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 (surrugates) 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 NMR;
- 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 (allogic homo = "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 as in relational databases, but at least contains tags/markers to separate semantic elements and enforce hierarchies of records and fields within the data, aka schematics 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 number of edges connecting to a node;

Adv. Organizer 1/3: A-G

Adv. Organizer 2/3: H-P

Adv. Organizer 3/3: R-V

Schedule

1. Intro: Computer Science meets Life Sciences, challenges, future directions
2. Back to the future: Fundamentals of Data, Information and Knowledge
3. Structured Data: Coding, Classification (ICD, SNOMED, MedP, UMLS)
4. Biomedical Databases: Accessible, Information Retrieval for a node
5. Semi structured and weakly structured data (structural homologies)
6. Multimedia Data Mining and Knowledge Discovery
7. Knowledge and Decision: Cognitive Science & Human-Computer Interaction
8. Biomedical Decision Making: Reasoning and Decision Support
9. Intelligent Information Visualization and Visual Analytics
10. Biomedical Information Systems and Medical Knowledge Management
11. Biomedical Data: Privacy, Safety and Security

WS 2015 11.11.2015
Glossary

- ANSI = American National Standards Institute
- CD = cardiac development
- CDAA = Clinical Document Architecture
- CHD = congenital heart disease
- CMRA = Correlated motif mining
- DR = Dossier Patient Integre’ = integrated patient record
- E = Edge
- EPR = Electronic Patient Record
- G Ex = Graph
- Gl = gastrointestinal
- HER = Electronic Health Record
- HLF = 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

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]
- Weakly structured data [4]


Slide 5-1: Mathematically seen our world is ...

Complex and High dimensional


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

Complexity Problem: Time versus Space

P versus NP and the Computational Complexity Zoo, please have a look at https://www.youtube.com/watch?v=XX40bAa4Chs

Slide 5-2: Remember: Standardization/Structurization

Slide 5-3: Example: Well-Structured Data

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

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

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

Welcome to the Network Science – Graph Theory lecture. Today we will explore the relationship between networks and graphs, examining key concepts such as connectivity, centrality, and the applications of graph theory in various fields including biology, social sciences, and computer science. We will also discuss the importance of graph theory in the field of bioinformatics, particularly in the context of data integration and genomics.
Slide 5-7: Complex Biological Systems key concepts

- In order to understand complex biological systems, the three following key concepts need to be considered:
  - **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;
  - **robustness**, biological systems maintain their main functions even under perturbations imposed by the environment; and
  - **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

\[ G(V, E) \text{ Graph } \]
\[ V \ldots \text{vertex } (a, b) \]
\[ E \ldots \text{edge } (a, b) \]
\[ a, b \in V; a = b \]

Slide 5-9: Computational Graph Representation

Adjacency (\( |A| \times |V| \)) matrix \( A = (d_{ij}) \):

\[ d_{ij} = \begin{cases} 1, & \text{if } (j, k) \in E \\ 0, & \text{otherwise} \end{cases} \]

Simple graph, symmetric, binary

Directed and weighted


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

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


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

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

\[ \sum_{ij} \]

Clustering coefficient \( k \) = 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:

\[ C = \frac{2k}{n(n-1)} \]

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

\[ l = \frac{1}{n(n-1)} \sum_{ij} \]

- **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: \(k_i = \sum a_{ij}\)

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-B).

**Slide 5-13: Network Topologies**

- **Regular**
- **Small-world**
- **Random**
- **Scale-free**
- **Modular**

**Slide 5-14: Small-World Networks**

Increasing randomness


**Slide 5-15 Graphs from Point Cloud Data Sets**


**Slide 5-16 Graphs from Images**

- a) quadtree tessellation
- b) RAG assoc. to the quadtree
- c) Watershed Algorithm
- d) SLIC superpixels


**Slide 5-17 Example Watershed Algorithm**

**Slide 5-19: Graphs from Images: Watershed + Centroid**

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


**Slide 5-21: Points -> Voronoi -> Delaunay**


**Slide 5-22: Example: Graph Entropy Measures**


**Slide 5-23: Example for a Medical Knowledge Space**


**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


Slide 5-26: Example for finding related structures

Relationship between Adrenaline (center black node) and Dobutamine (top left black node). Dark red: Contraindication; Light red: Condition; Green nodes (from dark to light): 1. Application (one or 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


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


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.


Slide 5-29: Example: Cell-based therapy

Lage et al. (2010)


Slide 5-30: Identifying Networks in Disease Research

Lage et al. (2010)
**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)

**Slide 5-32: Example Transcriptional Regulatory Network**


**Slide 5-33: Network Representations of Protein Complexes**


**Slide 5-34 Correlated Motif Mining (CMM)**


**Slide 5-35 Steepest Ascent Algorithm applied to CMM**

**Slide 5-36: Metabolic Network**

Boyen et al. (2011)

Metabolic networks are usually big ... big data ...

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

Using EPRs to Discover Disease Correlations

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

Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts.

Heatmap of disease-disease correlations (ICD)

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.
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://www.geneticseducation.nhs.uk

Appendix: clustering network motifs in integrated networks

Example from Immunology: Structural Homology


Klein Bottle

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

Example from Immunology: Structural Homology

... 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...

First yeast protein-protein interaction network (2001)

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

Let us collect n-dimensional observations: $x_i = [x_{i1}, ..., x_{in}]$

Example: To predict the folding of a protein

Source: Theoretical and computational Biophysics Group: http://www.ics.uci.edu/

Backup Slide: Overview Some Network Metrics

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**
Graphs from Images: Voronoi <-> Delauney

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

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

Example: Network Generated by Gene Duplication
