Special-topic lecture bioinformatics: Mathematics of Biological Networks

Leistungspunkte/Credit points: 5 (V2/Ü1)

This course is taught in English language.

The material (from books and original literature) are provided online at the course website:

http://gepard.bioinformatik.uni-saarland.de/teaching/ss-2014/stl-bioinformatics-mathcellnet-ss14

Topics to be covered:

This course will enter into details of selected topics on the topology of biological networks.

Tutorial

We will handout 6 **bi-weekly assignments**.

Groups of up to two students can hand in a solved assignment.

Send your **solutions** by e-mail to the responsible tutors :

Maryam Nazarieh (#1 - #3) and Thorsten Will (#4 - #6)

until the time+date indicated on the assignment sheet.

The weekly **tutorial** on Tuesday 12.45 am – 1.30 pm (same room) will discuss the assignment solutions.

On demand, the tutors may also give some advice for solving the new assignments.

Schein / certification conditions

The successful participation in the lecture course („Schein“)
will be certified upon fulfilling

- Schein condition 1 (> 50% of the points for the assignments)

- and upon passing the **final written exam** at the end of the semester

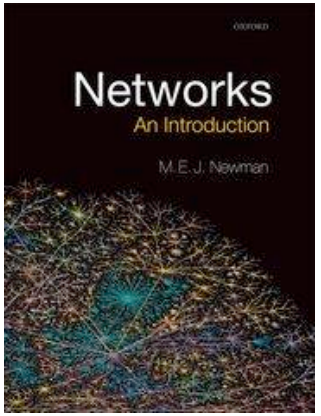
The **grade** on your „Schein“ equals that of your final exam.

Everybody who took the final exam (and passed it or did not pass it)
and those who have missed the final exam
can take the **re-exam** at the beginning of WS14/15.

The better grade counts! But there will no second re-exam.

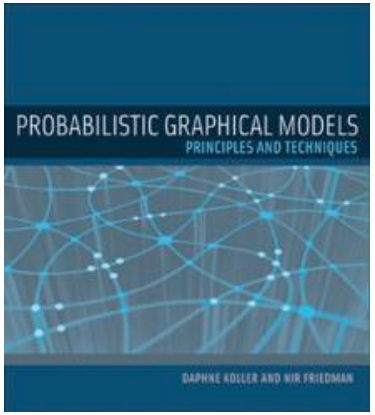
Lecture material

Lectures 1-6 follow this book by Mark Newman / Oxford Univ Press



- Chapter 7: measures and metrics
- Chapter 11: matrix algorithms and graph partitioning
- Chapter 17: epidemics on networks

Chapter 7-10/12 follow this book by Daphne Koller & Nir Friedman /MIT Press



- Chapter X:
- Chapter Y:
- Chapter Z:

You can find both books in the **CS library**.

Lectures 11/13-15 introduce modern methods
to reconstruct gene-regulatory networks

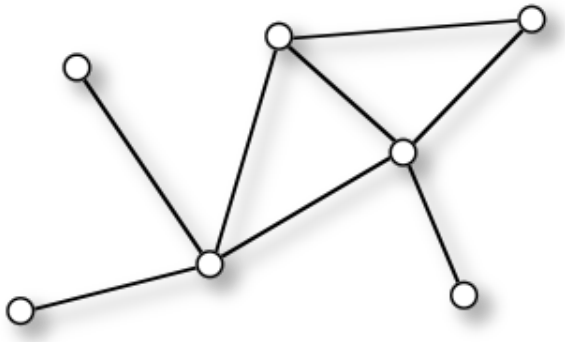
Some Graph Basics

Network <=> Graph

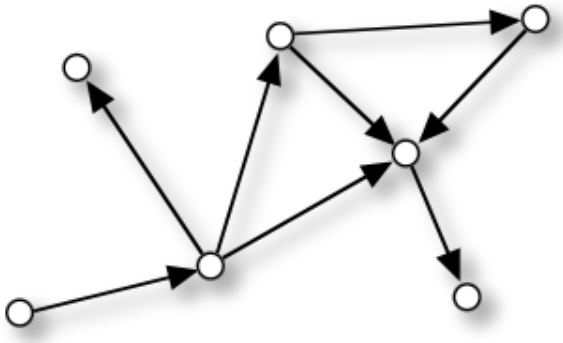
Formal **definition**:

A **graph** *G* is an ordered pair (*V, E*) of a set *V* of **vertices** and a set *E* of **edges**.

G = (V, E)



undirected graph



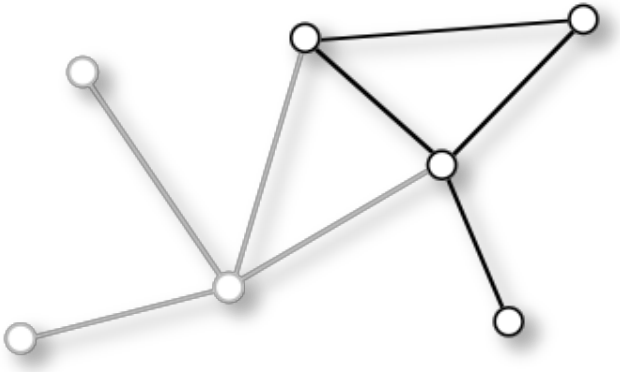
directed graph

If $E = V^{(2)}$ => fully connected graph

Graph Basics II

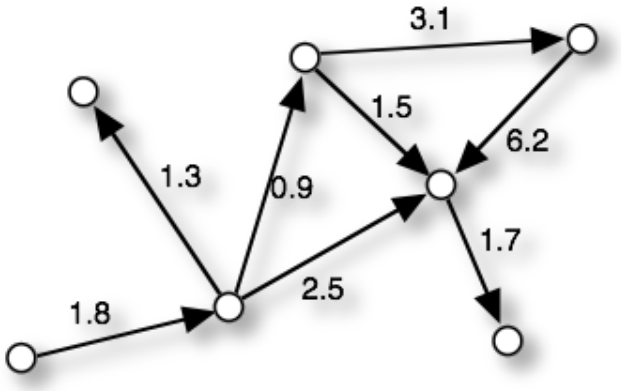
Subgraph:

