A graph theory approach to characterize the relationship between protein functions and structure of biological networks

Serene Wong

March 15, 2011

Hybrid system

Graph theory

graphlet representation

Biological discoveries

Infer protein functions

Understand underlying mechanisms of disease

http://www.toyota.ca/toyota/en/vehicles/prius/gallery

Outline

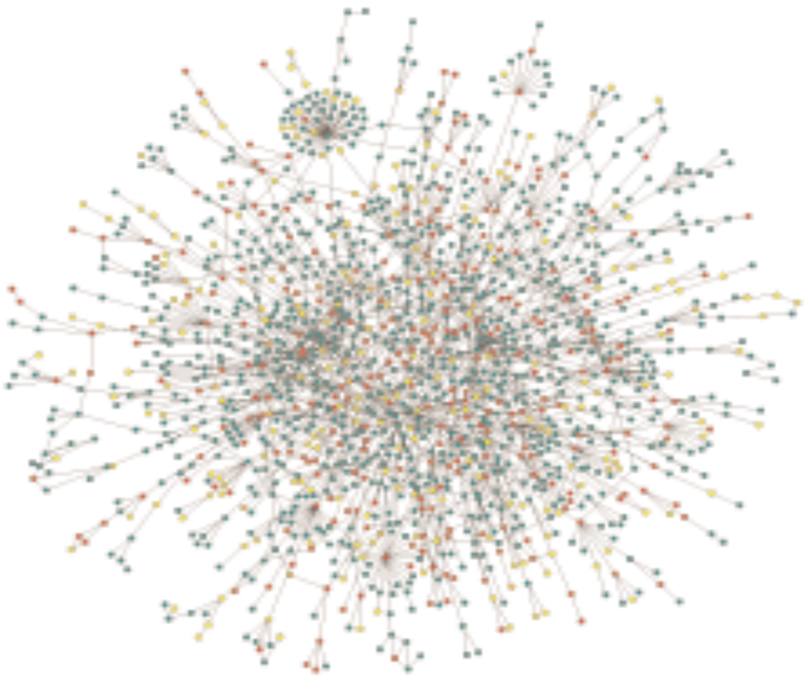
- Introduction
- Network properties
- An example of relationship between network properties and disease
- Biological network comparisons
- Uncovering biological network function
- Conclusion

Introduction

Biological networks

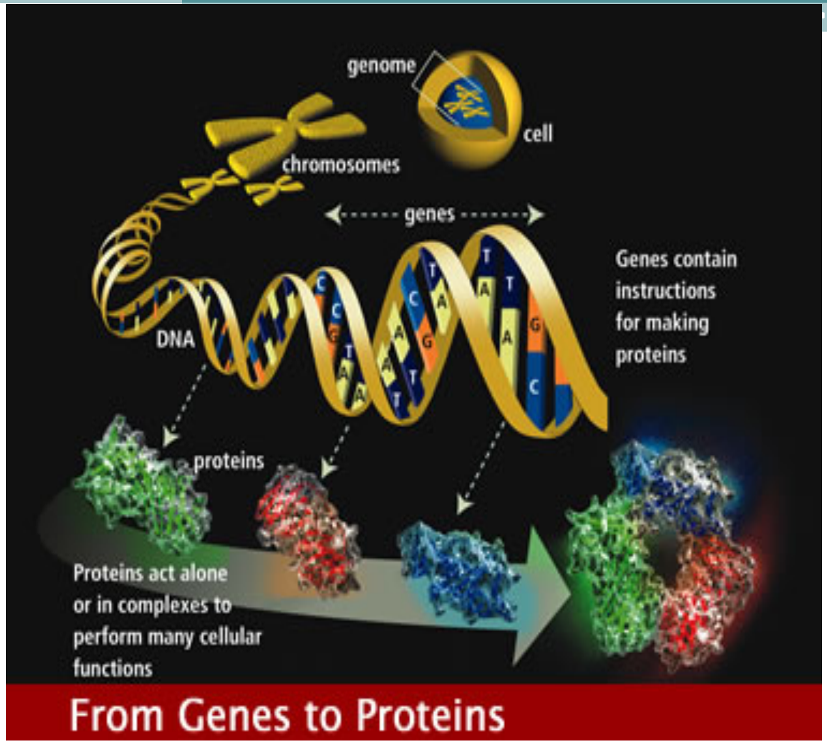
Vertices: proteins
Edges: physical interactions

- Traditionally, individual cellular components and their functions are studied
- most biological functions are due to interactions between different cellular constituents
- various networks have emerged including protein-protein interactions networks.



(H. Jeong et al., 2001) Lethality and centrality in protein networks, H.Jeong et al., 2001

Central dogma


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
From Genes to Proteins

http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml

DNA (segments of DNA, ‘genes’)

Transcription

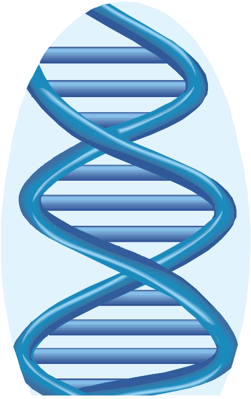
mRNA (mRNA abundance detected using microarrays)

Translation

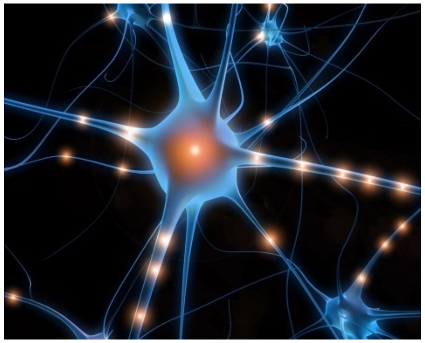
Protein

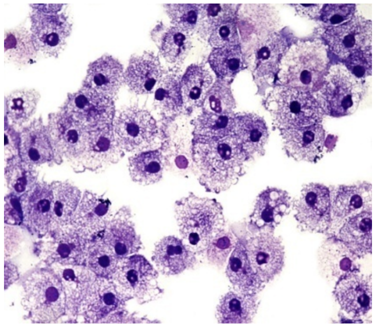
(H. C. Causton et al., 2003) *Figure 1.2 (partial) from Ch 1 of Microarray Gene Expression Data Analysis*

Gene expression studies



In general, each cell in the body has the same DNA





Different type of cells - difference is in the subset of genes that a cell expressed

- Different responses to stimuli can also lead to expressing different subsets of genes
- Gene expression studies enable the understanding of the mechanism in the molecular level

Definitions

- **Definition 1:**
  Let G(V,E) denotes a graph where V is the set of vertices, and E, E ⊆V x V, is the set of edges in G

- **Definition 2:**
  Let x and y be vertices from G. y is adjacent to x if there is an edge between x and y, and y is a neighbor of x. Let N(x) denote the set of vertices that are adjacent to x, and N(x) is the neighborhood of x

