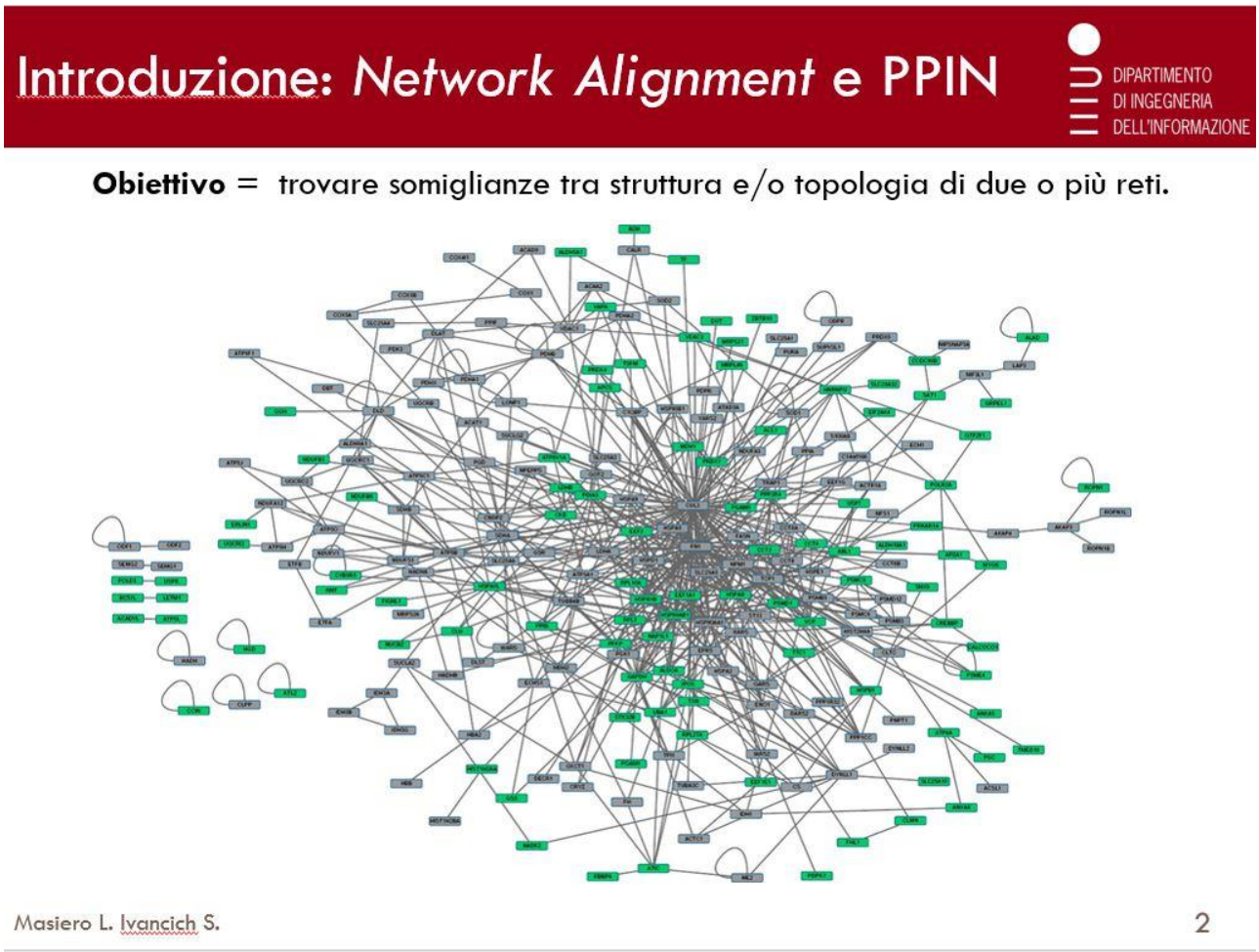


Buongiorno e benvenuti! Siamo Luca Masiero e Stefano Ivancich ed oggi vi parleremo del Network Alignment. Cominceremo presentandovi l'argomento in generale, così che abbiate familiarità con la terminologia utilizzata, per passare poi alla descrizione di quattro degli algoritmi più famosi che abbiamo studiato. Ok, cominciamo…

L'obiettivo del ***Network Alignment*** (che possiamo tradurre con *allineamento delle reti*) consiste nel trovare somiglianze tra la struttura e la topologia di due o più reti, rappresentate tramite grafi. Confrontare le reti di diversi organismi è, attualmente, uno dei problemi più importanti della Biologia.

LUCA 1

Introduzione: *Network Alignment* e PPIN

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Gli *allineamenti* di reti biologiche sono utili in molti contesti: avendo molte informazioni su alcuni nodi di una determinata rete G_1 e quasi nulla su nodi topologicamente simili in un'altra G_2, la conoscenza specialistica di uno di quei nodi può dirci qualcosa di nuovo sul corrispettivo.

Le *Protein-Protein Interaction Networks* (PPIN) sono strumenti validi per comprendere:
- funzioni delle cellule;
- malattie umane;
- design e riposizionamento dei farmaci;
- interattomi (insieme delle interazioni molecolari in una cellula).

Date le grandi dimensioni (migliaia di elementi), le reti PPI sono analizzate tramite l'identificazione di *sottoreti*, o **moduli**.

Modulo topologico = gruppo di nodi che hanno molte più connessioni con i nodi del gruppo piuttosto che con quelli esterni.
Modulo funzionale = gruppo di nodi che condividono una funzione biologica.

Masiero L. Ivancich S.

LUCA 2

Gli allineamenti di reti biologiche possono infatti risultare molto utili perché, avendo molte informazioni riguardo alcuni nodi di una determinata rete G_1 e quasi nulla su nodi *topologicamente simili* in un'altra rete G_2, ebbene, la conoscenza specialistica di uno dei nodi di G_1 può dirci *qualcosa di nuovo* sul corrispettivo in G_2. Gli allineamenti delle reti possono anche essere utilizzati per misurare la somiglianza globale tra reti complete di specie diverse.

Le ***Protein-Protein Interaction Networks*** (cioè le *reti di interazione proteina-proteina*) sono strumenti validi per comprendere le funzioni delle cellule, le malattie umane e il design di nuovi farmaci.

