Bioinformatik

Wintersemester 2019 / 2020, Uni Potsdam

Stefanie Hartmann

Networks

Jan 10, 2020

Klausuranmeldung in PULS

Klausur: 7. Februar 2020, 8:15h, 2.27.1.01
Nachklausur: 20. März 2020

Bioinformatik (4 LP)
Studiengang: BSBIWH20102

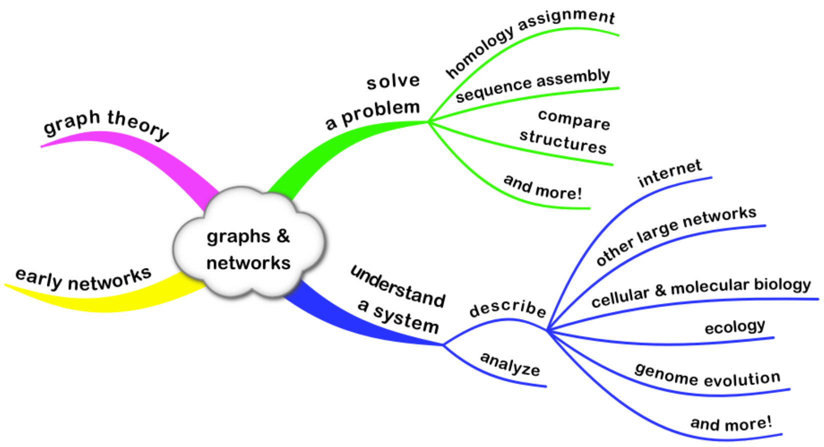
Bioinformatik (6 LP)
Studiengang: BSBIWH20172, BSICSH20132,
BSICSH20192, MSCOSH20132
(Bioinformatik + Molekulare Evolution)

course evaluation

- please complete the survey before Jan 29
- I will discuss results on Jan 31

- https://www.surveymonkey.de/r/CQ872K9

Overview



- literature on Moodle: network biology

Reading: network biology

- Introduction
- Basic network nomenclature
- Architectural features of cellular networks
- Motifs, modules, and hierarchical networks
- Network robustness
  - Topological robustness
- Beyond topology: characterizing the links

Box 1: Network measures

Box 2: Network models *(random and scale-free only)*

Bridges of Königsberg

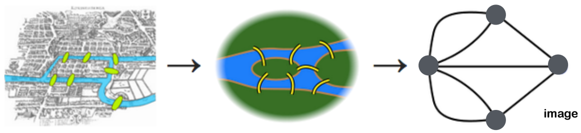
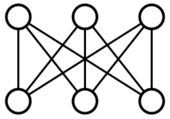


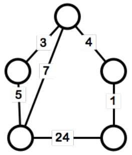
image taken from Wikipedia

- Leonhard Euler, 1735
- find a walk through the city that crosses each bridge exactly one time
- abstraction into a graph
  - land = circle = node
  - bridge = connection = edge
- Eulerian path
  - Königsberg? for networks with which properties?

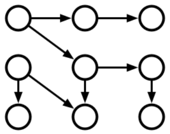
Graph theory



- a classical area in mathematics
- interaction with other and often distant areas

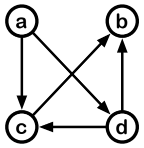


- graphs (networks)
  - consist of vertices (nodes) and edges (links)
  - can be undirected, directed, weighted, cyclic, acyclic



Graph representation

- graphs are excellent structures for storing, searching, and retrieving data
- conceptually, representation



graph

	a	b	c	d
a			1	1
b				
c		1		
d		1	1	

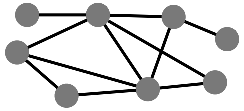
adjacency matrix

a	c, d
c	b
d	c, b

adjacency list

Early networks

networks (graphs): connected elements



mathematical description

- food webs
- social systems
  (insects, political parties, communities)
- road systems
- …

The internet!

early 1990

- world wide web, graphical web browsers
- connections in business, science, academia, homes