- **Definition 3:**
  A degree of a vertex, x, d(x) is the number of incident edges to x

- **Definition 4:**
  An induced subgraph, H, is a subgraph such that E(H) consists of all edges that are connected to V(H) in G

Network properties

Global network properties versus local network properties

Global network properties	Local network properties
• Look at the overall network • PPI networks are incomplete, and contain bias	• Focus on local structures or patterns • Can measure properties in local regions even though networks are incomplete

Global network properties

- Degree distribution, P(k)
  - is the probability in which any randomly selected vertex has degree k

- Diameter
  - the maximum shortest path length between any pair of vertices. Often, it is the average shortest path length between all pairs of vertices

- ▫ Centrality measures

Centrality measures – degree centrality

degree centrality of vertex *u*:

$$C_d(u) = d(u)$$

Centrality measures – closeness centrality

center of *G*:

$$Cen(G) = \{x \in V | e(x) = r(G)\}$$

excentricity of *x*:

$$e(x) = \max_{y \in V} d(x, y)$$

radius of *G*:

$$r(G) = \min_{x \in V} e(x)$$

Centrality measures – betweenness centrality

betweenness centrality of vertex *w*:

$$\{u, v, w \in V | u \neq v, v \neq w\}$$

$$BC(w) = \sum_{u,v \in V} \frac{S_{uv}(w)}{S_{uv}}$$

$S_{uv}(w)$ is the number of geodesic paths between *u* and *v* that pass through *w*

S_{uv} is the number of geodesic paths between *u* and *v*

Local network properties

Motifs

- Small subgraphs in a network whose patterns appear significantly more than in randomized networks
- Do not take into account patterns that appear with average or low frequency
- Depend on randomization scheme

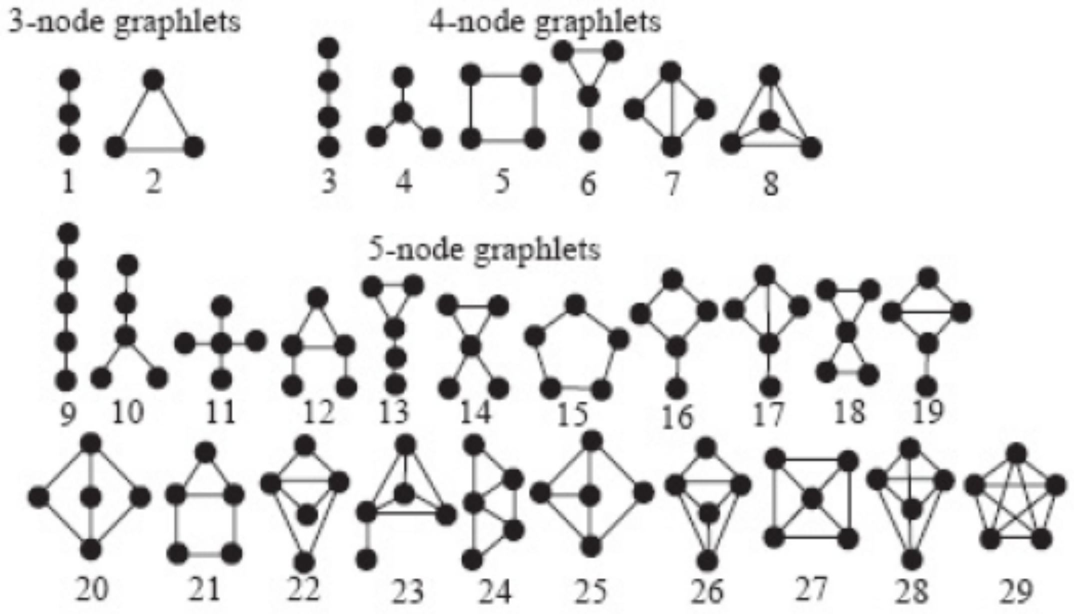
(R. Milo et al, 2002)

Graphlets

- All non-isomorphic connected induced graphs on a certain number of vertices
- Identify all structures, not only the over-represented ones

(N. Pr ˇzulj et al. 2004b)

Graphlets



Not limited to 3-5 node graphlets!

All 3 to 5 node graphlets, graphlet No. 1 to 29. Fig. 1 of Modeling interactome: scale-free or geometric.

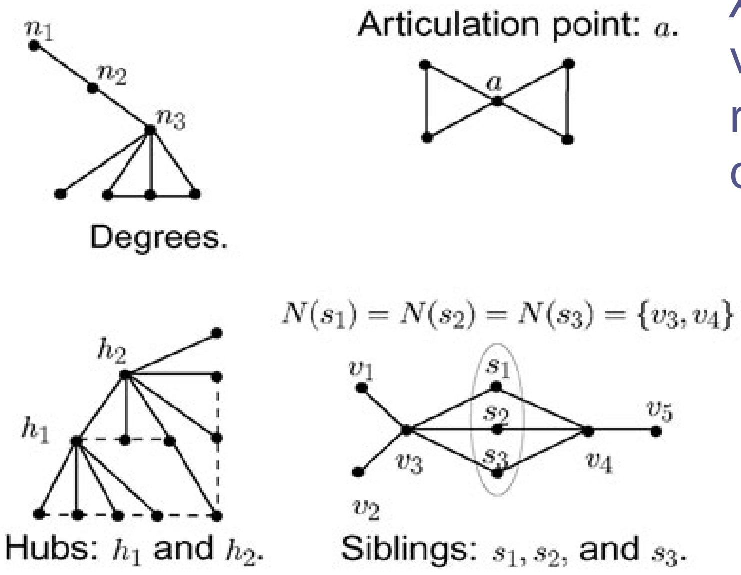
(N. Prˇzulj et al. 2004b)

An example of relationship between network properties and disease

Protein essentiality

Minimum spanning tree (MST): an acyclic connected subgraph that contains all the vertices of the graph, and the edges that give the minimum sum of edge weights

Hubs: highly connected vertices in the MST



(N. Prˇzulj et al., 2004) Graph theoretic properties. Partial Fig. 1B of Functional topology in a network of protein interactions