Negli anni sono stati proposti diversi algoritmi per l'interpretazione automatica delle PPI, in un primo momento considerando esclusivamente la *topologia* della rete, e successivamente integrando i *termini dell'Ontologia Genica* come attributi di somiglianza dei nodi (l'Ontologia Genica, indicata con l'acronimo *GO* è un progetto bioinformatico che ha l'obiettivo di unificare la descrizione delle caratteristiche genetiche di tutte le specie).

Nelle reti PPI, un sistema biologico è descritto in termini di *proteine*, che costituiscono i *nodi* del grafo, e le loro relazioni (cioè interazioni fisiche o funzionali) sono rappresentate dagli *archi* del grafo.

Date le grandi dimensioni (in genere vengono coinvolte migliaia di elementi), le reti PPI sono analizzate tramite l'identificazione di sottoreti, o **moduli**, che mostrano specifiche caratteristiche topologiche o funzionali. L'espressione *modulo topologico* si riferisce ad un gruppo di nodi che hanno molte più connessioni con i nodi del gruppo piuttosto che con quelli esterni. L'espressione *modulo funzionale* si riferisce invece ad un gruppo di nodi che condividono una precisa funzione biologica.

Introduzione: *Network Alignment* e PPIN

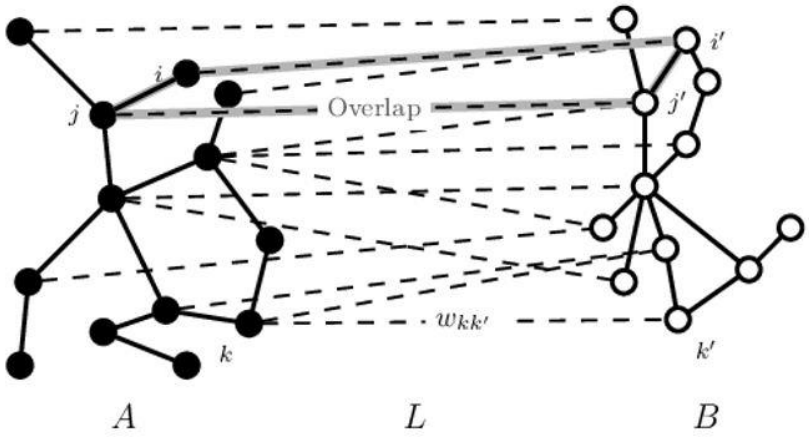
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Date due reti, **allinearle** significa trovare un *mapping* nodo-a-nodo (= *alignment*) tra le stesse in grado di ottimizzare due obiettivi:

(1) massimizzare il numero di proteine mappate (*nodi* del grafo) che sono correlate da un punto di vista funzionale;

(2) massimizzare il numero di interazioni comuni (*archi* del grafo) tra le reti.

Problema intrattabile dovuto all'NP-completezza sottostante al *sub-graph isomorphism problem* (Cook,1971).



Masiero L. Ivancich S. 4

Date due reti, *allinearle* significa trovare un **mapping nodo-a-nodo** tra le reti in grado di ottimizzare due obiettivi: il primo corrisponde a massimizzare il numero di proteine mappate (cioè i nodi nel grafo) che sono correlate da un punto di vista funzionale, mentre il secondo corrisponde a massimizzare il numero di interazioni comuni (gli archi) tra le reti.

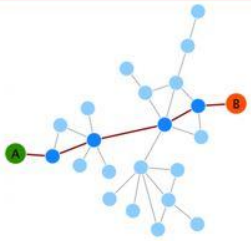
Il **Network Alignment** è un problema intrattabile dovuto all' NP-completezza del *sub-graph isomorphism problem*, individuato da Spephen Cook nel 1971. Si tratta di un problema computazionale nel quale, dati due grafi G e H in input, si vuole determinare se G contiene un *sottografo isomorfico* ad H, ci si interroga cioè sull'esistenza di una *corrispondenza biunivoca* tra gli elementi dei grafi).

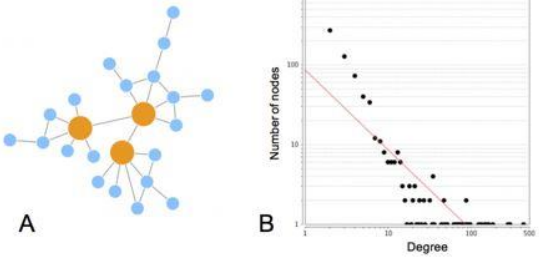
STEFANO 3

PPIN: proprietà fondamentali

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Effetto del piccolo mondo: tutte le reti complesse presenti in natura sono tali che due nodi qualsiasi possono essere collegati da un percorso costituito da un numero relativamente piccolo di collegamenti.

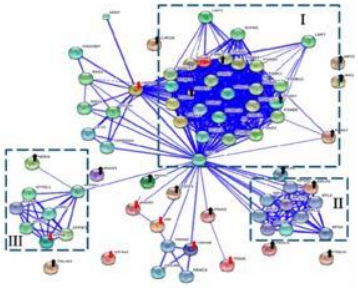




Scale-free networks: nodi con poche connessioni vs *hub*

PAROLE CHIAVE scalabilità, invarianza ai cambiamenti di scala, vulnerabilità agli attacchi mirati.

Transitività: misura la tendenza dei nodi a raggrupparsi. Utile per individuare *complessi proteici* (**moduli**).



Masiero L. Ivancich S. 5

Le PPIN sono soggette all'***effetto del piccolo mondo***, questo vuol dire che intercorre una grande connettività tra le proteine. Questo livello di connettività ha importanti conseguenze biologiche perché consente un flusso efficiente e rapido dei segnali all'interno della rete.

I sistemi biologici sono estremamente robusti e possono far fronte a una quantità relativamente elevata di perturbazioni in singole proteine. Per spiegare come ciò possa accadere, dobbiamo considerare un'altra proprietà fondamentale delle PPIN.

