Network Alignment

Studenti:

Luca Masiero

Stefano Ivancich

Supervisor:

Prof. Matteo Comin

DEPARTMENT OF INFORMATION ENGINEERING

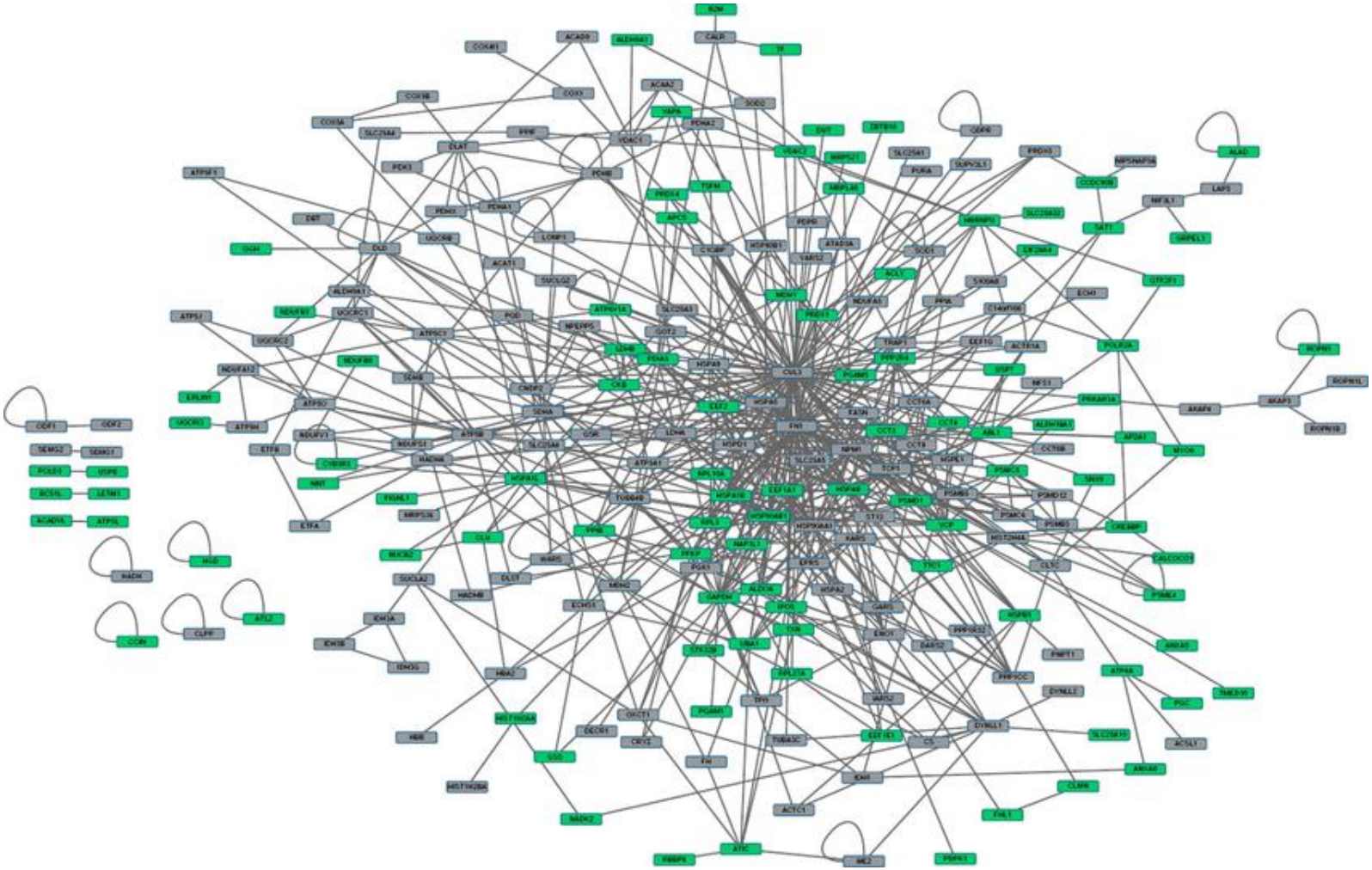
UNIVERSITY OF PADOVA

10 Giugno 2020

Introduzione: *Network Alignment* e PPIN

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Obiettivo = trovare somiglianze tra struttura e/o topologia di due o più reti.



Introduzione: *Network Alignment* e PPIN

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 **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.