Articulation point is a vertex that, if removed, results in a disconnected graph

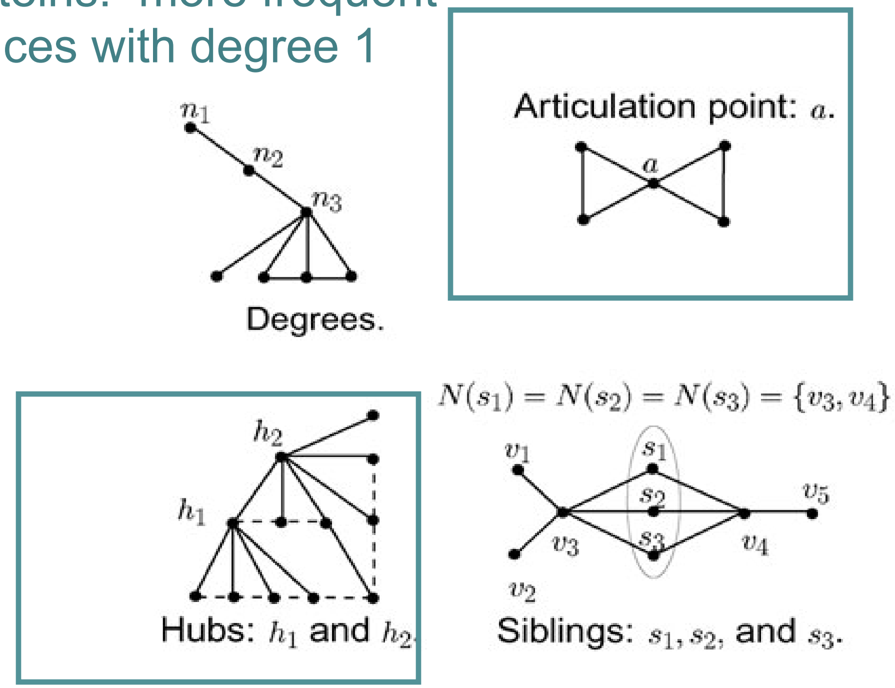
If 2 vertices have the same neighborhood, then they are *siblings*

(N. Prˇzulj et al., 2004)

Protein essentiality

Lethal proteins: more frequent in the top 3% of degree vertices
Viable proteins: more frequent in the vertices with degree 1

Lethal proteins were not only hubs, but they were articulation points



Viable proteins were more frequent in the group of vertices that belonged to the sibling group

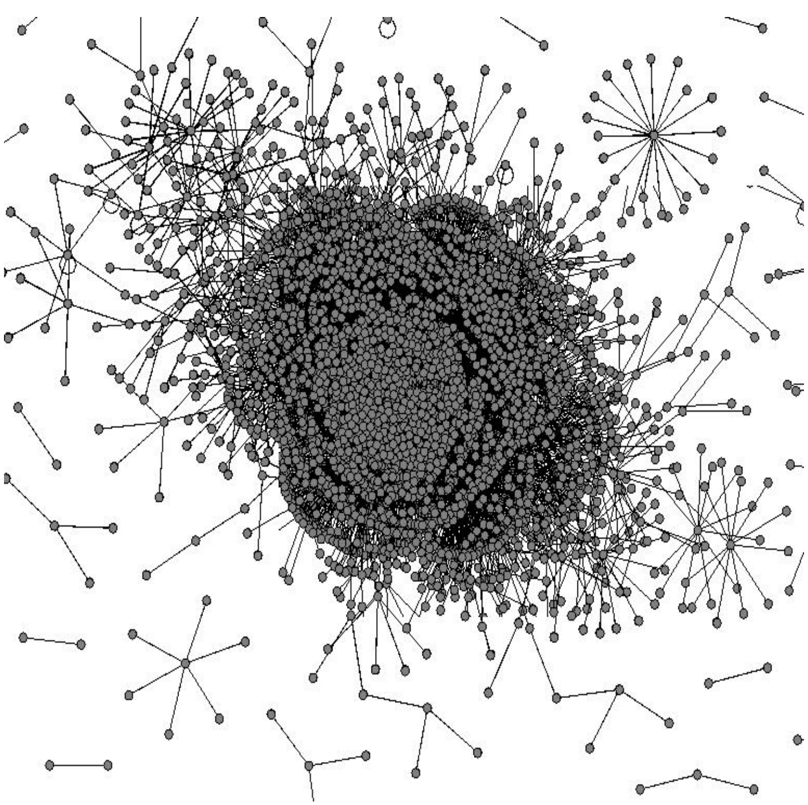
Graph theoretic properties. Partial Fig. 1B of Functional topology in a network of protein interactions

(N. Prˇzulj et al., 2004)

