# Lecture 5

## Semi-structured, Weakly Structured, and Unstructured Data

## 1 Learning Goals

At the end of this fourth lecture, you:

- would have acquired background knowledge on some issues in standardization and structurization of data;
- would have a general understanding of modeling knowledge in medicine and biomedical informatics;
- would get some basic knowledge on medical Ontologies and be aware of the limits, restrictions, and shortcomings of them;
- would know the basic ideas and the history of the International Classification of Diseases (ICD);
- would have a view on the Standardized Nomenclature of Medicine Clinical Terms (SNOMED CT);
- would have some basic knowledge on Medical Subject Headings (MeSH);
- would be able to understand the fundamentals and principles of the Unified Medical Language System (UMLS).

## 2 Advance Organizer

Abstraction — Process of mapping (biological) processes onto a series of concepts (expressed in mathematical terms)

Biological system — Collection of objects ranging in size from molecules to populations of organisms, which interact in ways that display a collective function or role

Coding — Any process of transforming descriptions of medical diagnoses and procedures into standardized code numbers, i.e., to track health conditions and for

A. Holzinger, *Biomedical Informatics*, DOI 10.1007/978-3-319-04528-3_5,

| | |
|---|---|
| | reimbursement, e.g., based on Diagnosis Related Groups (DRG) |
| Data model | Definition of entities, attributes, and their relationships within complex sets of data |
| DSM | Diagnostic and Statistical Manual for Mental Disorders, a multiaxial, multidimensional categorization of all (known) mental health disorders, used for clinical diagnostics |
| Extensible Markup Language (XML) | Set of rules for encoding documents in machine-readable form |
| GALEN | Generalized Architecture for Languages, Encyclopedias and Nomenclatures in Medicine, project aiming at the development of a reference model for medical concepts |
| ICD | International Classification of Diseases, the archetypical coding system for patient record abstraction (est. 1900) |
| Medical Classification | Provides the terminologies of the medical domain (or parts of it), 100+ classifications in use |
| MeSH | Medical Subject Headings is a classification to index the world medical literature and forms the basis for UMLS |
| Metadata | Data that describes the data |
| Model | A simplified representation of a process or object, which describes its behavior under specified conditions (e.g., conceptual model) |
| Nosography | Science of description of diseases |
| Nosology | Science of classification of diseases |
| Ontology | Structured description of a domain and formalizes the terminology (concepts-relations, e.g., IS-A relationship provides a taxonomic skeleton), e.g., gene ontology |
| Ontology engineering | Subfield of knowledge engineering, which studies the methods and methodologies for building ontologies |
| SNOMED | Standardized Nomenclature of Medicine, est. 1975, multiaxial system with 11 axes |
| SNOP | Systematic Nomenclature of Pathology (on four axes: topography, morphology, etiology, function), basis for SNOMED |
| System features | Static or dynamic; mechanistic or phenomenological; discrete or continuous; deterministic or stochastic; single-scale or multi-scale |
| Terminology | Includes well-defined terms and usage |
| UMLS | Unified Medical Language System is a long-term project to develop resources for the support of intelligent information retrieval |

## 3 Acronyms

| | |
|---|---|
| ACR | American College of Radiologists |
| API | Application Programming Interface |
| DAML | DARPA Agent Markup Language |
| DICOM | Digital Imaging and Communications in Medicine |
| DL | Description Logic |
| ECG | Electrocardiogram |
| EHR | Electronic Health Record |
| FMA | Foundational Model of Anatomy |
| FOL | First-order logic |
| GO | Gene Ontology |
| ICD | International Classification of Diseases |
| IOM | Institute of Medicine |
| KIF | Knowledge Interchange Format |
| LOINC | Logical Observation Identifiers Names and Codes |
| MeSH | Medical Subject Headings |
| MRI | Magnetic Resonance Imaging |
| NCI | National Cancer Institute (US) |
| NEMA | National Electrical Manufacturer Association |
| OIL | Ontology Inference Layer (description logic) |
| OWL | Ontology Web Language |
| RDF | Resource Description Framework |
| RDF Schema | A vocabulary of properties and classes added to RDF |
| SCP | Standard Communications Protocol |
| SNOMED CT | Systematized Nomenclature of Medicine—Clinical Terms |
| SOP | Standard Operating Procedure |
| UMLS | Unified Medical Language System |

## 4 Key Problems

**Slide 5-1: Mathematically Seen Our World Is Complex and High Dimensional**

Key problems in dealing with data in the life sciences include:

- Complexity of our world
- High-dimensionality (curse of dimensionality (Catchpoole et al. 2010))
- Most of the data is weakly structured and unstructured

(continued)

(continued)

A grand challenge in health care is the complexity of data, implicating two issues: structurization and standardization. As we have learned in Lecture 2, very little of the data is structured. Most of our data is weakly structured (Holzinger 2012). In the language of business there is often the use of the word "unstructured," but we have to use this word with care; unstructured would mean—in a strict mathematical sense—that we are talking about total randomness and complete uncertainty, which would mean noise, where standard methods fail or lead to the modeling of artifacts, and only statistical approaches may help. The correct term would be unmodeled data—or we shall speak about **unstructured information**. Please mind the differences.