Introduzione: *Network Alignment* e PPIN

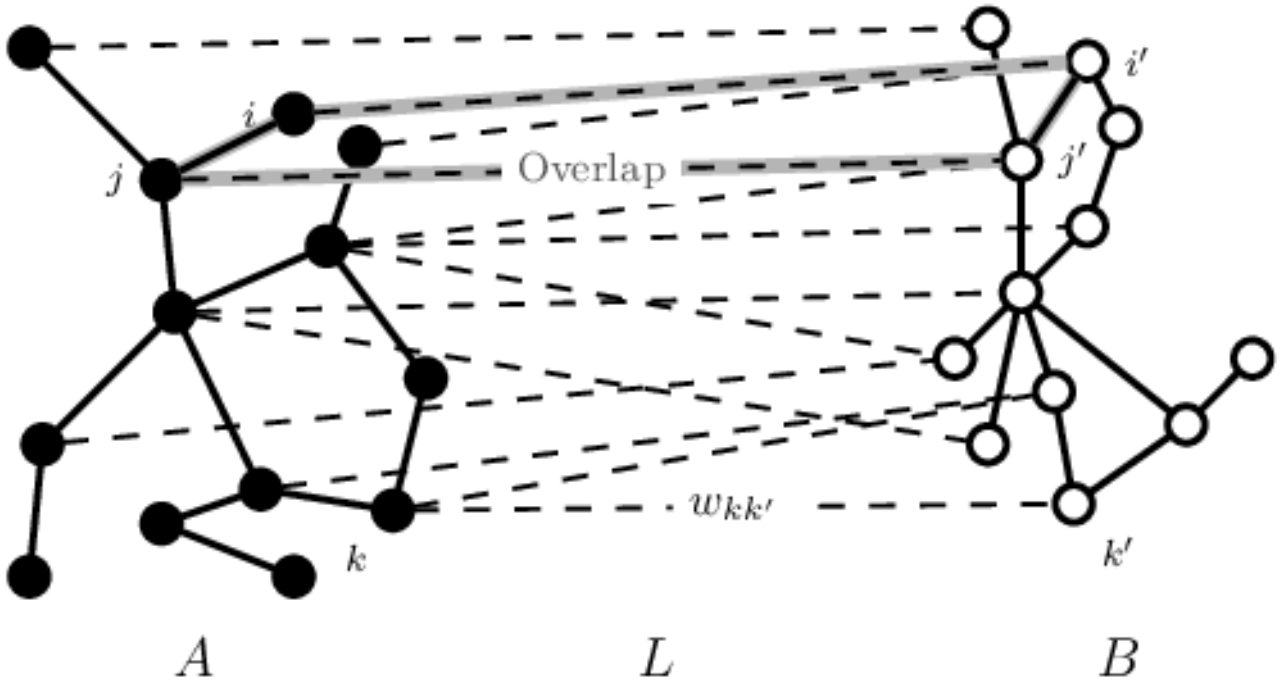
DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Date due reti, **allinearle** significa trovare un *mapping* nodo-a-nodo in grado di:

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

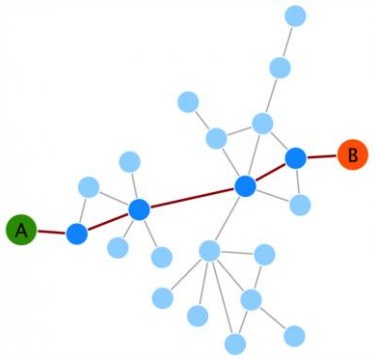
(2) massimizzare il numero di interazioni comuni (*archi*) tra le reti.

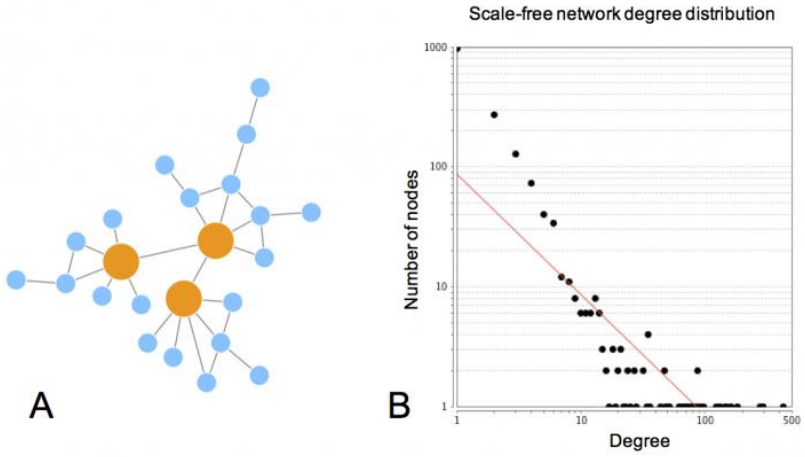
Problema intrattabile dovuto all'NP-completezza del *sub-graph isomorphism problem* (Cook,1971).