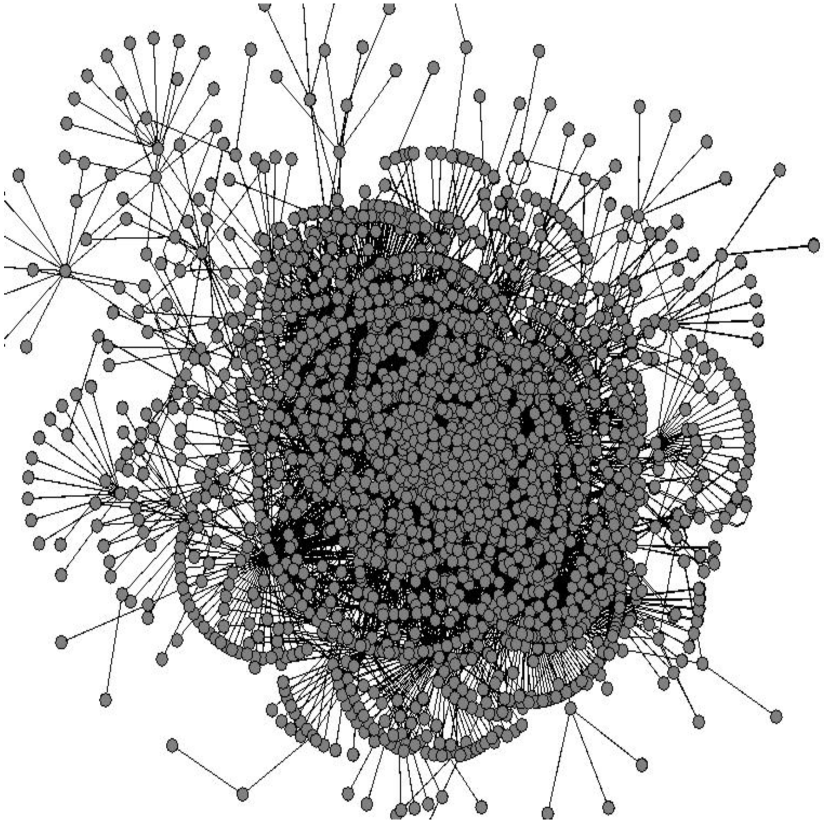
Grahplets

Biological network comparisons

Biological network comparisons



Network 1



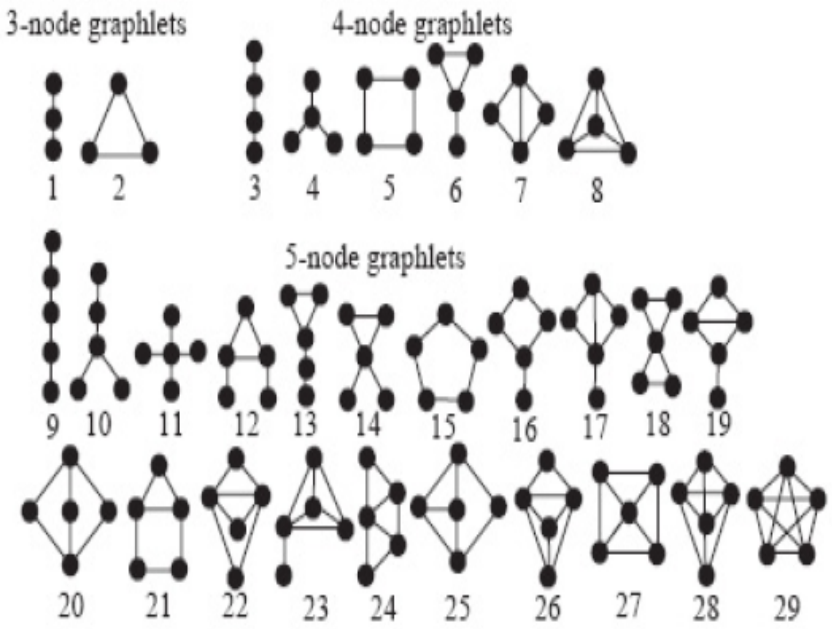
Network 2

Graphlets

- 2 local measures based on graphlets have developed
  - ▫ Relative graphlet frequency distance (RGF-distance)
  - ▫ Graphlet degree distribution agreement (GDD-agreement)

(N. Prˇzulj et al. 2004b, N. Prˇzulj 2007)

Graphlet frequency



All 3 to 5 node graphlets, graphlet No. 1 to 29. Fig. 1 of Modeling interactome: scale-free or geometric

- The count of how many graphlets of each type (ranging from 1 to 29)
- Not limited to 3 to 5 node graphlets
- If more graphlets can be computed, a greater number of local constrains are imposed on similarity measures

(N. Prˇzulj et al. 2004b)

Relative graphlet frequency

relative frequency of graphlets is defined to be: $\dfrac{N_i(G)}{T(G)}$

$N_i(G)$ is the number of graphlets of type i,
$i \in [1, ..., 29]$ in graph G

$T(G) = \sum_{i=1}^{29} N_i(G)$

(N. Prˇzulj et al. 2004b)

Relative graphlet frequency distance (RGF – distance)

relative graphlet frequency distance between graphs *G* and *H*, *D(G,H)*:

$$D(G,H) = \sum_{i=1}^{29} |F_i(G) - F_i(H)|,$$

where $F_i(G) = -log\frac{N_i(G)}{T(G)}$

(N. Pr ˇzulj et al. 2004b)

Graphlet degree distribution (GDD)

- Direct generalization of degree distribution
- Imposes *73* local constraints to the structure of networks
  - When used as similarity measure between networks, increases the possibility that the networks are indeed similar

(N. Prˇzulj, 2007)

Graphlet degree distribution (GDD)

- Direct generalization of degree distribution

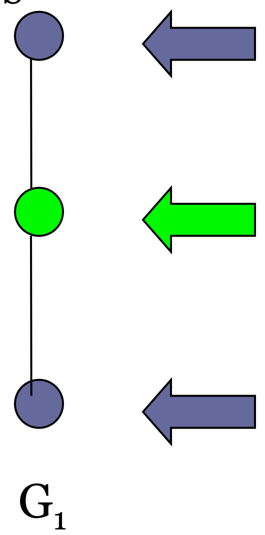
Degree distribution:
How many vertices 'touch' one G_0?
How many vertices 'touch' two G_0?

How many vertices 'touch' *k* G_0?
Graphlet degree distribution:
Apply the above also to the 29 graphlets G0, G1, ..., G29

- Imposes *73* local constrains to the structure of networks

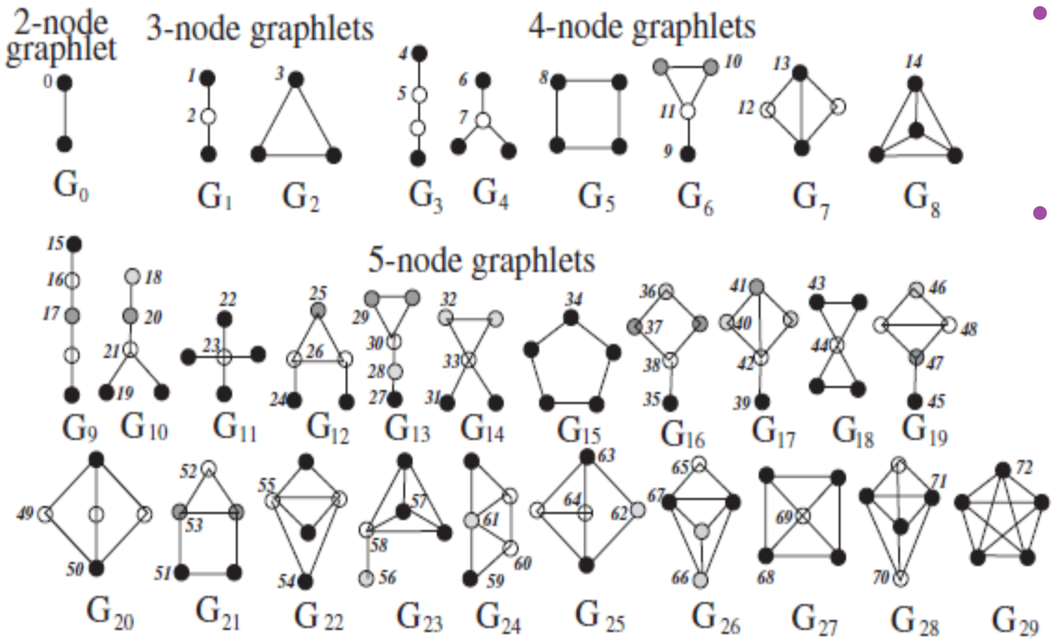
Topological Issue:

How many vertices 'touch' G_1?



(N. Prˇzulj, 2007)

Graphlet degree distribution 2



2-5 node graphlets with automorphism orbits 0 .. 72.
Fig. 1 of Biological network comparison.

- 73 graphlet degree distributions
- Each distributions answers questions such as
  - how many vertices touch 1 orbit 2 of G_1
  - How many vertices touch 2 orbit 2 of G_1
  - How many vertices touch k orbit 2 of G_1

(N. Pr ˇzulj, 2007)

GDD agreement measure

- To compare network similarity

- Reduce the 73 graphlet degree distributions into a scalar agreement between [0,1]
  - ▫ 0 – networks are far apart
  - ▫ 1 - the distributions of the 2 graphs are identical

(N. Prˇzulj, 2007)

GDD agreement - definitions

- Let *G* be a graph, and *j* be the *jth* orbit
  - ▫ d_G^j denotes the sample distribution for the graphlet with the *jth* orbit of *G*
  - ▫ $d_G^j(k)$ denotes the number of vertices that touch orbit *j* in *G k* times

(N. Prˇzulj, 2007)

