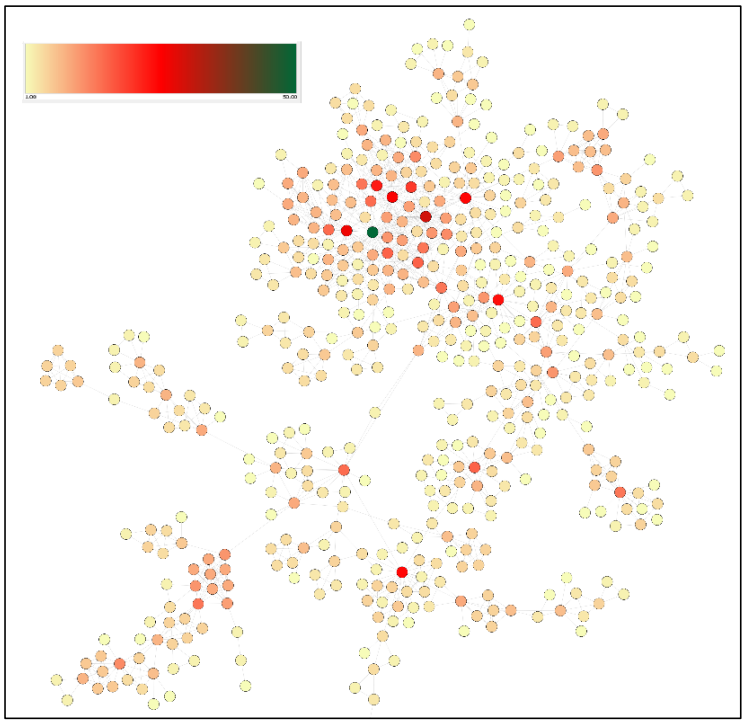

Figura 6 – proiezione della rete con nodi evidenziati in base al grado

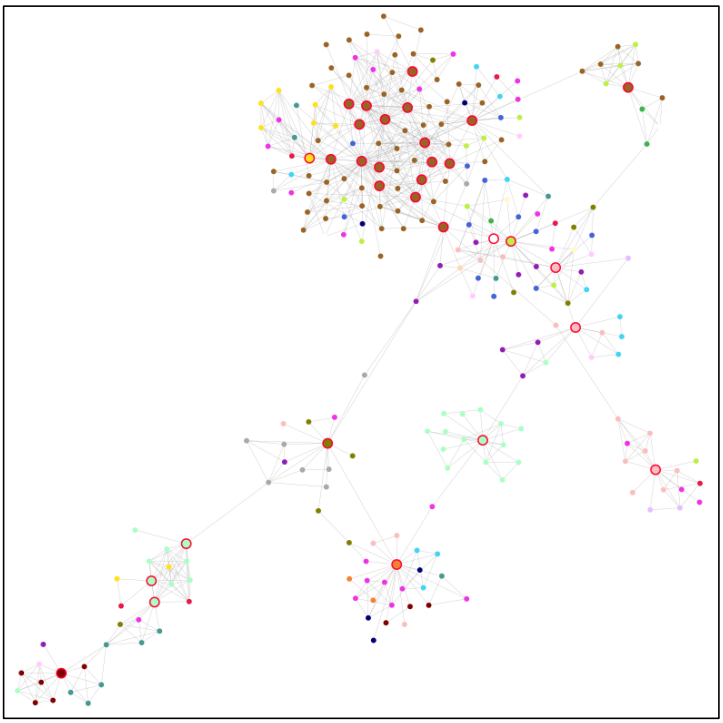

Figura 7 – rete con nodi HUB, medi e diretti vicini

La *figura 7* rappresenta una sottorete della HDN formata dai nodi HUB e da quelli di grado medio, entrambi evidenziati da un contorno rosso, a cui abbiamo aggiunto i diretti vicini (disegnati con circonferenza più piccola. Abbiamo analizzato nello specifico alcuni nodi (fra parentesi è indicato il grado):

- Diabates mellitus (24) – Figura A
- Alzheimer disease (15) – Figura B
- Deafness (25) – Figura C

Queste malattie sono gravi patologie dell'uomo che spesso sono diretta conseguenza delle complicazioni portate da altre malattie.