PPIN: proprietà fondamentali

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

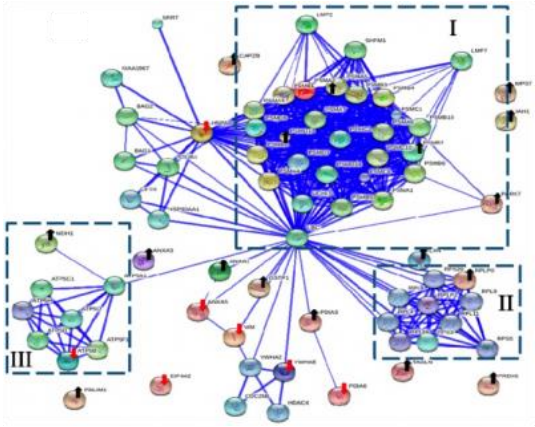


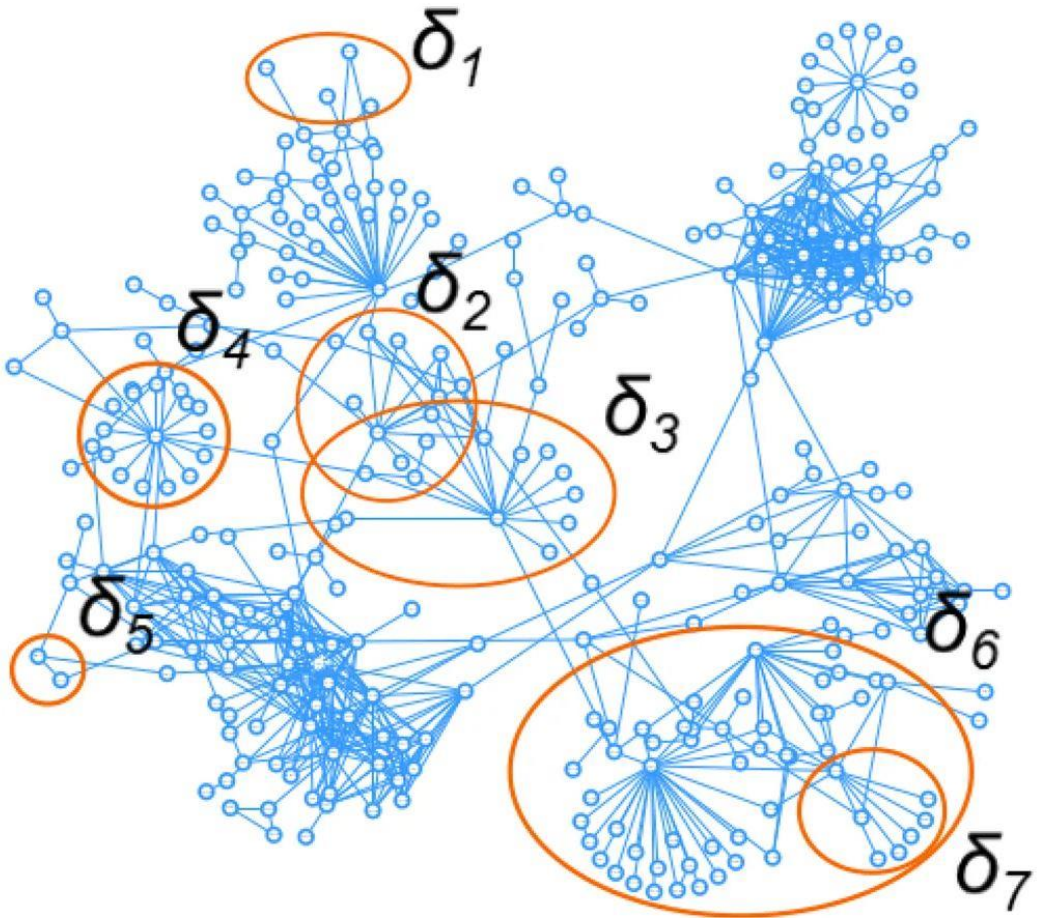


Scale-free networks: nodi con poche connessioni vs *hub*

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**).





Metodo *GO-based* per identificare moduli funzionali.

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

Overlapping e copertura totale della rete.

MTGO: Fasi

3 FASI

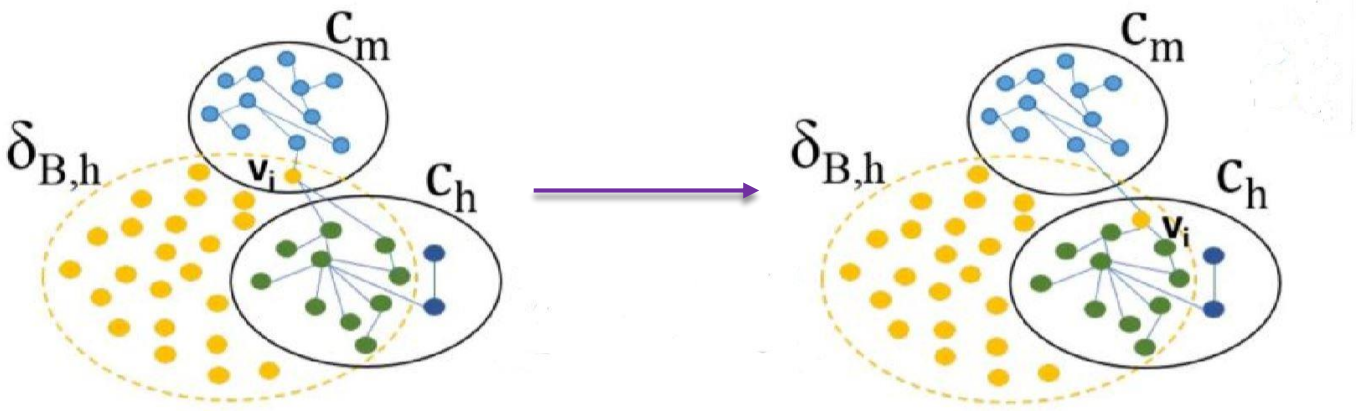
1) **Inizializzazione**: creazione delle partizioni e dei moduli