$G' = (V', E')$ is a subset of $G = (V, E)$



Weighted graph:

Weights assigned to the edges



Walk the Graph

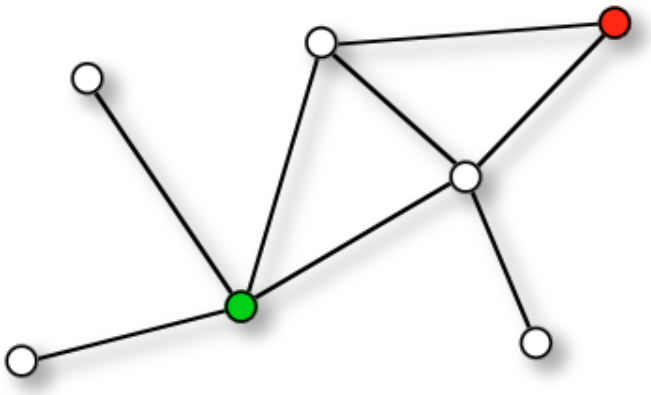
Path = sequence of connected vertices
start vertex => internal vertices => end vertex

Two paths are **independent** (internally vertex-disjoint),
if they have no internal vertices in common.

Vertices *u* and *v* are **connected**, if there exists a path from *u* to *v*.
otherwise they are disconnected

Trail = path, in which all edges are distinct

Length of a path = number of vertices || sum of the edge weights



There is an infinite number of paths connecting the green to the red vertex.

The shortest paths have length = 2.

Four trails go from the green to the red vertex.

Two of them are independent.

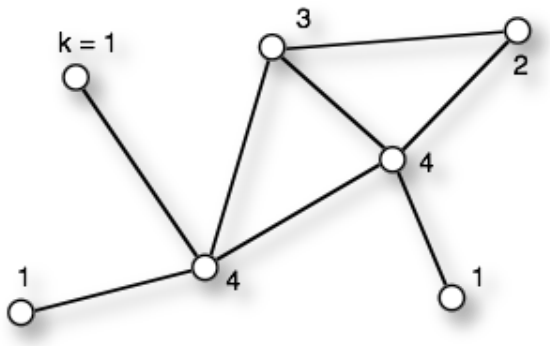
Local Connectivity: Degree

Degree *k* of a vertex = number of edges at this vertex

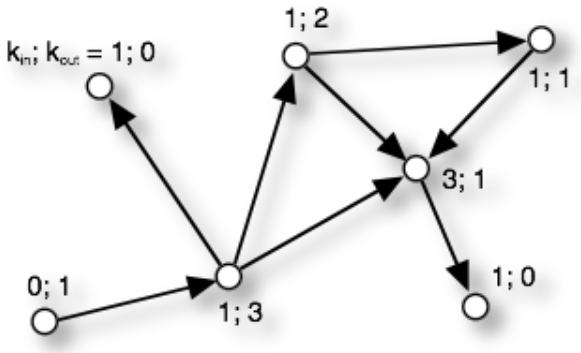
Directed graph => distinguish k_{in} and k_{out}

Degree distribution *P(k)* = fraction of nodes with *k* connections

$$P(k) = \frac{n_k}{N}$$



k	0	1	2	3	4
P(k)	0	3/7	1/7	1/7	2/7



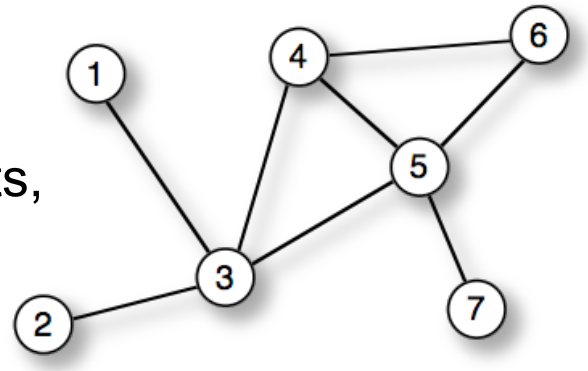
k	0	1	2	3
$P(k_{in})$	1/7	5/7	0	1/7
$P(k_{out})$	2/7	3/7	1/7	1/7

Graph Representation: e.g. by adjacency matrix

Adjacency matrix is a N x N matrix
with entries M_{uv}

M_{uv} = weight when edge between u and v exists,
0 otherwise

→ symmetric for undirected graphs



	1	2	3	4	5	6	7
1	–	0	1	0	0	0	0
2	0	–	1	0	0	0	0
3	1	1	–	1	1	0	0
4	0	0	1	–	1	1	0
5	0	0	1	1	–	1	1
6	0	0	0	1	1	–	0
7	0	0	0	0	1	0	–

+ fast *O(1)* lookup of edges
– large memory requirements
– adding or removing nodes is expensive

Note: very convenient in programming languages that support sparse multi-dimensional arrays
=> Perl

Measures and Metrics

" Which are the most important or central vertices in a network? "