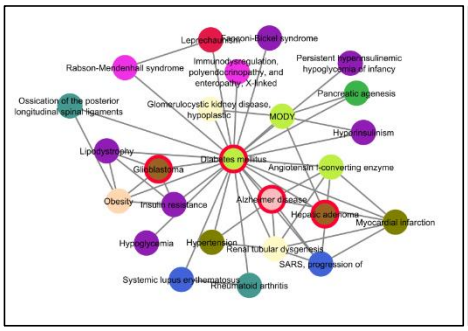

Figura A

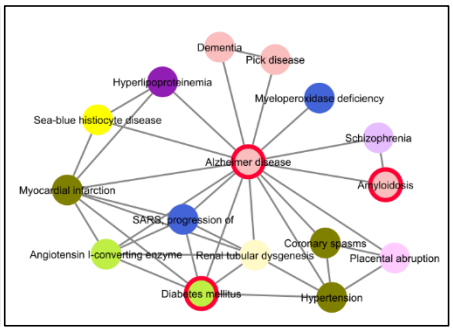

Figura B

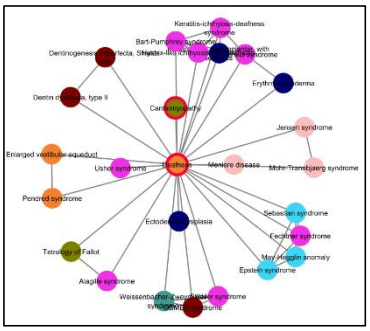

Figura C

Prendiamo per esempio la figura A, possiamo vedere come il Diabetes mellitus abbia una posizione centrale rispetto agli altri nodi, un nodo collegato è quello con il nominativo Insulino resistance e, come spiegato in nell'articolo "*Diabete. Scoperto il gene dell'insulino-resistenza?*"[2] di Maria Rita Montebelli, è un alto fattore di rischio per il diabete, stesso vale per la malattia denominata Pancreatic Agenesis come descritto nell'articolo "*Pancreatic Agenesis as Cause for Neonatal Diabetes Mellitus*"[3] e pubblicato sul portale

[2] https://www.quotidianosanita.it/scienza-e-farmaci/articolo.php?articolo_id=26801
[3] https://pubmed.ncbi.nlm.nih.gov/15770578/

National Library of Medicine. Concludiamo la serie di esempi lasciando altri due articoli per quanto riguarda la derivazione di malattie in alzheimer e in sordità: "*Coronary artery disease is associated with Alzheimer disease neuropathology in APOE4 carriers*"[4] pubblicato su National Center for Biotechnology Information e "*Deafness-dystonia-optic neuronopathy syndrome*"[5] riportato su U.S. National Library of Medicine.

Siamo arrivati alla conclusione che più il grado di un nodo è alto e più quest'ultimo rappresenta una complicanza verso uno stadio finale comune piuttosto che una base di partenza per l'evoluzione. La relazione fra due nodi HUB invece spesso indica che da ciascuna delle malattie può derivarne l'altra indistintamente.

4.2 QUALI MALATTIE PRESENTANO IL MAGGIOR DISTACCO GENETICO?

È Interessante individuare le malattie connesse che presentano il maggior distacco genetico. Abbiamo pensato che le malattie distanti geneticamente indicassero che le mutazioni dei geni potessero avvenire più velocemente e quindi che le malattie potessero evolversi in altre completamente diverse in pochi passaggi. Il ruolo di questi particolari nodi è dunque quello di tenere unita la rete e permettere il veloce spostamento al suo interno. Siamo andati così a calcolare la betweenness centrality di ciascun nodo.

