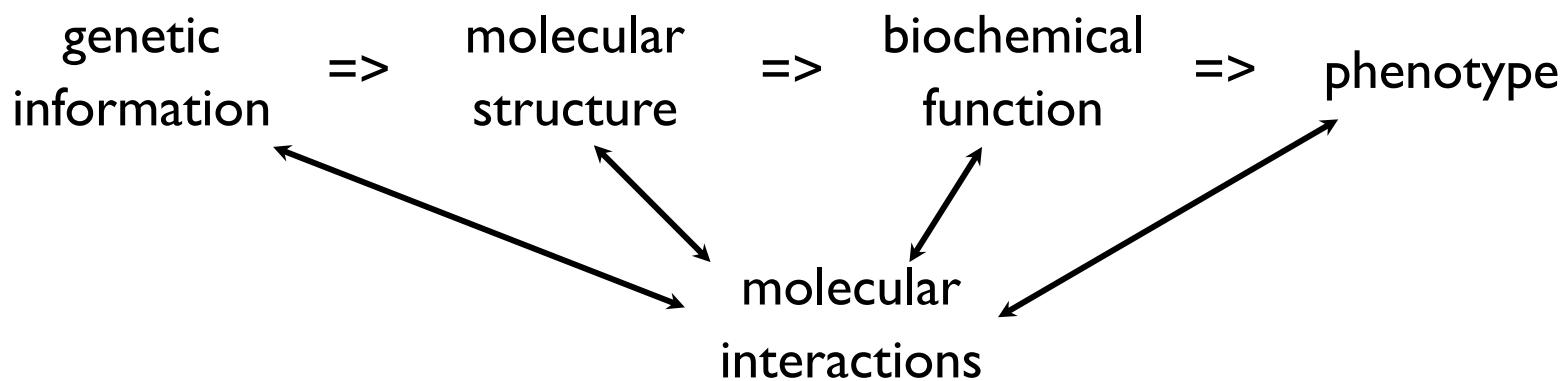


Bioinformatics III: Network View of Cell Biology

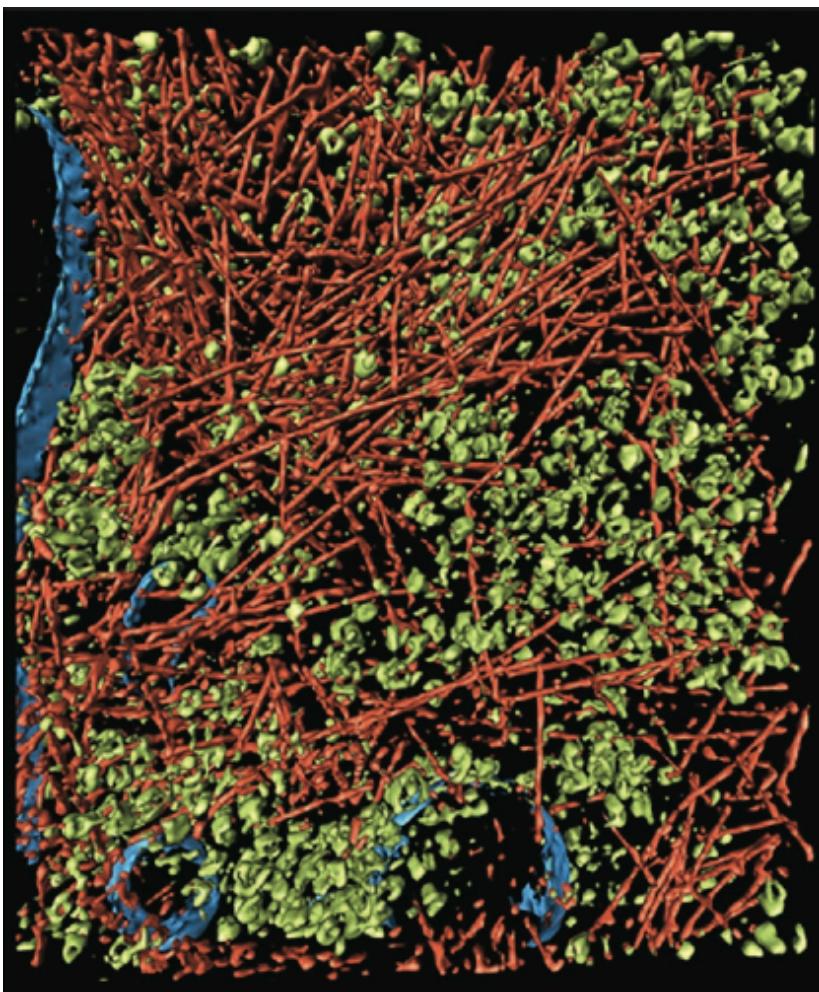
Molecular Systems Biology: "It's both + molecular interactions"



→ highly connected network of various interactions, dependencies

=> **study networks**

V1 - Introduction



Medalia et al, Science 298 (2002) 1209

A cell is a crowded environment
=> many different proteins,
metabolites, compartments, ...

At the microscopic level
=> direct two-body interactions

At the macroscopic level
=> complex behavior

Can we understand the behavior
from the interactions?

=> Connectivity

Lecture – Overview

Protein complexes: **spatial** structure

=> experiments, spatial fitting, docking

Protein association:

=> interface properties, spatial simulations

Protein-Protein-Interaction Networks: **pairwise** connectivity

=> data from experiments, quality check

PPI: static network **structure**

=> network measures, clusters, modules, ...

Gene regulation: cause and **response**

=> Boolean networks

Metabolic networks: steady state of **large networks**

=> FBA, extreme pathways

Metabolic networks / signaling networks: **dynamics**

=> ODEs, modules, stochastic effects

Lecture – Table of contents - Chapters

1. Introduction – Networks in Biological Cells
2. Structures of Protein Complexes and Subcellular Structures
3. Analysis of protein-protein binding
4. Algorithms on mathematical graphs
5. Protein-Protein Interaction Networks – Pairwise Connectivity
6. Protein-Protein Interaction Networks – Structural Hierarchies
7. Protein-DNA interactions
8. Gene Expression and Protein Synthesis
9. Gene Regulatory Networks
10. Regulatory Noncoding RNA
11. Computational Epigenetics
12. Metabolic Networks
13. Kinetic Modeling of Cellular Processes
14. Stochastic processes in biological cells
15. Integrated Cellular Networks

Lecture – type of mathematics

Mathematical concept	Object of Investigation	Analysis of Complexity	Time-dependent	Treated in Chapter #
Mathematical graphs	protein-protein networks; protein complexes; gene regulatory networks	Yes	no	5, 6, 9, 10
Stoichiometric analysis; matrix algebra	metabolic networks*	yes (count # of possible paths that connect two metabolites)	no	12
Differential Equations	signal transduction, energy transduction, gene regulatory networks	No	yes	9, 13
Equations of motion	individual proteins, protein complexes		yes	14, 15
Correlation functions, Fourier transformation	reconstruction of two- and three-dimensional structures of cellular structures and individual molecules	No	yes, when applied on time-dependent data	2
Statistical tests	Differential expression and methylation; enriched network motifs	No	yes, when applied on time-dependent data	8, 9, 10
Machine learning (linear regression, hidden Markov model)	Predict gene expression, classify chromatin states	No	no	8, 11

Appetizer: A whole-cell model for the life cycle of the human pathogen *Mycoplasma genitalium* (15.2)

Theory

A Whole-Cell Computational Model Predicts Phenotype from Genotype

Jonathan R. Karr,^{1,4} Jayodita C. Sanghvi,^{2,4} Derek N. Macklin,² Miriam V. Gutschow,² Jared M. Jacobs,² Benjamin Bolival, Jr.,² Nacyra Assad-Garcia,³ John I. Glass,³ and Markus W. Covert^{2,*}

¹Graduate Program in Biophysics

²Department of Bioengineering

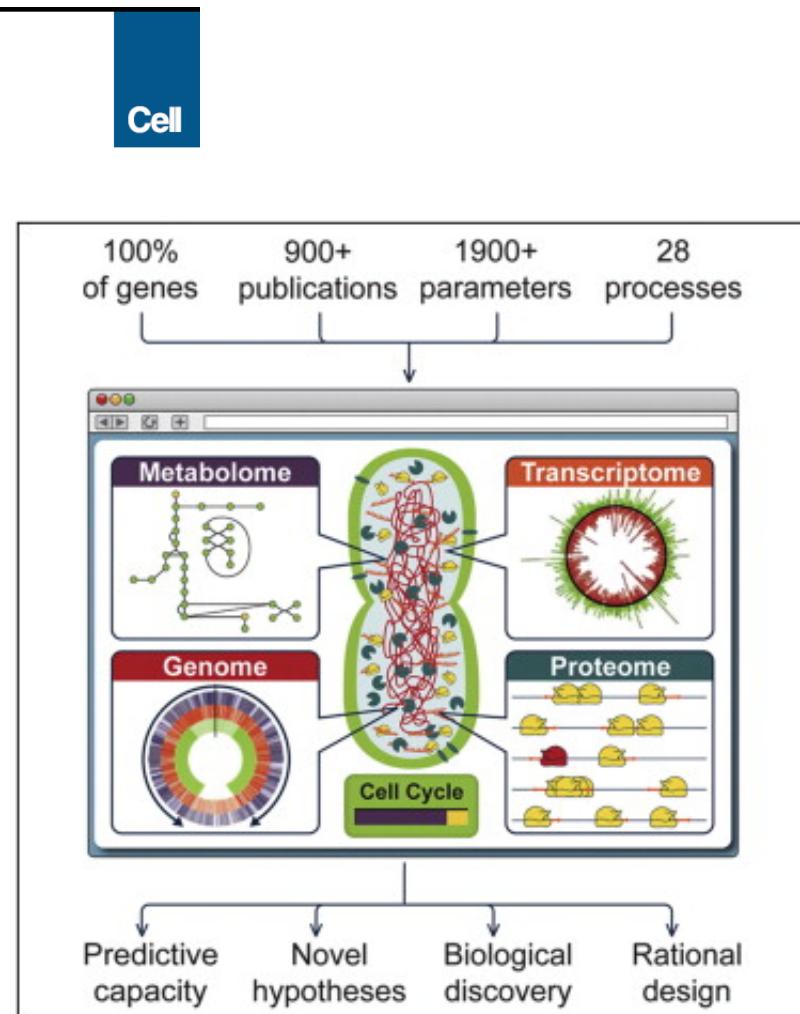
Stanford University, Stanford, CA 94305, USA