Le PPIN sono ***scale-free networks***. La maggior parte dei nodi nelle *scale-free networks* hanno solo *poche* connessioni con altri nodi, mentre altri (denominati *hub*) sono collegati a *molti altri nodi* della rete stessa.

La natura *scale-free* delle PPIN conferisce loro una serie di caratteristiche; in primis la stabilità: se i guasti si verificano in modo casuale e la maggior parte delle proteine costituisce un grado di connettività basso, la probabilità che un *hub* venga colpito è minima; se si verifica un *hub-failure*, la rete, in genere, non perde la sua connettività grazie agli *hub* che rimangono.

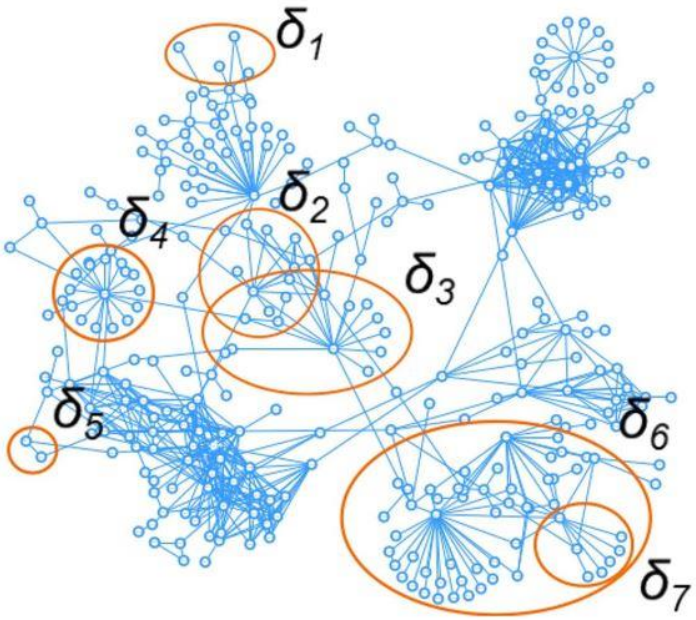
Un'altra caratteristica cruciale delle PPIN è la loro modularità. La *transitività* di una rete misura la tendenza dei nodi a raggrupparsi. Un'alta transitività significa che la rete contiene "comunità" (fare con le dita le "virgolette") o gruppi di nodi che sono densamente connessi.

Trovare queste comunità è molto importante perché può aiutare ad individuare **complessi proteici**. I complessi proteici possono essere considerati un tipo di modulo (quindi un'unità funzionale ed intercambiabile) in cui le proteine interagiscono in modo stabile, mantenendo una configurazione più o meno costante sia nel tempo che nello spazio. Lo studio dei moduli è utile anche per definire le **interazioni intermodulari** tra le proteine.

STEFANO 4

MTGO - *Module detection via Topological information and GO knowledge* -

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE



Metodo GO-*based* per identificare moduli funzionali.

Combinazione di informazioni provenienti dalla topologia delle reti con conoscenza biologica.

Overlapping e copertura totale della rete.

Masiero L. Ivancich S. 6

Il metodo MTGO costituisce un nuovo approccio di identificazione dei moduli funzionali nelle PPIN. Questo metodo combina le informazioni provenienti dalla topologia delle reti con la conoscenza biologica relativa alle proteine. MTGO utilizza partizioni ripetute della rete sfruttando la modularità del grafo. A differenza dei precedenti metodi allo stato dell'arte basati su GO, MTGO fornisce un unico termine GO che descrive al meglio la natura biologica di ogni modulo identificato.

MTGO è in grado di individuare moduli funzionali all'interno delle PPIN, prevede l'*overlapping* e la *copertura totale della rete*, due *features* davvero molto importanti per gli algoritmi di identificazione di moduli.

LUCA 5

MTGO: struttura

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

3 FASI

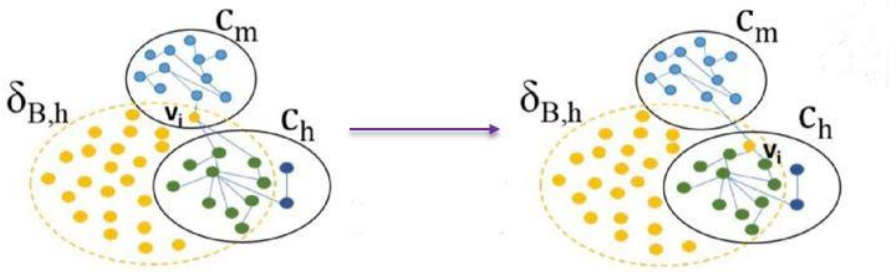
1) **Inizializzazione**: creazione delle partizioni e dei moduli

$$C = \{c_1, \ldots, c_h, \ldots, c_H\} \qquad \Phi = \{\phi_1, \ldots, \phi_h, \ldots, \phi_H\}$$

$$c_1 \cap c_2 \ldots \cap c_h \ldots \cap c_H \equiv \emptyset \qquad \phi_1 \cap \phi_2 \ldots \cap \phi_h \ldots \cap \phi_H \equiv \emptyset$$

$$c_1 \cup c_2 \ldots \cup c_h \ldots \cup c_H \equiv V \qquad \phi_1 \cup \phi_2 \ldots \cup \phi_h \ldots \cup \phi_H \subseteq V$$

2) **Iterazioni** in cui i nodi vengono ri-assegnati alle partizioni



Masiero L. Ivancich S. 7

L'obiettivo di MTGO è quello di processare il grafo in input per trovare gruppi di nodi che condividono sia proprietà topologiche, sia funzionali. L'insieme dei moduli topologici C (indica C) costituisce una partizione della rete (indicare formula) in modo che l'intersezione delle H componenti sia vuota e che la loro unione fornisca l'intero insieme dei nodi; l'insieme Φ invece descrive i moduli funzionali.

