Andreas Holzinger
VO 709.049 Medical Informatics
11.11.2015 11:15-12:45

Lecture 05

Semi structured, weakly structured data Graphs, Networks and Homologies

a.holzinger@tugraz.at
Tutor: markus.plass@student.tugraz.at
http://hci-kdd.org/biomedical-informatics-big-data

Schedule

Keywords of the 5th Lecture

- Big data pools
- Complex networks
- Computational graph representation
- Electronic patient record (EPR)
- Homology modeling
- Macroscopic structures
- Medical documentation
- Metabolic network
- Microscopic structures
- Network metrics
- Structural data dimension
- Topological structures

Advance Organizer (1/3) A-G

- **Adjacency matrix** = simplest form of computational graph representation, in which 0 or 1 denotes whether or not there is a directed edge from one node to another (in graph theory adjacent nodes in a graph are linked by an edge);
- **Artifacts** = not only a noise disturbance, which is contaminating and influencing the signal (surrogates) but also data which is wrong, however interpreted as to be reliable, consequently may lead to a wrong decision;
- **Computational graph representation** = e.g. by adjacency matrices
- **Data fusion** = data integration techniques that analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion (e.g. combined sensor data in Ambient Assisted Living);
- **Global Distance Test (GDT)** = a measure of similarity between two protein structures with identical amino acid sequences but different tertiary structures. It is most commonly used to compare the results of protein structure prediction to the experimentally determined structure as measured by X-ray crystallography or protein NMRM;
- **Graph theory** = study of mathematical structures to model relations between objects from a certain collection;
- **Graphs** = a hypothetical structure consisting of a series of nodes connected by weighted edges (graphs can be directed/undirected and stoichometric/non-stoichometric regarding interaction classes);

Advance Organizer (2/3) H-P

- **Homology** = in mathematics (especially algebraic topology and abstract algebra), it is (ὁμόιος homos = "identical") a certain general procedure to associate a sequence of Abelian groups (i.e. does not depend on their order) or modules with a given mathematical object such as a topological space or a group;
- **Homology modeling** = comparative modeling of protein, refers to constructing an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein (the "template"); in Bioinformatics, homology modeling is a technique that can be used in molecular medicine.
- **In silico** = via computer simulation, in contrast to in vivo (within the living) or in vitro (within the glass);
- **Multi-scale representation** = in a graph, nodes do not have to represent biological objects on the same scale, one node (e.g. a molecule) may have an edge connecting it to a node representing a cell or tissue (the edge indicates that the molecule exerts an effect on the cell/tissue);
- **Network** = graphs containing cycles or alternative paths;
- **Network analysis** = a set of techniques used to characterize relationships among discrete nodes in a graph or a network;
- **Network topology** = the shape or structure of a network;
- **Petri-Net** = a special class of graph, consisting of two general classes or node: place and transition nodes;
- **Predictive modeling** = a set of techniques in which a mathematical model is created or chosen to best predict the probability of an outcome (e.g. regression);
- **P-System** = addresses the slowness of Petri-nets

Advance Organizer (3/3) R-V

- **Radius of a graph** = average minimum path length (biological networks are not arranged in a regular or symmetrical pattern);
- **Scale-free Topology** = ensures that there are very short paths between any given pair of nodes, allowing rapid communication between otherwise distant parts of the network (e.g. the Web has such a topology);
- **Semi-structured data** = does not conform with the formal structure of tables/data models assoc. with relational databases, but at least contains tags/markers to separate semantic elements and enforce hierarchies of records and fields within the data; aka schemaless or self-describing structure; the entities belonging to the same class may have different attributes even though they are grouped together;
- **Spatial analysis** = a set of techniques, applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data set;
- **Structural homology** = similar structure but different function;
- **Supervised learning** = machine learning techniques that infer a function or relationship from a set of training data (e.g. classification and support vector machines);
- **Time series analysis** = set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data;
- **Time series forecasting** = use of a model to predict future values of a time series based on known past values of the same or other series (e.g. structural modeling); decomposition of a series into trend, seasonal, and residual components, which can be useful for identifying cyclical patterns in the data;
- **Unstructured data** = complete randomness, noise; (wrongly, text is called unstructured, but there is some structure, too, so text data is a kind of weakly structured data);
- **Vertex degree** = within a topology, the numbers of edges connecting to a node;

Glossary

- ANSI = American National Standards Institute
- CD = cardiac development
- CDA = Clinical Document Architecture
- CHD = congenital heart disease
- CMM = Correlated motif mining
- DPI = Dossier Patient Integre´ = integrated patient record
- E = Edge
- EPR = Electronic Patient Record
- G(V,E) = Graph
- GI = gastrointestinal
- HER = Electronic Health Record
- HL7 = Health Level 7
- KEGG = Kyoto Encyclopedia of Genes and Genomes
- NP = nondeterministic polynomial time
- OWL = Web Ontology Language
- PPI = Protein-Protein Interaction
- SGML = Standard Generalized Markup Language
- TF= Transcription factor
- TG = Target Gene
- V = Vertex
- XML = Extensible Markup Language

Learning Goals … at the end of the 5th lecture you …

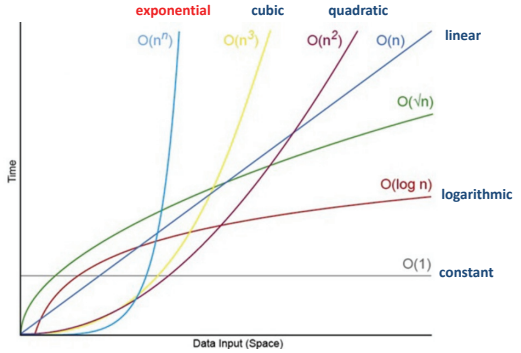
- … have an idea of the **complexity of data** in biomedical informatics
- … are aware of the various **contents** of Electronic Patient Records
- … have seen some application examples of **network structures** from both macro-cosmos and micro-cosmos and are fascinated about it;
- … have a rough overview about some basics of how to **get point clouds** out of data sets
- … have an understanding of the challenges of **network science**

Key Problems

- Automated Machine Learning algorithms need much training data – focus is on adjusting model parameters without fully **understanding the data** that the learning algorithm is modeling [1]
- Curse of dimensionality [2] – need for privacy and **anonymization** [3] (see lecture 11)
- **Weakly structured data** [4]

[1] Smith, M. R., Martinez, T. & Giraud-Carrier, C. 2014. An instance level analysis of data complexity. *Machine learning,* 95, (2), 225-256.
[2] Friedman, J. H. 1997. On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data mining and knowledge discovery, 1, (1), 55-77.
[3] Aggarwal, C. C. On k-anonymity and the curse of dimensionality. Proceedings of the 31st international conference on Very large data bases VLDB, 2005. 901-909
[4] Holzinger, A., Stocker, C. & Dehmer, M. 2014. Big Complex Biomedical Data: Towards a Taxonomy of Data. In: CCIS 455. Berlin Heidelberg: Springer pp. 3-18.