³J. Craig Venter Institute, Rockville, MD 20850, USA

⁴These authors contributed equally to this work

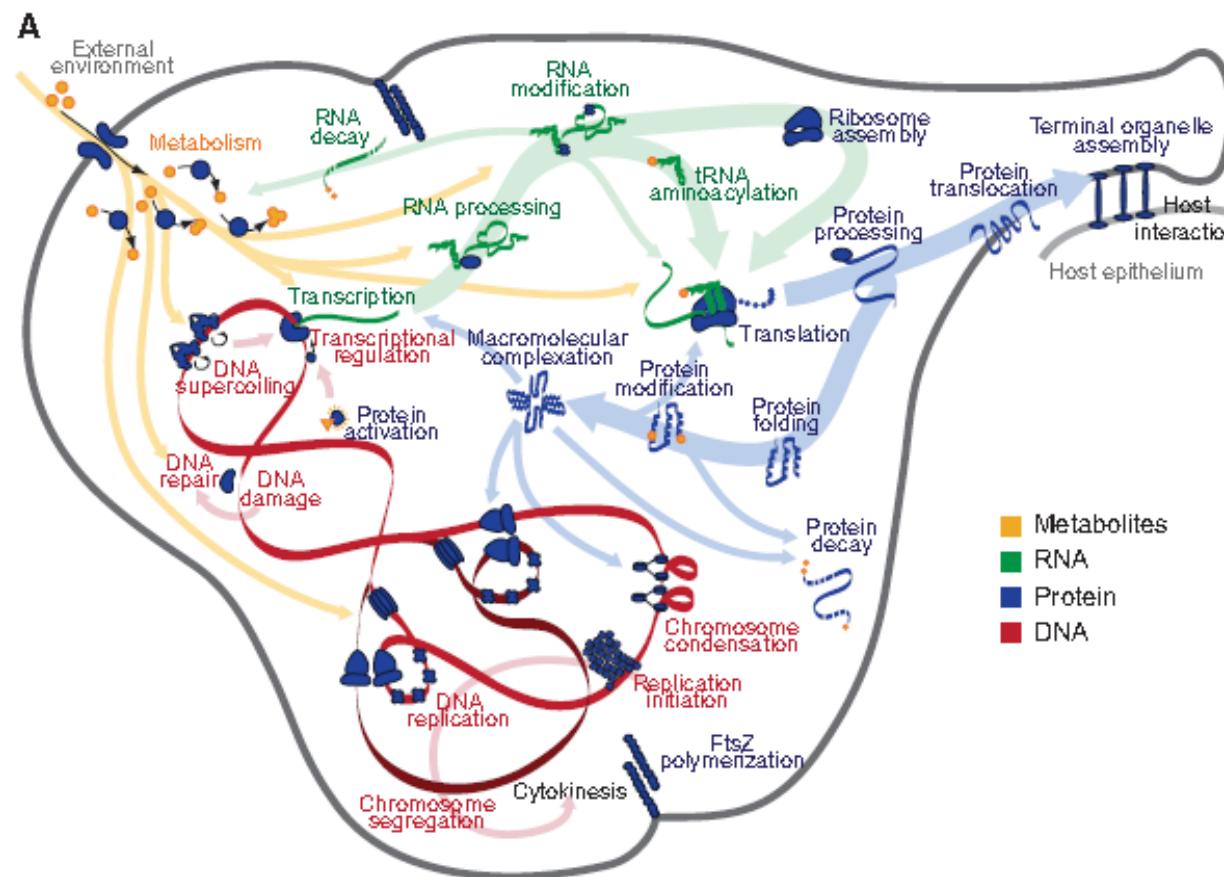
*Correspondence: mcovert@stanford.edu

<http://dx.doi.org/10.1016/j.cell.2012.05.044>



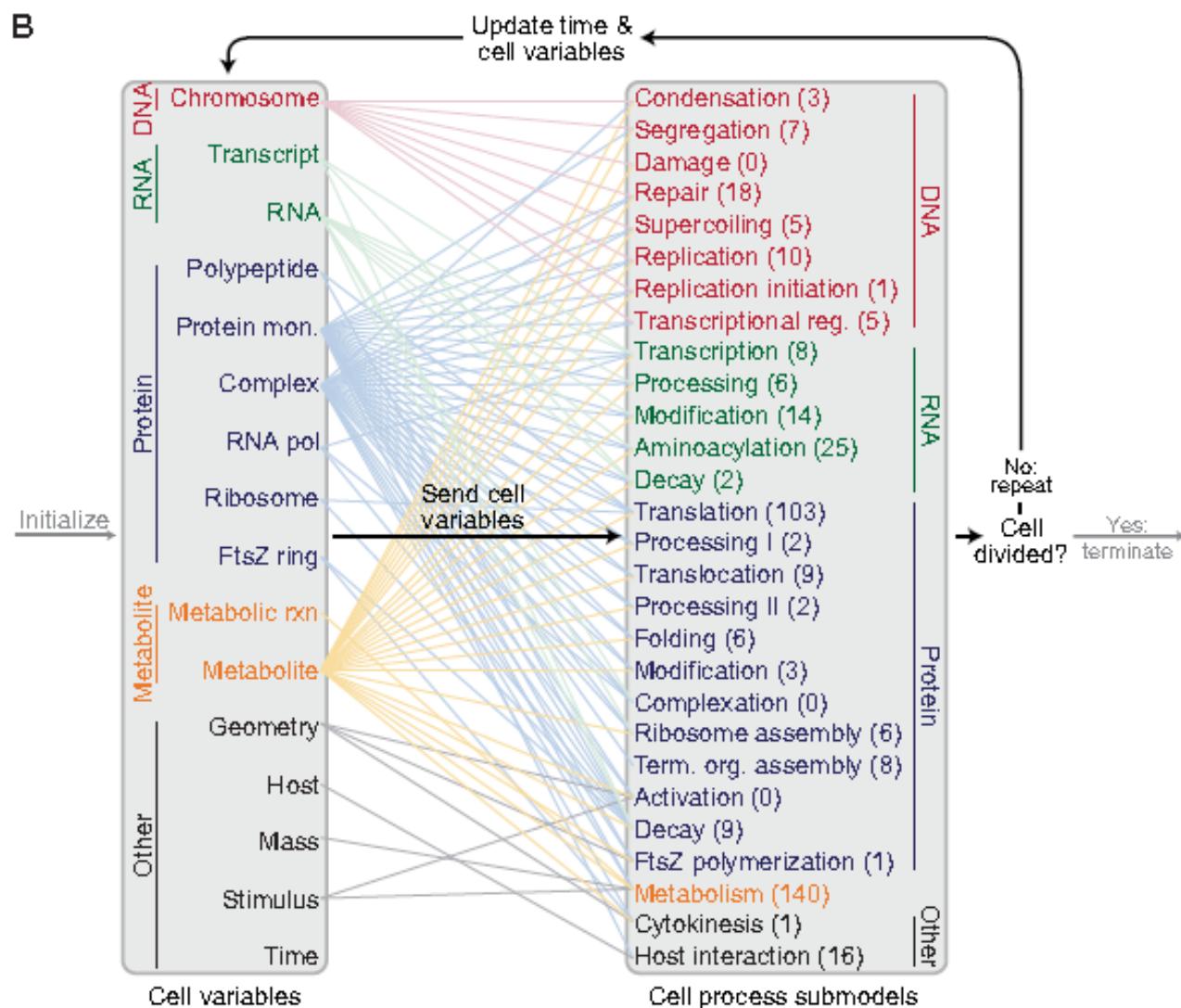
Cell 150, 389-401 (2012)

Divide and conquer approach (Caesar): split whole-cell model into 28 independent submodels