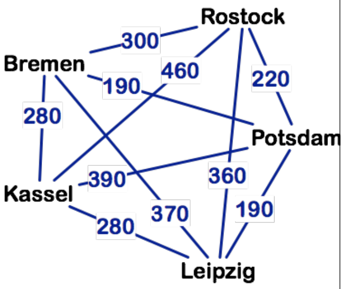
scientific interest in networks

- properties
- growth & dynamics
- effects
  - on markets
  - on communities
  - …

tools:
graph theory
mathematics
systems theory

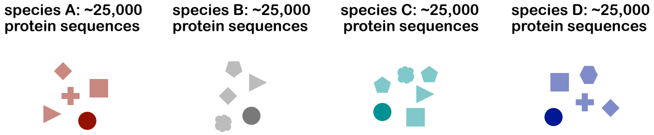
Graph algorithms

- the properties of graphs are well studied
- there are lots of problems in graph theory for which (efficient!) solutions exist
  - traversal algorithms
  - search algorithms
  - subgraph problems
  - clustering …



- state a biological problem in terms of graphs:
  ➜ algorithm available?

Large-scale homology assignment


species A: ~25,000 protein sequences
species B: ~25,000 protein sequences
species C: ~25,000 protein sequences
species D: ~25,000 protein sequences

data

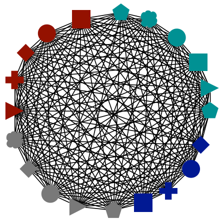
- all protein sequences from a number of species

goal

- identify all sets of homologs

approach

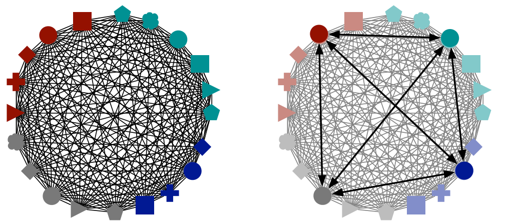
- compare all against all sequences
- identify significant matches



Large-scale homology assignment

a graph theoretic approach

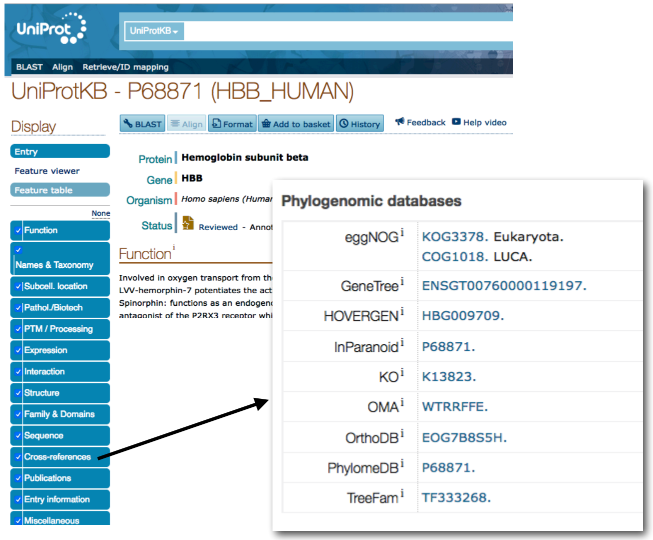
- nodes: DNA/protein sequences
- edges: sequence similarity values (e.g., E-values)
- group (cluster) sequences (nodes) by edge weights
  - efficient cluster algorithms exist for graphs!



cluster

low E-values within, higher E-values between

similar (homologous) sequences



UniProtKB - P68871 (HBB_HUMAN)

Protein | Hemoglobin subunit beta

Gene | HBB

Organism | *Homo sapiens (Human*

Status | Reviewed - Annot

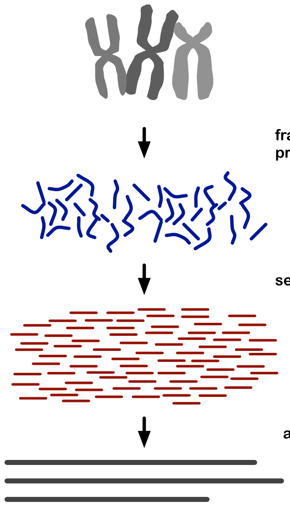
Function

Involved in oxygen transport from th
LVV-hemorphin-7 potentiates the act
Spinorphin: functions as an endogen
antagonist of the P2RX3 receptor whi

Phylogenomic databases

eggNOG	KOG3378. Eukaryota. COG1018. LUCA.
GeneTree	ENSGT00760000119197.
HOVERGEN	HBG009709.
InParanoid	P68871.
KO	K13823.
OMA	WTRRFFE.
OrthoDB	EOG7B8S5H.
PhylomeDB	P68871.
TreeFam	TF333268.