Examples of

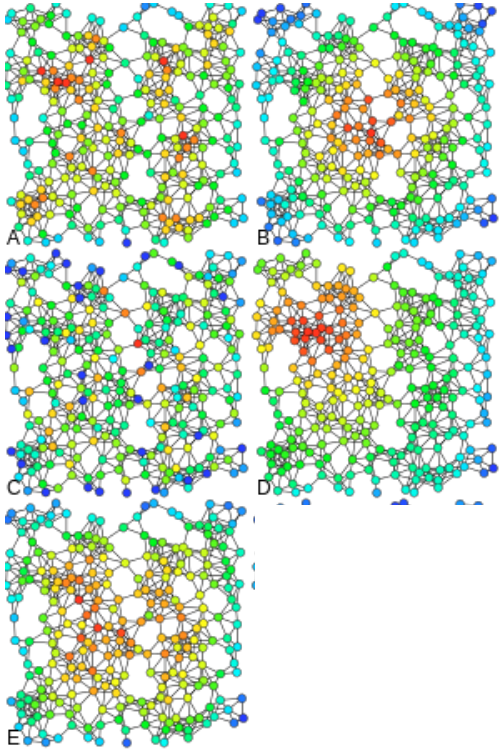
A) Degree centrality,

B) Closeness centrality,

C) Betweenness centrality,

D) Eigenvector centrality,

E) Katz centrality,



www.wikipedia.org

Degree centrality

Perhaps the simplest centrality measure in a network is the **degree centrality** that is simply equal to the **degree** of each vertex.

E.g. in a **social network**, individuals that have many connections to others might have

- more **influence**,
- more **access to information**,
- or more **prestige** than those individuals who have fewer connections.

A natural extension of the simple degree centrality is **eigenvector centrality**.

Towards Eigenvector Centrality

Let us start by defining the **centrality** of vertex x_i as the sum of the centralities of all its neighbors:

$$x_i' = \sum_j A_{ij} x_j$$

where A_{ij} is an element of the adjacency matrix.
(This equation system must be solved recursively until convergence.)

We can also write this expression in matrix notation as

x' = A x where **x** is the vector with elements x_i .

Repeating this process to make better estimates gives after *t* steps the following vector of centralities:

$$\mathbf{x}(t) = \mathbf{A}^t \mathbf{x}(0)$$

Eigenvector Centrality

Now let us write **x**(0) as a linear combination of the eigenvectors $\mathbf{v}_i$ of the (quadratic) adjacency matrix[1]

$$\mathbf{x}(0) = \sum_i c_i \mathbf{v}_i \qquad \text{with suitable constants } c_i$$

Then $\mathbf{x}(t) = A^t \sum_i c_i \mathbf{v}_i = \sum_i c_i \, k_i{}^t \, \mathbf{v}_i = k_1{}^t \sum_i c_i \left[\frac{k_i}{k_1}\right]^t \mathbf{v}_i$

where the k_i are the eigenvalues of **A** and k_1 is the largest of them.

(remember **A x** = λ **x** from linear algebra for each eigenvector **x**)

Since $k_i \,/\, k_1 < 1$ for all $i \neq j$, all terms in the sum decay exponentially as t becomes large.

In the limit $t \to \infty$, we get $\mathbf{x}(t) = c_1 \, k_1{}^t \, \mathbf{v}_1$

[1] Remember from linear algebra that a quadratic matrix with full rank can be diagonalized.

Eigenvector Centrality

This limiting vector of the eigenvector centralities is simply proportional to the leading eigenvector of the adjacency matrix.

Equivalently, we could say that the centrality **x** satisfies

$$\mathbf{A}\,\mathbf{x} = k_1\,\mathbf{x}$$

This is the **eigenvector centrality** first proposed by Bonacich (1987).

The centrality x_i of vertex i is proportional to the sum of the centralities of its neighbors:

$$x_i = {k_1}^{-1} \sum_j A_{ij} x_j$$

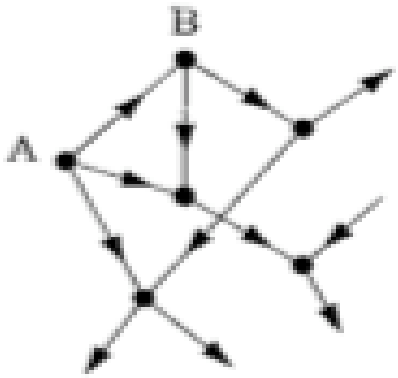
This has the nice property that the centrality can be large either because a vertex has many neighbors or because it has important neighbors (or both).

Problems of the Eigenvector Centrality

The eigenvector centrality works best for undirected networks.

For directed networks, certain complications can arise.

In the figure on the right,
vertex A will have eigenvector centrality zero.
Hence, vertex B will also have centrality zero.



Katz Centrality

One solution to the issues of the Eigenvector Centrality is the following:

We simply give each vertex a small amount of centrality "for free", regardless of its position in the network or the centrality of its neighbors.

→ we define $x_i = \alpha \sum_j A_{ij} x_j + \beta$ where α and β are positive constants.

In matrix terms, this can be written as $\mathbf{x} = \alpha \mathbf{A}\mathbf{x} + \beta \mathbf{1}$

where **1** is the vector $(1,1,1,\ldots)^T$. By rearranging for **x** we find

$$\mathbf{I}\,\mathbf{x} - \alpha\,\mathbf{A}\,\mathbf{x} = \beta\,\mathbf{1} \qquad \text{(where we used } \mathbf{I}\,\mathbf{x} = \mathbf{x}\text{)}$$
$$(\mathbf{I} - \alpha\,\mathbf{A})\,\mathbf{x} = \beta\,\mathbf{1}$$
$$(\mathbf{I} - \alpha\,\mathbf{A})^{-1}\,(\mathbf{I} - \alpha\,\mathbf{A})\,\mathbf{x} = (\mathbf{I} - \alpha\,\mathbf{A})^{-1}\,\beta\,\mathbf{1}$$
$$\mathbf{x} = \beta\,(\mathbf{I} - \alpha\,\mathbf{A})^{-1}\,\mathbf{1}$$

When setting β =1, we get the **Katz centrality** (1953) $\mathbf{x} = (\mathrm{I} - \alpha\,\mathbf{A})^{-1}\,\mathbf{1}$

Computing the Katz Centrality

The Katz centrality differs from the ordinary eigenvector centrality by having a **free parameter** α, which governs the balance between the eigenvector term and the constant term.