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

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

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

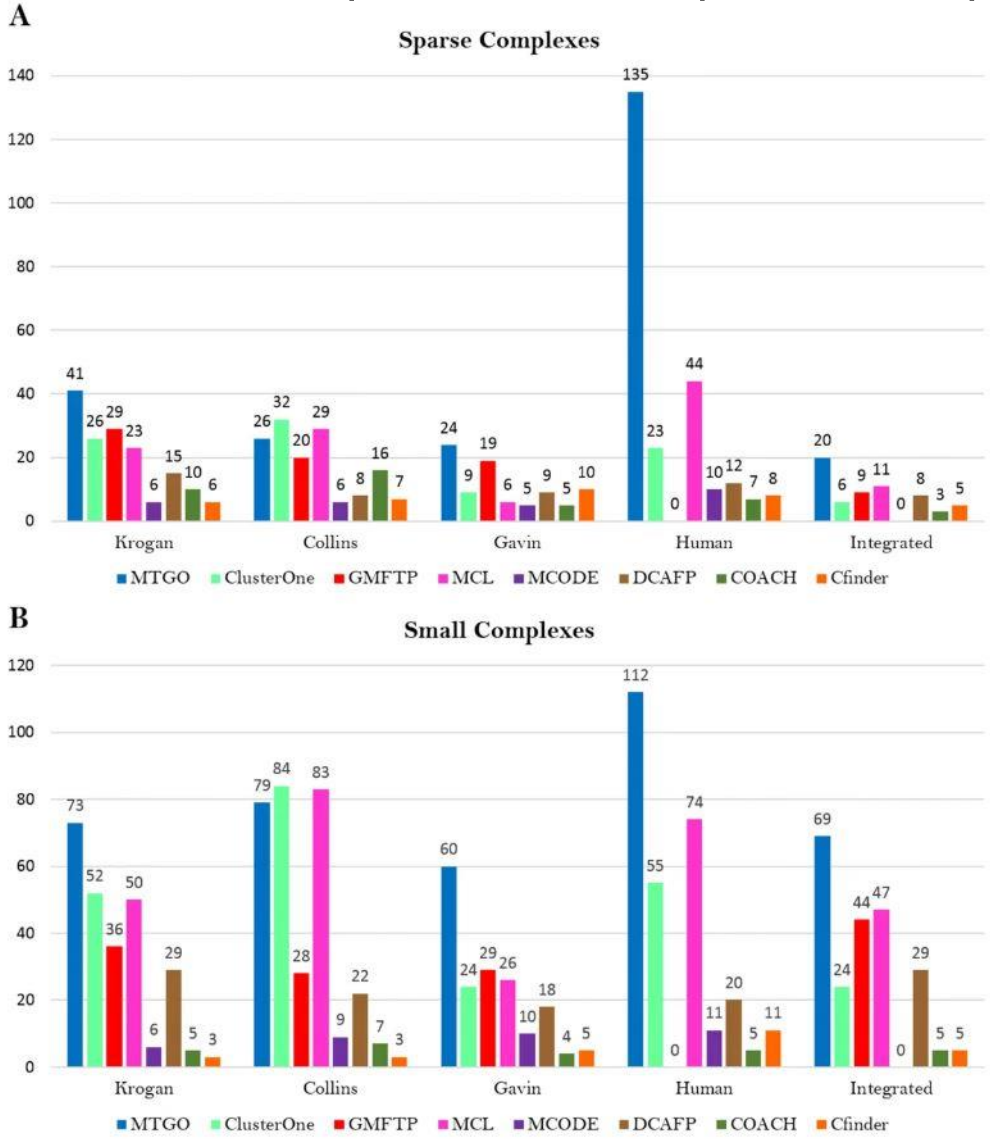
2) **Iterazioni** in cui i nodi vengono ri-assegnati alle partizioni



MTGO: Conclusione

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

IsoRank

Metodo per l'allineamento di più PPIN.