To the image in Slide 5-1: Advances in genetics and genomics have accelerated the discovery-based (=hypotheses generating) research that provides a powerful complement to the direct hypothesis-driven molecular, cellular, and systems sciences.

For example, genetic and functional genomic studies have yielded important insights into neuronal function and disease. One of the most exciting and challenging frontiers in neuroscience involves harnessing the power of large-scale genetic, genomic, and phenotypic datasets, and the development of tools for data integration and data mining (Geschwind and Konopka 2009).

## 5 Review on Data

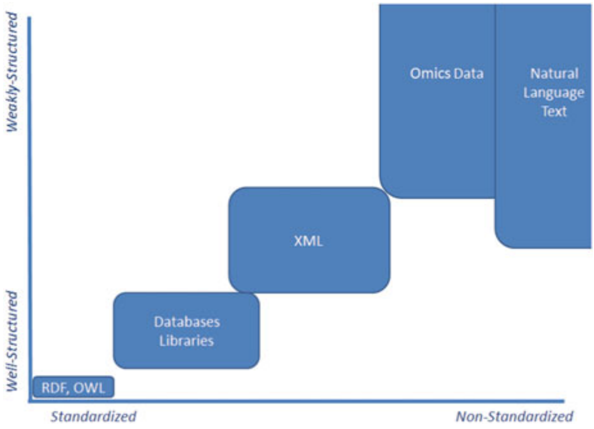


**Fig. 1** See Slide 5-2

### Slide 5-2: Remember: Standardization Versus Structurization

Before we proceed, please review the four different definitions of data in terms of structurization and standardization: well-structured and standardized data, semi-structured data (e.g., XML), weakly structured data (e.g., Omics data), and “unmodelled data”—unstructured information (text).

Do not confuse structure with standardization (see Slide 2-9). Data can be standardized (e.g., numerical entries in laboratory reports) and non-standardized. A typical example is non-standardized text—imprecisely called “Free-Text” or “unstructured data” in an electronic patient record (Kreuzthaler et al. 2011; Holzinger 2011; Holzinger et al. 2013c).

Well-structured data is the minority of data and an idealistic case when each data element has an associated defined structure, relational tables, or the resource description framework RDF, or the Web Ontology Language OWL (see Lecture 3).

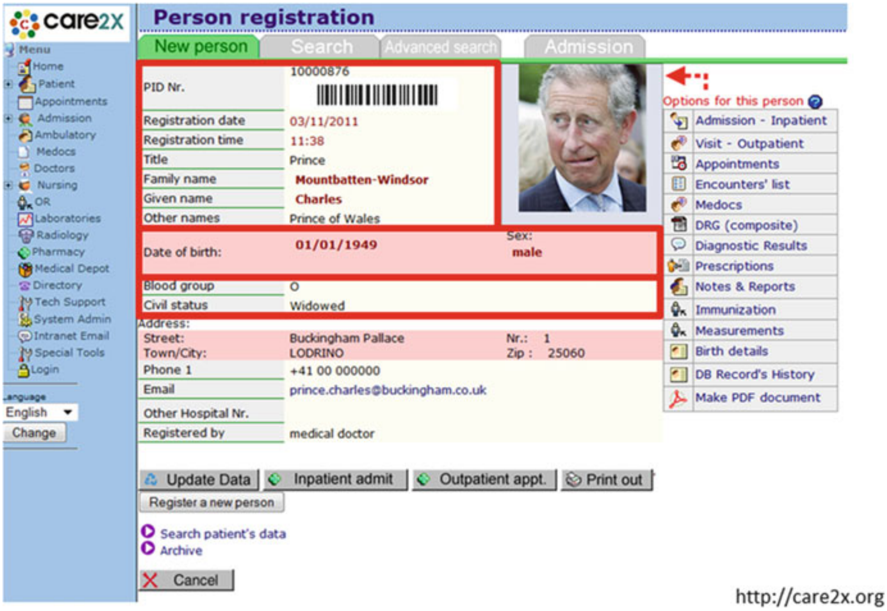
Note: Ill-structured is a term often used for the opposite of well-structured, although this term originally was used in the context of problem solving (Simon 1973).

