Data Analytics

Human Disease Network

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Informatica

Andrea **Ruggieri** *806808*
Francesco **Stranieri** *816551*

Human Disease Network – Obiettivi

A quali domande vogliamo rispondere?

La rete «**Human Disease Network**» (*2007*) rappresenta diverse malattie umane, con riferimento alle loro origini genetiche.

Una malattia genetica è un disordine causato da un cambiamento nella sequenza del DNA.

Tecniche basate su
- **Network Analysis**
- **Community Detection**

DOMANDA 1 Quali geni causano la maggior parte delle malattie? E quali tra queste sono associate al **gene "più rilevante"**?

DOMANDA 2 Come cambia la **rete dopo averla trasformata**? E cosa comporta questa trasformazione?

DOMANDA 3 Le **malattie** tendono ad **"essere legate"** tra di loro?

DOMANDA 4 Esistono **malattie "più rilevanti"** di altre?

DOMANDA 5 Quale algoritmo di **Community Detection** partiziona meglio la rete? Considerare questa come **pesata** ha effetto sulle partizioni?

DOMANDA 6 Esistono **gruppi di malattie "ben distinte"** tra di loro?

DOMANDA 7 E' possibile assegnare una categoria alle malattie classificate come **'Unclassified'**?

Human Disease Network – Obiettivi

A quali domande vogliamo rispondere?

La rete «**Human Disease Network**» (*2007*) rappresenta diverse malattie umane, con riferimento alle loro origini genetiche.

Una malattia genetica (disordine causato da cambiamento nella sequenza del DNA.

Una definizione su cosa si intende per **"più rilevante"**, **"essere legate"** e **"ben distinte"** verrà data successivamente.

basate su **ork Analysis** **nunity Detection**

DOMANDA 1 Quali geni causano la maggior parte delle malattie? E quali tra queste sono associate al **gene "più rilevante"**?

DOMANDA 2 Come cambia la **rete dopo averla trasformata**? E cosa comporta questa trasformazione?

DOMANDA 3 Le **malattie** tendono ad **"essere legate"** tra di loro?

DOMANDA 4 Esistono **malattie "più rilevanti"** di altre?

DOMANDA 5 Quale algoritmo di **Community Detection** partiziona meglio la rete? Considerare questa come **pesata** ha effetto sulle partizioni?

DOMANDA 6 Esistono **gruppi di malattie "ben distinte"** tra di loro?

DOMANDA 7 E' possibile assegnare una categoria alle malattie classificate come **'Unclassified'**?

Human Disease Network – Visualizzazione

Com’è strutturata la rete?