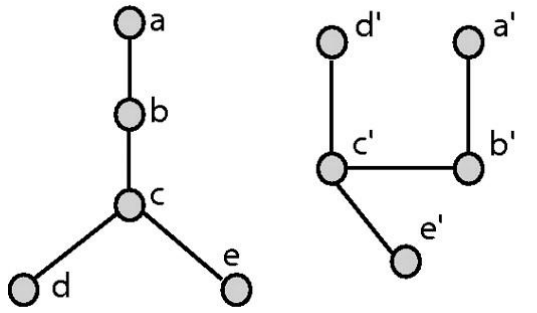
Intuizione = una proteina rappresenta una buona corrispondenza con una proteina in un'altra sequenza se le rispettive sequenze e i loro intorni topologici costituiscono una buona corrispondenza.

Rappresenta un *approccio di analisi comparativa* al GNA.

Caso di GNA a coppie

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

IsoRank: Fasi



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 (solo *scores* elevati) mantenendo la proprietà transitiva (**one-to-one** o **many-to-many**)

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

IsoRank: Conclusione

Il sottografo corrispondente all'allineamento globale possiede 1663 archi in comune ad almeno due PPIN e 157 archi in comune al almeno 3 PPIN.

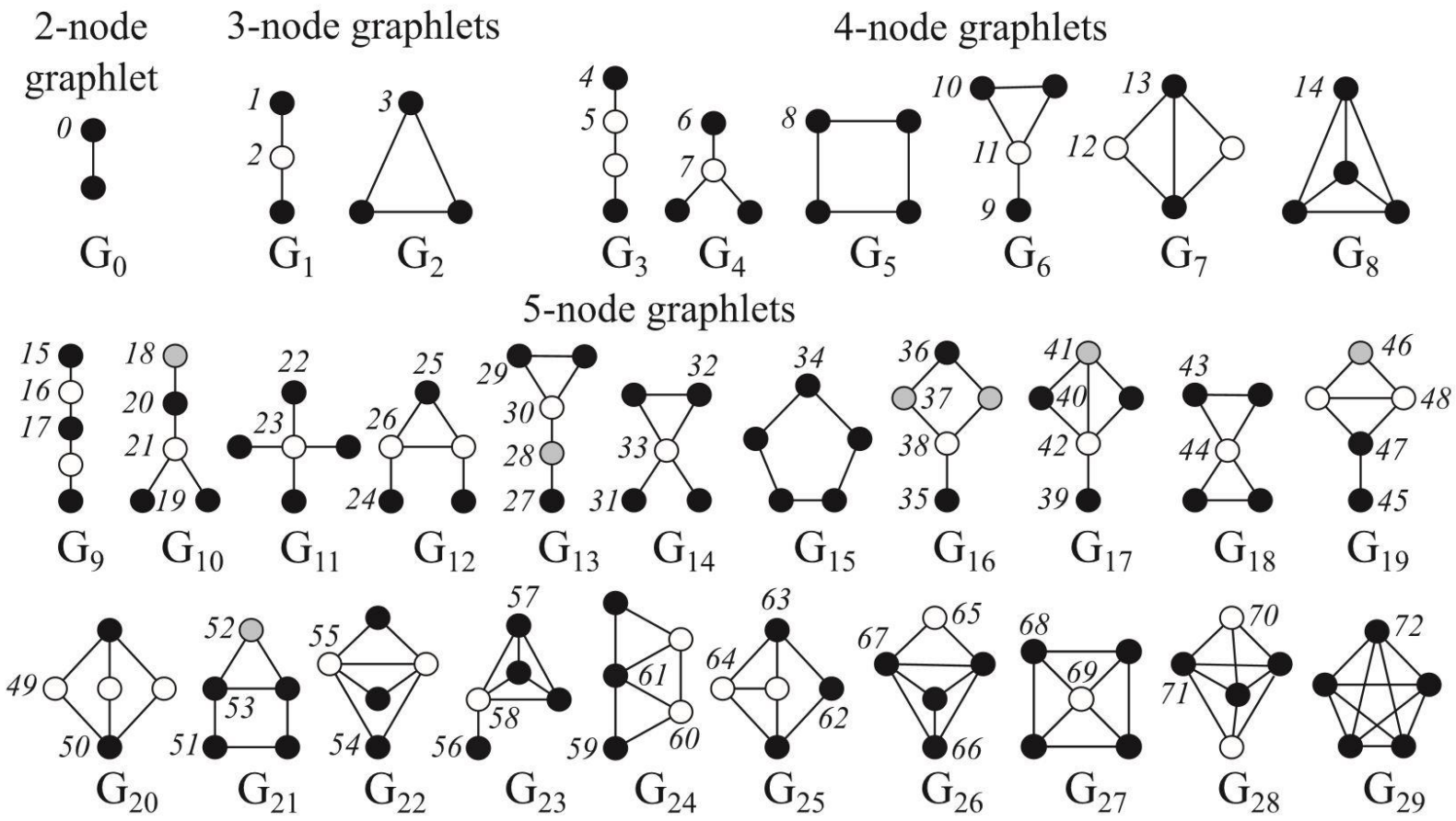
La dimensione del sottografo comune relativamente piccola (*overlap* con ≈ 5% della PPIN umana) a causa dell'incompletezza e della rumorosità dei dati.
All'aumentare della quantità e della qualità dei dati, l'*overlap* dovrebbe aumentare. Delle 86932 proteine provenienti dalle 5 specie, 59539 (68,5%) hanno ottenuto almeno un match con un'altra proteina di una rete diversa.