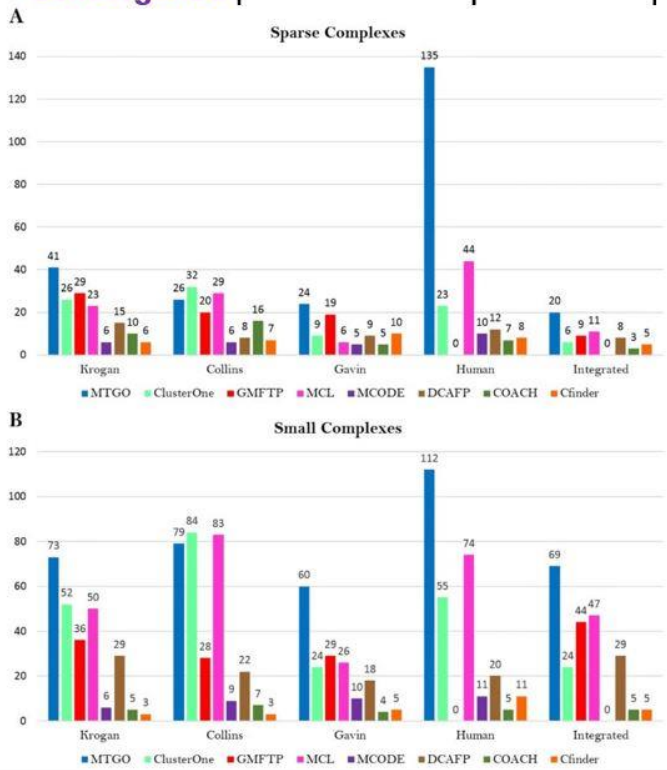
Il processo di assegnamento di un nodo ad una determinata partizione non è immediato e prevede riassegnamenti fino al raggiungimento della convergenza; sottolineiamo che MTGO tende ad assegnare i nodi di modo che i moduli topologici coincidano con quelli funzionali.

LUCA 6

MTGO: struttura

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

3) **Convergenza** per valutare la qualità della partizione finale e dell'*overlapping*



MTGO possiede l'abilità di individuare un insieme di termini GO fornendo un'interpretazione biologica significativa della PPIN, proprietà assente negli altri algoritmi allo stato dell'arte.

NUMERO DI CITAZIONI = 13
(Google Scholar)

8

Per valutare se la convergenza è stata raggiunta o meno si utilizzano due funzioni che valutano, rispettivamente, la qualità globale della partizione in output e il conseguente overlapping.

MTGO rimane, a quasi 3 anni dallo sviluppo, praticamente sconosciuto con solamente 13 citazioni registrate in *Google Scholar*. Dato comunque l'esiguo numero di citazioni, questo metodo costituisce un buon metro di paragone utilizzato da altri ricercatori nello sviluppo di efficienti algoritmi per l'identificazione di moduli funzionali nelle PPIN.

LUCA 7

IsoRank

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Metodo per l'allineamento di più PPIN.

Intuizione = una proteina rappresenta una buona corrispondenza con una proteina in un'altra sequenza se le rispettive sequenze e i loro intorni topologici costituiscono, a loro volta, una buona corrispondenza.

Rappresenta un approccio di analisi comparativa al GNA.

Consideriamo il caso di GNA a coppie.
Input: due PPIN G_1 e G_2 (ogni arco e può aver associato un peso $w(e)$, con $0 \leq w(e) \leq 1$) e *similarity measure* tra i nodi delle due reti.

IsoRank è un metodo per l'allineamento globale di più PPIN. L'intuizione da tener presente è che una proteina in una rete PPI produce una buona corrispondenza (un *match*) con una proteina in un'altra rete se le loro rispettive sequenze e i loro intorni topologici costituiscono una buona corrispondenza.

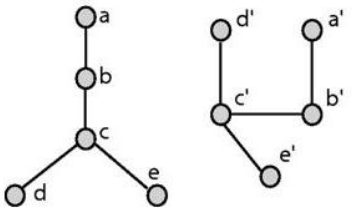
IsoRank rappresenta un approccio di analisi comparativa tra PPIN con l'obiettivo di trovare una soluzione al problema di allineamento ottimo *globale* tra due o più PPIN, mirando ad identificare la corrispondenza tra i nodi e gli archi delle reti in input che massimizzi il match totale tra le reti.

Consideriamo un semplice caso di Global Network Alignment a coppie. L'input consiste in due PPIN G_1 e G_2 (ogni arco *e* può aver associato un peso *w(e)*, con $0 \leq w(e) \leq 1$) e di una *similarity measure* tra i nodi delle due reti (per esempio BLAST). L'output desiderato è un mapping tra i nodi delle due reti che massimizza la combinazione convessa delle seguenti funzioni obiettivo: (1) la dimensione del grafo in comune in seguito al mapping e (2) la somiglianza tra le sequenze dei nodi mappati gli uni negli altri.

STEFANO 8

IsoRank

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE



R

	a'	b'	c'	d'	e'
a	0.0312		0.0937		
b		0.1250		0.0625	0.0625
c	0.0937		0.2812		
d		0.0625		0.0312	0.0312
e		0.0625		0.0312	0.0312

$R_{aa'} = \frac{1}{4} R_{bb'}$

$R_{bb'} = \frac{1}{3} R_{ac'} + \frac{1}{3} R_{a'c} + R_{aa'} + \frac{1}{9} R_{cc'}$

$R_{dd'} = \frac{1}{9} R_{cc'}$

$R_{cc'} = \frac{1}{4} R_{bb'} + \frac{1}{2} R_{be'} + \frac{1}{2} R_{bd'} + \frac{1}{2} R_{eb'} + \frac{1}{2} R_{db'} + R_{ee'} + R_{ed'} + R_{de'} + R_{dd'}$

Obiettivo: trovare sottografo comune alle reti in input.