Semi-structured is a form of structured data that does not conform with the strict formal structure of tables and data models associated with relational databases but contains tags or markers to separate structure and content, i.e., they are schema-less or self-describing; a typical example is a markup language such as XML (see Lectures 3 and 4).

Weakly structured data is the most of our data in the whole universe, whether it is in macroscopic (astronomy) or microscopic structures (biology)—see Lecture 5.

Non-structured data or unstructured data is an imprecise definition used for information expressed in natural language, when no specific structure has been defined. This is an issue for debate: Text has also some structure: words, sentences, paragraphs. If we are very precise, unstructured data would meant that the data is complete randomized—which is usually called noise and is defined by (Duda et al. 2000) as any property of data which is not due to the underlying model but instead to randomness (either in the real world, from the sensors or the measurement procedure).

## 5.1 Well-Structured Data



Person registration

| | | | |
|---|---|---|---|
| PID Nr. | 10000876 | | |
| Registration date | 03/11/2011 | | |
| Registration time | 11:38 | | |
| Title | Prince | | |
| Family name | Mountbatten-Windsor | | |
| Given name | Charles | | |
| Other names | Prince of Wales | | |
| Date of birth: | 01/01/1949 | Sex: | male |
| Blood group | O | | |
| Civil status | Widowed | | |
| Address: | | | |
| Street: | Buckingham Pallace | Nr.: | 1 |
| Town/City: | LODRINO | Zip : | 25060 |
| Phone 1 | +41 00 000000 | | |
| Email | prince.charles@buckingham.co.uk | | |
| Other Hospital Nr. | | | |
| Registered by | medical doctor | | |

http://care2x.org

**Fig. 2** See Slide 5-3

### Slide 5-3: Example: Well-Structured Data

A look on the typical view of a hospital information system shows us the organization of well-structured data: Standardized and well-structured data is the basis for accurate communication. In the medical domain, many different people work at different times in various locations. Data standards can ensure that information is interpreted by all users with the same understanding. Moreover, standardized data facilitate comparability of data and interoperability of systems. It supports the reusability of the data, improves the efficiency of health-care services and avoids errors by reducing duplicated efforts in data entry. Remember: Data standardization refers to (a) the data content; (b) the terminologies that are used to represent the data; (c) how data is exchanged; and (d) how knowledge, e.g., clinical guidelines, protocols, decision support rules, checklists, standard operating procedures are represented in the health information system.

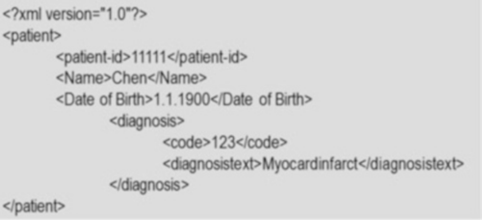
Note: The opposite, i.e., non-standardized data is the majority of data and inhibit data quality, data exchange, and interoperability.

Remark: Care2x is an Open Source Information System, see: http://care2x.org

See Lecture 10 for more details.

## 5.2 *Semi-structured Data*

**Fig. 3** See Slide 5-4



```
<?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>
```

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

This is a Medical example for semi-structured data in XML (Holzinger 2003). The eXtensible Markup Language (XML) is a flexible text format recommended by the W3C for data exchange and derived from SGML (ISO 8879), (Usdin and Graham 1998).

XML is often classified as semi-structured; however, this is in some way misleading, as the data itself is still **structured**, but in a flexible rather than a static way (Forster and Vossen 2012). Such data does not conform to the formal structure of tables and data models as for example in relational databases, but at least contains tags/markers to separate semantic elements and enforce hierarchies of records and fields within these data.

### Slide 5-5: Example: Generic XML Template for a Medical Report

This example by Rassinoux et al. (2003) shows how XML can be used in the hospital information system: The structure of any new document edited in the Patient Record (here: DPI) is based on a template defined in XML format (left). These templates play the role of DTDs or XML schemas as they precisely define the structure and content type of each paragraph, thus validating the document at the application level. Such a structure embeds a <HEADER> and a <BODY>. The header encapsulates the properties that are inherent to the new document and that will be useful to further classify it, according to various criteria, including the patient identification, the document type, the identifier of its redactors and of the hospitalization stay, or ambulatory consultation to which the document will be attached in the patient

(continued)

(continued)

trajectory, etc. The body encapsulates the content, and is divided into two parts: The <STRUCDOC> part describes the semantic entities that compose the document. The <FULLDOC> part embeds the document itself with its page layout information, which can be stored either as a draft, a temporary text or as a definitive text. This format guarantees the storage of dynamic and controlled fields for data input, thus allowing the combination of free text and structured data entry in the document. Once the document is no longer editable, it is definitively saved into the RTF format. A CDATA section is utilized for storing the rough document whatever its format, as it permits to disregard blocks of text containing characters that would otherwise be regarded as markup (Rassinoux et al. 2003).