However, inverting a matrix on a computer has a complexity of $O(n^3)$ for a graph with n vertices.

This becomes prohibitively expensive for networks with more than 1000 nodes or so.

It is more efficient to make an initial guess of x and then repeat

$$\mathbf{x}' = \alpha \mathbf{A}\mathbf{x} + \beta \mathbf{1}$$

many times. This will converge to a value close to the correct centrality.

A good test for convergence is to make two different initial guesses and run this until the resulting centrality vectors agree within some small threshold.

Towards PageRank

The Katz centrality also has one feature that can be undesirable.

If a vertex with high Katz centrality has edges pointing to many other vertices, then all those vertices also get high centrality.

E.g. if a Wikipedia page points to my webpage, my webpage will get a centrality comparable to Wikipedia!

But Wikipedia of course also points to many other websites, so that its contribution to my webpage “should” be relatively small because my page is only one of millions of others.

-> we will define a variation of the Katz centrality in which the centrality I derive from my network neighbors is proportional to their centrality divided by their out-degree.

PageRank

This centrality is defined by

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

At first, this seems problematic if the network contains vertices with zero outdegree.

However, this can easily be fixed by setting $k_j^{out} = 1$ for all such vertices.

In matrix terms, this equation becomes

$$\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{D}^{-1}\, \mathbf{x} + \beta\, \mathbf{1}$$

where **1** is the vector $(1,1,1,\ldots)^T$ and **D** the diagonal matrix with $D_{ij} = max(k_j^{out}, 1)$

PageRank

By rearranging we find that

$$\mathbf{x} = \beta\,(\mathbf{I} - \alpha\,\mathbf{A}\,\mathbf{D}^{-1})^{-1}\,\mathbf{1}$$

Because β plays the same unimportant role as before, we will set $\beta = 1$.

Then we get $\mathbf{x} = (\mathbf{I} - \alpha\,\mathbf{A}\,\mathbf{D}^{-1})^{-1}\,\mathbf{1} = \mathbf{D}\,(\mathbf{D} - \alpha\,\mathbf{A})^{-1}\,\mathbf{1}$

This centrality measure is commonly known as **PageRank**,
using the term used by Google.

PageRank is one of the ingredients used by Google to determine the ranking of the answers to your queries.

α is a free parameter and should be chosen less than 1. (Google uses 0.85).

Hubs and Authorities

So far we have considered measures that assign high centrality to a vertex if those vertices that point to it have high centrality too.

However, in some networks it is appropriate also to accord a vertex high centrality if it **points** to others with high centrality.

E.g. a review article pointing at many important papers in one research field may be a useful source of information.

Authorities are nodes that contain useful information on a topic of interest.

Hubs are nodes that tell us where the best authorities can be found.

An authority may also be a hub, and vice versa.

Hubs and Authorities

Kleinberg developed this into a centrality algorithm called Hyperlink-induced topic search (HITS).

The HITS algorithm gives each vertex *i* in a network an **authority centrality** x_i and a **hub centrality** y_i .

A vertex with high authority centrality is pointed to by many hubs, i.e. by many other vertices with high hub centrality.

A vertex with high hub centrality points to many vertices with high authority centrality.

Thus, an important scientific paper (in the authority sense) would be one that is cited in many important reviews (in the hub sense).

An important review is one that cites many important papers.

Authority and Hub Centralities

Kleinberg defined the **authority centrality** of a vertex to be proportional to the sum of the hub centralities of the vertices that point to it

$x_i = \alpha \sum_j A_{ij} y_j$ where α is a constant.

Similarly the **hub centrality** of a vertex is proportional to the sum of the authority centralities of the vertices it points to:

$y_i = \beta \sum_j A_{ji} x_j$ with another constant β

Note that the indices of the matrix element A_{ji} are swapped around in this second equation.

These equations can be written as $\mathbf{x} = \alpha\, \mathbf{A}\, \mathbf{y}$ and $\mathbf{y} = \beta\, \mathbf{A}^t\, \mathbf{x}$

Or, combining the two, $\mathbf{A}\, \mathbf{A}^t\, \mathbf{x} = \lambda\, \mathbf{x}$, $\mathbf{A}^t\, \mathbf{A}\, \mathbf{y} = \lambda\, \mathbf{y}$

Closeness centrality

An entirely different measure of centrality is provided by the **closeness centrality.**

Suppose d_{ij} is the length of a geodesic path (i.e. the shortest path) from a vertex *i* to another vertex *j*.
Here, length means the number of edges along the path.

Then, the mean **geodesic distance** from *i,* averaged over all vertices *j* in the network is

$$l_i = \frac{1}{n}\sum_j d_{ij}$$

The mean distance l_i is not a centrality measure in the same sense as the other centrality measures.

It gives *low* values for more central vertices and high values for less central ones.

Closeness centrality

The inverse of l_i is called the **closeness centrality** C_i

$$C_i = \frac{1}{l_i} = \frac{n}{\sum_j d_{ij}}$$

It has become popular in recent years to rank film actors according to their closeness centrality in the network of who has appeared in films with who else.

Using data from www.imdb.com the largest component of the network includes more than 98 % of about half a million actors.

Closeness centrality

The highest closeness centrality of any actor is 0.4143 for Christopher Lee.

The second highest centrality has Donald Pleasence (0.4138).

The lowest value has the Iranian actress Leia Zanganeh (0.1154).