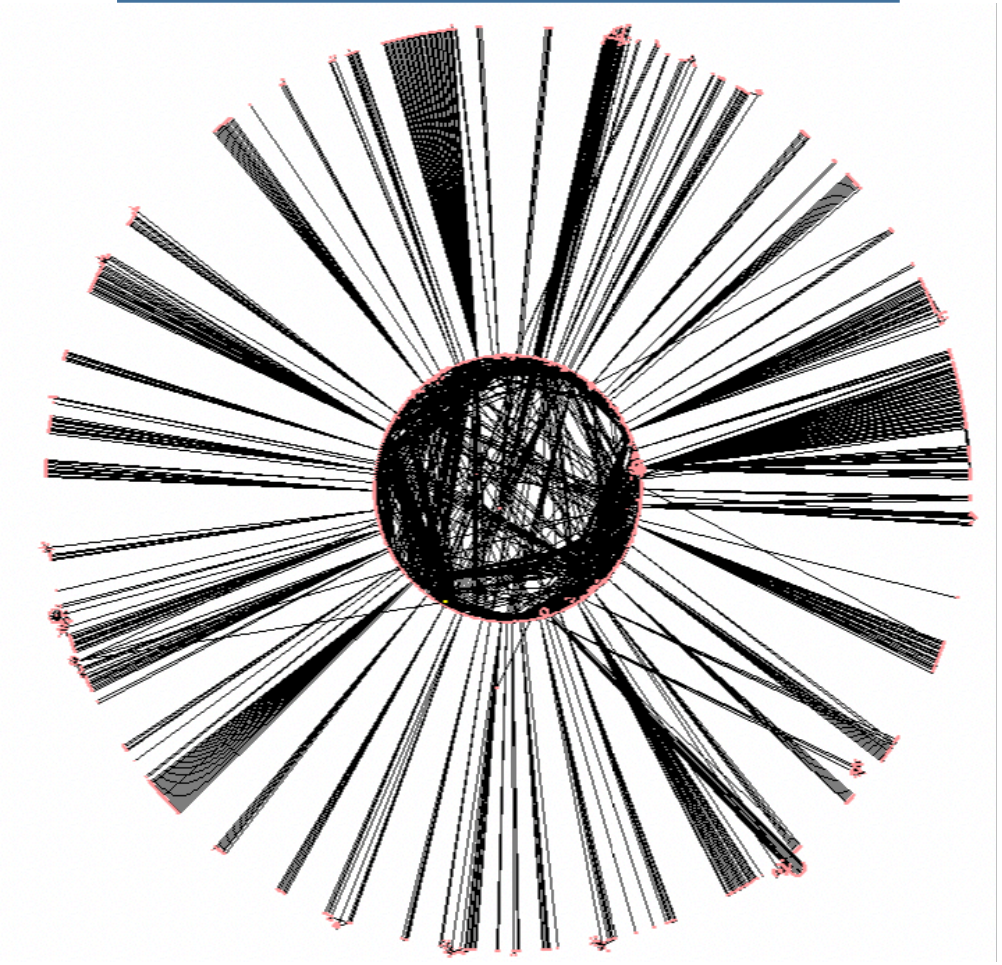
HDN come un **grafo bipartito**

- Geni
- Malattie

- Archi **orientati** e con un peso pari a *1.*
- Arco da malattia a gene indica che la malattia è provocata dalla mutazione di quel determinato gene.
- Arco tra due malattie indica che condividono un gene.

Ogni malattia ha associata una **categoria**.

iGraph e Cytoscape.



1419 nodi e *3926* archi

Human Disease Network – Network Analysis

1) Quali geni causano la maggior parte delle malattie? E quali tra queste sono associate al gene più rilevante?

Disease	Degree
TP53	**11**
PAX6	10
FGFR2	9
PTEN	9
FGFR3	8
MEN1	8
MSH2	8

Definiamo il **gene più rilevante**, come quel gene che presenta il più alto numero di link entranti (**indegree**), corrispondenti al numero di malattie associate.

TP53 è associato alle seguenti patologie*: Breast Cancer, Colon Cancer, Hepatic Adenoma, Histiocytoma, Li-Fraumeni Syndrome, Nasopharyngeal Carninoma, Osteosarcoma, Pancreatic Cancer, Thyroid Carcinoma, Adnrenal Corical Carcinoma, Multiple Malignancy Syndrome*.
Notiamo come tutte queste malattie risultano essere associate alla categoria *Cancer*.