2 FASI:

1) Assegnazione di *functional similarity scores*;

2) *Mapping* per il GNA, considerando solo scores elevati, mantenendo la proprietà transitiva (**one-to-one** oppure **many-to-many**)

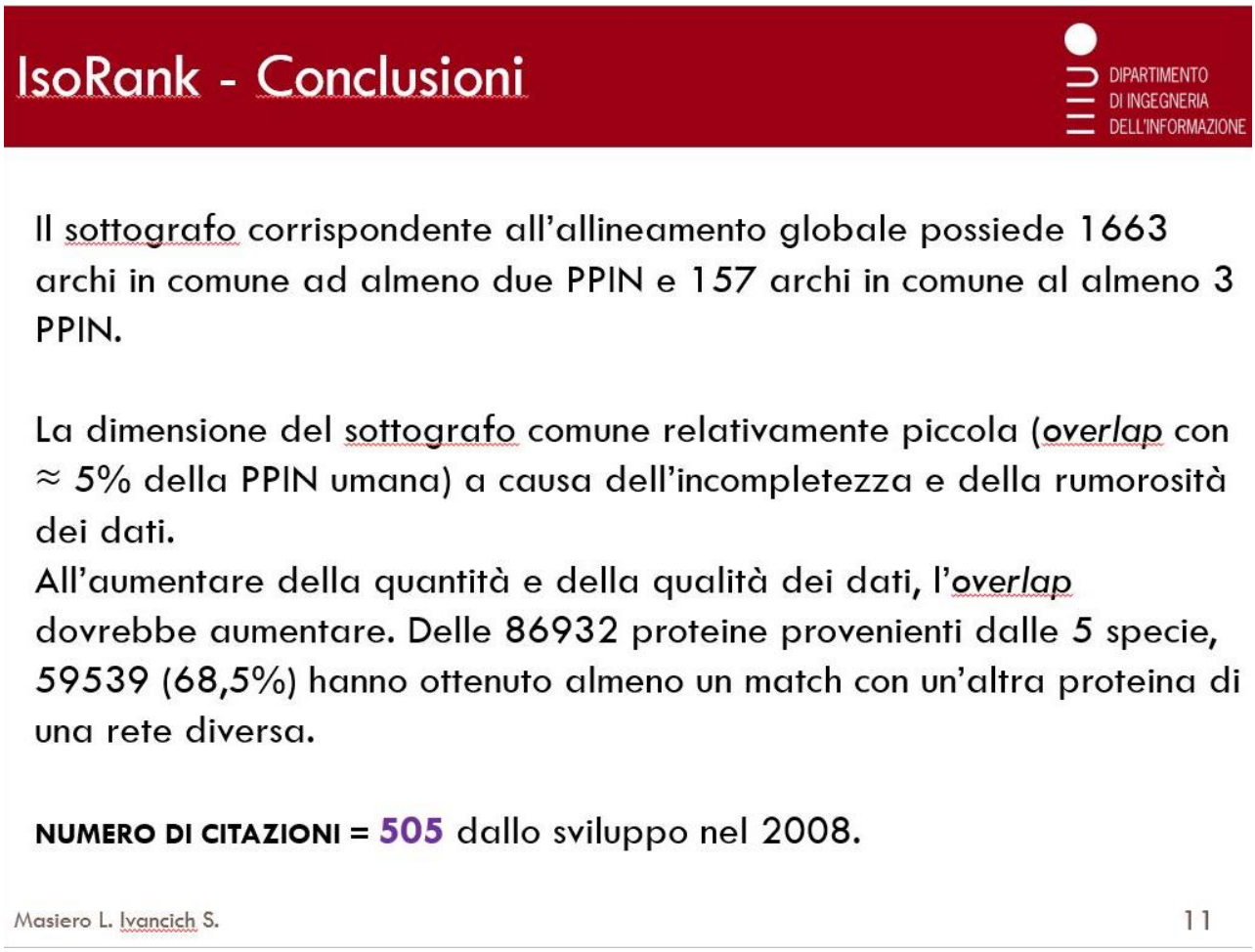
$$\boldsymbol{R} = \sum R_{ij} = \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{1}{|N(i)||N(j)|} R_{ij} \text{ con } i \in V_1,\ j \in V_2$$

Masiero L. Ivancich S. 10

L'algoritmo prevede **due fasi**. Nella prima fase associa un *functional similarity score* ad ogni possibile match tra i nodi delle due reti. Sia R_{ij} lo score per la coppia di proteine *(i,j)* dove i proviene dalla rete G_1, mentre j da G_2. La seconda fase costruisce la mappatura per il Global Network Alignment estraendo un insieme di score elevati, in accordo con **R**, il vettore di tutti gli R_{ij} (indica la formula di **R**). Per calcolare il *functional similarity score Rij* consideriamo la coppia (i,j) un "buon match" se le sequenze di i e di j sono allineate e i loro "vicini" costituiscono a loro volta un buon match gli uni con gli altri. I nodi che hanno una buona corrispondenza avranno *score* R_{ij} più alti.

A questo punto bisogna assicurarsi che il mapping mantenga la proprietà di transitività. Il *mapping* si può ottenere in due modi: one-to-one (ogni nodo viene mappato in al massimo un altro nodo, per specie), oppure many-to-many (un nodo può essere mappato in più di un nodo in un'altra specie).

STEFANO 9



Concludiamo proponendo l'analisi del sottografo comune ottenuto dall'allineamento di cinque specie (tra cui la specie umana e il *mus musculus* – il topo comune -). Il sottografo corrispondente all'allineamento globale possiede 1663 archi in comune ad almeno due PPIN e 157 archi in comune al almeno 3 PPIN. La dimensione del sottografo comune è relativamente piccola (abbiamo un *overlap* solamente con circa il 5% della PPIN umana) a causa delle probabili incompletezza e rumorosità dei dati. All'aumentare della quantità e della qualità dei dati, l'*overlap* dovrebbe aumentare sensibilmente. Delle quasi 90.000 proteine provenienti dalle 5 specie, quasi il 70% ha ottenuto almeno un match con un'altra proteina di una rete diversa.

IsoRank è stato citato più di 500 volte dallo sviluppo nel 2008; proposto in moltissime varianti, costituisce un "baluardo" per il GNA.

STEFANO 10

struc2vec

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

La *structural identity* corrisponde ad un concetto di *simmetria* nel quale i nodi di una rete vengono identificati in base alla struttura della rete stessa e tramite relazioni con altri nodi.

struc2vec:
1) è un framework flessibile per l'apprendimento di *latent representations* per l'identità strutturale dei nodi;
2) Utilizza un grafo multi-livello;
3) È molto performante.