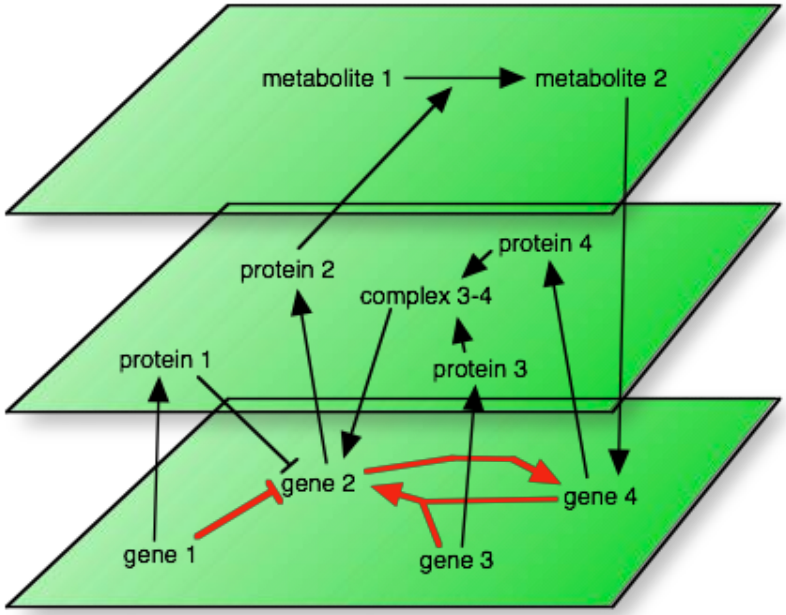
→ the closeness centrality values are crammed in a very small interval [0,0.4143]

Other centrality measures including degree centrality and eigenvector centrality typically don‘t suffer from this problem. They have a wider dynamic range.

Pictures from wikipedia

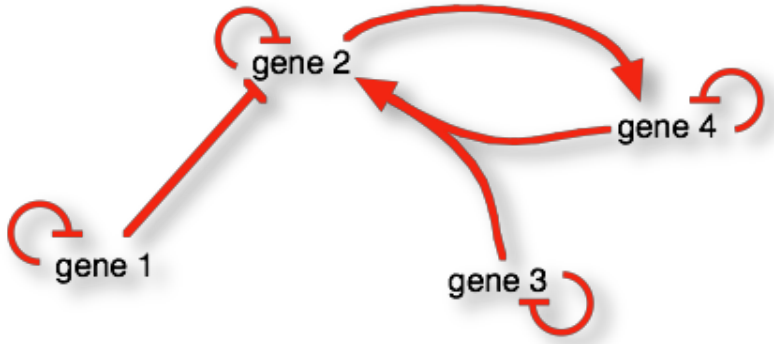
Gene-regulatory networks (GRNs)

Biological regulation
via proteins and metabolites



<=> Projected gene-regulatory network

<=>



Remember:
genes do not interact directly

Centrality of Genes in Gene Regulatory Networks

Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks

Dirk Koschützki[1,2] and Falk Schreiber[1,3]

Gene Regulation and Systems Biology 2008:2 193–201

Falk Schreiber

Authors analyzed centralities within the gene regulatory network (GRN) of *Escherichia coli*.

The GRN network was constructed based on the transcriptional regulatory interactions of genes in **RegulonDB**, Version 5.5 (Salgado et al. (2006)).

Genes are represented by **vertices** and transcriptional **regulatory interactions** between genes are modelled as **edges**, a common approach to model GRNs.

The interactions between genes represent transcriptional control of transcription factors on the transcription of regulated genes.

The resulting network consisted of 1250 vertices and 2515 edges.

Subgraph motifs in biological networks

Several **motifs** (overrepresented subgraphs) have been identified in all kinds of biological networks.

The best studied motif is the feed-forward loop (FFL) motif. Its functional properties have been analyzed in detail theoretically and experimentally especially in gene regulatory networks (Shen-Orr et al. (2002).

Different motifs occurring in a human cellular signalling network were analysed by Awan et al. (2007).
They discovered that genes which are related to cancer are enriched in the target vertices of several motifs and that cell mobility genes are enriched in the source vertices of motifs.

Motif-based centrality

Given: a graph G, a motif M and
the corresponding motif match set MS_G.

Define the motif-based centrality C_{mb} that assigns
to every vertex v the number of matches
the vertex v occurs in.

E.g. the vertex $v01$ in the graph shown in Fig. 2
occurs in 2 matches of the FFL motif shown
in Fig. 3. Therefore $C_{mb}(v01) = 2$.

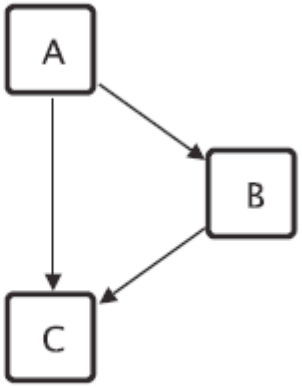


Figure 3. The FFL motif with roles.