NUMERO DI CITAZIONI = **505** dallo sviluppo nel 2008.

Google Scholar

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

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



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).

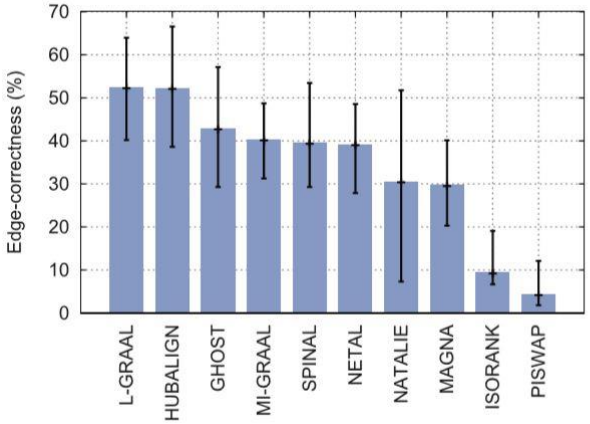
FASI

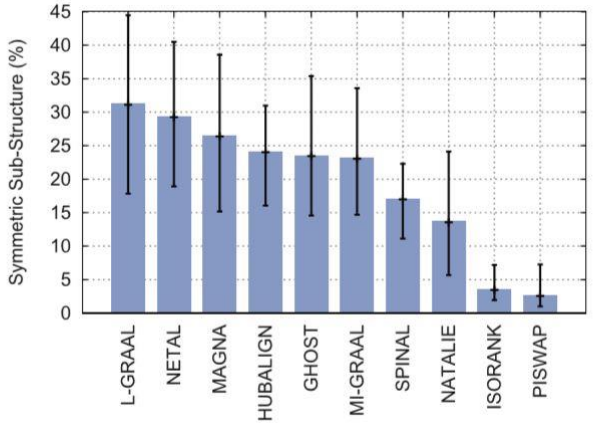
1) *Similarity scores* e definizione della *topological similarities;*
2) Risoluzione dell'equazione

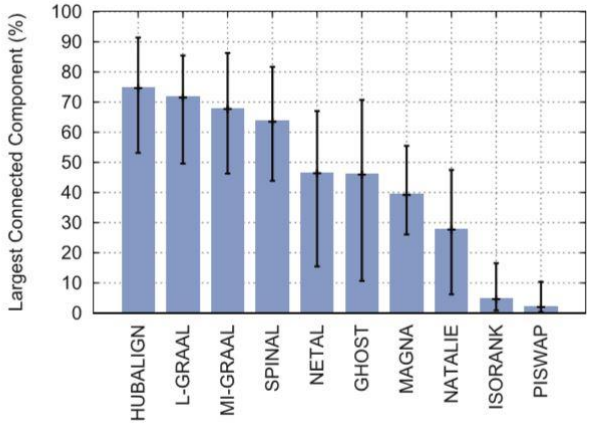
$$\mathrm{IP} = max_{x,y}\left(\alpha \sum n(i,k) \times x_{ik} + (1-\alpha) \sum e(i,j,k,l) \times y_{ijkl}\right)$$

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

Problema NP-completo.







struc2vec

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.

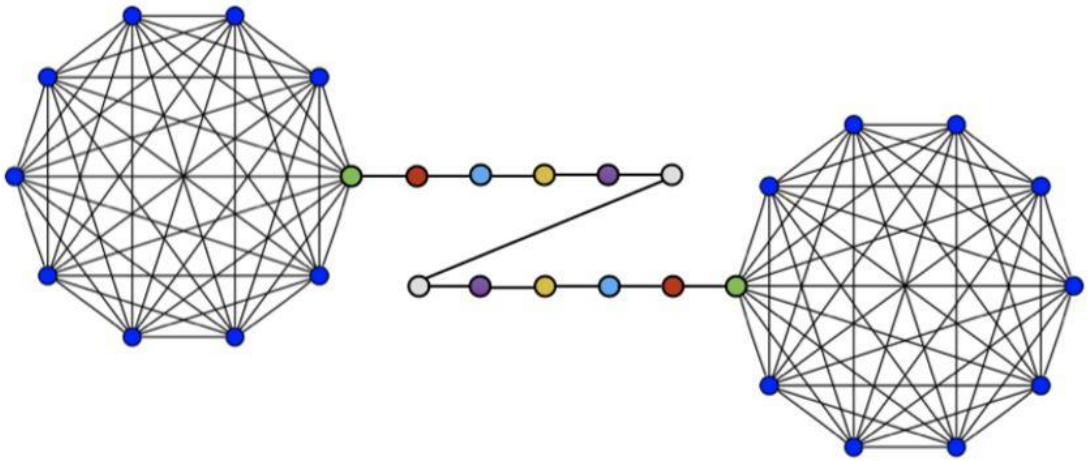
struc2vec

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.