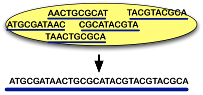
Sequencing genomes



fragment, prepare for sequencing

sequence

assemble



Assemble sequence reads, conceptually

the data: many sequence fragments (reads)

CTAGCGC

TCGCATC TCCATCG

GCATCGC CGATCCA

the approach: identify and merge overlaps

```
CTAGCGC               TCGCATC?
     GCATCGC       TCCATCG
        TCGCATC

                   TCCATCG
                CGATCCA
```

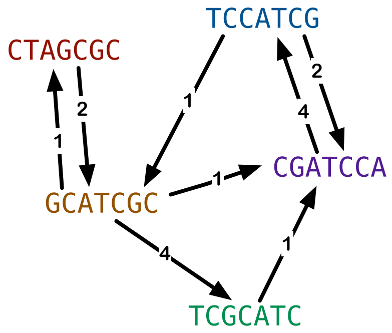
the result: genome, transcripts, etc

```
CTAGCGC          TCCATCG
     GCATCGC  CGATCCA
        TCGCATC
CTAGCGCATCGCATCGATCCATCGC
```

Assemble sequence reads, in practice

use a graph theoretic approach

- 1. construct overlap graph from the reads



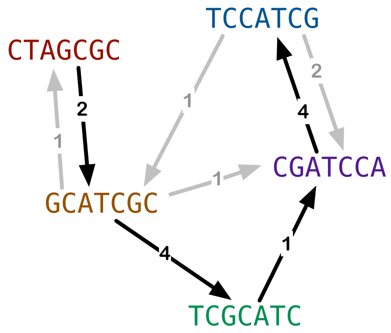
```
CTAGCGC
     GCATCGC

      CTAGCGC
GCATCGC
```

Assemble sequence reads, in practice

use a graph theoretic approach

- 2. find the shortest path with the highest weights that passes each node (read!) exactly once



```
CTAGCGC          TCCATCG
     GCATCGC  CGATCCA
        TCGCATC
CTAGCGCATCGCATCGATCCATCGC
```

path length: 25
path weight:11

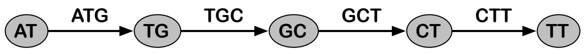
Assemble sequence reads, in practice

use a graph theoretic approach

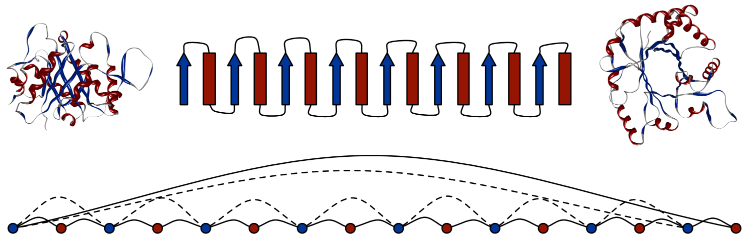
- 2. find the shortest path with the highest weights that passes each node (read!) exactly once
  - Hamiltonian path
  - heuristics are needed
  - optimal solution not guaranteed
  - used for long reads (Sanger, PacBio)
  - not feasible for short read data (Illumina)

Sequence assembly of short reads

- also uses graph theoretic approaches
- different type of graph, different type of path
  - de Bruijn graph
  - Eulerian path
  - much more efficent
  - does not guarantee the correct sequence, either!
- many different variations & implementations exist



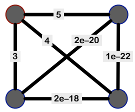
Comparison of protein structures



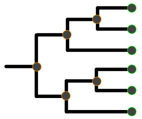
comparison of

- positions in a 3D coordinate system
- arrangement of secondary structure elements
- secondary structure graph