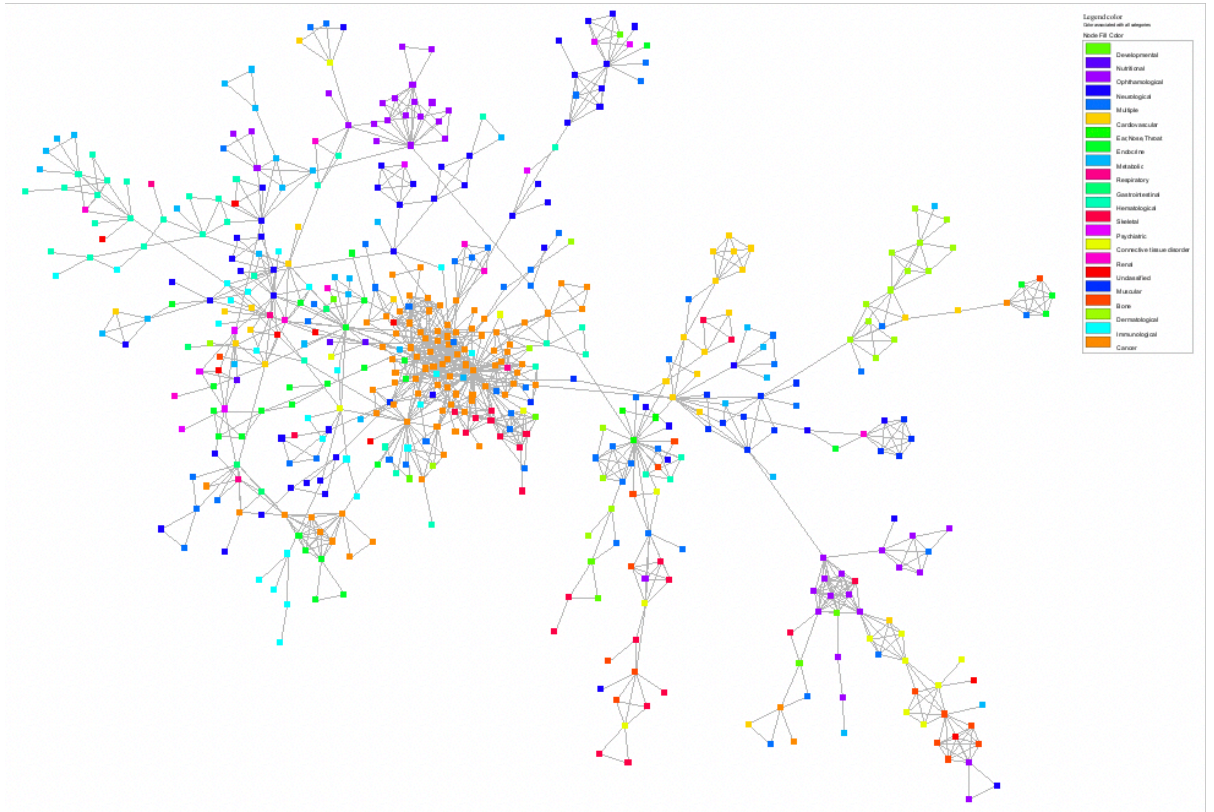
Il **TP53** è un gene che istruisce la cellula a produrre la proteina del tumore (p53). Le mutazioni ereditate o somatiche in TP53 possono provocare la perdita di controllo del ciclo cellulare. Approssimativamente, il 40% dei tumori al seno hanno mutazioni somatiche **TP53** *[Sim18]*.

Human Disease Network – Trasformazione del Grafo

2) Come cambia la rete dopo averla trasformata? E cosa comporta questa trasformazione?

- Aggiungiamo un peso per ogni arco, raffigurante il numero dei geni in comune tra due malattie.
- Rimuoviamo i nodi relativi ai geni.
- Rimuoviamo la direzione degli archi.