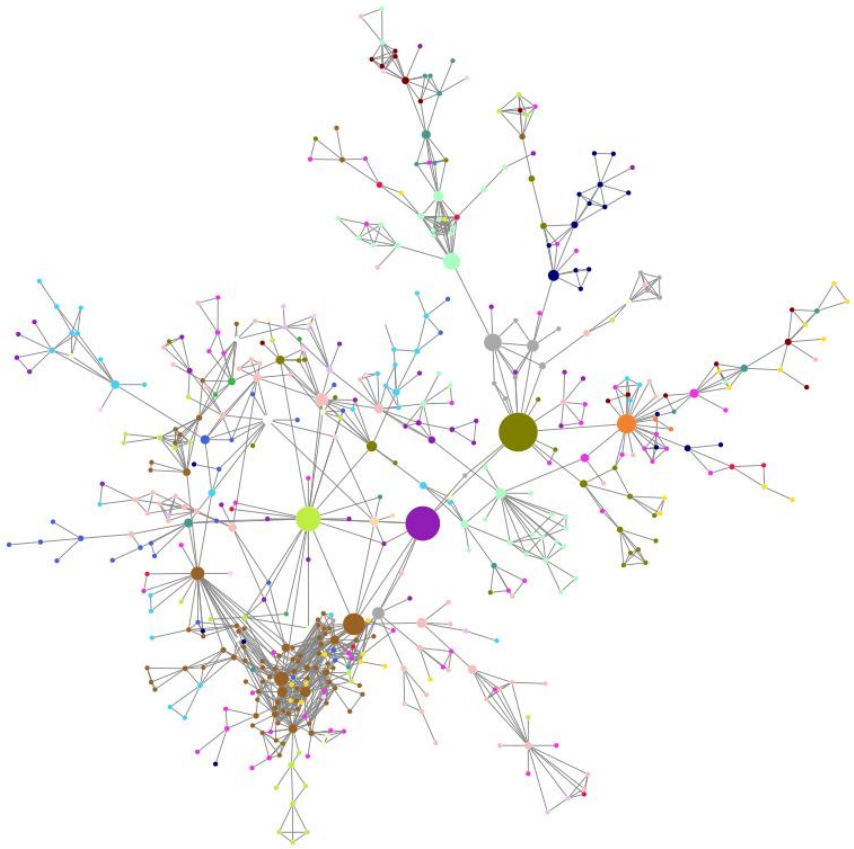

Figura 8 – proiezione della rete in cui viene evidenziata la betweenness

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163092/
[5] https://ghr.nlm.nih.gov/condition/deafness-dystonia-optic-neuronopathy-syndrome

Grazie a questa misura di centralità siamo riusciti ad individuare quei nodi che agiscono da ponte il maggior numero di volte fra le malattie e che dunque corrispondono alle malattie che mutano per il maggior numero di geni. I casi evidenziati sono:

- Cardiomyopathy (*nodo verde scuro*)
- Lipodystrophy (*nodo viola*)
- Diabetes mellitus (*nodo verde chiaro*)
- Glioblastoma (*nodo marrone*)
- Deafness (*nodo arancio*)

I nodi "Deafness" e "Diabetes Mellitus" sono nodi che avevamo precedentemente analizzato ed individuato come HUB, in quanto tali e data la struttura disassortativa della rete il valore di betwennes alto potrebbe essere falsato e non andare ad indicare un vero e proprio distacco genetico ma essere dovuto al fatto che questi nodi facciano parte dello shortest path della maggior parte dell'elevato numero di nodi collegati.

4.3 QUALI SONO LE MALATTIE PIÙ IMPORTANTI?

Abbiamo cercato di individuare le malattie più influenti e importanti all'interno del grafo, per fare ciò abbiamo calcolato l'EigenVector centrality (*figura 9*), una misura che dà importanza ad un nodo sulla base dell'importanza dei nodi vicini.

Figura 9 – proiezione della rete con focus sull' eigenvector centrality

Dall'immagine si può notare come i nodi di colore marrone (neoplasie) sono particolarmente in evidenza, la rilevanza di questi nodi è dovuta soprattutto all'alta connettività che caratterizza la tipologia. Questo concetto spiega inoltre l'elevato degree posseduto dai membri del gruppo descritto durante la prima analisi. Per evidenziare ulteriormente questa condizione abbiamo utilizzato un algoritmo di individuazione delle cliques presenti all'interno del grafo.