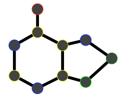
Graphs in biology



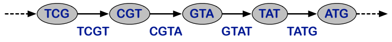
sequence similarity



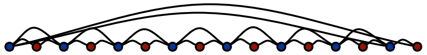
(phylogenetic) trees



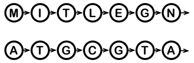
chemical compounds



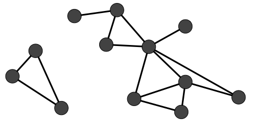
overlap of sequence fragments



protein structures

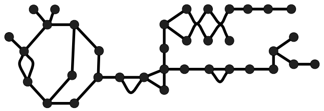


DNA, protein sequence

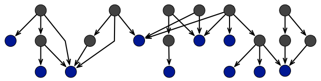


animal movement network

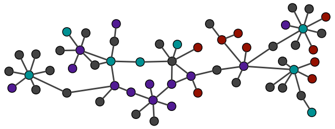
Graphs in biology



metabolic networks



gene regulation



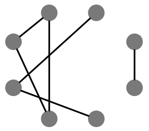
protein-protein interactions

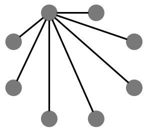
Understanding biological systems

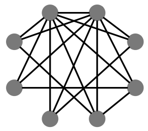
describing graphs

- type, number of nodes, number of connections (total, per node), shortest paths, subgraphs, ...
  - degree: num. of links one node has
  - diameter: max distance between any two nodes
  - components: disconnected sets of nodes
  - clustering coefficient: measures presence of subsets
  - ...

specific terms are used







Understanding biological systems

describing graphs

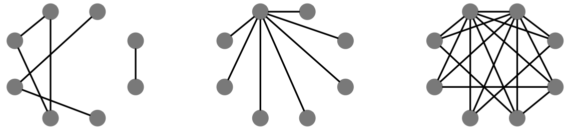
- type, number of nodes, number of connections (total, per node), shortest paths, subgraphs, …

specific terms are used

analyzing graphs

- search for substructures or paths
- modelling flow through the network

efficient algorithms exist

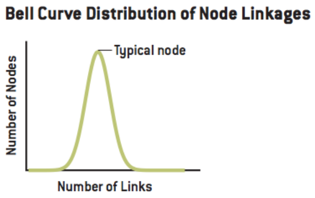


Early network theory (~1960-2000)

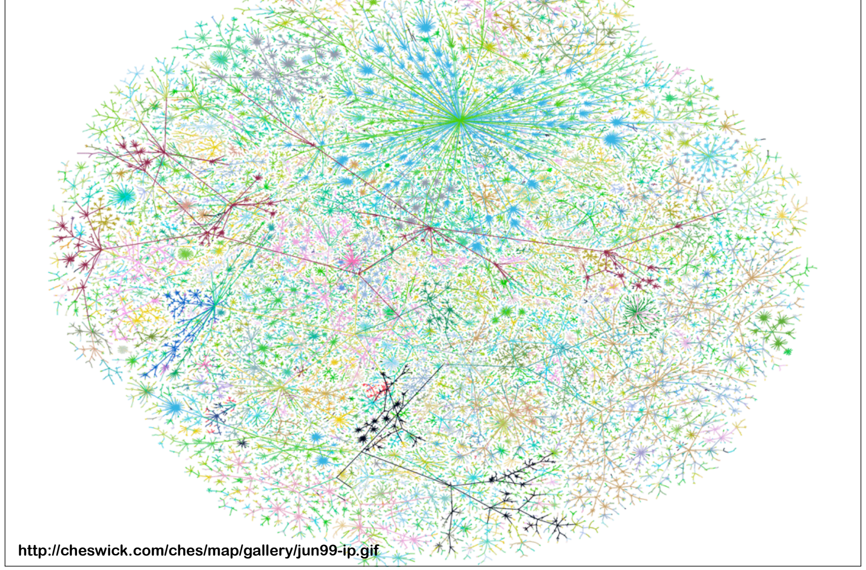
random networks

- edges are placed randomly
- most nodes have about the same number of edges
- example: US highway system

images: Scientific American, 2003



A (partial) map of the internet (1999)



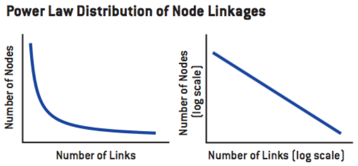
http://cheswick.com/ches/map/gallery/jun99-ip.gif

Scale-free networks

scale-free network

- large!
- a few nodes have lots of edges (→ hubs)
- most nodes have very few edges
- example: US airline system