Clustering Coeff.	Value
Before Transformation	0.25
After Transformation	**0.43**

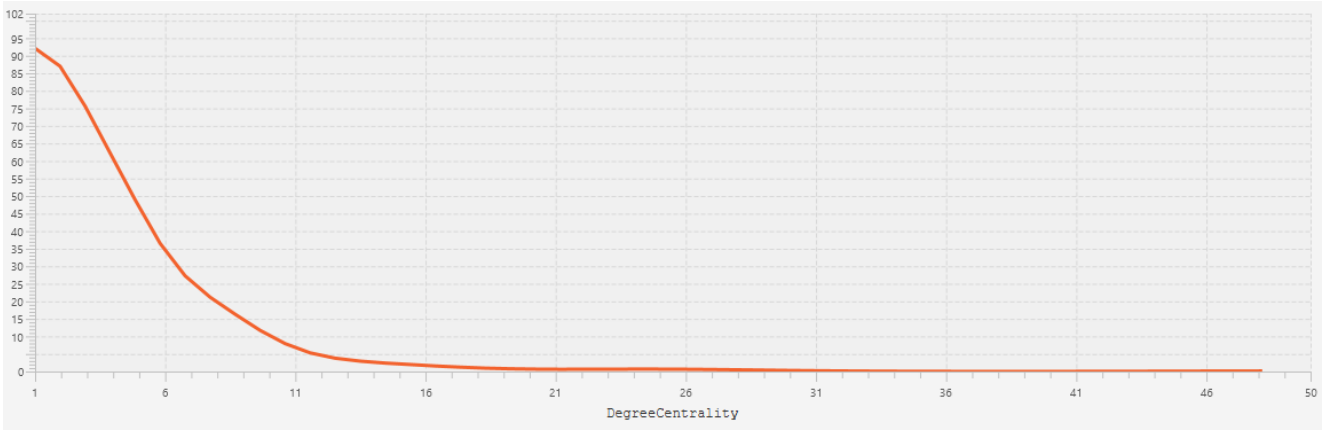


516 nodi e *2376* archi

Human Disease Network – Network Analysis

3) Le malattie tendono ad "essere legate" tra di loro?

Disease	Degree
Colon Cancer	**50**
Breast Cancer	30
Gastric Cancer	27
Thyroid Carcinoma	26
Leukemia	26



La patologia che risulta avere un degree maggiore risulta essere *Colon Cancer*, con un valori pari a *50*.
Tale risultato suggerisce che le mutazioni genetiche che causano *Colon Cancer* possono essere associate ad altre *50* diverse malattie.

La maggior parte dei nodi esibisce un grado minore di *11*. 214 nodi (41% dei nodi totali) esibiscono un grado a ≤ 5

Funzione di **HUB**

Human Disease Network – Network Analysis

4) Esistono malattie "più rilevanti" di altre?

Disease	Degree
Colon Cancer	**50**
Breast Cancer	30
Gastric Cancer	27
Thyroid Carcinoma	26
Leukemia	26

La patologia che risulta avere un degree maggiore risulta essere *Colon Cancer*, con un valori pari a *50*.
Tale risultato suggerisce che le mutazioni genetiche che causano *Colon Cancer* possono essere associate ad altre *50* diverse malattie.

Disease	Eigenvector
Colon Cancer	**1**
Breast Cancer	0.80
Ovarian Cancer	0.61
Lymphoma	0.47
Pancreatic Cancer	0.44

Le malattie con la maggiore eigenvector, risultano appartenere tutte alla categoria Cancer.

Colon Cancer risulta essere, per noi, il nodo **più rilevante** all'interno della rete.

Human Disease Network – Community Detection

5) Quale algoritmo di Community Detection partiziona meglio la rete? Considerare questa come pesata ha effetto sulle partizioni?

Il nostro studio non vuole concentrarsi nell'implementazione di svariati algoritmi presenti in letteratura, ma bensì su una più ristretta e dettagliata analisi che coinvolge i seguenti algoritmi:

- **Label Propagation**
- **Vertex Partition**
- **Infomap**

Sono stati scelti questi particolari algoritmi poiché risultano essere efficienti, facili da studiare, ampiamente utilizzati in letteratura ed appartenenti a tre diverse categorie, con caratteristiche differenti.

L'algoritmo che meglio partiziona la rete è quello che meglio massimizza le **misure di performance**, rispetto alla **ground truth**.

Precision | Recall | F1 | Rand Index

Human Disease Network – Community Detection

5) Quale algoritmo di Community Detection partiziona meglio la rete? Considerare questa come pesata ha effetto sulle partizioni?

Label Propagation

E' importante notare come *Label Propagation* non fornisce alcuna garanzia sul poter ottenere le stesse partizioni al termine di ogni esecuzione.

Per ovviare a questo problema l'algoritmo è stato eseguito per un totale di *100 esecuzioni*, riportando alla fine i **valori medi** ottenuti con la corrispettiva deviazione standard.