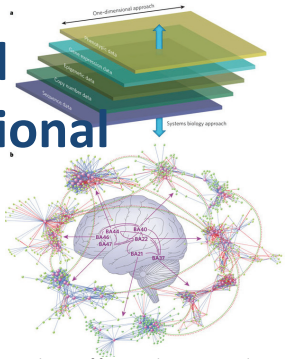
Complexity Problem: Time versus Space



P versus NP and the Computational Complexity Zoo, please have a look at https://www.youtube.com/watch?v=YX40hbAHx3s

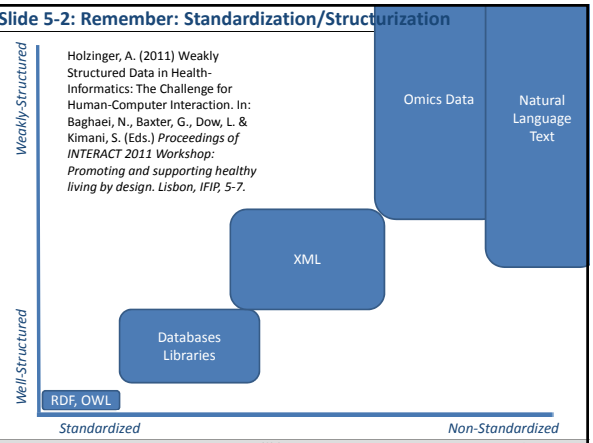
Slide 5-1: Mathematically seen our world is …

Complex and High dimensional



Geschwind, D. H. & Konopka, G. 2009. Neuroscience in the era of functional genomics and systems biology. *Nature, 461, (7266), 908-915.*

Slide 5-2: Remember: Standardization/Structurization



Slide 5-3: Example: Well-Structured Data

http://care2x.org

Slide 5-4: Example: Semi-structured Data: XML

```
<?xml version="1.0"?>
<patient>
        <patient-id>11111</patient-id>
        <Name>Chen</Name>
        <Date of Birth>1.1.1900</Date of Birth>
                <diagnosis>
                        <code>123</code>
                        <diagnosistext>Myocardinfarct</diagnosistext>
                </diagnosis>
</patient>
```

Holzinger, A. (2003) *Basiswissen IT/Informatik. Band 2: Informatik. Das Basiswissen für die Informationsgesellschaft des 21. Jahrhunderts. Wuerzburg, Vogel Buchverlag.*

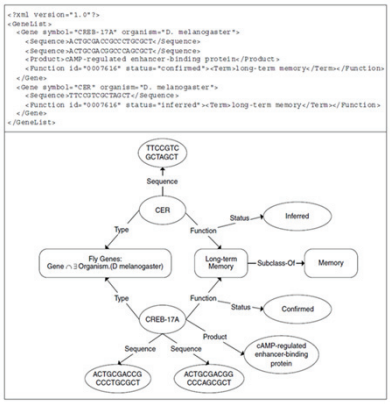
Slide 5-5 Example: Generic XML template for a med. report

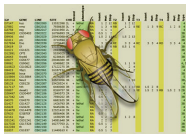
```
<DOC>
 <HEADER>
   <DESCRIPTION> Short label for the
                 document </DESCRIPTION>
   <DOCTYPE> Document type </DOCTYPE>
   <AUTHORID> Author </AUTHOR.ID>
   <DOCID> Unique identifier of the
           document </DOCID>
    … Any other relevant tags
 </HEADER>

 <BODY>
   <STRUCDOC>
     <PARAS>
       <PARA CONTENT="Content type"
             STATUS="Present or optional">
         <TITLE>Paragraph title</TITLE>
         <CONTENT>Textual content</CONTENT>
       </PARA>
         … see examples of paragraphs below
       <PARA CONT="Text" STATUS="Present">
         <PTITLE>MOTIF D'HOSPITALISATION
         </PTITLE>
       </PARA>
       <PARA CONT="Text" STATUS="Present">
         <PTITLE>ANTECEDENTS PERSONNELS
         </PTITLE>
       </PARA>
       <PARA CONT="Text" STATUS="Present">
         <PTITLE>ANAMNESE ACTUELLE
         </PTITLE>
       </PARA>
       …
     </PARAS>
   </STRUCDOC>
   <FULLDOC>
     <![CDATA[……]]>
   </FULLDOC>
 </BODY>
</DOC>
```

DPI = Dossier Patient Integre´ = integrated patient record

Rassinoux, A.-M., Lovis, C., Baud, R. & Geissbuhler, A. (2003) XML as standard for communicating in a document-based electronic patient record: a 3 years experiment. *International Journal of Medical Informatics, 70, 2-3, 109-115.*

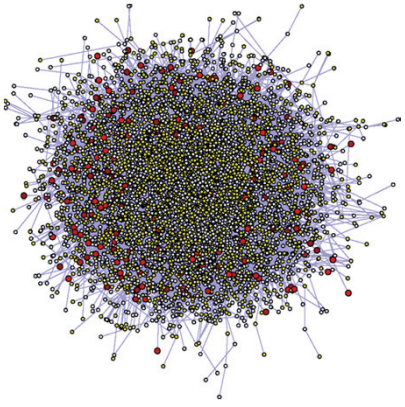
Slide 5-6 Comparison of XML - RDF/OWL in Bioinformatics





Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. & Tarczy-Hornoch, P. 2007. Data integration and genomic medicine. *Journal of Biomedical Informatics, 40, (1), 5-16.*

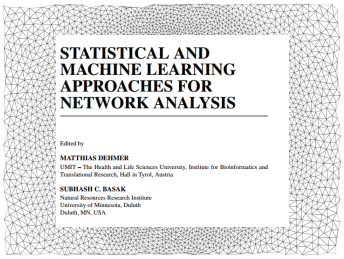
Slide 5-6 Example: Weakly structured data set - PPI

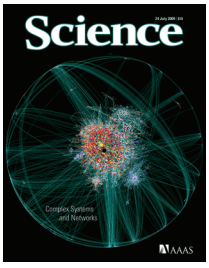


Kim, P. M., Korbel, J. O. & Gerstein, M. B. 2007. Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. Proceedings of the National Academy of Sciences, 104, (51), 20274-20279.

Network Science – Graph Theory

Networks = Graphs





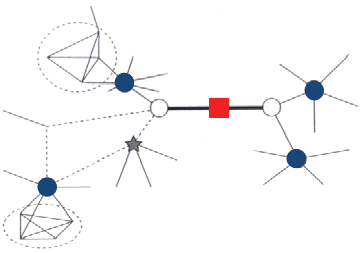
http://www.wired.com/tag/network-science/

Slide 5-7: Complex Biological Systems key concepts

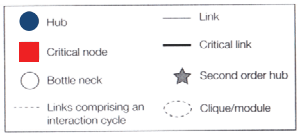
- In order to understand complex biological systems, the three following key concepts need to be considered:
- (i) **emergence,** the discovery of links between elements of a system because the study of individual elements such as genes, proteins and metabolites is insufficient to explain the behavior of whole systems;
- (ii) **robustness,** biological systems maintain their main functions even under perturbations imposed by the environment; and
- (iii) **modularity,** vertices sharing similar functions are highly connected.
- Network theory can largely be applied for biomedical informatics, because many tools are already available

Slide 5-8: Networks on the Example of Bioinformatics

$G(V, E)$ *Graph*
V … *vertex*
E … *edge* $\{a, b\}$
$a, b \in V; a \neq b$



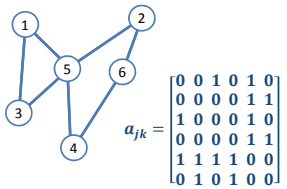
Hodgman, C. T., French, A. & Westhead, D. R. (2010) *Bioinformatics. Second Edition. New York, Taylor & Francis.*



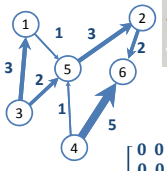
Slide 5-9: Computational Graph Representation