28 submodels are built / parametrized / iterated independently

Cell variables



System state is described by 16 cell variables

Colored lines: cell variables affected by individual submodels

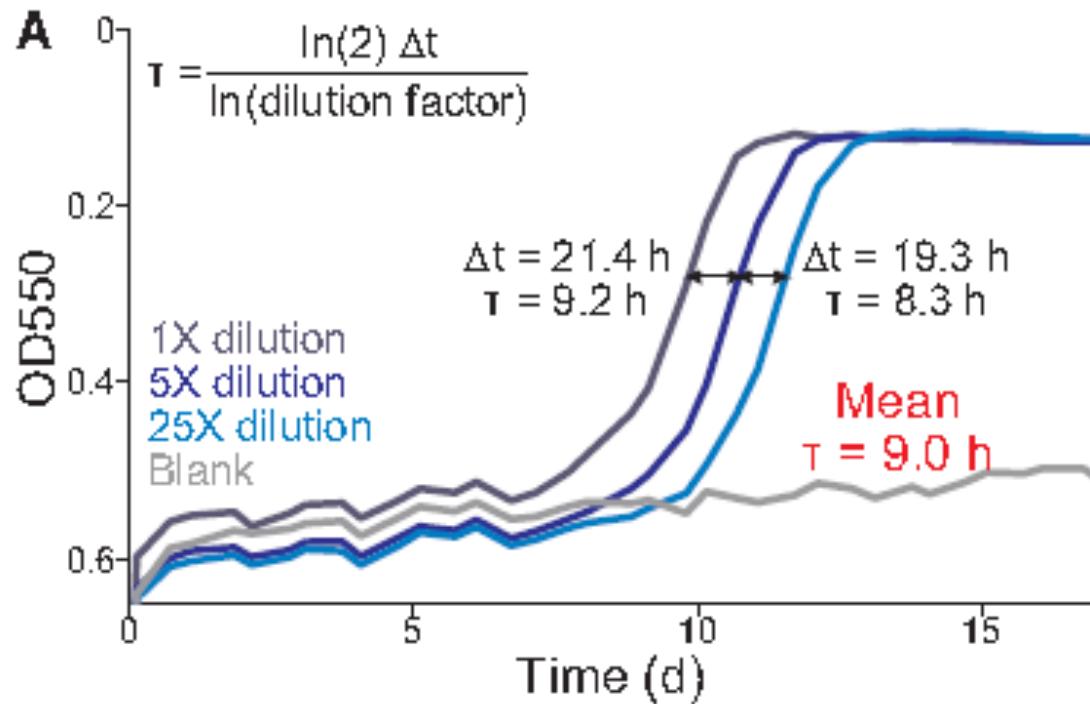
Mathematical tools:

- Differential equations
- Stochastic simulations
- Flux balance analysis

List S1. Primary sources of the *M. genitalium* reconstruction.

Data source	Content
Bernstein <i>et al.</i> , 2002 ⁶⁰²	mRNA half-lives
BioCyc ⁶	Genome annotation, metabolic reactions
BRENDA ⁵⁷⁰	Reaction kinetics
CMR ¹⁶⁸	Genome annotation
Deuerling <i>et al.</i> , 2003 ³⁸⁸	Chaperone substrates
DrugBank ⁹⁴⁷	Antibiotics
Eisen <i>et al.</i> , 1999 ⁸⁹¹	DNA repair
Endo <i>et al.</i> , 2007 ³⁹¹	Chaperone substrates
Feist <i>et al.</i> , 2007 ⁵⁵⁸	Metabolic reactions
Glass <i>et al.</i> , 2006 ¹⁹³	Gene essentiality
Güell <i>et al.</i> , 2009 ⁴¹⁸	Transcription unit structure
Gupta <i>et al.</i> , 2007 ²⁸⁰	N-terminal methionine cleavage
KEGG ¹¹³	Genome annotation, orthology
Kerner <i>et al.</i> , 2005 ³⁸⁹	Chaperone substrates
Krause <i>et al.</i> , 2004 ⁴⁰⁹	Terminal organelle assembly
Lindahl <i>et al.</i> , 2000 ⁴⁶²	DNA damage
Morowitz <i>et al.</i> , 1962 ⁸⁷⁰	Cell chemical composition
NCBI Gene ^{61,777}	Genome annotation
Neidhardt <i>et al.</i> , 1990 ³⁹³	Cell chemical composition
Peil, 2009 ¹⁰⁵	RNA modification
PubChem ⁵⁸⁷	Metabolite structures
SABIO-RK ¹⁰⁰	Reaction kinetics
Solabia ^{754–759}	Media chemical composition
Suthers <i>et al.</i> , 2009 ⁶¹⁰	Metabolic reactions
UniProt ⁹⁶	Genome annotation
Weiner <i>et al.</i> , 2000 ⁴¹¹	Promoters
Weiner <i>et al.</i> , 2003 ⁵⁶⁹	mRNA expression

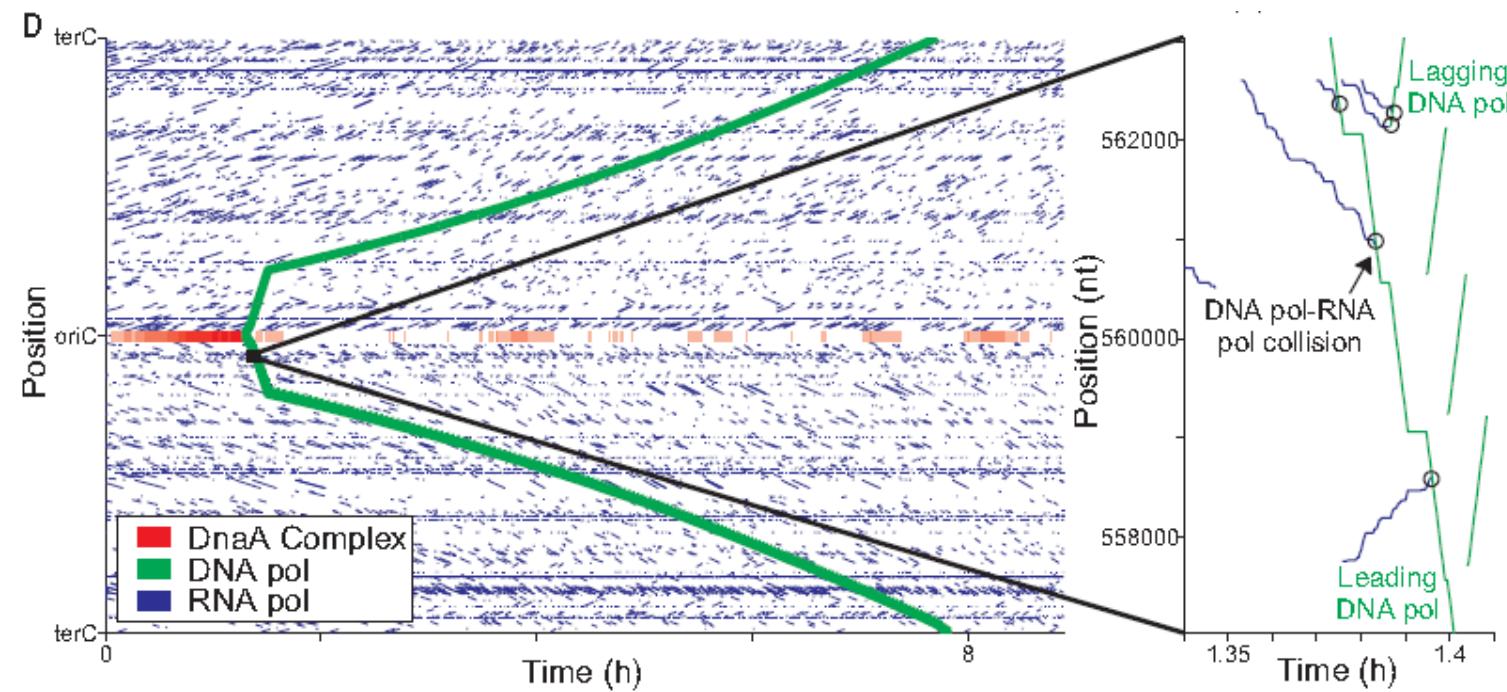
Growth of virtual cell culture



Growth of three cultures (dilutions indicated by shade of blue) and a blank control measured by OD₅₅₀ of the pH indicator phenol red. The doubling time, t , was calculated using the equation at the top left from the additional time required by more dilute cultures to reach the same OD₅₅₀ (black lines).

The model calculations were consistent with the observed doubling time!