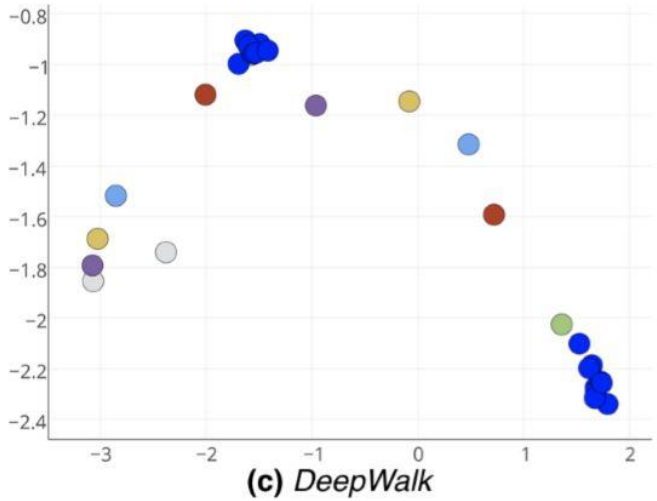
4 FASI

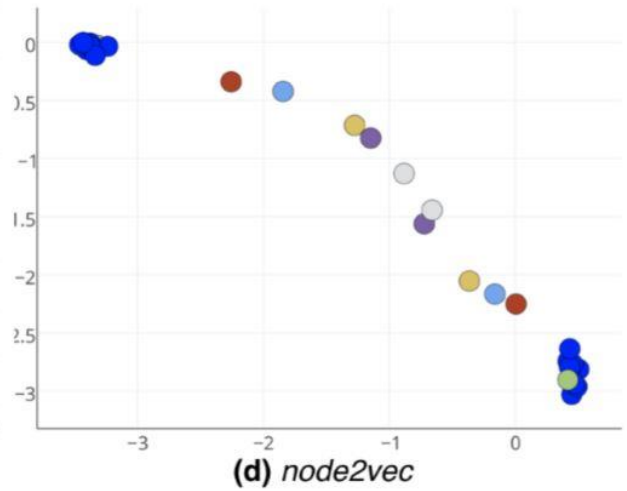
1) Determinazione dell'*identità strutturale* tra due nodi;
2) Calcola $R_k(u)$: insiemi dei nodi a distanza $k \geq 0$ da $u \in G$;
3) Compara sequenze con DTW e costruisce un grafo multi-livello pesato;
4) Utilizza una tecnica di *unsupervised learning* per imparare le *latent representations*.