Adjacency (ə-ˈjā-sᵊn(t)-sē) **Matrix** $A = (a_{jk})$

$$a_{jk} = \begin{cases} 1, & if\ \{j,k\} \in E \\ 0, & otherwise \end{cases}$$



$$a_{jk} = \begin{bmatrix} 0&0&1&0&1&0 \\ 0&0&0&0&1&1 \\ 1&0&0&0&1&0 \\ 0&0&0&0&1&1 \\ 1&1&1&1&0&0 \\ 0&1&0&1&0&0 \end{bmatrix}$$



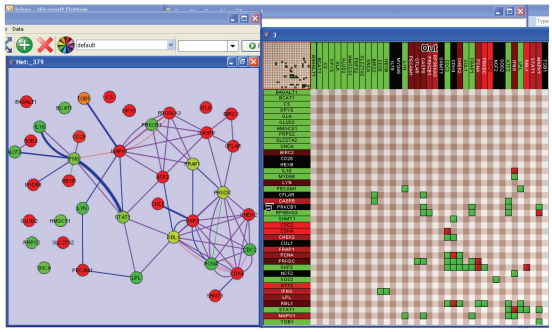
$$w_{jk} = \begin{bmatrix} 0&0&-3&0&1&0 \\ 0&0&0&0&0&3 \\ 3&0&0&0&2&0 \\ 0&0&0&0&1&-5 \\ 1&-2&0&0&-1&0 \\ 0&0&0&0&5&0 \end{bmatrix}$$

Simple graph, symmetric, binary

Directed and weighted

For more information: Diestel, R. (2010) *Graph Theory, 4th Edition. Berlin, Heidelberg, Springer.*

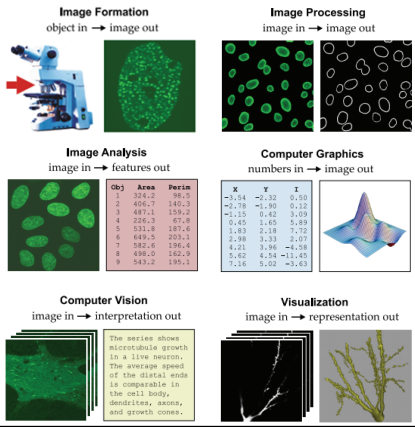
Slide 5-10: Example: Tool for Node-Link Visualization



Jean-Daniel Fekete http://wiki.cytoscape.org/InfoVis_Toolkit

Fekete, J.-D. The infovis toolkit. Information Visualization, INFOVIS 2004, 2004. IEEE, 167-174.

Excursus: Do *not* mix up Image Processing with Visualization (see L 09)

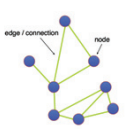


Meijering, Erik & Cappellen, Gert (2006) Biological Image Analysis Primer, available via http://www.imagescience.org/meijering/publications/1009/ Erasmus University Medical Center

Slide 5-11: Some Network Metrics (1/2)

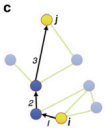
Order = total number of nodes n; **Size** = total number of links (a):

$$\sum_i \sum_j a_{ij}$$



Clustering Coefficient (b) = the degree of concentration of the connections of the node's neighbors in a graph and gives a measure of local inhomogeneity of the link density:

$$C_i = \frac{2t_i}{k(k_i - 1)} \qquad C = \frac{1}{n}\sum_i C_i$$



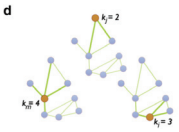
Path length (c) = is the arithmetical mean of all the distances:

$$l = \frac{1}{n(n-1)}\sum_{i \neq j} d_{ij}$$

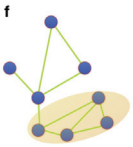
Costa, L. F., Rodrigues, F. A., Travieso, G. & Boas, P. R. V. (2007) Characterization of complex networks: A survey of measurements. *Advances in Physics, 56, 1, 167-242.*

Slide 5-12: Some Network Metrics (2/2)

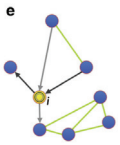
- **Centrality** (d) = the level of "betweenness- centrality" of a node I ("hub-node in Slide 28);



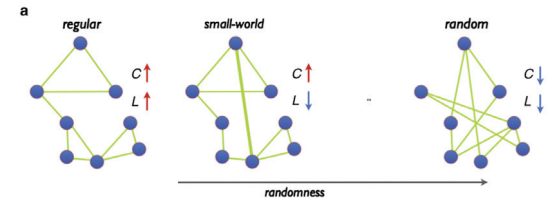
- **Nodal degree** (e) = number of links connecting *i* to its neighbors: $k_i = \sum_i a_{ij}$

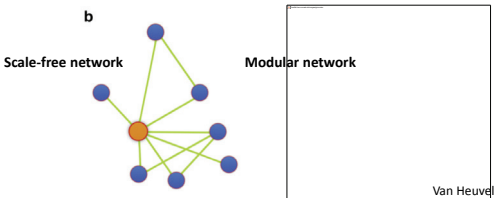


Modularity (f) = describes the possible formation of communities in the network, indicating how strong groups of nodes form relative isolated sub-networks within the full network (refer also to Slide 5-8).



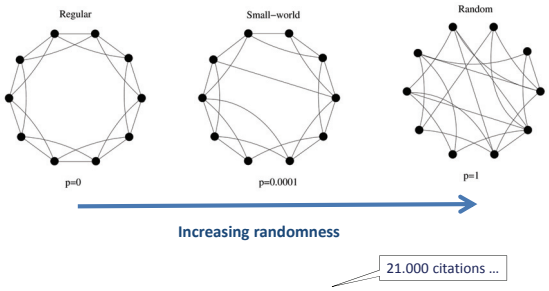
Slide 5-13: Network Topologies





Van Heuvel & Hulshoff (2010)

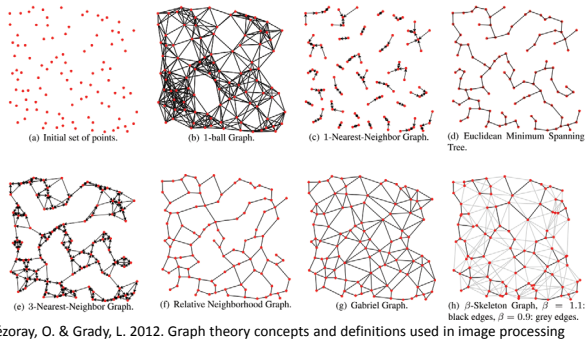
Slide 5-14: Small-World Networks



Watts, D. J. & Strogatz, S. (1998) Collective dynamics of small-world networks. *Nature, 393, 6684, 440-442.*

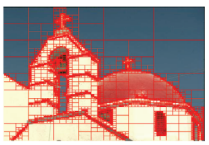
Milgram, S. 1967. The small world problem. *Psychology today, 2, (1), 60-67.*

Slide 5-15 Graphs from Point Cloud Data Sets



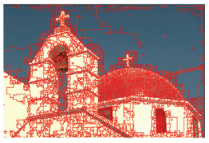
Lézoray, O. & Grady, L. 2012. Graph theory concepts and definitions used in image processing and analysis. *In: Lézoray, O. & Grady, L. (eds.) Image Processing and Analysing With Graphs: Theory and Practice. Boca Raton (FL): CRC Press, pp. 1-24.*

Slide 5-16 Graphs from Images



a) quadtree tessellation

b) RAG assoc. to the quadtree



c) Watershed Algorithm

d) SLIC superpixels

Lézoray, O. & Grady, L. 2012. Graph theory concepts and definitions used in image processing and analysis. *In: Lézoray, O. & Grady, L. (eds.) Image Processing and Analysing With Graphs: Theory and Practice. Boca Raton (FL): CRC Press, pp. 1-24.*

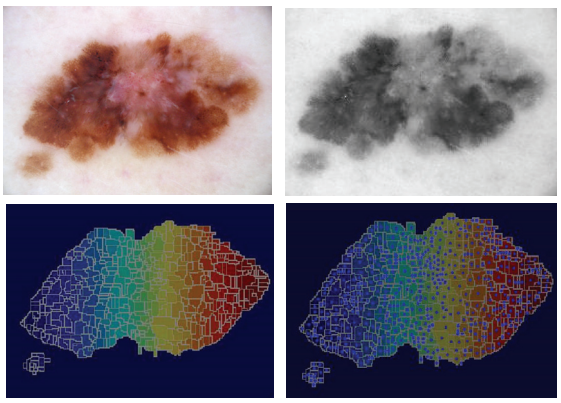
Slide 5-18 Example Watershed Algorithm

Algorithm 4.2 Watershed transform w.r.t. topographical distance based on image integration via the Dijkstra-Moore shortest paths algorithm.