GDD agreement

$$S_G^j(k) = \frac{d_G^j(k)}{k}$$

Is scaled in order to decrease the effect on large *k*s

$$T_G^j = \sum_{k=1}^{\infty} S_G^j(k)$$

Total area

$$N_G^j(k) = \frac{S_G^j(k)}{T_G^j}$$

Is normalized with respect to total area

(N. Prˇzulj, 2007)

Distance

Let *H* be another graph. The distance of the *j* orbit between two graphs, *G* and *H* is defined to be:

$$D^j(G,\ H) = \frac{1}{\sqrt{2}} \left(\sum_{k=1}^{\infty} \left[N_G^j(k) - N_H^j(k) \right]^2 \right)^{\frac{1}{2}}$$

The *j*th GDD *agreement* is defined to be:

$$A^j(G,\ H) = 1 - D^j(G,\ H)\ ,\ for\ j \in \{0, 1, \ldots, 72\}$$

(N. Prˇzulj, 2007)

GDD Agreement

The *agreement* for graph *G* and *H* can be defined as the arithmetic mean over $A^j(G;H)$ for all *j*:

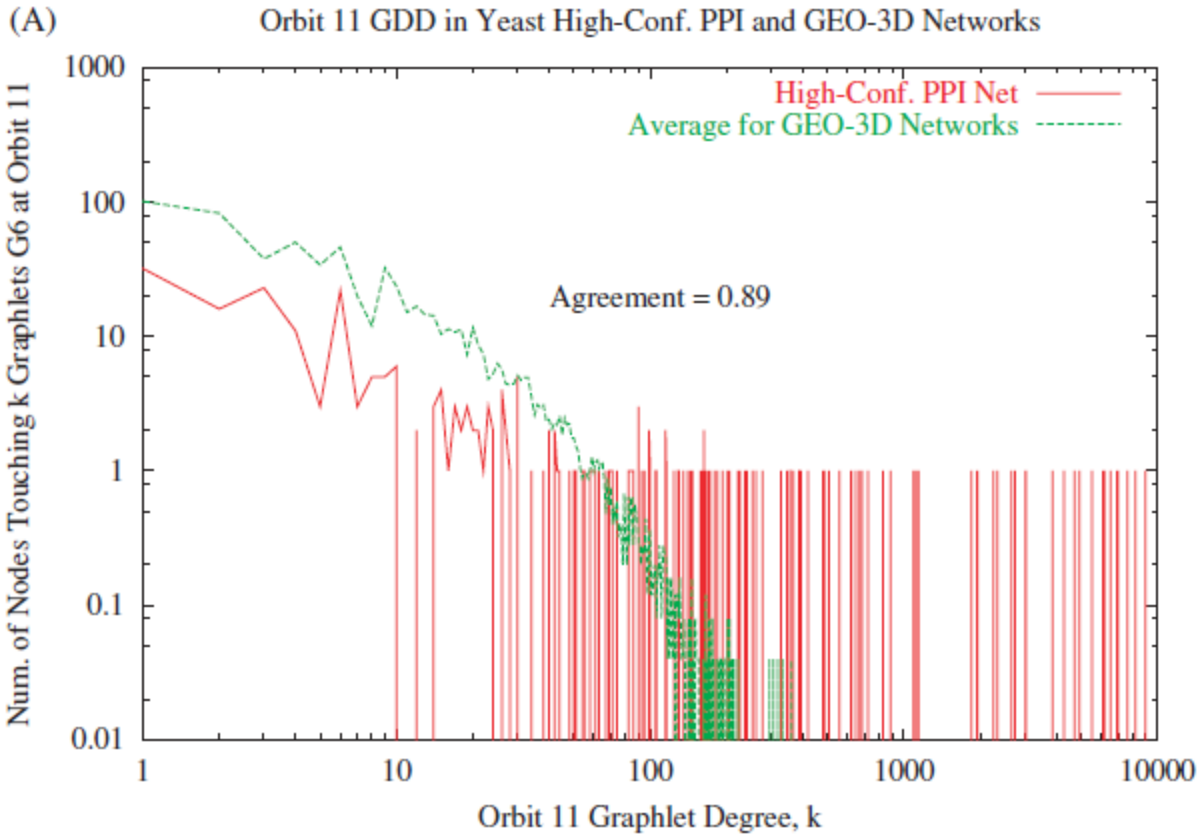
$$A_{arith}(G,H) = \frac{1}{73} \sum_{j=0}^{72} A^j(G,H)$$

or the geometric mean over $A^j(G;H)$ for all *j*:

$$A_{geo}(G,H) = (\prod_{j=0}^{72} A^j(G,H))^{1/73}$$

(N. Prˇzulj, 2010)

Example of graphlet degree distribution & agreement



(N. Prˇzulj, 2007)

Uncovering biological network function

Uncovering Biological Network Function

- Using neighborhood of proteins to infer protein functions
  - Majority rules
- Graphlets
  - Clustering method on node signatures
  - Nodes in a cluster do not need to be connected or in the same neighborhood

(T. Milenkovic, 2008)

1 objective

- Look for proteins with common biological processes, cellular components, tissue expressions in a cluster

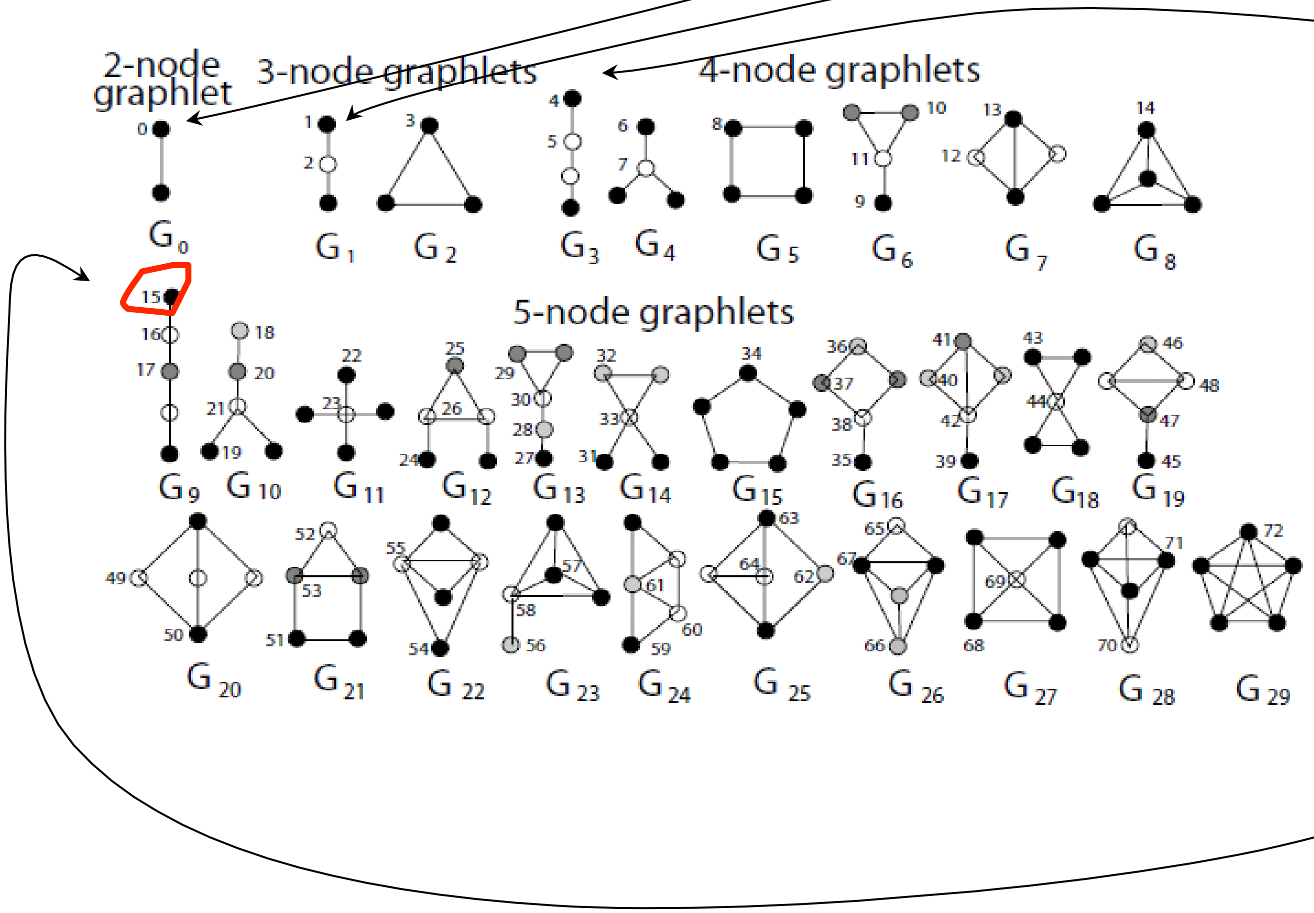
(T. Milenkovic, 2008)