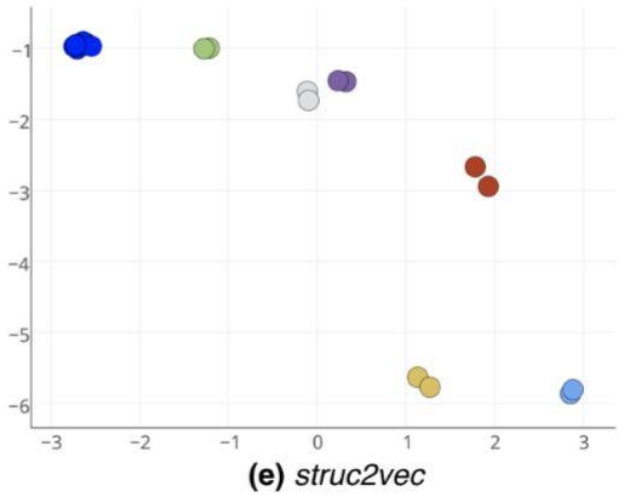
struc2vec vs DeepWalk e node2vec



Barbell graph

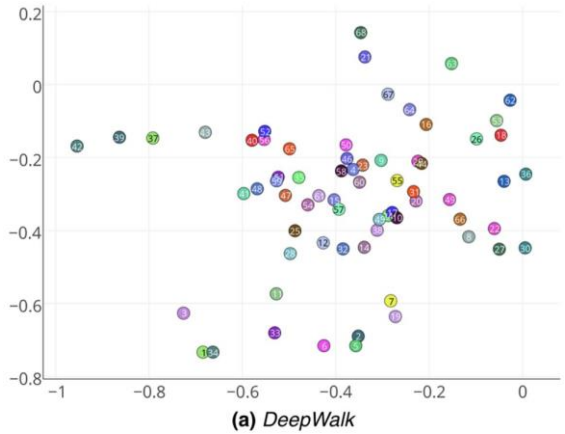




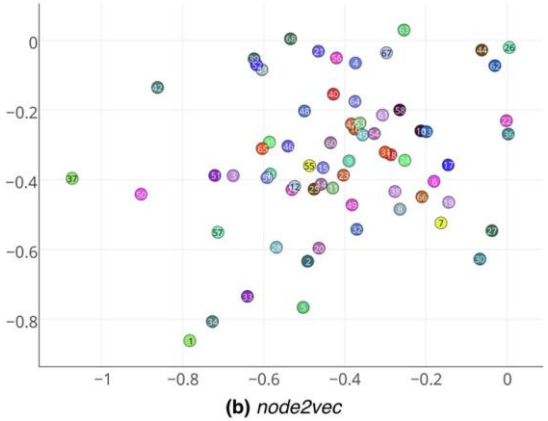


struc2vec vs DeepWalk e node2vec

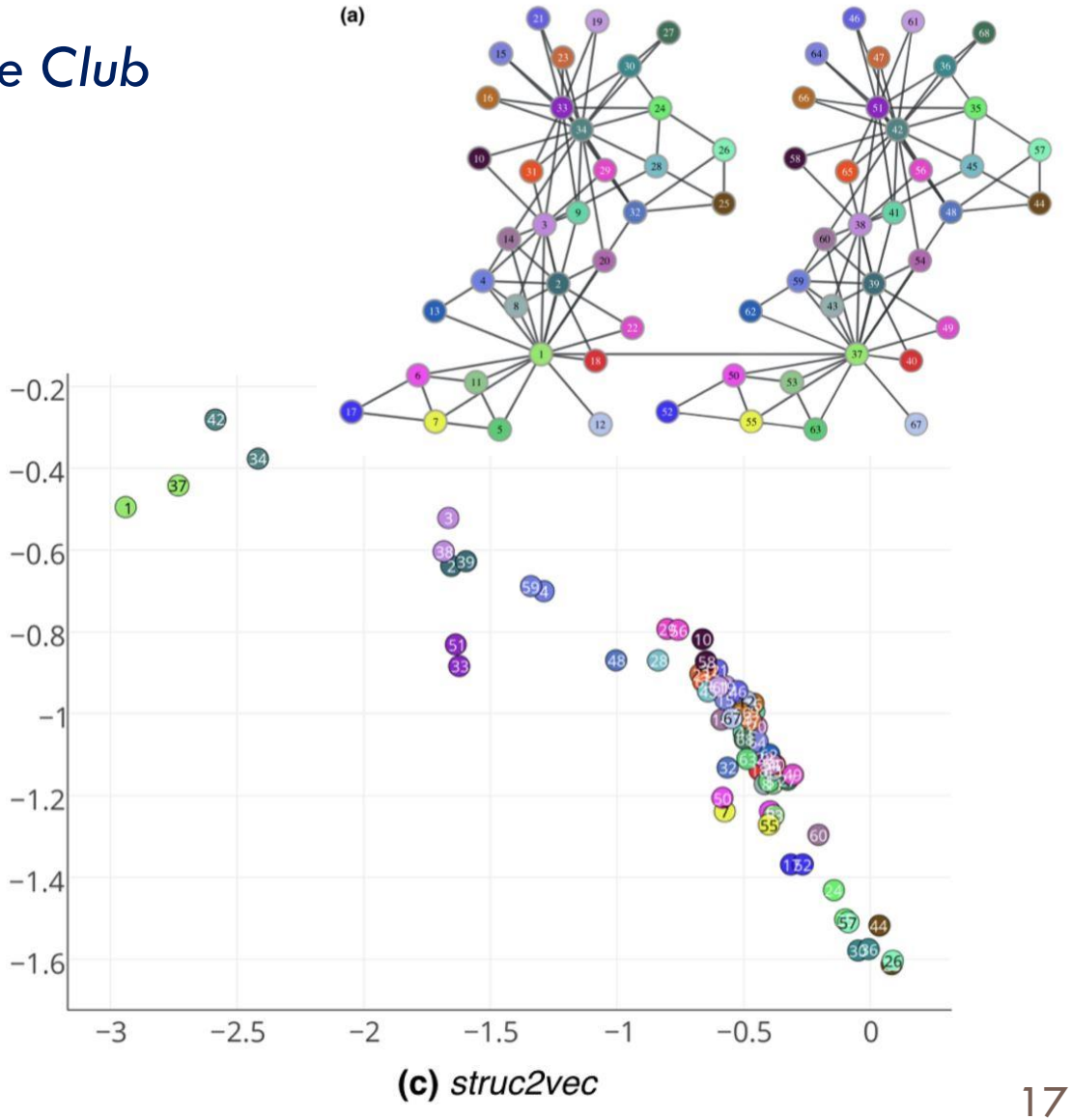
Zachary's Karate Club



(a) *DeepWalk*



(b) *node2vec*



(c) *struc2vec*

Conclusioni

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 le reti permettono di scoprire informazioni su complessi proteici che fino a pochi anni fa erano sconosciute.

Molte sfide sono ancora aperte e molte frontiere devono ancora essere esplorate.

Grazie per l'attenzione!