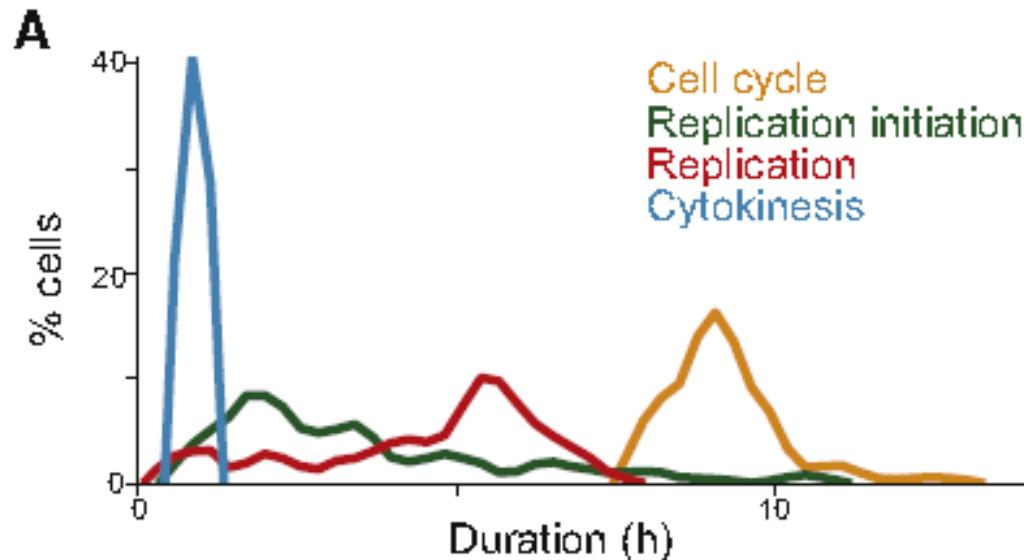
DNA-binding and dissociation dynamics



DNA-binding and dissociation dynamics of the oriC DnaA complex (red) and of RNA (blue) and DNA (green) polymerases for one *in silico* cell. The oriC DnaA complex recruits DNA polymerase to the oriC to initiate replication, which in turn dissolves the oriC DnaA complex. RNA polymerase traces (blue line segments) indicate individual transcription events. The height, length, and slope of each trace represent the transcript length, transcription duration, and transcript elongation rate, respectively.

Inset : several predicted collisions between DNA and RNA polymerases that lead to the displacement of RNA polymerases and incomplete transcripts.

Predictions for cell-cycle regulation



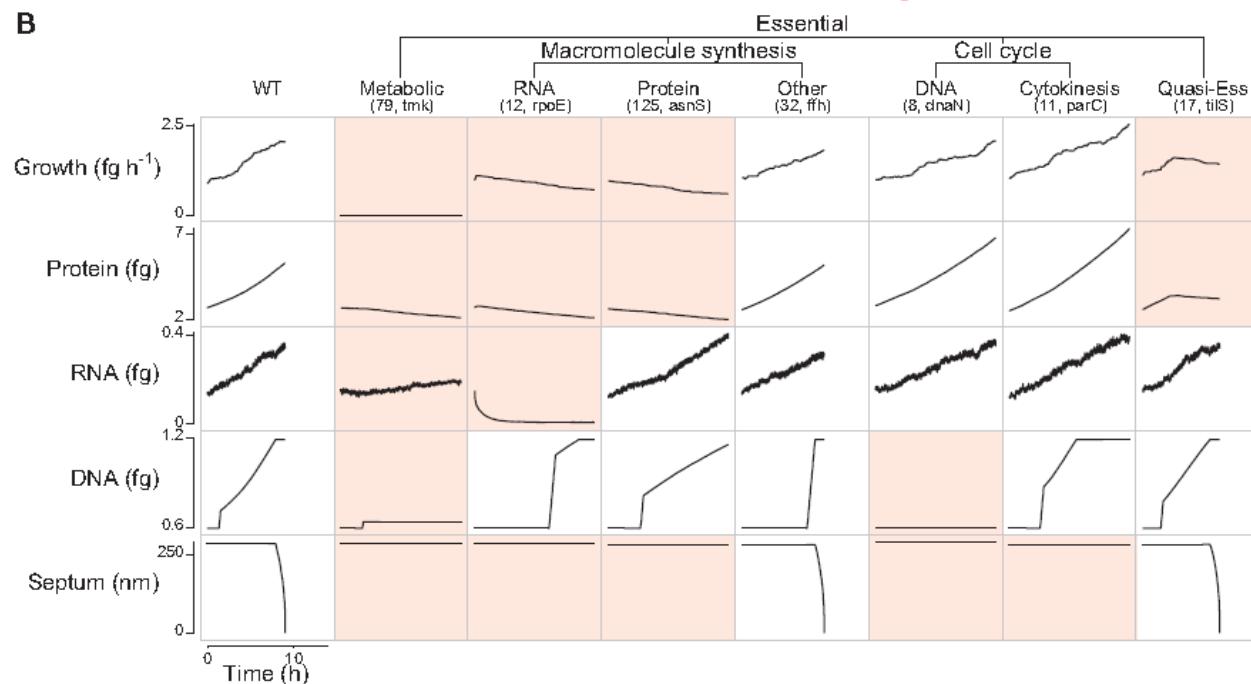
Distributions of the duration of three cell-cycle phases, as well as that of the total cell-cycle length, across 128 simulations.

There was relatively more cell-to-cell variation in the durations of the replication initiation (64.3%) and replication (38.5%) stages than in cytokinesis (4.4%) or the overall cell cycle (9.4%).

This data raised two questions:

- (1) what is the source of duration variability in the initiation and replication phases; and
- (2) why is the overall cell-cycle duration less varied than either of these phases?

Single-gene knockouts : essential vs. non-essential genes



Single-gene disruption strains grouped into phenotypic classes (columns) according to their capacity to grow, synthesize protein, RNA, and DNA, and divide (indicated by septum length).

Each column depicts the temporal dynamics of one representative *in silico* cell of each essential disruption strain class.

Dynamics significantly different from wild-type are highlighted in red.

The identity of the representative cell and the number of disruption strains in each category are indicated in parenthesis.

Literature

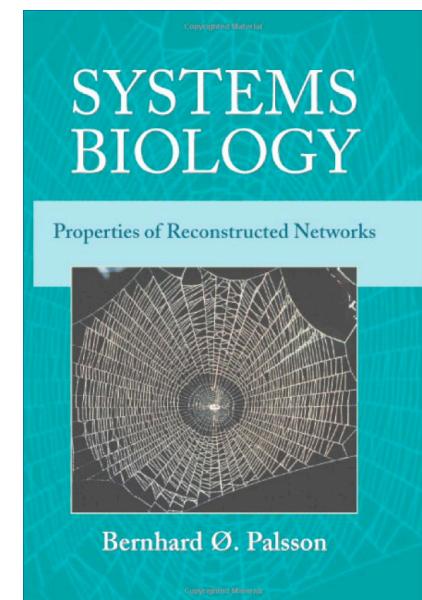
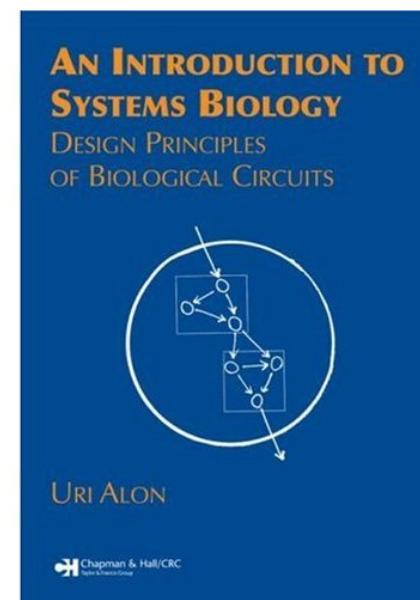
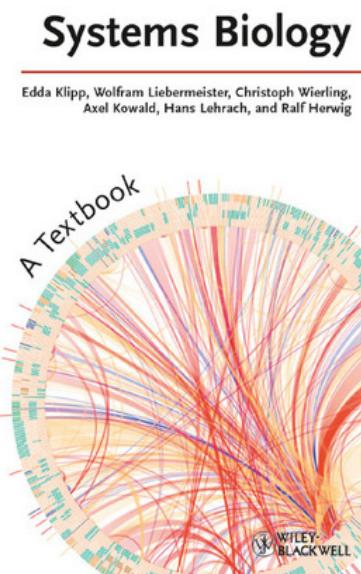
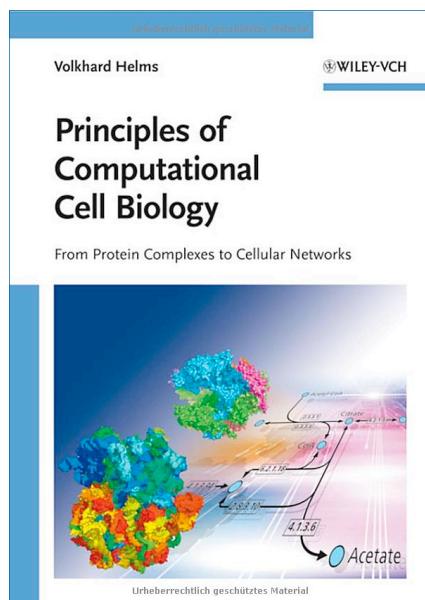
Lecture **slides** — available before the lecture

Suggested **reading**

=> check our web page

[http://gepard.bioinformatik.uni-saarland.de/teaching/...](http://gepard.bioinformatik.uni-saarland.de/teaching/)

Textbooks



=> check computer science library

How to pass this course

Schein = you need to qualify for the **final exam** and pass it

Final exam:

written test of 180 min length about selected parts of the lecture
(slides will be defined 2 weeks before exam) AND about selected assignments

requirements for participation in final exam:

- 50% of the points from the **assignments**
- one assignment task presented @ blackboard in tutorial

Final exam will take place at the end of the semester.

In case you are sick (final exam) you should submit a medical certificate
to take the written re-exam (then this will be counted as first exam).

Re-exam: will take place in first week of the winter term 2018/19.
Everybody can take the re-exam (first exam failed or passed).