Clustering

- For each vertex *u* in the network
  - Vertex *v* belongs to the cluster if the signature similarity metric for *u*, *v* > threshold

(T. Milenkovic, 2008)

Signature of a node



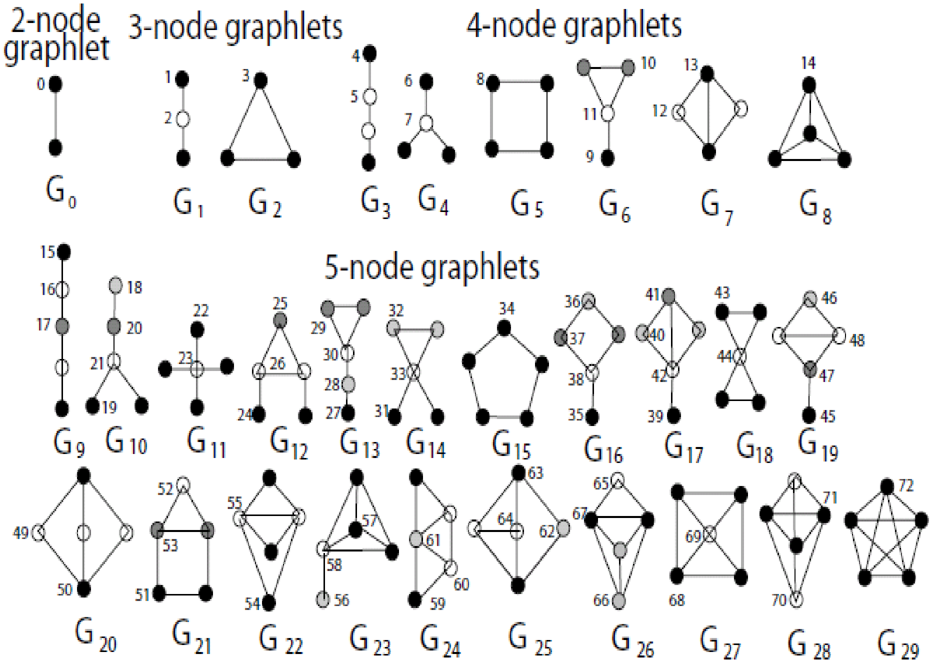
0 .. 72

0	1
1	1
2	0
3	0
4	1
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	1
16	0
•	
•	
•	
72	0

0

(T. Milenkovic, 2008)

Weight vector

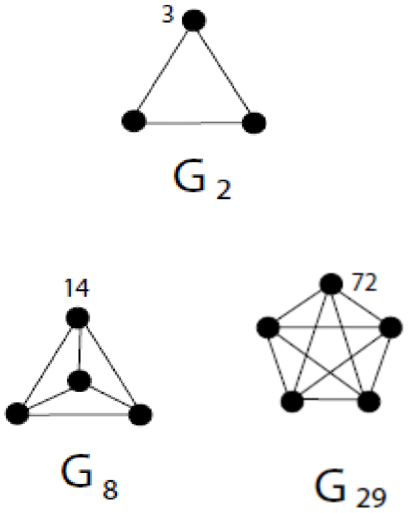


weight vector

$i \in \{0, \ldots, 72\}$

0	
1	
2	
3	
4	
5	
6	
7	
8	
i	w_i
72	

- Remove redundancy



E.g. difference in orbit 3 will affect difference in orbits such as 14, 72

(T. Milenkovic, 2008)

Weight

- Weight ($w_i \in [0, 1]$)
  - higher to important orbits (orbits that do not depend on a lot on other orbits)
  - lower to less important orbits (orbits that depend on lots of other orbits)
- Computed as

$$w_i = 1 - \frac{log(o_i)}{log(73)}.$$

where o_i is the count of orbits that affect i

  - E.g. o_{15} =4, orbit 15 is affected by 0, 1, 4, 15

(T. Milenkovic, 2008)

Distance

- Distance for orbit *i* between node *u* and *v*

$$D_i(u,v) = w_i \times \frac{\left|log\left(u_i+1\right) - log\left(v_i+1\right)\right|}{log\left(max\left\{u_i,v_i\right\}+2\right)}.$$

u$_i$– number of times node *u* touches orbit *i*

- Distance between node *u* and *v*

$$D(u,v) = \frac{\sum_{\mathrm{i=0}}^{72} \mathrm{D_i}}{\sum_{i=0}^{72} w_i}.$$

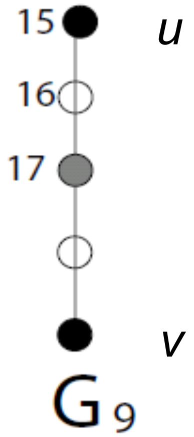
(T. Milenkovic, 2008)