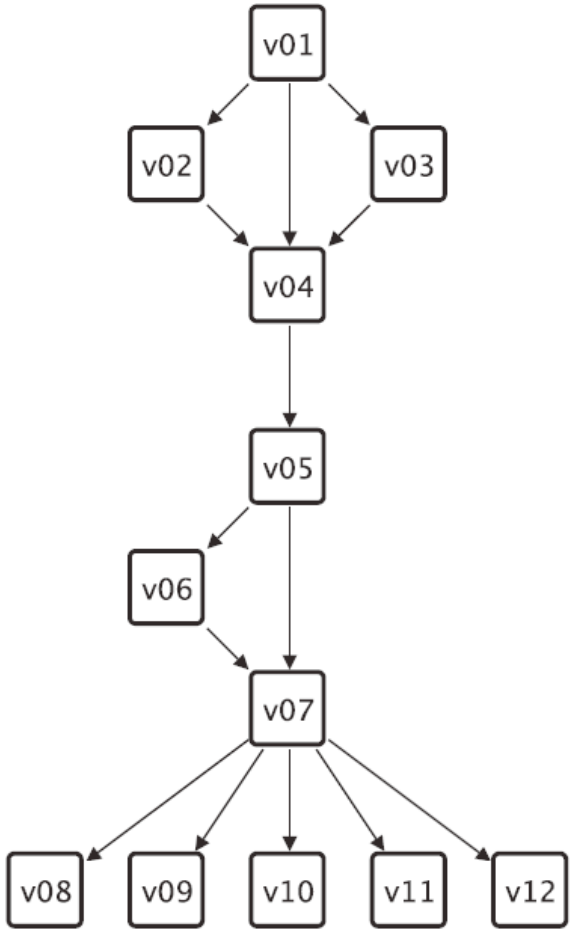


Figure 2. An example graph used to explain different centrality measures.

Gene Regulation and Systems Biology 2008:2 193–201

Motif-based centralities

Two extensions of this motif-based centrality exist:
- motif-based centrality with roles and
- motif-based centrality with classes.

Vertices of motifs may represent different functions.

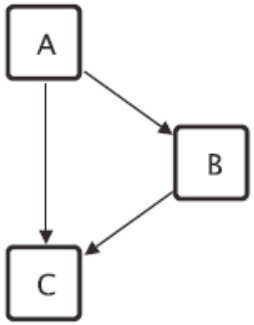


Figure 3. The FFL motif with roles.

E.g. in the gene regulatory network context 3 different functions of the vertices of the feed forward loop (FFL) motif can be identified:

(1) the vertex at the top is the **master regulator**, this vertex regulates the other two vertices;

(2) the vertex on the right side is the **intermediate regulator**, it is regulated by the master regulator and itself regulates together with the master regulator the vertex at the bottom

Motif-based centralities

(3) the vertex at the bottom of the drawing is regulated by both other vertices and is therefore called the **regulated vertex**.

Such different functions of vertices within motifs are called **roles** and 3 roles can be assigned to the vertices of the FFL motif.

The **motif-based centrality with roles** C_{mbr} restricts the number of counted matches to those matches where the vertex occurs in the match with the role under consideration.

Chain of motifs

Using the previously introduced concepts we can extend the motif-based centrality method further.

By assigning the same role to similar vertices of a group of similar motifs we can establish a centrality based on a class (or group) of motifs.

Consider, for example, a group of **chains** (see Fig. 4), where all vertices at the start of such chains have a similar characteristic (no incoming edges) and all vertices at the end have another similar characteristic (no outgoing edges).

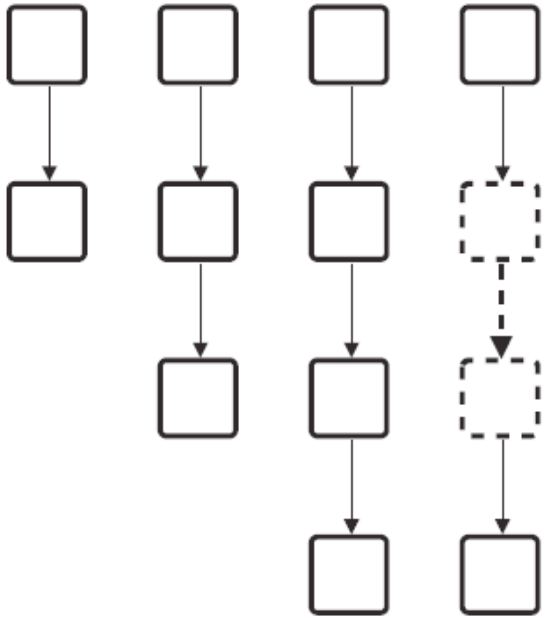


Figure 4. Several motifs of the chain motif class.

Motif classes

For gene regulatory networks several motif classes are known.

For example, the **regulatory chain motif class**, as in the example above, consists of a set of chains of 3 or more regulators in which one regulator regulates another regulator, which in turn regulates a third one and so forth.

In the motif class **single input motif** (SIM) a set of vertices is exclusively regulated by a single vertex.

The motif-based centrality with classes C_{mbc} therefore is the sum of motif-based centralities with roles C_{mbr} for the same role in similar or related motifs.

Radiality and integration centralities

These two centrality measures are related to the closeness centrality.

Given the distance matrix $D = (dist(i,j))$ between all vertices,
one can define the **reverse distance matrix**

$$RD_{ij} = diameter(G) + 1 - D_{ij}$$

where *diameter(G)* is the highest distance value of the graph.

On the basis of this, the **radiality** is defined as

$$C_{rad}(i) = \frac{\sum_{i \neq j} RD_{ij}}{n-1}$$

and **integration** is defined as $C_{int}(j) = \frac{\sum_{i \neq j} RD_{ij}}{n-1}$

A vertex with high radiality value can easily reach other vertices.
A vertex with a high integration value is easily reachable from other vertices.
Both measures are shortest-pathway based measures.