Assignments

Tutors: Thorsten Will, Duy Nguyen, Daria Gaidar ,
Markus Hollander

Tutorial: Tue, 12:15–13:45, E2 I, room 007

10 assignments with 100 points each

Assignments are part of the course material (not everything is covered in lecture)

- => **one** solution for **two** students (or one)
- => hand-written or one **printable** PDF/PS file per email
- => content: data analysis + interpretation — **think!**
- => no 100% solutions required!!!
- => attach the source code of the programs for checking (no suppl. data)
- => present one task at the **blackboard**

Hand in at the following Fri electronically until 13:15 or
printed at the start of the lecture.

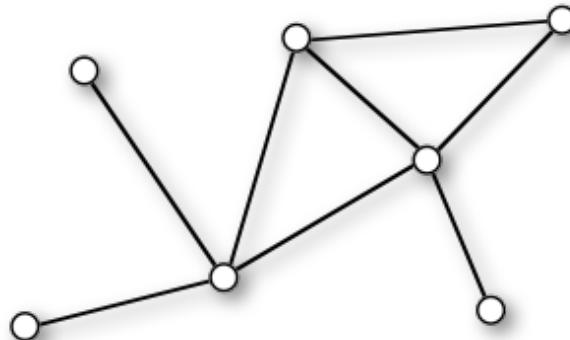
Some Graph Basics

Network <=> Graph

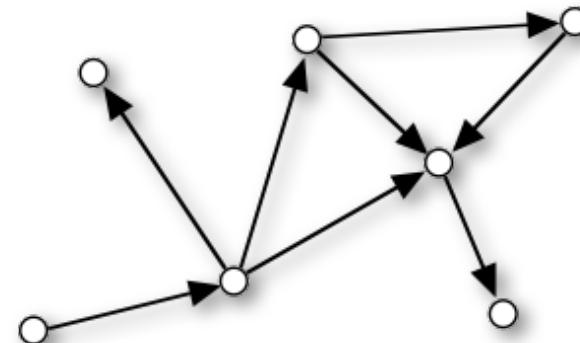
Formal definition:

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



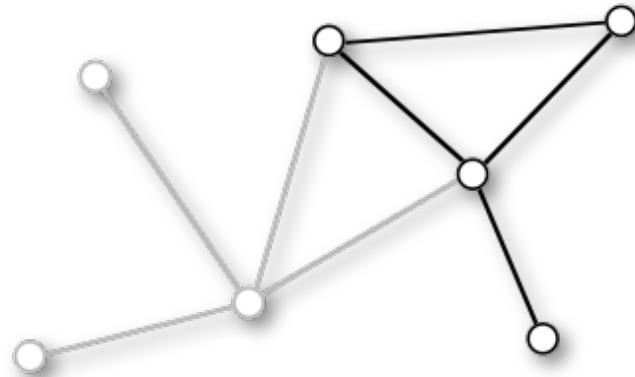
directed graph

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

Graph Basics II

Subgraph:

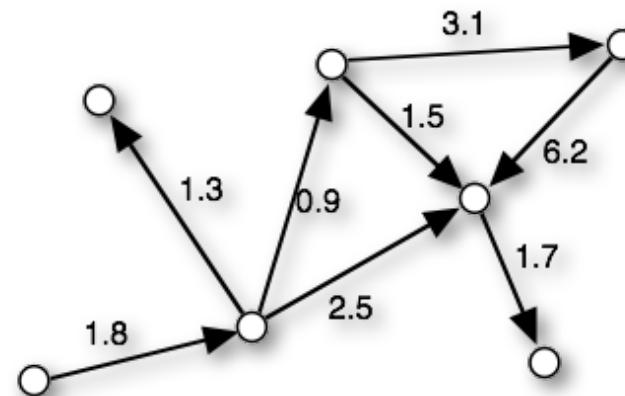
$G' = (V', E')$ is a subset of $G = (V, E)$



Practical question: how to define useful subgraphs?

Weighted graph:

Weights assigned to the edges



Note: no weights for vertices

Walk the Graph

Path = sequence of connected vertices

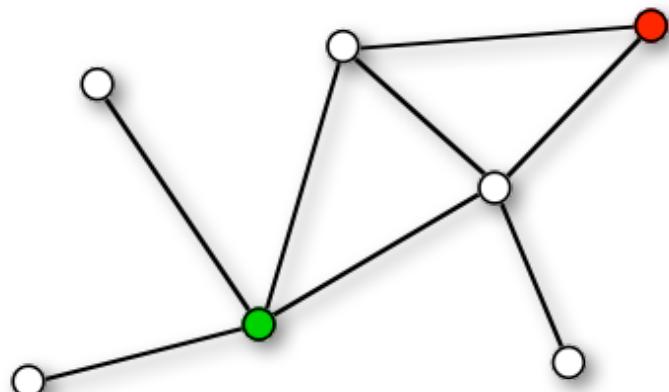
start vertex => internal vertices => end vertex

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: disconnected

Trail = path, in which all edges are distinct

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



How many paths connect the green to the red vertex?

How long are the shortest paths?

Find the four trails from the green to the red vertex.

How many of them are independent?

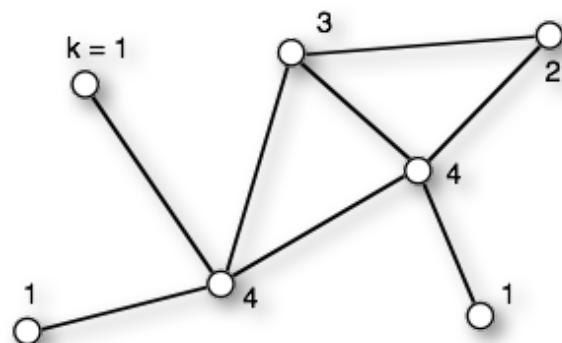
Local Connectivity: Degree/Degree Distribution

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

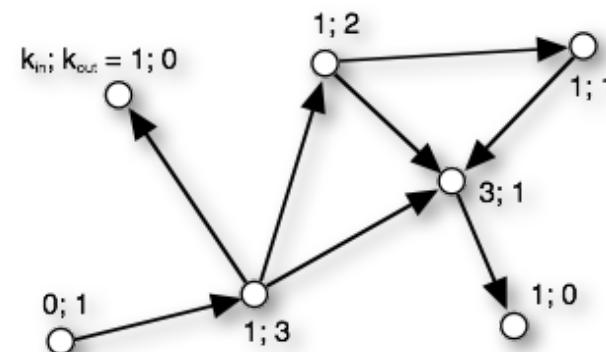
Directed graph => distinguish k_{in} and k_{out}

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 $N \times N$ matrix

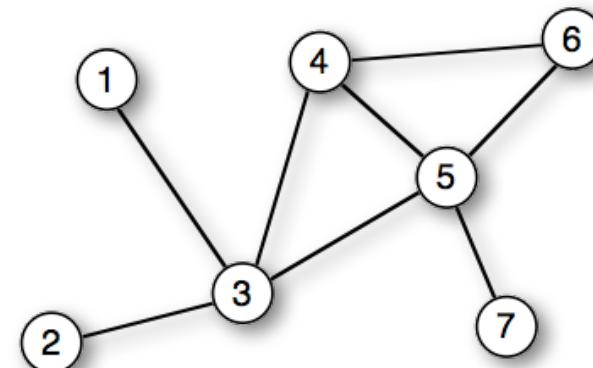
with entries M_{uv}

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

→ symmetric for undirected graphs

- + 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



	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	—

Measures and Metrics

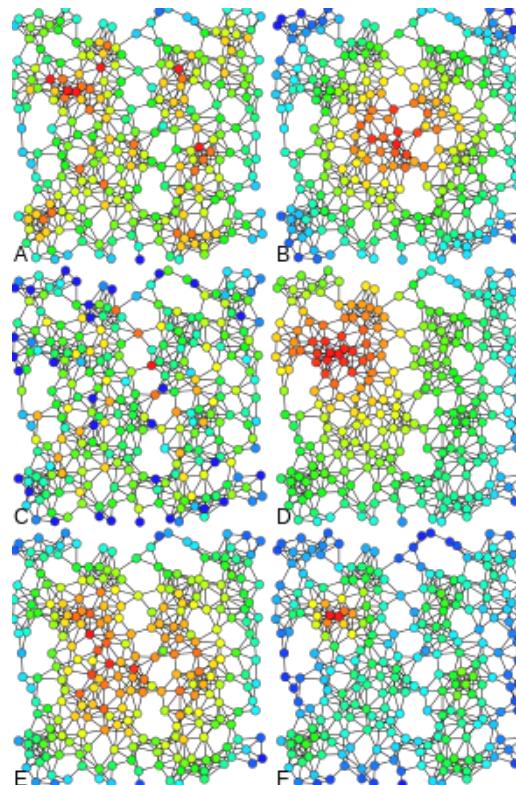
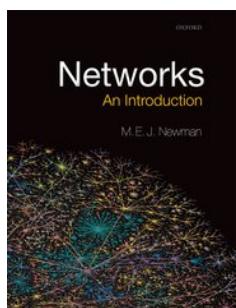
“Which are the most important or central vertices in a network? “

Examples of

A) Degree
centrality,

C) Betweenness
centrality,

E) Katz centrality,



B) Closeness centrality,

D) Eigenvector centrality,

F) Alpha centrality of the
same graph.