Masiero L. Ivancich S. 12

struc2vec è un framework flessibile per l'apprendimento di *latent representations* (= tutte le informazioni importanti necessarie per rappresentare i dati originali) per l'identità strutturale dei nodi. *struc2vec* utilizza una gerarchia, definita dalla sequenza ordinata dei gradi dei nodi, per misurare la *similarity* dei nodi stessi e costruisce un grafo multi-livello per codificare le somiglianze strutturali.

Sviluppato nel 2017, questo metodo presenta prestazioni elevate e supera i limiti raggiunti dagli approcci precedenti. Gli esperimenti indicano che struc2vec migliora le prestazioni su attività di classificazione che dipendono principalmente dall'identità strutturale; struc2vec eccelle anche quando la rete originale è soggetta a forti rumori casuali (per esempio la rimozione casuale di archi dal grafo).

STEFANO 13

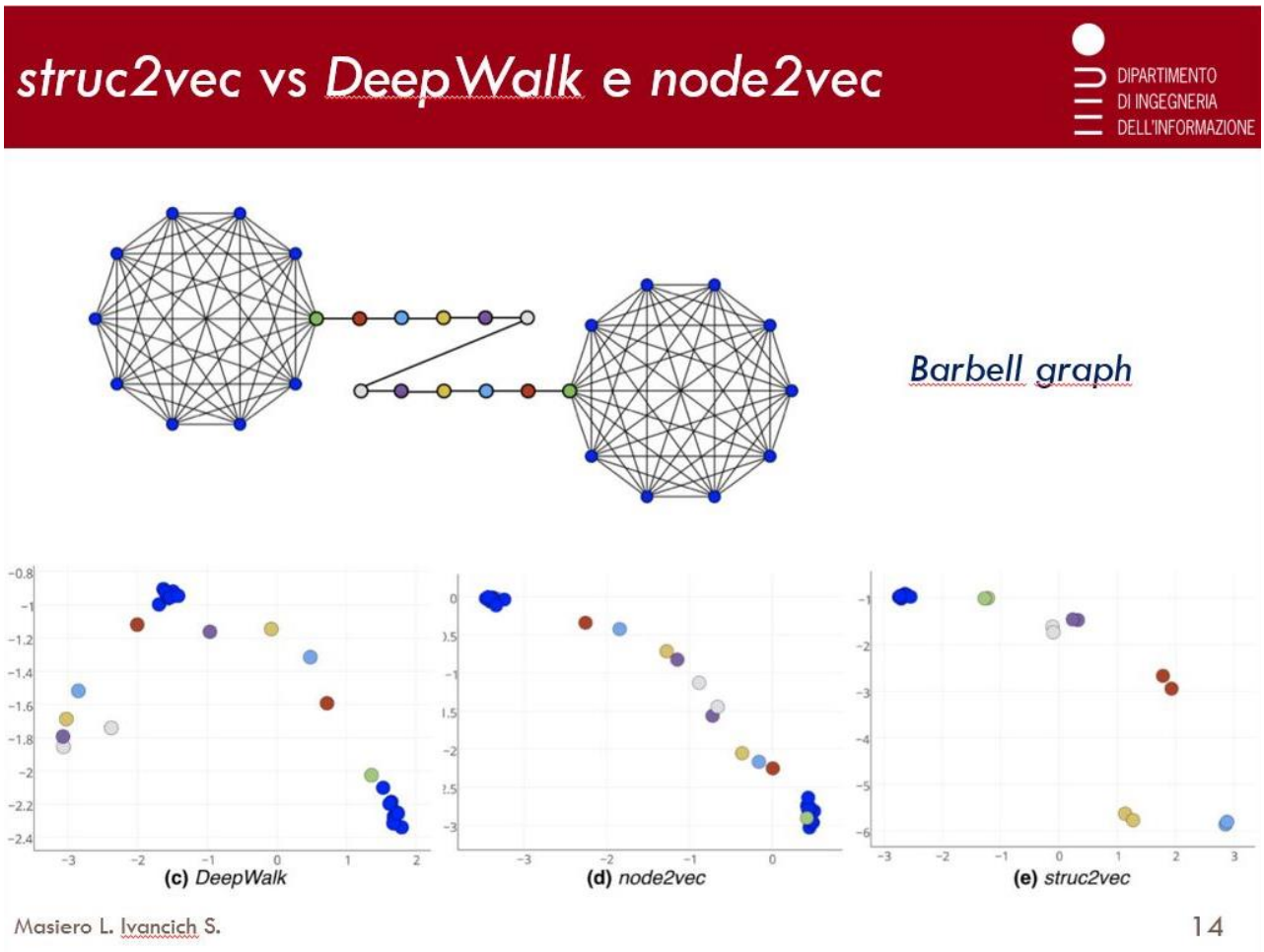
4 FASI

1) Determinazione dell'*identità strutturale* tra due nodi;
2) Calcola $R_k(u)$: insiemi dei nodi a distanza $k \geq 0$ da $u \in G$;
3) Compara sequenze con DTW e costruisce un grafo multi-livello pesato;
4) Utilizza una tecnica di *unsupervised learning* per imparare le *latent representations*.

Il primo passo compiuto dall'algoritmo consiste nel determinare l'identità strutturale tra due nodi senza utilizzare attribuiti di nodi o di archi. Intuitivamente, due nodi che hanno lo stesso grado sono strutturalmente simili, ma se i loro vicini hanno anch'essi lo stesso grado, allora sono ancora di più strutturalmente simili.

Consideriamo un grafo G non orientato e non pesato. $R_k(u)$ denota l'insieme dei nodi a distanza esattamente k ≥ 0 da *u* in G. (come se fosse un *anello*). Il metodo poi compara le *sequenze di gradi ordinate* con il Dynamic Time Warping (un algoritmo che permette di determinare la distanza tra due sequenze). Utilizza infine *Skip-Gram*, una tecnica di *unsupervised learning* per il Natural Language Processing, per identificare o derivare le *latent representations* dalle sequenze ottenute.