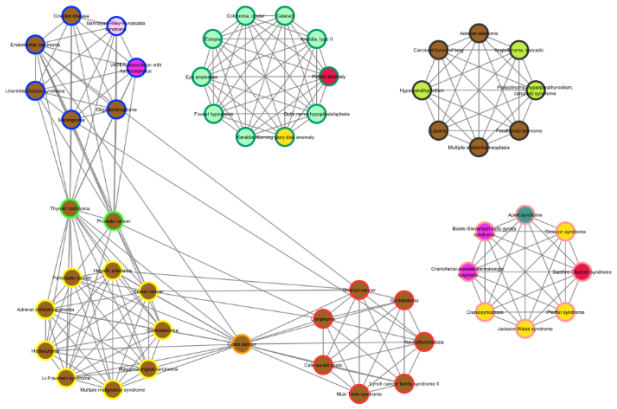

Figura 10 – cliques della rete

Le cliques ci mostrano l'elevata connettività dei nodi Cancer, 4 su 6 cliques totali sono formate dalla prevalenza, se non dalla totalità, di nodi marroni. Notiamo inoltre come gli HUB Colon cancer, Prostate cancer e Thyroid carcinoma appartengano a più di una clique contemporaneamente, spiegazione aggiuntiva dell'elevata importanza e dell'elevato grado assegnati a questi HUB.

4.4 QUALI MALATTIE SI SOMIGLIANO GENETICAMENTE? SONO DELLO STESSO TIPO?

L'immagine *figura 2*, presente nel primo paragrafo del paper dedicato alla descrizione della rete, rappresenta la distribuzione dei nodi secondo la propria tipologia. I gruppi più numerosi sono:

1. Cancer
2. Neurological
3. Multiple
4. Opthamological

Abbiamo utilizzato diversi algoritmi per l'individuazione di community, l'idea è che le community vadano a raggruppare nodi geneticamente simili che condividono parte dei propri geni. Gli algoritmi utilizzati appartengono al pacchetto ClusterMaker2 di Cytoscape, in particolare sono:

1. Community Detection GLay [6]
2. MCL [7]
3. Spectral clustering[8]

[6] https://academic.oup.com/bioinformatics/article/26/24/3135/289729
[7] https://micans.org/mcl/
[8] https://pdfs.semanticscholar.org/040d/02525b07bf807d7efa05d3556431de99282b.pdf