www.wikipedia.org

book by Mark Newman / Oxford Univ Press

- Chapter 7: measures and metrics
- Chapter 11: matrix algorithms and graph partitioning

Degree centrality

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

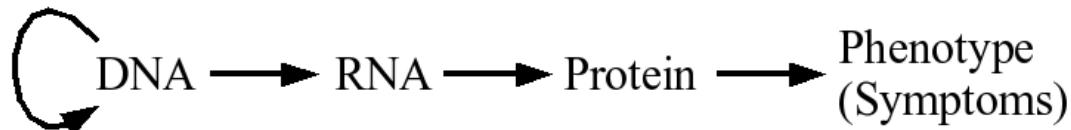
E.g. in a **social network**, individuals that have many connections to others might have

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



1.2 Biological Background

Central Paradigm of Molecular Biology

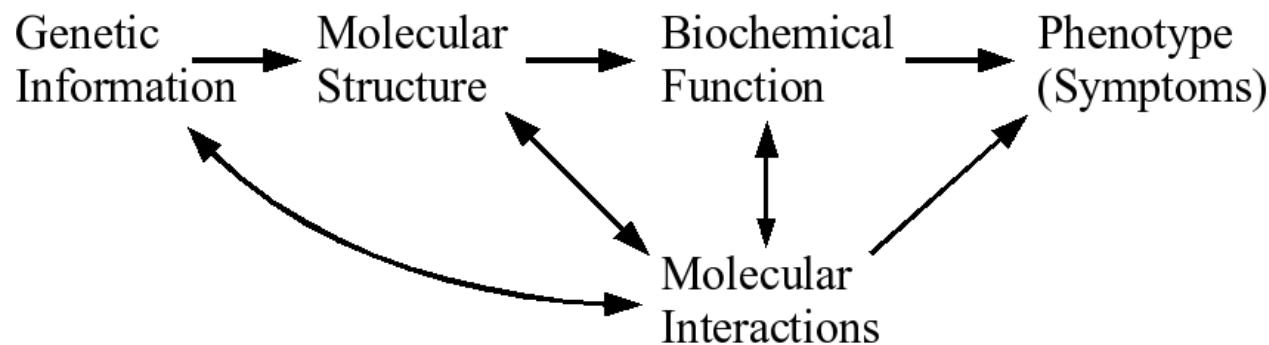


However, there exist feedback loops (transcription factors, microRNAs)

Central Paradigm of Structural Biology



Central Paradigm of Molecular Systems Biology



1.2 Cellular components

We will mostly use 3 different levels of description:

Inventory lists and lists of processes:

- Proteins in particular compartments
- Proteins forming macromolecular complexes
- Biomolecular interactions
- Regulatory interactions
- Metabolic reactions

Structural descriptions:

- Structures of single proteins
- Topologies of protein complexes
- Subcellular compartments

Dynamic descriptions:

Cellular processes ranging from nanosecond dynamics for the association of two biomolecules up to processes occurring in seconds and minutes such as the cell division of yeast cells.

1.2 Biomolecules

Macromolecules: proteins
nucleic acids
polysaccharides
lipids.

Building blocks of macromolecules:

sugars as the precursors of poly-saccharides
amino acids as the building blocks of proteins
nucleotides as the precursors of DNA and RNA
fatty acids which are incorporated into lipids.

Interestingly, in biological cells, only a small number of the theoretically synthesizable macromolecules exist at a given point in time.

At any moment during a normal cell cycle, many new macromolecules need to be synthesized from their building blocks and this is meticulously controlled by the complex **gene expression machinery**.

Even during a steady-state of the cell, there exists a constant **turn-over** of macromolecules.

1.2 Biomolecules

Metabolic intermediates (**metabolites**):

The molecules in a cell have complex chemical structures and must be synthesized step-by-step beginning with specific starting materials that may be taken up as energy source.

In the cell, connected chemical reactions are often grouped into metabolic pathways.

Molecules of **miscellaneous function**:

vitamins

steroid or amino acid hormones

molecules involved in energy storage (e.g. ATP)

regulatory molecules (e.g. cyclic AMP)

metabolic waste products such as urea.

1.2 Compartments

Organization into various compartments greatly simplifies the temporal and spatial process flow in eukaryotic cells.

At each time point during a cell cycle only a small subfraction of all potential proteins are being synthesized (and not yet degraded).

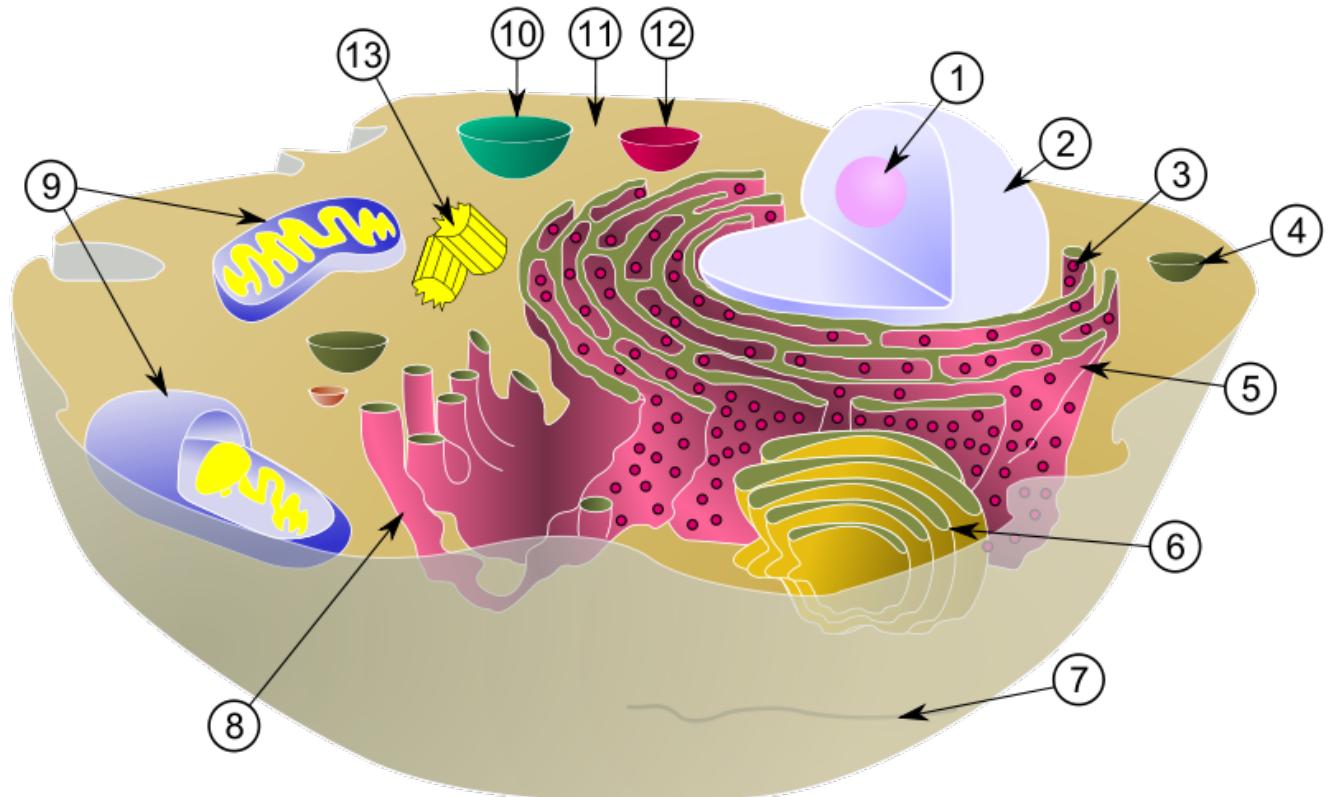
Many proteins are only available in very **small concentrations**, possibly with only a few copies per cell.

However, due to **localizing** these proteins to particular spots in the cell, e.g. by attaching them to the cytoskeleton or by partitioning them into lipid rafts, their **local concentrations** may be much higher.