Comparison of centrality measures

chains: motif-based centrality for the chain class

fflA, fflB, fflC: motif-based centralities for the FFL motif with roles

Table 1. The centrality values that are discussed in this paper computed for the example graph in Figure 2.

	ideg	odeg	par	parR	kat	katR	spb	int	rad	chains	fflA	fflB	fflC	fflSum
v01	0.00	3.00	0.04	0.19	0.00	37.64	0.00	0.00	2.18	47.00	2.00	0.00	0.00	2.00
v02	1.00	1.00	0.05	0.07	0.95	12.32	0.00	0.36	1.45	15.00	0.00	1.00	0.00	1.00
v03	1.00	1.00	0.05	0.07	0.95	12.32	0.00	0.36	1.45	15.00	0.00	1.00	0.00	1.00
v04	3.00	1.00	0.12	0.16	4.66	11.97	24.00	1.09	1.82	14.00	0.00	0.00	2.00	2.00
v05	1.00	2.00	0.14	0.16	5.37	11.60	28.00	1.18	2.09	13.00	1.00	0.00	0.00	1.00
v06	1.00	1.00	0.10	0.08	6.05	5.46	0.00	1.18	1.73	6.00	0.00	1.00	0.00	1.00
v07	2.00	5.00	0.18	0.14	12.75	4.75	30.00	1.55	1.82	5.00	0.00	0.00	1.00	1.00
v08	1.00	0.00	0.07	0.03	13.07	0.00	0.00	1.36	0.00	0.00	0.00	0.00	0.00	0.00
v09	1.00	0.00	0.07	0.03	13.07	0.00	0.00	1.36	0.00	0.00	0.00	0.00	0.00	0.00
v10	1.00	0.00	0.07	0.03	13.07	0.00	0.00	1.36	0.00	0.00	0.00	0.00	0.00	0.00
v11	1.00	0.00	0.07	0.03	13.07	0.00	0.00	1.36	0.00	0.00	0.00	0.00	0.00	0.00
v12	1.00	0.00	0.07	0.03	13.07	0.00	0.00	1.36	0.00	0.00	0.00	0.00	0.00	0.00

Abbreviations: chains: motif-based centrality for the chain class; fflA, fflB and fflC: motif-based centrality for the FFL motif with roles (different roles A, B, C; see Figure 3); fflSum: motif-based centrality for the FFL motif without roles; ideg: in-degree; int: integration; kat: Katz status index; katR: Katz status index for the reversed graph; odeg: out-degree; par: PageRank; parR: PageRank for the reversed graph; rad: radiality; spb: shortest-path betweenness.

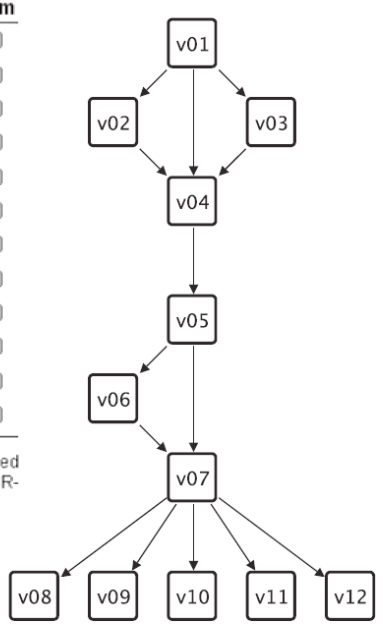


Figure 2. An example graph used to explain different centrality measures.

ideg: in-degree

odeg: out-degree

par: PageRank

parR: PageRank for the reversed graph

kat: Katz status index, katR: reversed g.

spb: shortest-path betweenness

int: integration

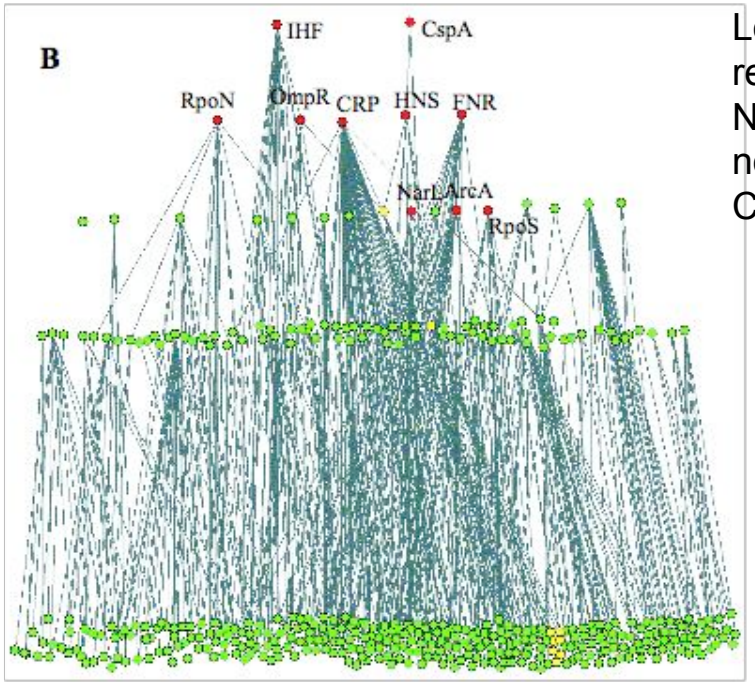
rad: radiality

Gene Regulation and Systems Biology 2008:2 193–201

Background: Hierarchical GRN of *E.coli*

Largest WCC: 325 operons
(3/4 of the complete network)

WCC = weakly connected component (ignore directions of regulation)



Lowest level: operons that code for TFs with only auto-regulation, or no TFs
Next layer: delete nodes of lower layer, identify TFs that do not regulate other operons in this layer (only lower layers)
Continue …

→ Network with all regulatory edges pointing downwards

→ a few global regulators (•) control all the details

Ma et al., *BMC Bioinformatics* **5** (2004) 199

Most central genes in *E. coli* GRN

Table 2. Names of the top 25 genes (top 2% of all genes) according to 8 best centrality measures, i.e. centralities which find a high number of global regulators within the top 2% of all genes. Global regulators according to Martínez-Antonio and Collado-Vides (2003) are highlighted in bold face. Note that in few cases were genes with the same centrality value occur they are ranked in alphabetical order. For each centrality the last row of the table shows the number of global regulators identified within the top 2% of all genes.