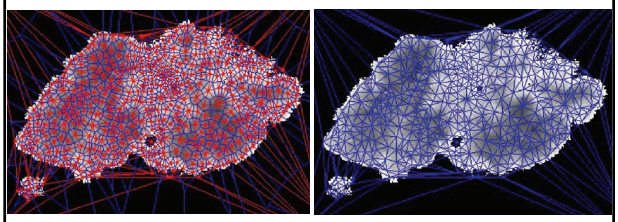
```
1: procedure ShortestPathWatershed;
2: INPUT: lower complete digital grey scale image G = (V, E, im) with cost function cost.
3: OUTPUT: labelled image lab on V.
4: #define WSHED 0                    (* label of the watershed pixels *)
5: (* Uses distance image dist. On output, dist[v] = im[v], for all v ∈ V. *)
6:
7: for all v ∈ V do                   (* Initialize *)
8:   lab[v] ← 0 ; dist[v] ← ∞
9: end for
10: for all local minima m_i do
11:   for all v ∈ m_i do
12:     lab[v] ← i ; dist[v] ← im[v]   (* initialize distance with values of minima *)
13:   end for
14: end for
15: while V ≠ ∅ do
16:   u ← GetMinDist(V)               (* find u ∈ V with smallest distance value dist[u] *)
17:   V ← V \ {u}
18:   for all v ∈ V with (u, v) ∈ E do
19:     if dist[u] + cost[u, v] < dist[v] then
20:       dist[v] ← dist[u] + cost(u, v)
21:       lab[v] ← lab[u]
22:     else if lab[v] ≠ WSHED and dist[u] + cost[u, v] = dist[v] and lab[v] ≠ lab[u] then
23:       lab[v] = WSHED
24:     end if
25:   end for
26: end while
```

Meijster, A. & Roerdink, J. B. A proposal for the implementation of a parallel watershed algorithm. Computer Analysis of Images and Patterns, 1995. Springer, 790-795.

Slide 5-19 Graphs from Images: Watershed + Centroid



Slide 5-20 Graphs from Images: Voronoi <> Delaunay

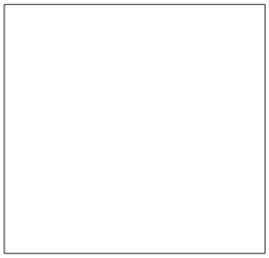


Holzinger, A., Malle, B. & Giuliani, N. 2014. On Graph Extraction from Image Data. In: Slezak, D., Peters, J. F., Tan, A.-H. & Schwabe, L. (eds.) Brain Informatics and Health, BIH 2014, Lecture Notes in Artificial Intelligence, LNAI 8609. Heidelberg, Berlin: Springer, pp. 552-563.

For Voronoi please refer to: Aurenhammer, F. 1991. Voronoi Diagrams - A Survey of a fundamental geometric data structure. *Computing Surveys,* 23, (3), 345-405.

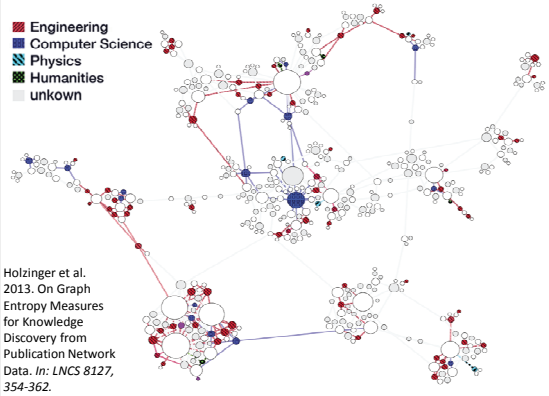
For Delaunay please refer to: Lee, D.-T. & Schachter, B. J. 1980. Two algorithms for constructing a Delaunay triangulation. Intl. Journal of Computer & Information Sciences, 9, (3), 219-242.

Slide 5-21 Points -> Voronoi -> Delaunay



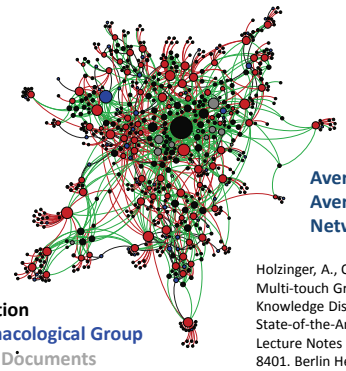
Kropatsch, W., Burge, M. & Glantz, R. 2001. Graphs in Image Analysis. *In: Kropatsch, W. & Bischof, H. (eds.) Digital Image Analysis. Springer New York, pp. 179-197.*

Slide 5-22 Example: Graph Entropy Measures



Holzinger et al. 2013. On Graph Entropy Measures for Knowledge Discovery from Publication Network Data. *In: LNCS 8127, 354-362.*

Slide 5-23: Example for a Medical Knowledge Space



 # Nodes: 641
 # Edges: 1250

Average Degree: 3,888
Average Path Length: 4.683
Network Diameter: 9

Agent
Condition
Pharmacological Group
Other Documents

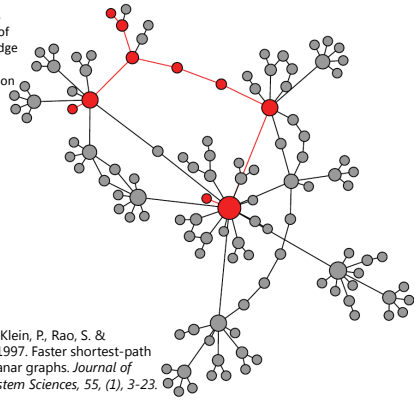
Holzinger, A., Ofner, B. & Dehmer, M. 2014. Multi-touch Graph-Based Interaction for Knowledge Discovery on Mobile Devices: State-of-the-Art and Future Challenges. In: Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 241-254

Slide 5-24: Medical Details of the Graph

- **Nodes**
  - **drugs**
  - clinical guidelines
  - **patient conditions** (indication, contraindication)
  - **pharmacological groups**
  - tables and calculations of medical scores
  - algorithms and other medical documents
- **Edges**: 3 crucial types of relations inducing medical relevance between two active substances
  - pharmacological groups
  - indications
  - contra-indications