1.2 Compartments

Compartments of a typical animal cell:

- (1) nucleolus
- (2) nucleus
- (3) ribosome
- (4) vesicle
- (5) rough endoplasmic reticulum (ER)
- (6) Golgi apparatus
- (7) Cytoskeleton
- (8) smooth ER
- (9) mitochondria
- (10) vacuole
- (11) cytoplasm
- (12) lysosome
- (13) centrioles

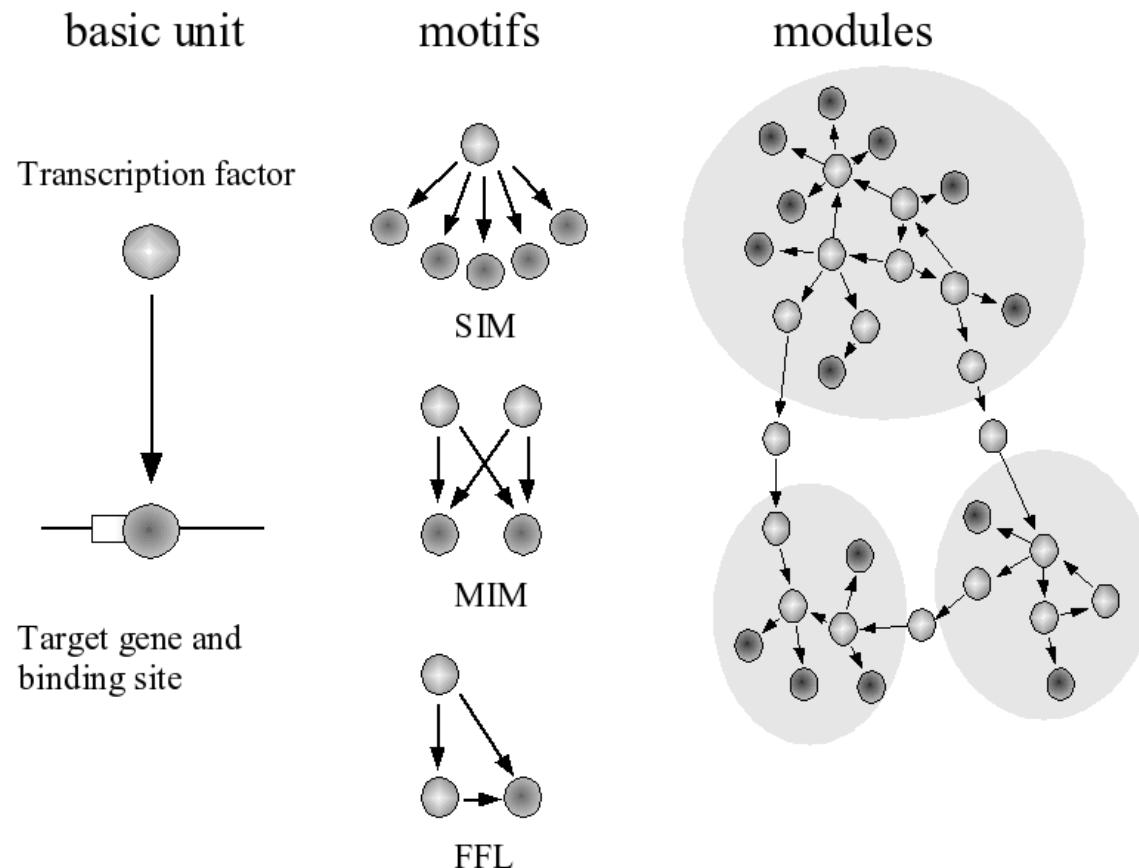


www.wikipedia.org

1.2 Organisms

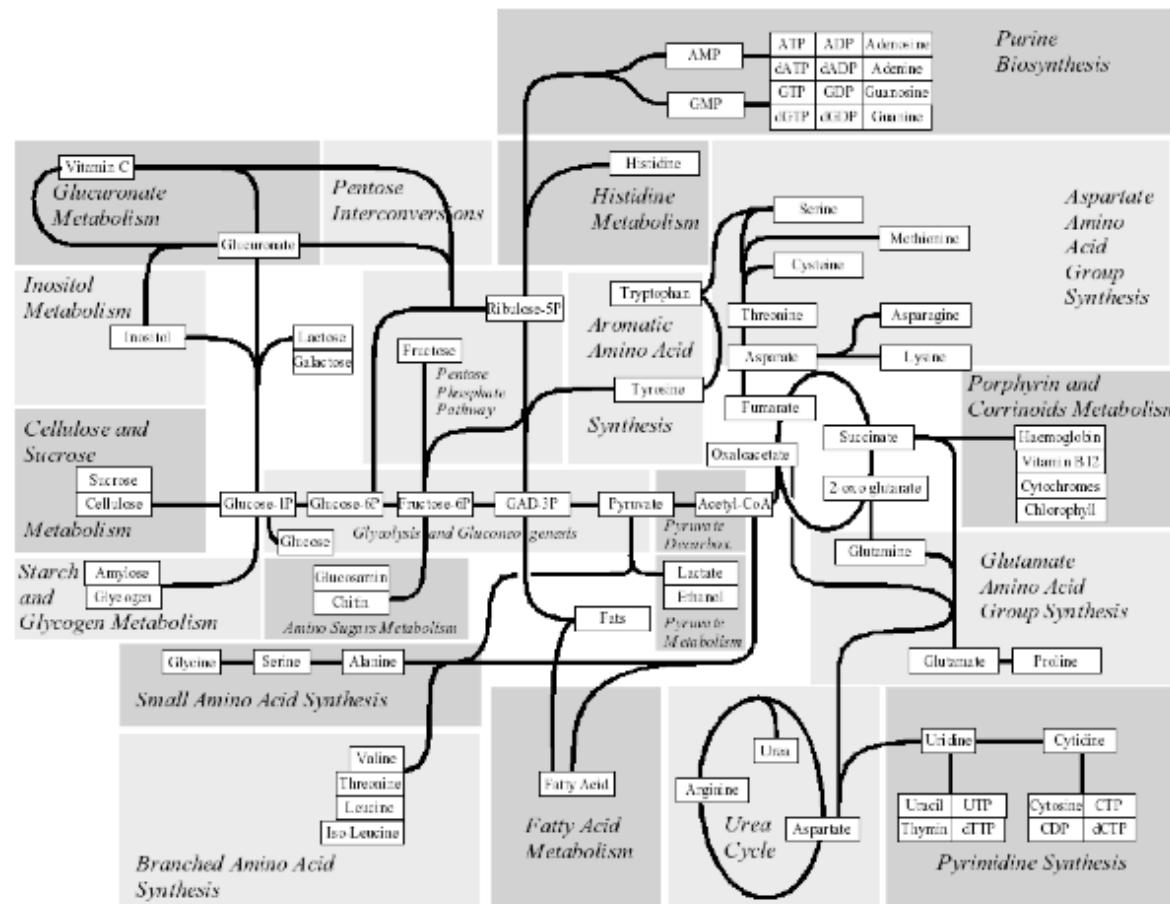
Organism	Length of genome [Mb]	Number of protein- coding genes	Number of RNA genes	Number of transporter proteins
Prokaryotes	0.6	476	43	53
<i>Mycoplasma genitalium G37</i>				
<i>Bacillus subtilis BSN5</i>	4.2	4145	113	552
<i>Escherichia coli</i> (<i>E. coli</i>) APEC01	4.6	4890	93	665
Eukaryotes	1.3	6002	425	341
<i>Saccharomyces Cerevisiae S288C</i>				
<i>Drosophila melanogaster</i>	12	13929	3209	662
<i>Caenorhabditis elegans</i>	100.2	20093	24969	669
<i>Homo sapiens</i>	3150	20338	19201	1467

1.3 Cellular Pathways



After Babu (2004)