Distance 2

- Signature similarity

$$S(u, v) = 1 - D(u, v).$$

- For example



- D(u,v)= 0 (same signatures)
- S(u,v) = 1

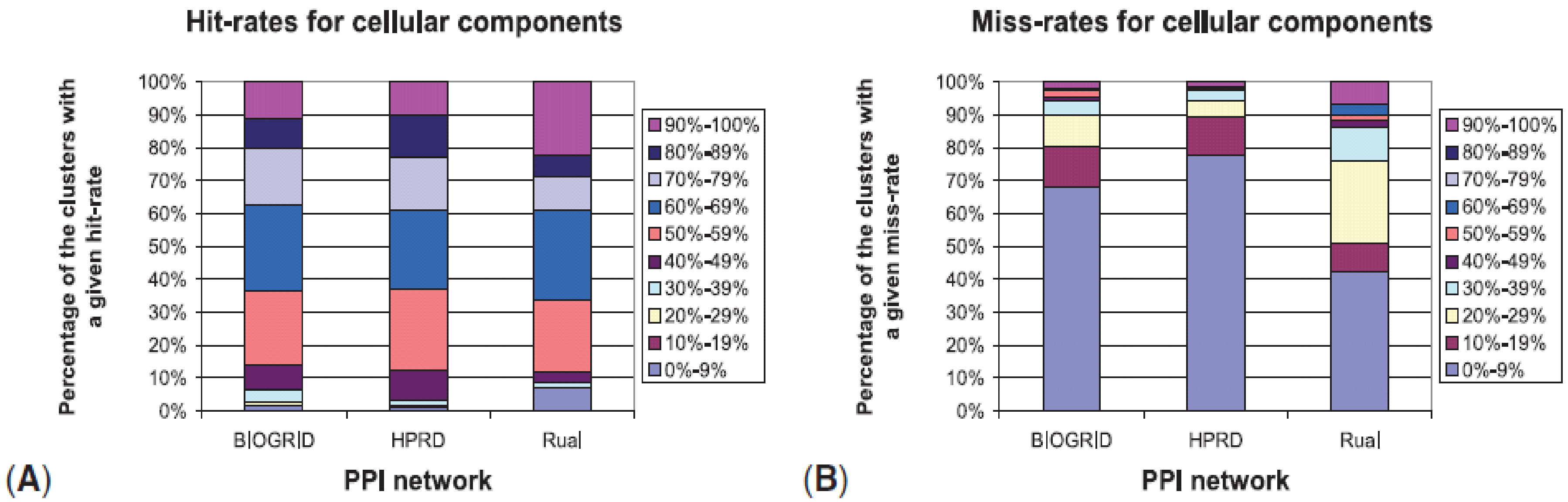
(T. Milenkovic, 2008)

Evaluation method

- Hit-rate of cluster *C*
  $Hit(C) = max\ N_p/N$
  - ▫ *Np - number of vertices in C with protein property p*
  - ▫ *N - total number of vertices in C*
- Miss-rate of cluster *C*
  $Miss(C) = U_{p/N}$
  - ▫ *Up - number of vertices in C that do not share their protein properties p with any other vertices in C*
  - ▫ *N - total number of vertices in C*

(T. Milenkovic, 2008)

Results



- Cellular components
  - Hit-rates
    - All 3 networks, 86% of clusters have hit-rates > 50%
  - Miss-rates
    - BIOGRID, HPRD, 68% of clusters have miss-rates < 10%
    - Rual, 76% of clusters have miss-rates < 29%

(T. Milenkovic, 2008)

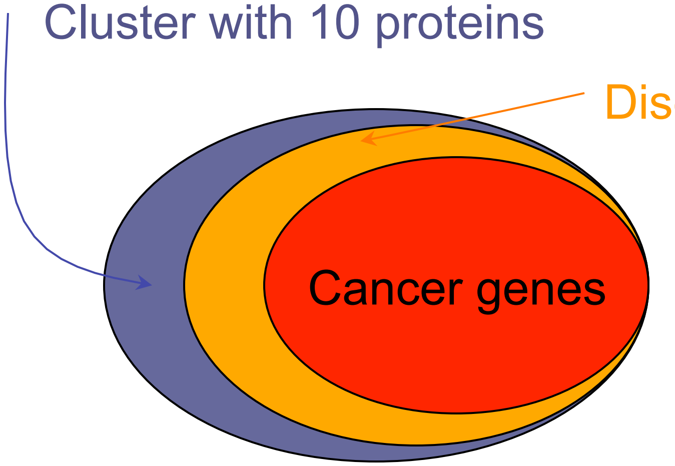
Disease genes

- Hypothesis:
  - If the topology of a network is related to function, then cancer genes might have similar graphlet degree signatures

(T. Milenkovic, 2008)

Cancer genes

- Protein of interest
  - TP53
- Look for proteins with signature similarity >= 0.95
- Resulting cluster



Cluster with 10 proteins

Disease genes: 8

Cancer genes: 6 (TP53, EP300, SRC, BRCA1, EGFR, AR)

(T. Milenkovic, 2008)