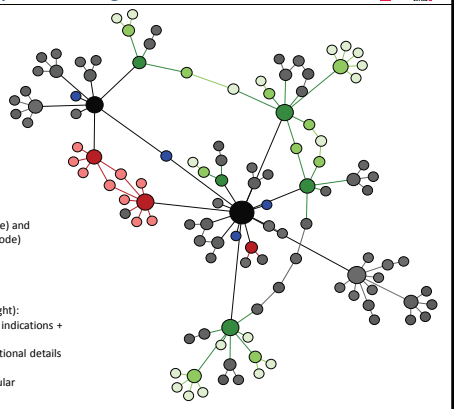
Slide 5-25: Example for the shortest path

Holzinger, A., et al. 2013. Constraints of List-based Knowledge Interaction. In: Medicine 2.0 London



Henzinger, M. R., Klein, P., Rao, S. & Subramanian, S. 1997. Faster shortest-path algorithms for planar graphs. *Journal of Computer and System Sciences, 55, (1), 3-23.*

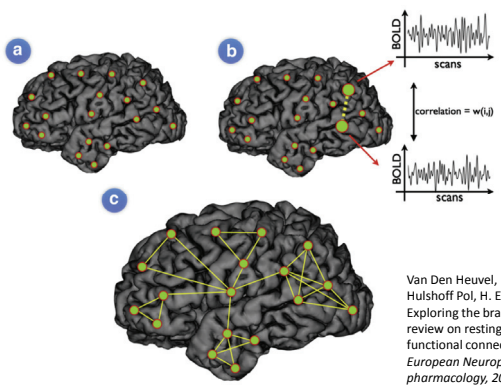
Slide 5-26: Example for finding related structures



Relationship between Adrenaline (center black node) and Dobutamine (top left black node)
Blue: Pharmacological Group
Dark red: Contraindication;
Light red: Condition

Green nodes (from dark to light):
1. Application (one ore more indications + corresponding dosages)
2. Single indication with additional details (e. g. "VF after 3rd Shock")
3. Condition (e.g. VF, Ventricular Fibrillation)

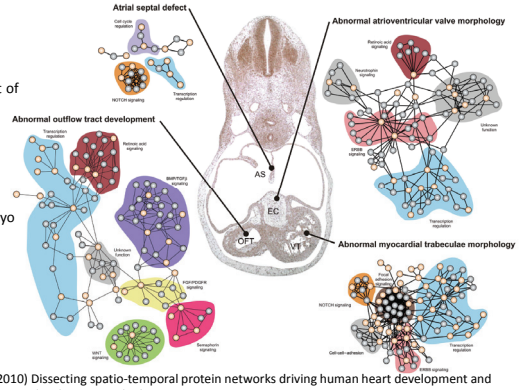
Slide 5-27: Example: The brain is a complex network



Van Den Heuvel, M. P. & Hulshoff Pol, H. E. (2010) Exploring the brain network: a review on resting-state fMRI functional connectivity. *European Neuropsycho-pharmacology, 20, 8, 519-534.*

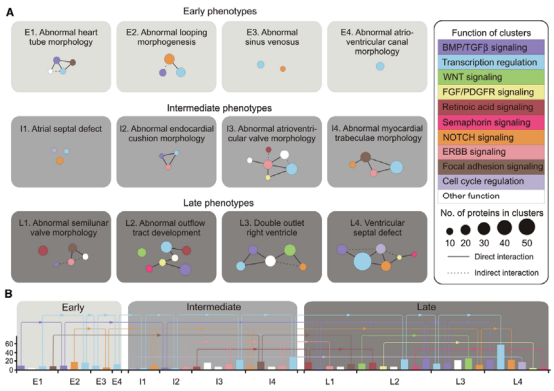
Slide 5-28: Representative Examples of disease complexes

Examples of 4 functional networks driving the development of different anatomical structures in the human heart of a 37-day old human embryo



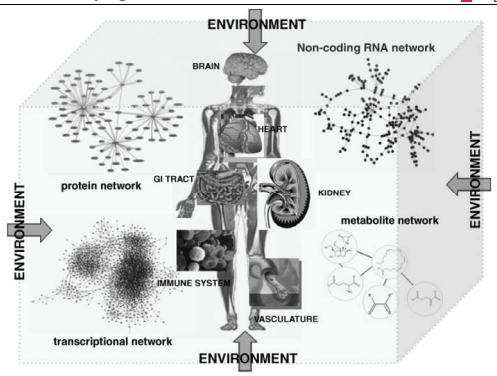
Lage, K. et. al (2010) Dissecting spatio-temporal protein networks driving human heart development and related disorders. *Molecular systems biology, 6, 1, 1-9.*

Slide 5-29: Example: Cell-based therapy



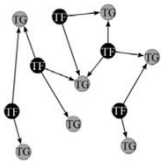
Lage et. al (2010)

Slide 5-30: Identifying Networks in Disease Research

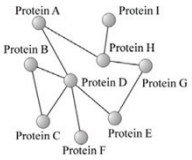


Schadt, E. E. & Lum, P. Y. (2006) Reverse engineering gene networks to identify key drivers of complex disease phenotypes. *Journal of lipid research, 47, 12, 2601-2613.*

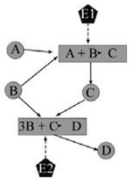
Slide 5-31: Three main types of biomedical networks



Transcriptional regulatory network with two components:
TF = transcription factor
TG = target genes
(TF regulates the transcription of TG)



Protein-Protein interaction network

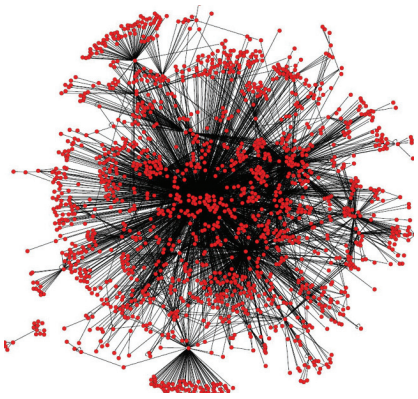


Metabolic network (constructed considering the reactants, chemical reactions and enzymes)

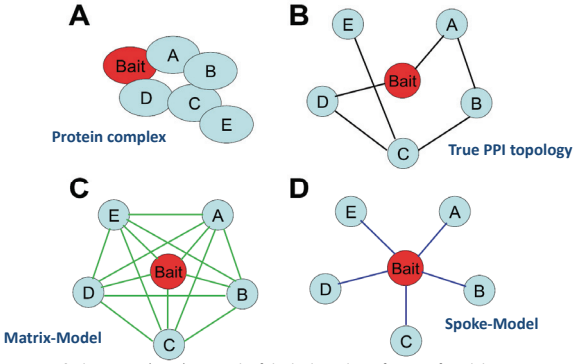
Costa, L. F., Rodrigues, F. A. & Cristino, A. S. (2008) Complex networks: the key to systems biology. *Genetics and Molecular Biology, 31, 3, 591–601.*

Slide 5-32: Example Transcriptional Regulatory Network

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Peralta-Gil, M., Peñaloza-Spínola, M. I., Martínez-Antonio, A., Karp, P. D. & Collado-Vides, J. 2006. The comprehensive updated regulatory network of Escherichia coli K-12. *BMC bioinformatics, 7, (1), 5.*

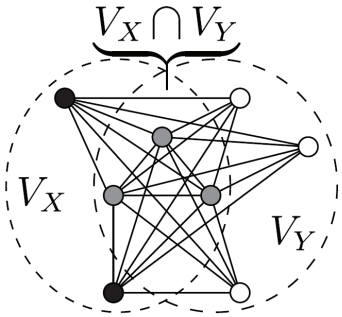


Slide 5-33: Network Representations of Protein Complexes



Wang, Z. & Zhang, J. Z. (2007) In search of the biological significance of modular structures in protein networks. *PLoS Computational Biology, 3, 6, 1011-1021.*