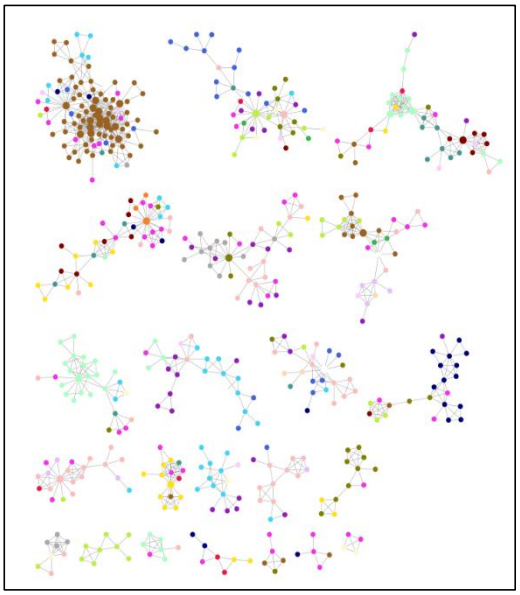

Figura 11 - Community Detection GLay

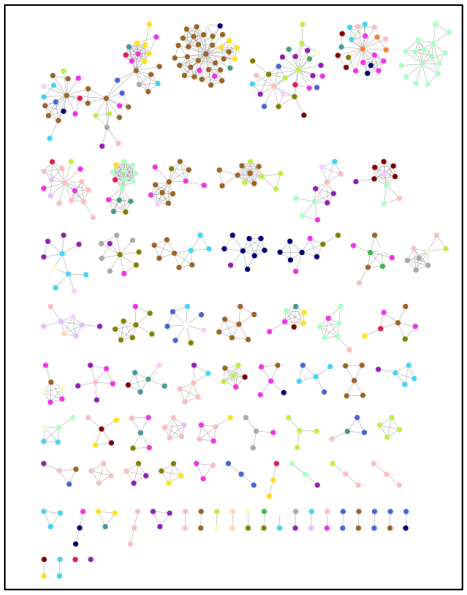

Figura 12 – MCL

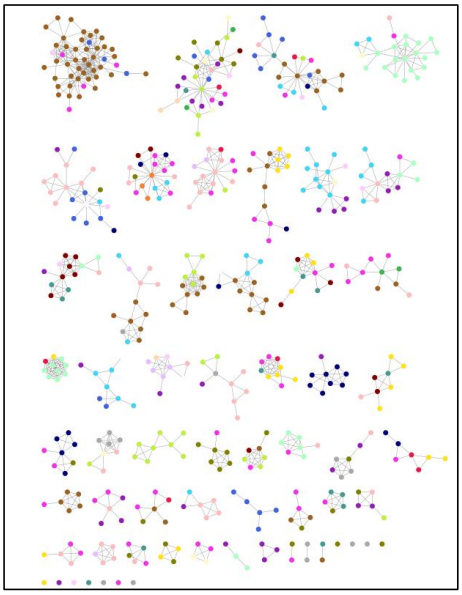

Figura 13 – Spectral clustering

Il primo algoritmo ha individuato nella rete 22 community, MCL ben 56 (non sono state considerate community i cluster formati da meno di 3 nodi) e infine lo spectral clustering 42 community. Durante l'analisi delle community è stata posta particolare attenzione alle tipologie di nodi Mutiple ("multiplo"), Unclassified ("Non classificato") e Developmental ("dello sviluppo") considerate "Jolly" in quanto comprendono malattie che potenzialmente possono colpire diversi apparati fisiologici umani e dunque essere una combinazione di tipi.

Dalle figure si vede come, per tutti i diversi algoritmi utilizzati, si genera una maggioranza di community dal carattere eterogeneo. Come già detto esse corrispondono a gruppi di malattie con geni simili fra i propri membri, il fatto che ci sia una maggioranza eterogenea è dovuto alla proprietà che uno specifico gene codifica una specifica proteina e, quest'ultima, viene utilizzata da diversi apparati del corpo umano e non solo da uno specifico.

Per quanto riguarda la presenza di community con una maggioranza di nodi omogenea rispetto al tipo, sono state evidenziate le seguenti tipologie:

- Cancer (*marrone*)
- Ophthamological (*azzurro*)
- Dermatological (*blu*)
- Cardiovascular (*verde scuro*)

La spiegazione è data dal fatto che queste particolari tipologie posseggono nel codice genetico una parte di geni "core" e quindi condivisi da tutti i nodi di quella tipologia e altri specifici che mutando vanno a generare le diverse malattie. Ulteriore conferma è data dalla precedente analisi delle cliques che, per esempio come descritto in una sezione del portale Cancer.Org[9], evidenziano la particolare connettività di alcuni gruppi tra cui le neoplasie.

[9] https://www.cancer.org/cancer/cancer-causes/genetics/genes-and-cancer/gene-changes.html

5 CONCLUSIONE

La Human Disease Network svolge un ruolo fondamentale per lo studio e l'analisi della genetica delle malattie. I nodi HUB, presenti in questa rete disassortativa, rappresentano malattie con ormai una genetica avanzata e una sorta di stato finale, diversamente dai nodi con basso grado che sono maggiormente propensi alla mutazione e all'evoluzione.

La genetica delle malattie è molto varia, esistono malattie con genetiche simili, come evidenziato dallo studio delle community e nuovamente dall'individuazione delle cliques, che però colpiscono apparati organici completamente diversi e questo dovuto al fatto che il compito dei geni è quello di codificare specifiche proteine che poi vengono utilizzate in diversi organi dal corpo umano. Abbiamo inoltre individuato come i nodi Cardiomyopathy e Lipodystrophy siano di fondamentale importanza per mantenere unita la HDN e come i nodi della tipolgia Cancer (Neoplasie) siano particolarmente uniti e di importante rilevanza per la rete.

6 EVENTUALI SVILUPPI

Per eventuali sviluppi futuri si consiglia di analizzare la rete dando un peso ad ogni arco secondo il numero di geni condivisi e di utilizzare un algoritmo per la ricerca di community che permetta l'overlapping fra le community[10].

10 https://www.researchgate.net/publication/5500337_Fuzzy_communities_and_the_concept_of_bridgeness_in_complex_networks