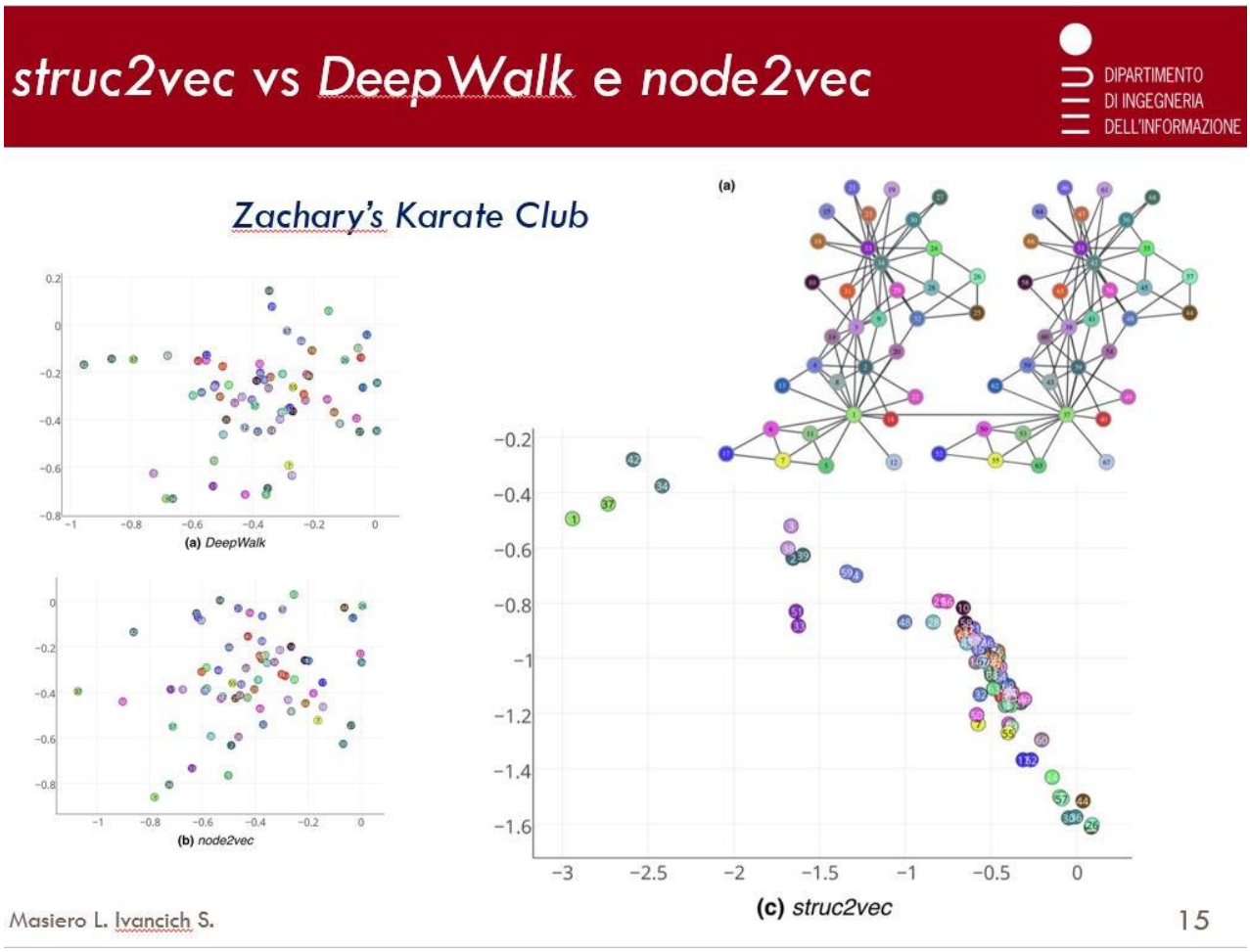
STEFANO 14



struc2vec è stato testato in diversi scenari e confrontato con gli algoritmi allo stato dell'arte (*DeepWalk* e *node2vec*) per l'apprendimento di *latent representations*. Il primo test ha previsto la costruzione di un particolare tipo di grafo, il *barbell graph*, costituito da due grafi completi connessi da un path graph (indicare il collegamento). Ogni coppia di nodi che è strutturalmente equivalente dovrebbe avere *latent representations* simili (che a loro volta dovrebbero essere in grado di descrivere, nel miglior modo possibile, la gerarchia strutturale).

Anche in seguito ad un tuning dei parametri, *DeepWalk* e *node2vec* falliscono nell'individuare le equivalenze strutturali; *struc2vec* invece individua le *latent representations* posizionando i nodi strutturalmente equivalenti gli uni vicino agli altri.

STEFANO 15



Un secondo test ha previsto l'utilizzo della *Zachary's Karate Club network*, una rete composta da 34 nodi e 78 archi, nella quale ogni nodo rappresenta un membro del club e gli archi denotato un'interazione (esterna al club) fra due membri (cioè una relazione di "amicizia"). La rete è stata duplicata (mostrare in alto a destra) in due grafi G1 e G2 nei quali ogni nodo in G1 possiede un corrispettivo **specchio** in G2. I due grafi sono stati connessi tramite un arco fra i nodi 1 e 37 (indicare il collegamento).

Anche in questo caso *DeepWalk* e *node2vec* falliscono nell'individuare le *latent representations* di nodi strutturalmente equivalenti (inclusi i nodi specchio), mentre *struc2vec* fornisce i risultati migliori.