Slide 5-34 Correlated Motif Mining (CMM)



Boyen, P., Van Dyck, D., Neven, F., van Ham, R. C. H. J. & van Dijk, A. (2011) SLIDER: A Generic Metaheuristic for the Discovery of Correlated Motifs in Protein-Protein Interaction Networks. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 8, 5, 1344-1357.*

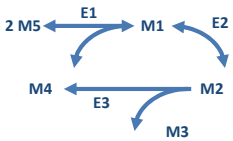
Slide 5-35 Steepest Ascent Algorithm applied to CMM

Input: PPI-network $G = (V, E, \lambda)$, $\ell, d \in \mathbb{N}$, $d < \ell$
Output: $\{X^*, Y^*\}$ best correlated motif pair found in G

1: $\{X^*, Y^*\} \leftarrow$ randomMotifPair()
2: $maxsup \leftarrow f(\{X^*, Y^*\}, G)$
3: $sup \leftarrow -\infty$
4: **while** $maxsup > sup$ **do**
5: $\{X, Y\} \leftarrow \{X^*, Y^*\}$
6: $sup \leftarrow maxsup$
7: **for all** $\{X', Y'\} \in N(\{X, Y\})$ **do**
8: **if** $f(\{X', Y'\}, G) > maxsup$ **then**
9: $\{X^*, Y^*\} \leftarrow \{X', Y'\}$
10: $maxsup \leftarrow f(\{X', Y'\}, G)$

Boyen et al. (2011)

Slide 5-36: Metabolic Network



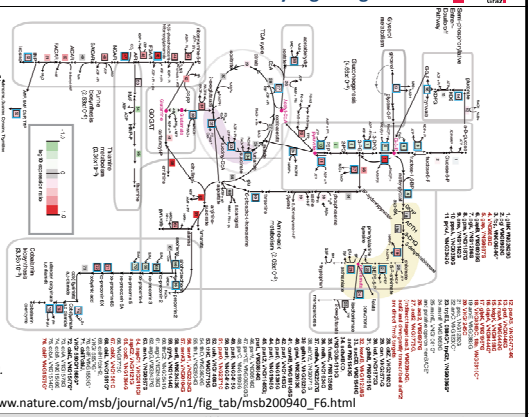
	M1	M2	M3	M4	M5
M1	0	1	0	1	1
M2	1	0	1	1	0
M3	0	0	0	0	0
M4	1	0	0	0	0
M5	1	0	0	0	0

M1	M2
M1	M4
M1	M5
M2	M1
M2	M3
M2	M4
M4	M1
M5	M1

Matrix contains many sparse elements - In this case it is computationally more efficient to represent the graph as an adjacency list

Hodgman, C. T., French, A. & Westhead, D. R. (2010) *Bioinformatics. Second Edition. New York, Taylor & Francis.*

Slide 5-37 Metabolic networks are usually big … big data …

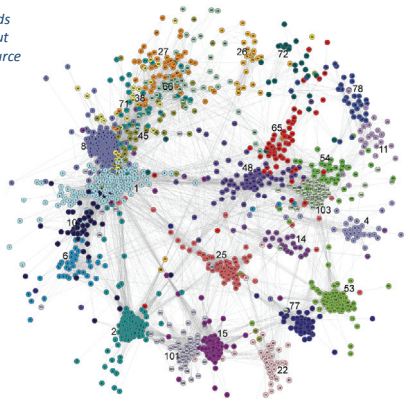


Schmid, A. K., Reiss, D. J., Pan, M., Koide, T. & Baliga, N. S. (2009) A single transcription factor regulates evolutionarily diverse but functionally linked metabolic pathways in response to nutrient availability. *Molecular Systems Biology, 5, 1-9.*

http://www.nature.com/msb/journal/v5/n1/fig_tab/msb200940_F6.html

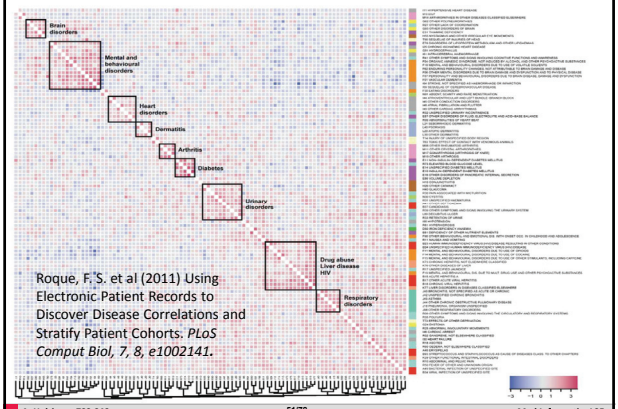
Slide 5-38 Using EPRs to Discover Disease Correlations

Electronic patient records remain a unexplored, but potentially rich data source for example to discover correlations between diseases.



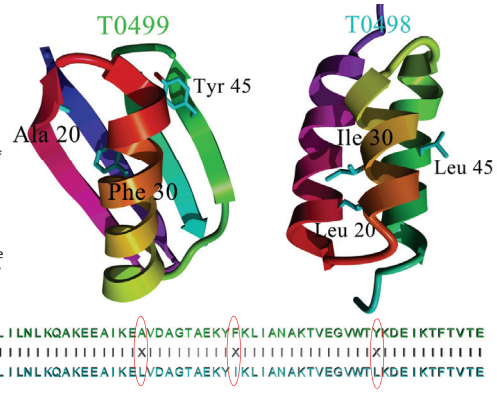
Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., Jensen, L. J. & Brunak, S. (2011) Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Computational Biology, 7, 8, e1002141.*

Slide 5-39: Heatmap of disease-disease correlations (ICD)



Roque, F. S. et al (2011) Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts. *PLoS Comput Biol, 7, 8, e1002141.*

Slide 5-40: Example: ὁμολογέω (homologeo)

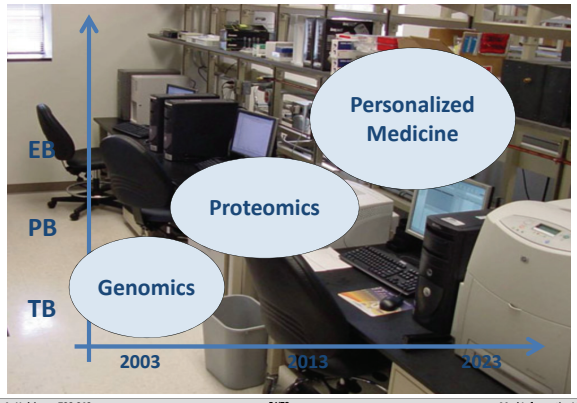


He, Y., Chen, Y., Alexander, P., Bryan, P. N. & Orban, J. (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. Proceedings of the National Academy of Sciences, 105, 38, 14412.