Signature vectors

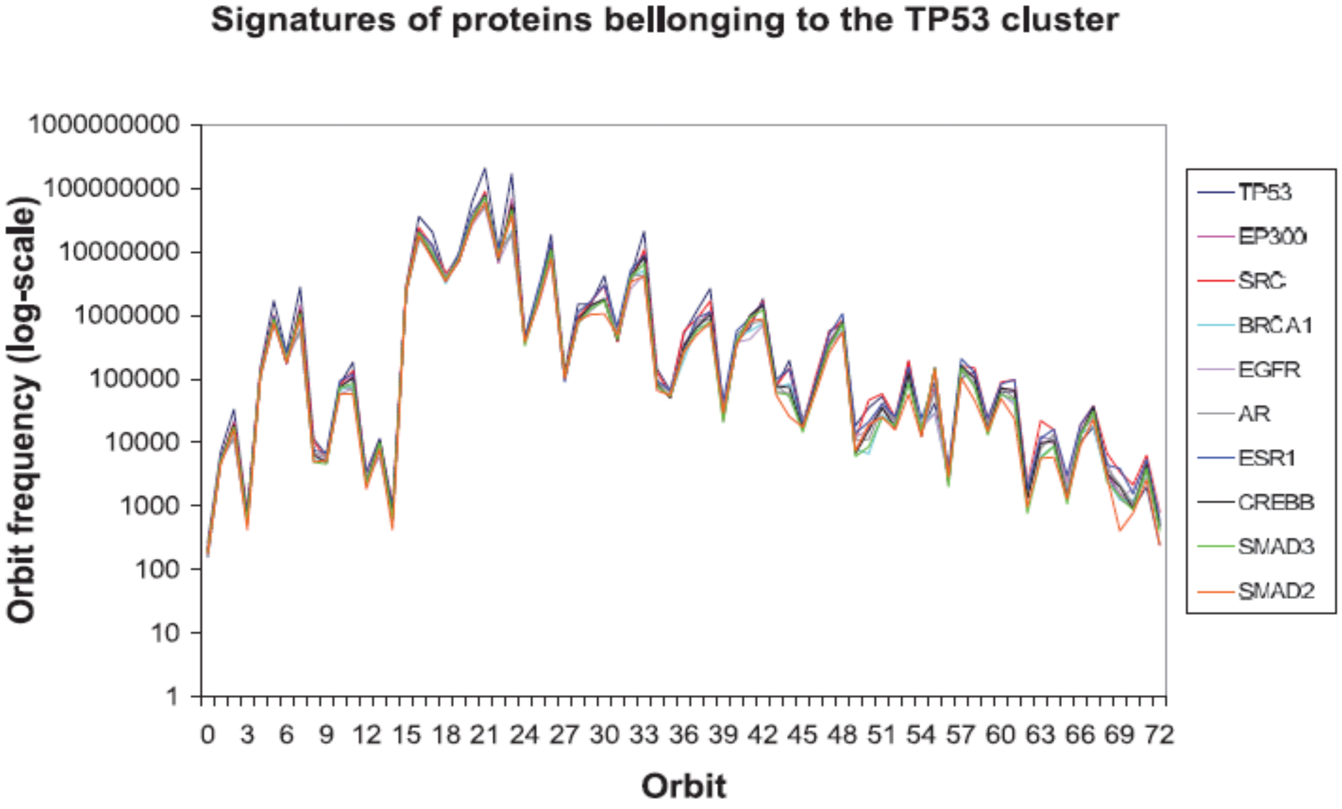


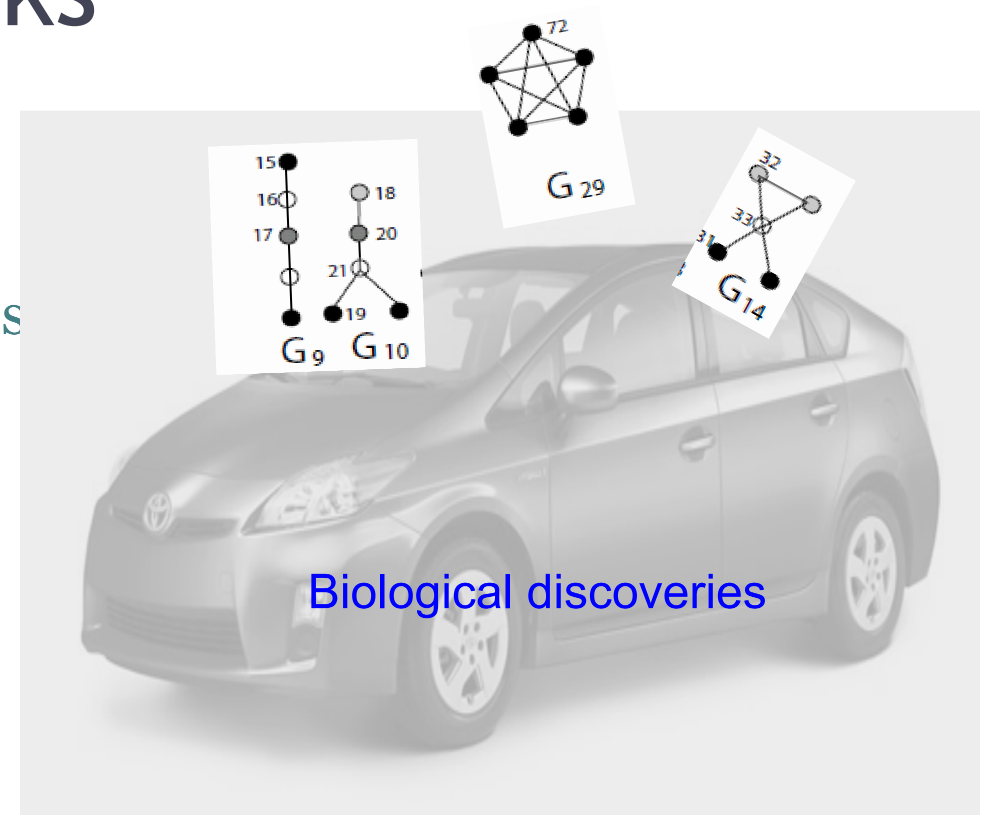
Figure 6. Signature vectors of proteins belonging to the TP53 cluster. The cluster is formed using the threshold of 0.95.

(T. Milenkovic, 2008)

Conclusion

Concluding remarks

- Graphlets can be used to
  - Compare networks
  - To infer protein functions
  - Characterize the relationship between disease and structure of networks



Biological discoveries

http://www.toyota.ca/toyota/en/vehicles/prius/gallery

References

- (H. Jeong et al., 2001) H. Jeong, S. P. Mason, A.-L. Barab´asi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature Brief Communications*, 411:41–42, May 2001.
- (R. Milo et al, 2002) R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science, 298* (5594):824–827, 2002.
- (H. C. Causton et al., 2003) H. C. Causton, J. Quackenbush, and A. Brazma. *Microarray Gene Expression Data Analysis, chapter Introduction. Blackwell Publishing,* 2003.
- (A.-L. Barab´asi et al., 2004) A.-L. Barab´asi and Z. N. Oltvai. Network biology: understanding the cell' s functional organization. Nature Reviews Genetics, 5:101–113, February 2004.
- (N. Prˇzulj et al., 2004) N. Prˇzulj, D. Wigle, and I. Jurisica. Functional topology in a network of protein interactions. *Bioinformatics, 20(3):340–348, 2004.*
- (N. Prˇzulj et al. 2004b) N. Prˇzulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics, 20(18):3508–3515, 2004.*

References 2

- (N. Prˇzulj, 2005) N. Prˇzulj. *Analyzing large biological networks: protein-protein interactions example. PhD thesis, University of Toronto, 2005.*
- (N. Prˇzulj, 2006) N. Przulj. *Knowledge discovery in proteomics, chapter Graph theory* analysis of protein-protein interactions. Chapman and Hall CRC. Mathematical biology and medicine series. CRC Press Taylor and Francis. Group, 2006.
- (O. Mason et al., 2007) O. Mason and M. Verwoerd. Graph theory and networks in biology. *Systems biology, IET, 1(2):89–119, March 2007.*
- (N. Prˇzulj, 2007) N. Prˇzulj. Biological network comparison using graphlet degree distribution. *Bioinformatics, 23(2):e177–e183, 2007.*
- (N. Prˇzulj, 2010) N. Prˇzulj. Erratum Biological network comparison using graphlet degree distribution. *Bioinformatics, 26(6):853-854, 2010.*
- (M. T. Landi et al., 2008) M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen, S. E. Murphy, P. Yang, A. C. Pesatori, D. Consonni, P. A. Bertazzi, S. Wacholder, J. H. Shih, N. E. Caporaso, and J. Jen. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS one, 3(2),* 2008.

References 3

- (T. Milenkovic, 2008) T. Milenkovic and N. Przulj . Uncovering Biological Network Function via Graphlet Degree Signatures. *Cancer Informatics, 6 257-273, 2008.*