Label Propagation				
Description	**N° of communities**	**Rand Index**	**Precision**	**Recall**
Weighted	57 [±3] (Min 49, Max 64)	0.89 [±0.003]	0.54 [±0.01]	0.56 [±0.01]
Not Weighted	57 [±3] (Min 48, Max 64)	0.88 [±0.003]	0.53 [±0.002]	0.55 [±0.01]

Human Disease Network – Community Detection

5) Quale algoritmo di Community Detection partiziona meglio la rete? Considerare questa come pesata ha effetto sulle partizioni?

Vertex Partition

Vertex Partition				
Description	**N° of communities**	**Rand Index**	**Precision**	**Recall**
Weighted	14	0.85	0.4	0.47
Not Weighted	14	0.84	0.41	0.47

Human Disease Network – Community Detection

5) Quale algoritmo di Community Detection partiziona meglio la rete? Considerare questa come pesata ha effetto sulle partizioni?

Infomap

Infomap				
Description	**N° of communities**	**Rand Index**	**Precision**	**Recall**
Weighted	56	0.89	0.55	0.57
Not Weighted	55	0.89	0.55	0.57

Human Disease Network – Community Detection

5) Quale algoritmo di Community Detection partiziona meglio la rete? Considerare questa come pesata ha effetto sulle partizioni?

- **Label Propagation e Infomap risultato essere gli algoritmi maggiormente affidabili.**
- **Non esistono differenze significative, in termini prestazionali, tra le versioni pesate e non pesate.**

Label Propagation		Vertex Partition		Infomap	
F1	Partition Mod.	F1	Partition Mod.	F1	Partition Mod.
0.53 [±0.01]	0.76 [±0.008]	0.41	0.83	0.55	0.78

Osservando i valori della **Partition Modularity**, l'algoritmo *Vertex Partition* risulta offrire le migliori prestazioni.
Tale risultato risulta essere comunque poco attendibile, dal momento che *Vertex Partition* identifica solamente *14* comunità, rispetto a *Label Propagation* e *Infomap* che ne identificano più di *50*, e risente quindi meno del **limite di risoluzione**.

Human Disease Network – Ulteriori Analisi

6) Esistono gruppi di malattie "ben distinte" tra di loro?

Famiglia degli algoritmi di Community Detection conosciuti come *Node-Centric*.

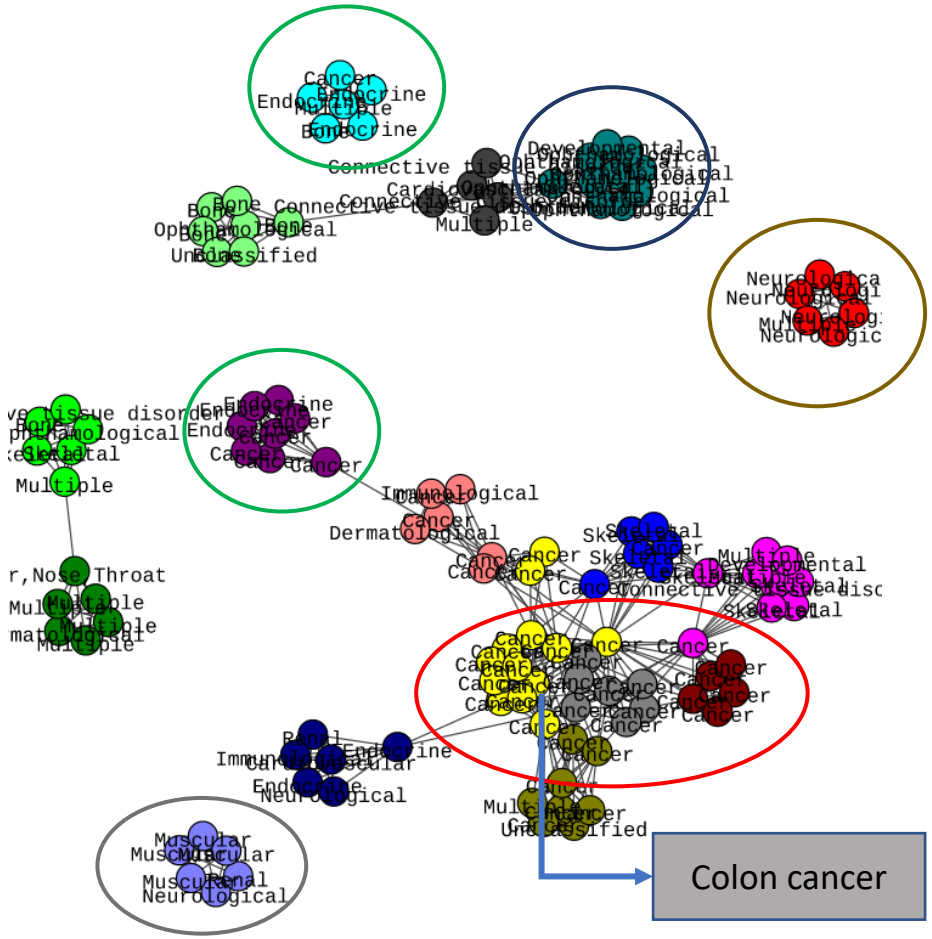
Clique di un grafo.