Slide 5-41 Conclusion

- Homology modeling is a knowledge-based prediction of protein structures.
- In homology modeling a protein sequence with an unknown structure (the target) is aligned with one or more protein sequences with known structures (the templates).
- The method is based on the principle that homologue proteins have similar structures.
- **Homology modeling will be extremely important to personalized and molecular medicine in the future.**

Slide 5-42: Future Outlook



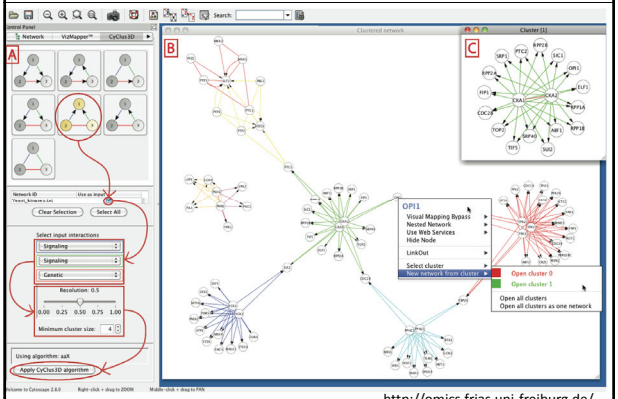
Sample Questions

- Which are the four main "big data" pools in the health care domain and what problems involved?
- What is the main problem in medical documentation?
- What is the advantage of an integrated Patient record?
- What are the advantages/disadvantages of XML/OWL for data in bioinformatics?
- What are the three key concepts in order to understand complex biological systems?
- What are the main symbols describing a network as used in Bioinformatics?
- How can networks represented computationally effectively?
- What are the main network metrics?
- What are the main network topologies used in Biomedical informatics?
- What is the Small-World Theory?
- Why is the study of networks relevant for medical professionals?
- Which are the three main types of biomedical networks?
- What is a Motif?
- What benefits can we gain from Correlated Motif Mining (CMM)?
- What is more efficient if a matrix contains many sparse elements?
- Why are structural homologies interesting for biomedical informatics?

Some Useful Links

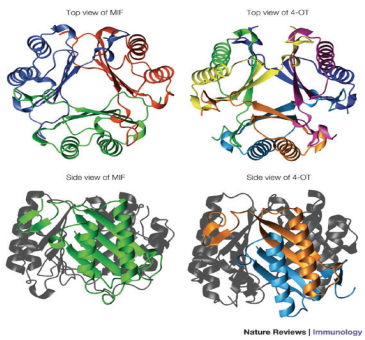
- http://www.cdisc.org
- http://www.w3.org/Math/
- http://www.sgpp.org/structures.shtml
- http://salilab.org/modeller
- http://swissmodel.expasy.org
- http://www.expasy.org/tools
- http://www.geneticseducation.nhs.uk

Appendix: clustering network motifs in integrated networks



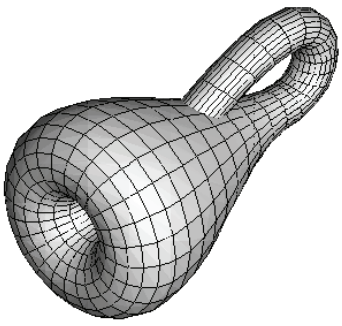
http://omics.frias.uni-freiburg.de/

Example from Immunology: Structural Homology



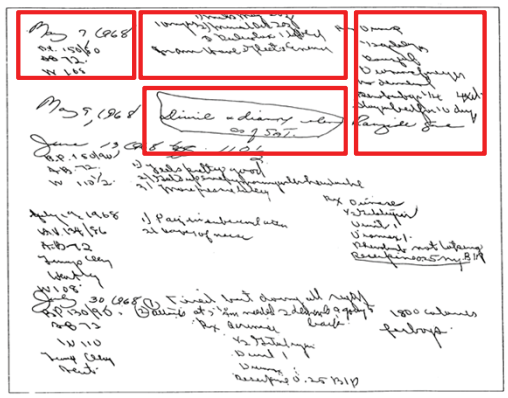
Calandra, T. & Roger, T. 2003. Macrophage migration inhibitory factor: a regulator of innate immunity. *Nat Rev Immunol, 3, 791-800.*

Klein Bottle



http://www.maa.org/cvm/1998/01/tprppoh/article/Pictures/KleinBottle.gif

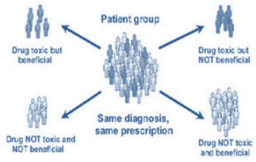
Medical Documentation – Patient Record



Medical Documentation - Electronic Patient Record

http://care2x.org

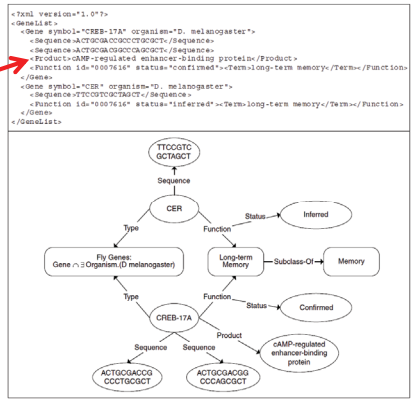
Challenge is in Genomic medicine …



- … to integrate and analyze these diverse and voluminous data sources to elucidate both normal and disease physiology.
- XML is suited for describing semi-structured data including a natural modeling of biological entities, because it allows features as e.g. nesting …

Example: Comparison of XML and OWL data in bioinformatics

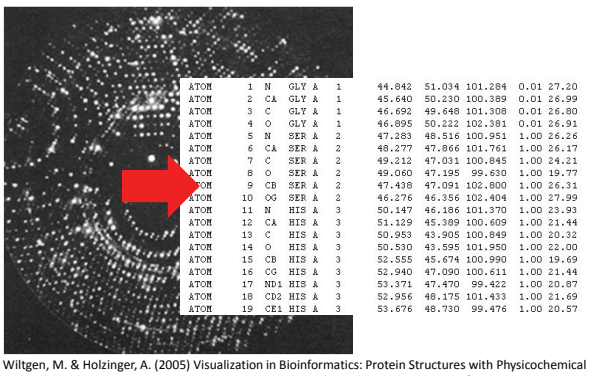
difficulty of modeling many-to-many relationships, such as the relationship between genes and functions



Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. & Tarczy-Hornoch, P. (2007) Data integration and genomic medicine. *Journal of Biomedical Informatics, 40, 1, 5-16.*

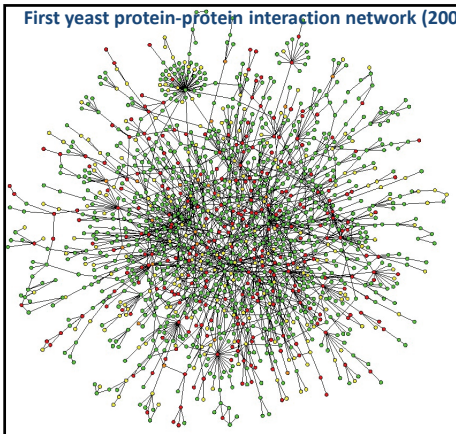
On time and space of data …

… to microscopic atomistic structures



Wiltgen, M. & Holzinger, A. (2005) Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: *Central European Multimedia and Virtual Reality Conference. Prague, Czech Technical University (CTU), 69-74*

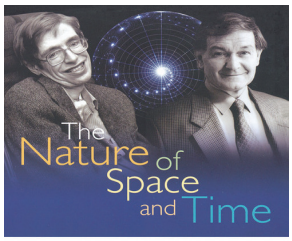
First yeast protein-protein interaction network (2001)



Nodes = proteins
Links = physical interactions (bindings)
Red Nodes = lethal
Green Nodes = non-lethal
Organge = slow growth
Yellow = not known

Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature, 411, 6833, 41-42.*

The Nature of Space and Time

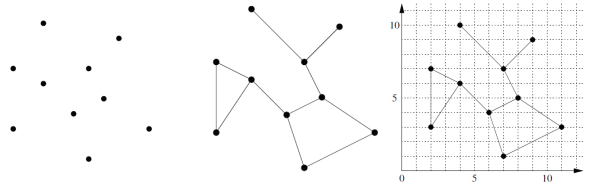


with a new afterword by the authors

STEPHEN HAWKING AND ROGER PENROSE

Clouds of data – unordered sequence of points in n-dim

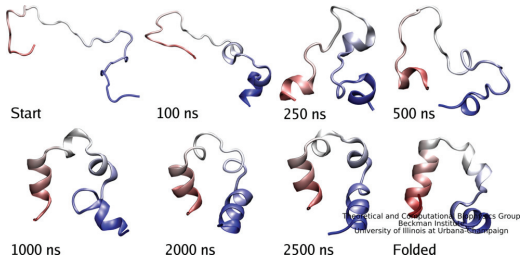
Let us collect n-dimensional i observations: $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{in}]$



Point cloud in $\mathbb{R}^2$ **topological space** **metric space**

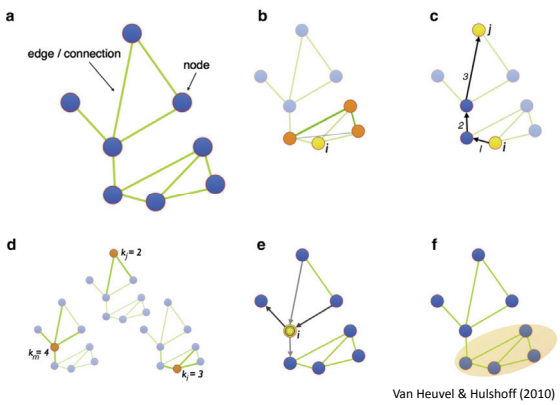
Zomorodian, A. J. 2005. *Topology for computing, Cambridge (MA), Cambridge University Press.*

Example: To predict the folding of a protein



Source: Theoretical and computational Biophysics Group: http://www.ks.uiuc.edu/

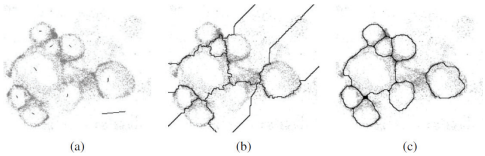
Backup Slide: Overview Some Network Metrics



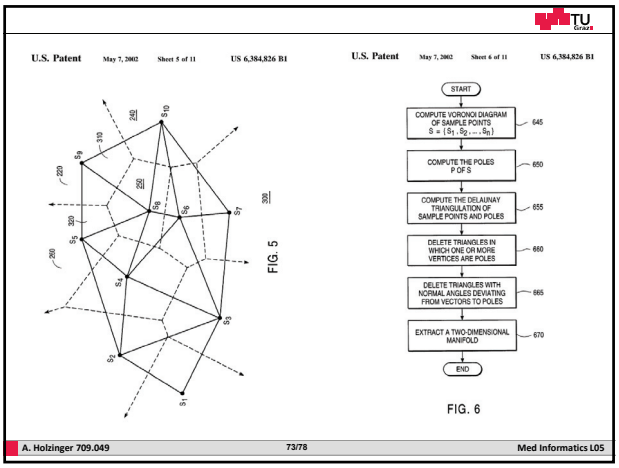
Van Heuvel & Hulshoff (2010)

Slide 5-19: Watershed Principle

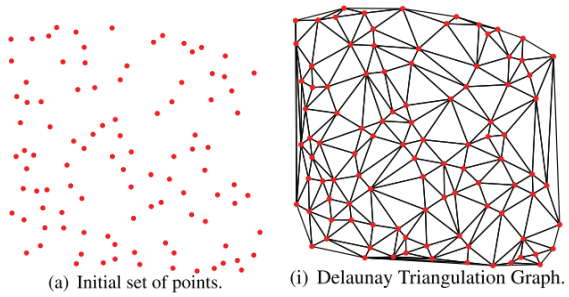
- Catchment basins:
  - treating an image as a height field or landscape, regions where the rain would flow into the same lake

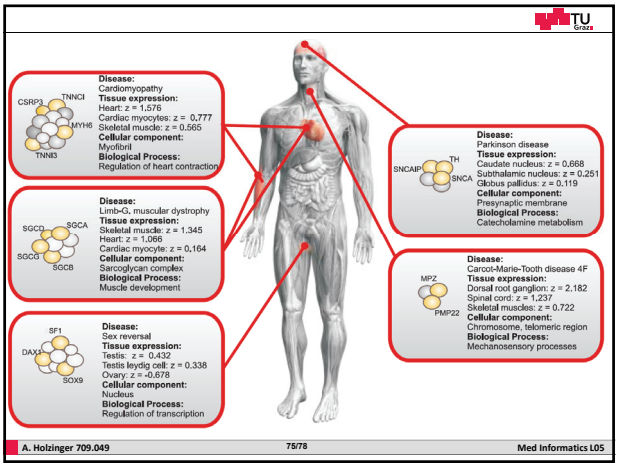


- Start flooding from local minima, and label ridges wherever differently evolving components meet

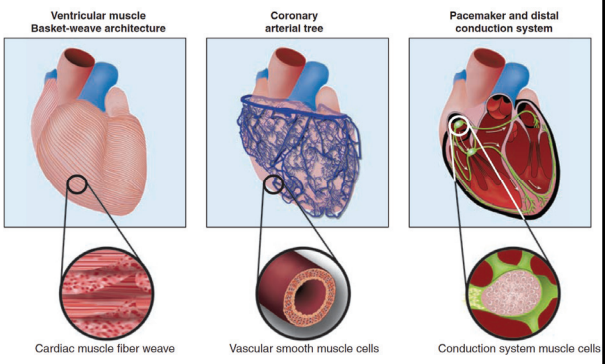


Slide 5-15 Graphs from Images: Voronoi <> Delauney



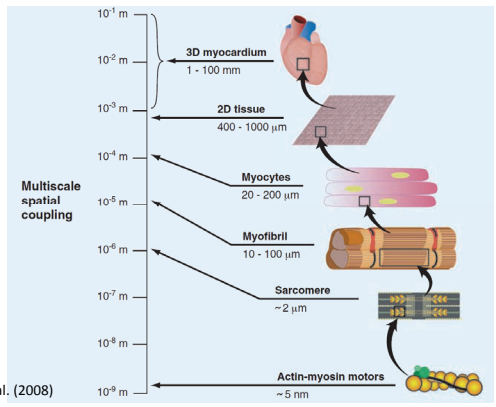


Example: Cell based therapy (1) (Heart transplantation)



Chien, K. R., Domian, I. J. & Parker, K. K. (2008) Cardiogenesis and the complex biology of regenerative cardiovascular medicine. *Science, 322, 5907, 1494.*

Example: Cell based therapy (2) (Heart transplantation)

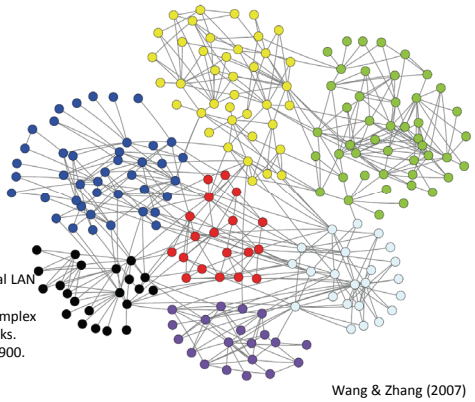


Chien et al. (2008)

Example: Network Generated by Gene Duplication

High Modularity (Modularity = 0.6717, Scaled Modularity = 29); Different colors represent different modules identified by Guimera and Amaral's algorithm [28].

Guimera R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433: 895–900.



Wang & Zhang (2007)