1.3 Major Metabolic Pathways



static
connectivity

\Leftrightarrow

dynamic response to
external conditions

\Leftrightarrow

different states during
the cell cycle



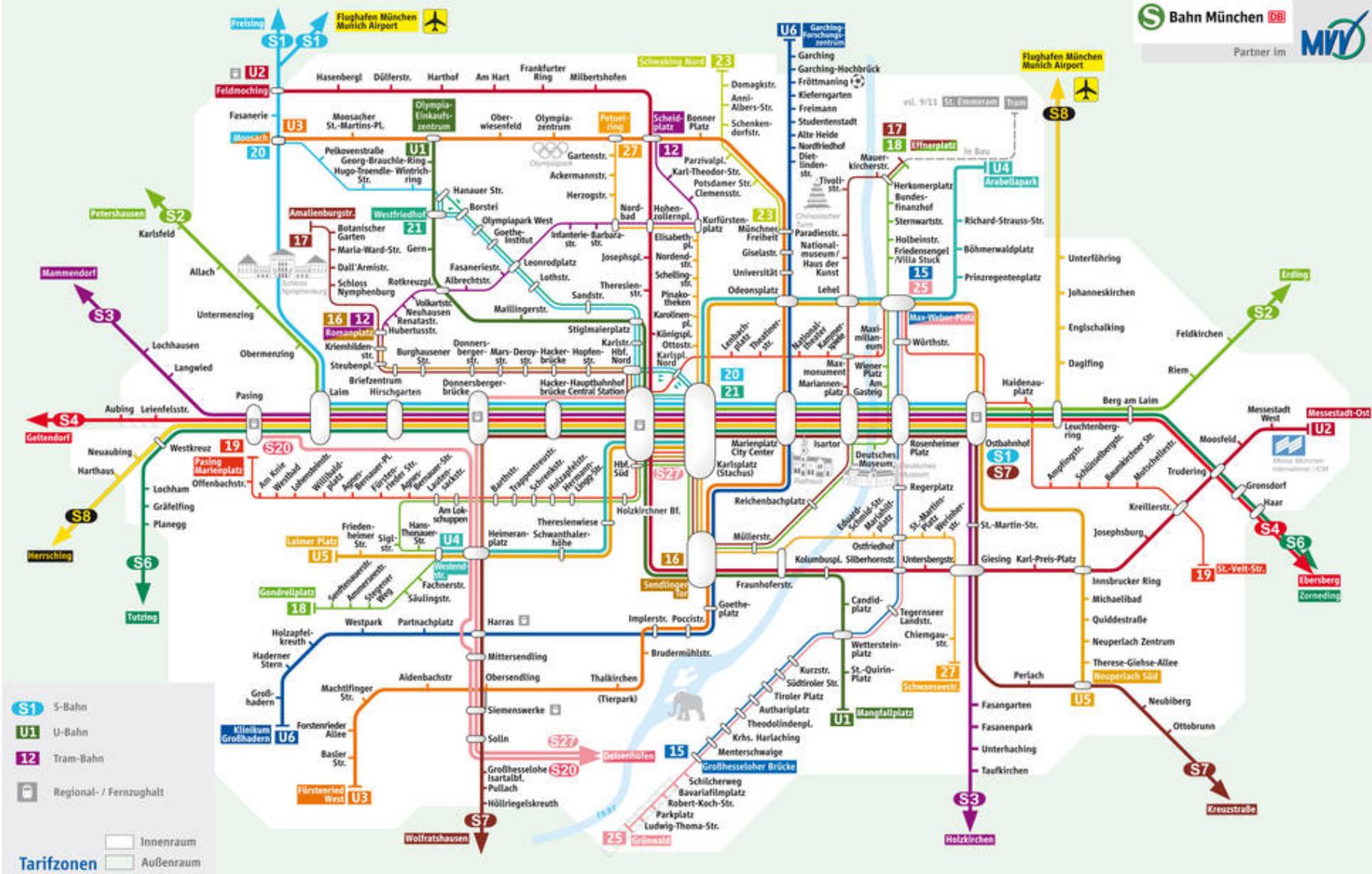
Schienennetzplan München



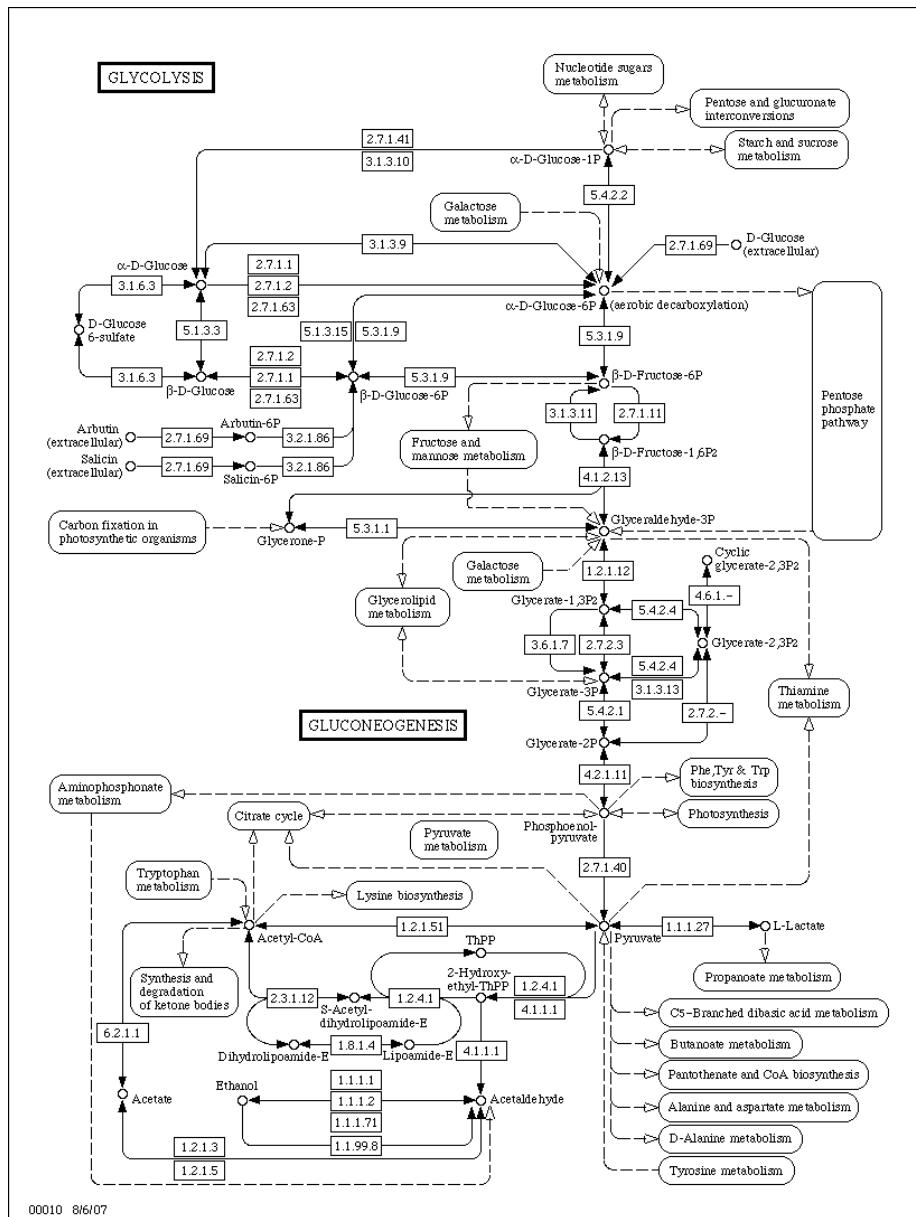
S Bahn München DB

The logo consists of the letters "MVV" in a blue sans-serif font, with a green checkmark symbol positioned to the right of the "V".

Partner im



1.3 Cellular Pathways



Glycolysis pathway of *E. coli* KEGG database

1.4 Ontologies & Databases

- Gene Ontology (chapter 8)
- KEGG (<http://www.genome.jp/kegg/>)
- Reactome
- BRENDA (<https://www.brenda-enzymes.org/>)
- DAVID (<https://david.ncifcrf.gov/>)
- Protein Databank (www.rcsb.org)

Summary

What you learned **today**:

- => **networks** are everywhere
- ⇒ how to get the "**Schein**" for BI3
- ⇒ What is the lecture content
- ⇒ Basic biological background

Next lectures:

- Random graphs vs. scale-free networks (assignments 1 + 2)
- Structures of protein complexes

Additional slides (not used)

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.)

Remember the multiplication of a matrix with a vector below ...

$$\mathbf{A} = \begin{pmatrix} a & b & c \\ p & q & r \\ u & v & w \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

$$\mathbf{AB} = \begin{pmatrix} a & b & c \\ p & q & r \\ u & v & w \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} ax + by + cz \\ px + qy + rz \\ ux + vy + wz \end{pmatrix},$$

en.wikipedia.org

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.

We can also write this expression in matrix notation as

$$\mathbf{x}' = \mathbf{A} \mathbf{x} \quad \text{where } \mathbf{x} \text{ is the vector with elements } x_i.$$

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 $\mathbf{x}(0)$ as a linear combination of the eigenvectors \mathbf{v}_i of the (quadratic) adjacency matrix¹

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

Then

$$\mathbf{x}(t) = A^t \sum_i c_i \mathbf{v}_i$$

Because \mathbf{v}_i are eigenvectors of A , $A \mathbf{v}_i = k_i \mathbf{v}_i$ with the eigenvalue k_i .

Let k_1 be the largest eigenvector.

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

Here, we divide by k_1 and multiply by k_1 in the front.

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

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

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 \mathbf{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:

$$x_i = k_1^{-1} \sum_j A_{ij} x_j \quad \text{Divide above eq. by } k_1$$

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

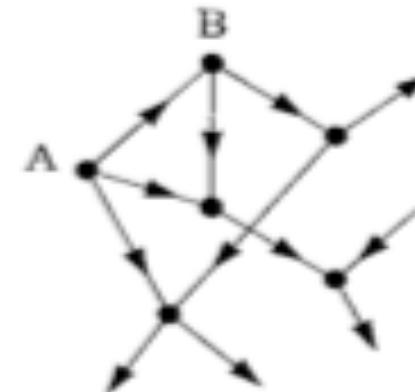


Figure 7.1: A portion of a directed network. Vertex A in this network has only outgoing edges and hence will have eigenvector centrality zero. Vertex B has outgoing edges and one incoming edge, but the incoming one originates at A, and hence vertex B will also have centrality zero.

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 $\mathbf{x} = \alpha \mathbf{Ax} + \beta \mathbf{I}$

where \mathbf{I} is the vector $(1, 1, 1, \dots)^T$. By rearranging for \mathbf{x} we find

$$\mathbf{I} \mathbf{x} - \alpha \mathbf{A} \mathbf{x} = \beta \mathbf{I} \quad (\text{where we used } \mathbf{I} \mathbf{x} = \mathbf{x})$$

$$(\mathbf{I} - \alpha \mathbf{A}) \mathbf{x} = \beta \mathbf{I}$$

$$(\mathbf{I} - \alpha \mathbf{A})^{-1} (\mathbf{I} - \alpha \mathbf{A}) \mathbf{x} = (\mathbf{I} - \alpha \mathbf{A})^{-1} \beta \mathbf{I} \quad (\text{multiply both sides with } (\mathbf{I} - \alpha \mathbf{A})^{-1})$$

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

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

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

$$\mathbf{x}' = \alpha \mathbf{Ax} + \beta \mathbf{I}$$

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

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

where \mathbf{I} is the vector $(1, 1, 1, \dots)^T$ and \mathbf{D} the diagonal matrix with $D_{jj} = \max(k_j^{out}, 1)$

PageRank

By rearranging we find that

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

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

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

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).

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 (0.11154).

→ 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