Questi algoritmi risultano essere molto onerosi dal punto di vista computazionale, pertanto sono generalmente usati su reti di piccole dimensioni, generalmente con al più *50* nodi.

Abbiamo quindi deciso di ricorrere ad un approccio *greedy*, tramite l'**algoritmo di Pruning.**

Human Disease Network – Ulteriori Analisi

6) Esistono gruppi di malattie "ben distinte" tra di loro?



Nella clique di dimensione maggiore, la maggior parte delle malattie risultano essere associate alla categoria *Cancer*.

Le malattie appartenenti alle categorie Cancer, Endocrine, Neurological, Muscolar e Ophthamological risultano essere **ben distinte** tra loro.

Pruning		
Description	**Precision**	**Recall**
Not Weighted	0.65	0.7

Human Disease Network – Ulteriori Analisi

7) E' possibile assegnare una categoria alle malattie classificate come '*Unclassified*'?

Disease	Categories			Description	Most Possible Classes
	Label Propagation	Infomap	Vertex Partition		
Alcohol Dependence	Psychiatric	Psychiatric	Psychiatric	Ricerca compulsiva di bevande alcoliche. Il gene **ADH1B**, viene ritenuto responsabile della velocità con cui vengono assorbiti i liquori nell'organismo [Min17].	Psychiatric
Carpal tunnel syndrome; familial	Metabolic	Metabolic	Metabolic	Compressione del nervo mediano che causa intorpidimento, formicolio della mano [BU07].	Neurological
Aquaporin-1 Deficiency	Hematological	Hematological	Hematological	Associata ad una malattia del sangue [19a].	Hematological, Immunological
Bannayan-Riley-Ruvalcaba Syndrome	Cancer	Cancer	Cancer	E' una malattia congenita rara con poliposi intestinale, lipomi, macrocefalia e lentigginosi genitale. Associata alla mutazione del gene **PTEN** [Gon+13].	Cancer
Benzene Toxicity	Cancer	Cancer	Cancer	Si tratta di una patologia che provoca il tumore al seno e la leucemia [19b].	Cancer
Van Buchem Disease	Bone	Bone	Bone	E' un'iperostosi cranio-tubolare rara, caratterizzata da iperostosi del cranio, della mandibola, della clavicole, delle costole, delle diafisi delle ossa lunghe e delle ossa tubolari delle mani e dei piedi [17].	Bone, Skeletral