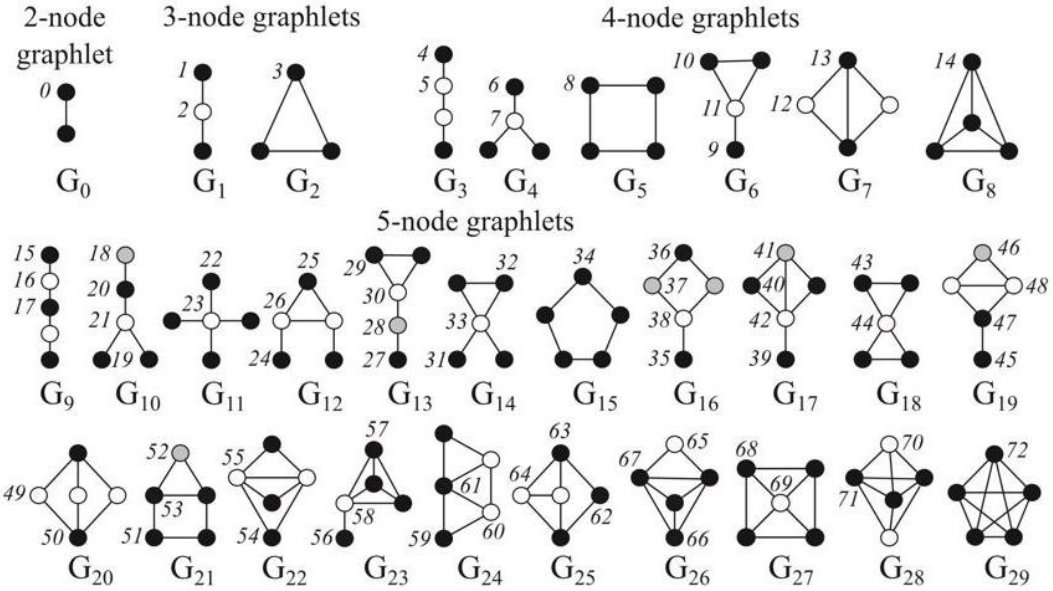
Questo metodo è stato citato oltre 300 volte e per gli ambiti più differenti. È stato confrontato con un metodo sviluppato l'anno successivo, Deep Recursive Network Embedding, sullo stesso dataset; le prestazioni rimangono tutt'ora molto elevate.

STEFANO 16

L-GRAAL

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Idea: mappare insieme nodi che costituiscono un pattern (*graphlet*) con molte interazioni condivise.



Sviluppato nel 2015, L-GRAAL è un metodo basato sull'idea di mappare insieme nodi che costituiscono un *pattern* (in questo caso dei *sottografi* chiamati **graphlet**) definito da una grande quantità di interazioni condivise.

LUCA 11

L-GRAAL

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Funzione obiettivo: fonde le informazioni dalle sequenze di proteine con le interazioni tra i *graphlet* (si risolve con la Programmazione Intera).

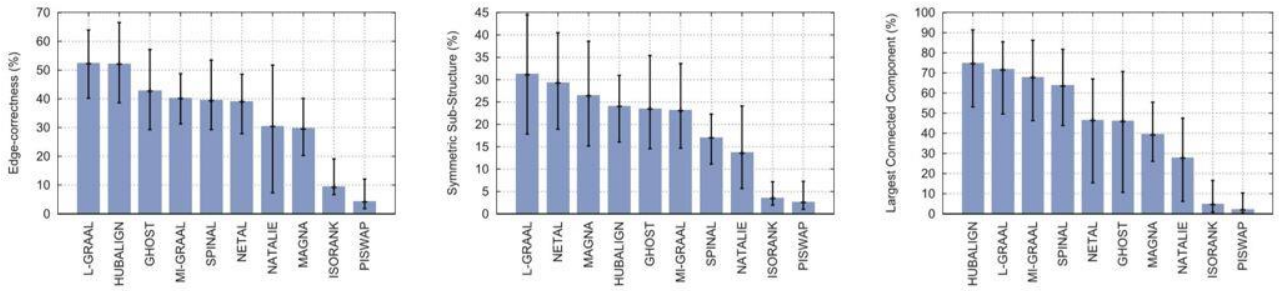
FASI

1) *Similarity scores* e definizione della *topological similarities;*
2) Risoluzione dell'equazione

$$\text{IP} = max_{x,y}\left(\alpha \sum n(i,k) \times x_{ik} + (1-\alpha) \sum e(i,j,k,l) \times y_{ijkl}\right)$$

in tempo $O(|V|^3 + |V|^2 \times d^3)$.

Problema NP-completo.



L-GRAAL è in grado di individuare l'*overlap* tra le reti e fornisce risultati migliori di tutti gli altri metodi GO-based a livello di mapping delle proteine e delle interazioni tra le stesse.

Questo metodo ottimizza una funzione obiettivo (indicare IP), che fonde le informazioni derivanti dalle sequenze di proteine con le interazioni tra i vari graphlet. Questa funzione viene risolta con la Programmazione Intera in O(blabla) dove |V| indica il numero di nodi nelle reti e *d* è il valore di grado massimo. Dal momento che con la sola Programmazione Intera otteniamo una soluzione parziale, è necessario ricondursi alla *formulazione duale* del problema ed utilizzare la tecnica del *gradient descent*.

Sfortunatamente, anche questo problema è NP-completo e, in pratica, si risolve fermando, di fatto, l'algoritmo dopo un determinato limite temporale o dopo un numero di iterazioni fissato.

In tutti i test svolti L-GRAAL (citato 95 volte in Google Scholar negli ultimi anni 5 anni) ha mostrato una percentuale di successo non indifferente, superiore a tutti gli altri metodi con cui è stato confrontato, come IsoRank (indica nel grafico).

LUCA 12

Conclusioni

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Negli ultimi anni, il corpus di dati PPI è cresciuto esponenzialmente. Scoprire e capire i pattern all'interno delle PPIN è un problema centrale in Biologia.

Gli allineamenti tra queste reti permettono di scoprire informazioni su complessi proteici che fino a pochi anni fa non erano note.

Molte sfide sono ancora aperte e molte frontiere devono ancora essere esplorate; con questo progetto abbiamo dato solamente una vaga idea della vastità dell'argomento, di cui si è appena iniziato a parlare.

Negli ultimi anni, il corpus di dati PPI è cresciuto esponenzialmente e il rapido ritmo di accumulo dati continua imperterrito tutt'oggi. L'obiettivo di questo progetto è stato di far capire la struttura delle reti di interazione proteina-proteina e le implicazioni dal punto di vista biologico.

Gli allineamenti tra queste reti permettono di scoprire informazioni su complessi proteici che fino a pochi anni fa non erano note.

Molte sfide sono ancora aperte e molte frontiere devono ancora essere esplorate; con questo progetto abbiamo solamente dato una vaga idea della vastità dell'argomento, ma speriamo comunque di aver stimolato l'interesse nei confronti dell'argomento. Grazie per l'attenzione!

LUCA 17