images: Scientific American, 2003



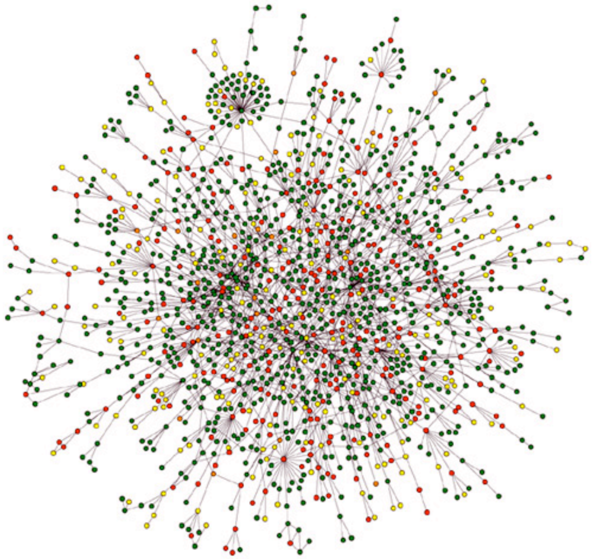
Power Law Distribution of Node Linkages
Number of Nodes
Number of Links
Number of Nodes (log scale)
Number of Links (log scale)

Large networks

Network	Nodes	Links	Directed / Undirected	nodes	edges (links)	‹K›
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorships	Undirected	23,133	93,437	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	10.43
E.coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

http://barabasi.com/networksciencebook/

Protein-protein interaction network (yeast)



Scale-free networks

- WWW, internet
- social networks
- scientific publications / authors
- network of Hollywood actors
- cellular metabolism
  - most molecules participate in 1-2 reactions
  - a few molecules participate in most reactions
- protein-protein interactions
  - most proteins interact with 1-2 other proteins
  - a few proteins interact with many proteins

Scale-free networks

properties

- robust!
- random / accidental failures generally don’t bring down the entire system
- vulnerable when hubs fail

implications & applications (biology!)

- random mutations → unaffected proteins allow the network to keep working
- basic science: understand complex systems
- applied science: drug development

Example: Modelling metabolic fluxes

- reconstruction of metabolic network
  - key pathways? species-specific differences? alternative pathways? hubs?
- analyze structure, resources, adaptability
  - differential equations, matrix operations
  - model network behavior under different conditions
- verification with experimental data (isotopes)
- detailed dynamic analysis of kinetics

→ systems biology

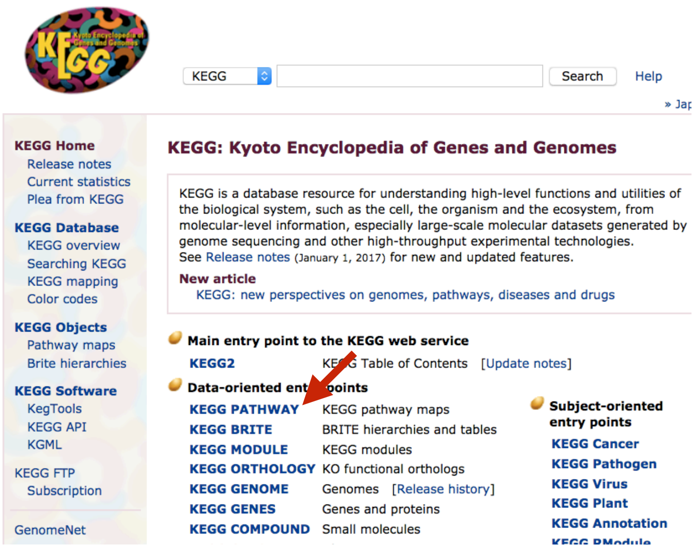
Network science

- interdisciplinary
  - different networks, but same general properties, same language, same general tasks
  - cross-fertilization despite different goals, details, challgenges
- data-driven
  - based on graph theory, but based on real data
- quantitative
  - graph theory, mathematics, statistics, engineering, etc
- computational

Today's computer lab

KEGG: Kyoto Encyclopedia of Genes and Genomes

- database of metabolic networks
- generic networks
- species-specific networks



KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
See Release notes (January 1, 2017) for new and updated features.