Disease	Categories			Description	Most Possible Classes
	Label Propagation	Infomap	Vertex Partition		
Placental Abruption	Cardiovascular	Cardiovascular	Bone	Complicazione rara ma grave che può manifestarsi durante la gravidanza e causa emorragie [And17].	Cardiovascular, Endocrine, Hematological
Beta-2-adrenoreceptor Agonist	Immunological	Neurological	Immunological	Disturbo che coinvolge le cellule e in particolare la produzione di amminoacidi [07].	Neurological, Metabolic
Aneurysm, Familial Arterial	Connective tissue disorder	Bone	Bone	E' una malattia ereditaria rara che colpisce l'aorta e il tessuto connettivo [19c].	Cardiovascular, Connective tissue disorder

Human Disease Network – Conclusioni

A quali domande abbiamo risposto?

La rete «*Human Disease Network*» (HDN) rappresenta diverse malattie umane, con riferimento alle loro origini genetiche.

Una malattia genetica è un disordine causato da un cambiamento nella sequenza del DNA.

Tecniche basate su
- **Network Analysis**
- **Community Detection**

DOMANDA 1 Un primo studio di Network Analysis sulla rete originale ha permesso di individuare il **gene TP53 come il più rilevante**.

DOMANDA 2 La rimozione delle informazioni superflue ha permesso di raddoppiare il valore del **coefficiente di clustering**.

DOMANDA 3 Una nuova attività di Network Analysis ha permesso di verificare come **poche malattie tendono ad essere legate** tra di loro.

DOMANDA 4 Sulla base del valore dell'Eigenvector Centrality, **Colon Cancer risulta essere la malattia più rilevante** .

DOMANDA 5 **Label Propagation ed Infomap risultano partizionare meglio la rete. La pesatura del grafo non sembra influire.**

DOMANDA 6 Attraverso la tecnica del pruning è stato possibile verificare che **esistono gruppi di malattie che ben si distinguono dalle altre**.

DOMANDA 7 **E' stato invece possibile assegnare una categoria a tutte quelle malattie che ne erano inizialmente sprovviste.**

Human Disease Network – Referenze

- *[Sim18]* Hannah Simmons. Che cosa è TP53? Aug 23, 2018. url: https : / / www . news - medical.net/life-sciences/What-is-TP53-(Italian).aspx.
- *[BU07]* K. Balci e U. Utku. «Carpal tunnel syndrome and metabolic syndrome». In:ActaNeurologica Scandinavica116.2 (2007), pp. 113–117.doi:10.1111/j.1600-0404.2007.00797.x.
- *[17]* Hsu, Shang-Fu, and Chen-Chun Lin. "Van Buchem disease: First case report in Tai-wan." Medicine vol.2017.doi:"doi:10.1097/MD.0000000000009209".
- *[07]* I.P. Hall, in Encyclopedia of Respiratory Medicine, 2006. 2007. url:https : / /www.sciencedirect.com/topics/neuroscience/beta-2-adrenergic-receptor.
- *[Gon+13]* Gabriela Maria Abreu Gontijo et al. «Bannayan-Riley-Ruvalcaba syndrome with defor-ming lipomatous hamartomas in infant - Case report». en. In:Anais Brasileiros de Der-matologia88 (dic. 2013), pp. 982–985.issn: 0365-0596.url:http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0365-05962013000600982&nrm=iso
- *[And17]* Ananth CV. Hansen AV Williams MA. Nybo Andersen. «Cardiovascular Disease inRelation to Placental Abruption: A Population-Based Cohort Study from Denmark».In:Paediatr Perinat Epidemiol. Vol. 31(3):209-218. 2017.doi:10.1111/ppe.12347.
- *[Min17]* State of Mind.Alcol. 2017.url:https://www.stateofmind.it/tag/alcool/#:~:text=La%20dipendenza%20alcolica%2C%20o%20alcolismo,sempre%20maggiori%20di%20bevande%20alcoliche).
- *[19a]* OMIM. 2019.url:https://omim.org/entry/107776.
- *[19b]* OMIM. 2019.url:https://omim.org/entry/125860.
- *[19c]* OMIM. 2019.url:https://omim.org/entry/607086.