position	odeg	parR	katR	spb	rad	chains	ffIA	ffISum
1	**crp**	**crp**	**crp**	**hns**	**crp**	**crp**	**crp**	**crp**
2	**fnr**	**ihfAB**	**fnr**	*gadX*	**ihfAB**	**ihfAB**	**fnr**	**fnr**
3	**ihfAB**	**fnr**	**arcA**	*flhD*	**fnr**	**arcA**	**ihfAB**	**arcA**
4	**fis**	**arcA**	**ihfAB**	**fur**	**arcA**	**fnr**	**arcA**	**fis**
5	**arcA**	**phoB**	**fis**	*gadE*	**fis**	**fis**	**fis**	**narL**
6	**narL**	*lexA*	**hns**	**fis**	*gadE*	*evgA*	*modE*	**ihfAB**
7	**hns**	**cpxR**	*gadE*	**lrp**	**hns**	*ydeO*	**soxS**	**hns**
8	**fur**	**soxR**	*gadX*	*rcsAB*	**fur**	*gadE*	**hns**	**fur**
9	**lrp**	**fis**	**cspA**	**soxS**	**soxS**	**soxR**	**cpxR**	*gadX*
10	*glnG*	*evgA*	*evgA*	**fnr**	*evgA*	**soxS**	*fhlA*	*hyfR*
11	*narP*	*cysB*	*ydeO*	**cspA**	*ydeO*	*torR*	*gadE*	*marA*
12	**cpxR**	*argR*	*torR*	*caiF*	*oxyR*	*gadW*	**rob**	*flhD*
13	**phoB**	*phoP*	*gadW*	**purR**	*gadX*	*cspE*	*gadX*	*nagC*
14	*fruR*	**fur**	*cspE*	**narL**	**cspA**	**cspA**	*galR*	**soxS**
15	*modE*	*allR*	**soxS**	*marA*	**narL**	*gadX*	**fur**	*modE*
16	*fhlA*	*glnG*	**soxR**	*metJ*	*modE*	**hns**	*gntR*	*tdcA*
17	*lexA*	*sdaR*	**rob**	*malT*	**soxR**	*oxyR*	*oxyR*	*yiaJ*
18	*flhD*	*trpR*	*marA*	**arcA**	*torR*	**fur**	*tdcR*	*gutM*
19	*gadE*	*agaR*	*marR*	*glnG*	*gadW*	*modE*	*gutM*	**ompR**
20	**purR**	*gadE*	*oxyR*	**ompR**	*cspE*	**narL**	*nagC*	*srlR*
21	**soxS**	**soxS**	**fur**	*Nac*	**lrp**	**lrp**	**narL**	*galS*
22	*argR*	**hns**	*modE*	*oxyR*	*glnG*	*glnG*	**ompR**	*idnR*
23	*cysB*	**lrp**	*gutM*	*hupAB*	**phoB**	**ompR**	*srlR*	*caiF*
24	*marA*	*tyrR*	*srlR*	*argP*	*narP*	**phoB**	*argP*	*chbR*
25	*nagC*	*torR*	**narL**	*dnaA*	**ompR**	**cpxR**	*cysB*	**cpxR**
#global regs.	13	12	12	11	14	15	12	11

Abbreviations: see Table 1.

Gene Regulation and Systems Biology 2008:2 193–201

Correlation between results for different centralities

Table 3. Kendall's correlation coefficients for the centralities used in the analysis of the ***E. coli*** network.

	odeg	parR	katR	spb	rad	chains	fflA	fflSum
odeg	1	0.97	0.93	0.49	0.98	0.98	0.47	0.17
parR	0.97	1	0.92	0.48	0.96	0.96	0.46	0.16
katR	0.93	0.92	1	0.47	0.95	0.95	0.46	0.14
spb	0.49	0.48	0.47	1	0.49	0.49	0.43	0.22
rad	0.98	0.96	0.95	0.49	1	1	0.48	0.18
chains	0.98	0.96	0.95	0.49	1	1	0.48	0.18
fflA	0.47	0.46	0.46	0.43	0.48	0.48	1	0.29
fflSum	0.17	0.16	0.14	0.22	0.18	0.18	0.29	1

Abbreviations: see Table 1.

Some centralities correlate with values above 0.9 to other centralities (out-degree, PageRank, Katz status index, radiality, motif-based centrality with chain classes (chains)).

These high coefficients can be easily explained:

1101 out of 1250 vertices have an out-degree of zero. All these vertices are assigned the same centrality of nearly zero for Katz, PageRank, and the value zero for the radiality and chains.

Gene Regulation and Systems Biology 2008:2 193–201

Centralities of vertices with non-zero outdegree

Table 4 shows the pairwise correlation coefficients for the centrality values of the vertices which have a non-zero out-degree.

These coefficients show a different picture: all 5 centralities rank the remaining 149 genes differently.
Only the centrality radiality and Katz status index achieve a considerable high correlation to each other and to chains.

Table 4. Kendall's correlation coefficient for the dataset with the zero out-degree vertices removed.

	odeg	rad	katR	parR	chains
odeg	1	0.75	0.7	0.52	0.72
rad	0.75	1	0.94	0.51	0.96
katR	0.7	0.94	1	0.48	0.97
parR	0.52	0.51	0.48	1	0.5
chains	0.72	0.96	0.97	0.5	1

In conclusion, the centralities applied to the GRN rank the genes quite differently.

The motif-based centrality with chain classes is able to rank the highest number (15) of interesting genes (18 global regulators identified by Martínez-Antonio and Collado-Vides (2003)) within the top 2% of all genes.

Summary

The analysis of network topology is of interest in many different disciplines, e.g. social networks.

There exist different sorts of networks for biological cells:
Protein-protein interaction networks, gene-regulatory networks, metabolic networks, …

For the gene regulatory network of *E. coli* motif-based centrality outperforms other methods in terms of identifying the key regulatory genes.