New article
KEGG: new perspectives on genomes, pathways, diseases and drugs

Main entry point to the KEGG web service

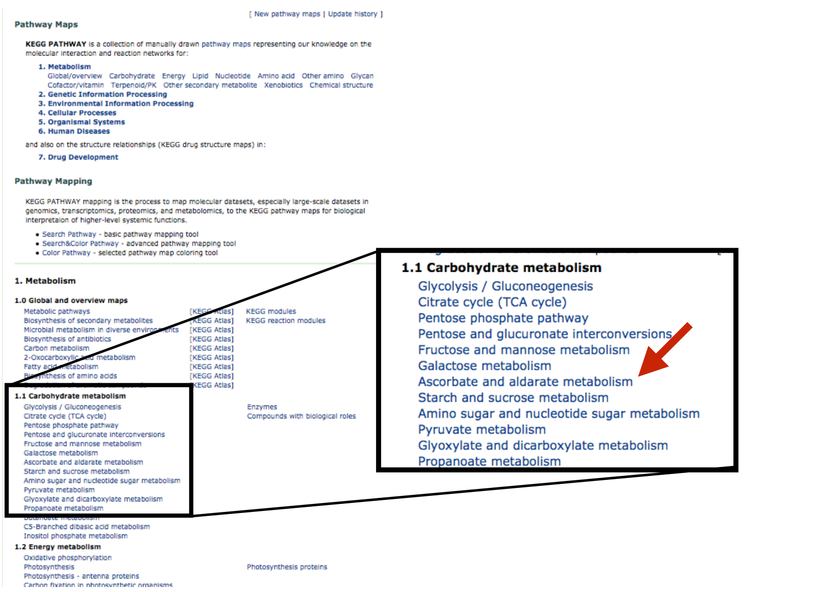
KEGG2	KEGG Table of Contents [Update notes]

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps
KEGG BRITE	BRITE hierarchies and tables
KEGG MODULE	KEGG modules
KEGG ORTHOLOGY	KO functional orthologs
KEGG GENOME	Genomes [Release history]
KEGG GENES	Genes and proteins
KEGG COMPOUND	Small molecules

Subject-oriented entry points

- **KEGG Cancer**
- **KEGG Pathogen**
- **KEGG Virus**
- **KEGG Plant**
- **KEGG Annotation**
- **KEGG RModule**



[New pathway maps | Update history]

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:

1. **Metabolism**
   Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan
   Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human Diseases**

and also on the structure relationships (KEGG drug structure maps) in:

7. **Drug Development**

Pathway Mapping

KEGG PATHWAY mapping is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretaion of higher-level systemic functions.

- Search Pathway - basic pathway mapping tool
- Search&Color Pathway - advanced pathway mapping tool
- Color Pathway - selected pathway map coloring tool

1. Metabolism

1.0 Global and overview maps

Metabolic pathways	[KEGG Atlas]	KEGG modules
Biosynthesis of secondary metabolites	[KEGG Atlas]	KEGG reaction modules
Microbial metabolism in diverse environments	[KEGG Atlas]	
Biosynthesis of antibiotics	[KEGG Atlas]	
Carbon metabolism	[KEGG Atlas]	
2-Oxocarboxylic acid metabolism	[KEGG Atlas]	
Fatty acid metabolism	[KEGG Atlas]	
Biosynthesis of amino acids	[KEGG Atlas]	
Degradation of aromatic compounds	[KEGG Atlas]	

1.1 Carbohydrate metabolism

Glycolysis / Gluconeogenesis	Enzymes
Citrate cycle (TCA cycle)	Compounds with biological roles
Pentose phosphate pathway	
Pentose and glucuronate interconversions	
Fructose and mannose metabolism	
Galactose metabolism	
Ascorbate and aldarate metabolism	
Starch and sucrose metabolism	
Amino sugar and nucleotide sugar metabolism	
Pyruvate metabolism	
Glyoxylate and dicarboxylate metabolism	
Propanoate metabolism	
Butanoate metabolism	
C5-Branched dibasic acid metabolism	
Inositol phosphate metabolism	

1.2 Energy metabolism

Oxidative phosphorylation	
Photosynthesis	Photosynthesis proteins
Photosynthesis - antenna proteins	
Carbon fixation in photosynthetic organisms	

1.1 Carbohydrate metabolism
Glycolysis / Gluconeogenesis
Citrate cycle (TCA cycle)
Pentose phosphate pathway
Pentose and glucuronate interconversions
Fructose and mannose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Starch and sucrose metabolism
Amino sugar and nucleotide sugar metabolism
Pyruvate metabolism
Glyoxylate and dicarboxylate metabolism
Propanoate metabolism

Terms and concepts

- graph, node, vertex, edge, link, Eulerian path, (un)directed graph, (un)weighted graph, adjacency matrix/list, hub

- graphs in biology (examples, nodes, edges, types)
- using graphs to solve problems
  - homolog assignment, sequence assembly
- using graphs to understand biological systems
  - description:
    random vs. scale free network
    (examples, properties, implications)
  - analysis & modelling not covered!