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

On top in this slide you can see a sample XML describing genes from Drosophila melanogaster involved in long-term memory. Nested within the gene elements, are sub-elements related to the parent. The first gene includes two nucleic acid sequences, a protein product, and a functional annotation. Additional information is provided by attributes, such as the organism. This example illustrates the difficulty of modeling many-to-many relationships, such as the relationship between genes and functions. Information about functions must be repeated under each gene with that function. If we invert the nesting, then we must repeat information about genes with more than a single function. Below the XML we see the information about genes using both RDF and OWL. Both genes are instances of the class Fly Gene, which has been defined as the set of all Genes for the organism D. melanogaster. The functional information is represented using a hierarchical taxonomy, in which Long-Term Memory is a subclass of Memory (Louie et al. 2007).

Remark: Drosophila melanogaster is a model organism and shares many genes with humans. Although Drosophila is an insect whose genome has only about 14,000 genes (half of humans), a remarkable number of these have very close counterparts in humans; some even occur in the same order in the fly's DNA as in our own. This, plus the organism's more than 100-year history in the lab, makes it one of the most important models for studying basic biology and disease (see for example http://www.lbl.gov/Science-Articles/Archive/sabl/2007/Feb/drosophila.html).

Note: The relational data model requires preciseness: The data must be regular, complete and structured. However, in Biology the relationships are

(continued)

(continued)

mostly un-precise. Genomic medicine is extremely data intensive and there is an increasing diversity in the type of data: DNA sequence, mutation, expression arrays, haplotype, proteomic, etc. In bioinformatics many heterogeneous data sources are used to model complex biological systems (Rassinoux et al. 2003; Achard et al. 2001). The challenge in genomic medicine is to integrate and analyze these diverse and huge data sources to elucidate physiology and in particular disease physiology. XML is suited for describing semi-structured data, including a kind of natural modeling of biological entities, because it allows features as for example nesting (see Slide 5-6 on top). Still a key limitation of XML is that it is difficult to model complex relationships; for example, there is no obvious way to represent many-to-many relationships, which are needed to model complex pathways. On top in Fig. 5-9 we can see a sample XML, describing genes involved in the long-term memory of a sample specimen d. melanogaster. Nested within the gene elements, are sub-elements related to the parent. The first gene includes two nucleic acid sequences, a protein product, and a functional annotation. Additional information is provided by attributes, such as the organism. This example illustrates the difficulty of modeling many-to-many relationships, such as the relationship between genes and functions. Information about functions must be repeated under each gene with that function. If we invert the nesting (i.e., nesting genes inside function elements), then we must repeat information about genes with more than a single function. At the bottom in Slide 5-6 we see the same information about genes, but using RDF and OWL. Both genes are instances of the class Fly Gene, which has been defined as the set of all Genes for the organism D. melanogaster. The functional information is represented using a hierarchical taxonomy, in which Long-Term Memory is a subclass of Memory (Louie et al. 2007).

## 5.3 *Weakly Structured Data*

### Slide 5-7: Example: Weakly Structured Data—Protein–Protein Interactions

Here we see a human protein interaction network and its connections: Proteins likely to be under positive selection are colored in shades of red (light red, low likelihood of positive selection; dark red, high likelihood). Proteins estimated not to be under positive selection are in yellow, and proteins for which the likelihood of positive selection was not estimated are in white (Kim et al. 2007).

## *5.4 On the Topology of Data*

Data has shape!

**Slide 5-8: On the Topology of Data: Data Has Shape!**

Such data does not conform to the formal structure of tables and data models as for example in relational databases, but at least contains tags/markers to separate semantic elements and enforce hierarchies of records and fields within these data.

**Slide 5-9: Again: What Is a Mathematical, What Is a Physical Space?**

Such data does not conform to the formal structure of tables and data models as for example in relational databases, but at least contains tags/markers to separate semantic elements and enforce hierarchies of records and fields within these data.

There are many different types of topology:

**Point-set topology**, aka general topology, studies properties of spaces and the structures defined on them, where the spaces may be very general, and do not have to be similar to manifolds (a manifold of dimension n is a topological space that near each point resembles an n-dimensional Euclidean space). General topology provides the most general framework where fundamental concepts of topology such as open/closed sets, continuity, interior/exterior/boundary points, and limit points can be defined (Gaal 1966).

**Algebraic topology** uses tools from abstract algebra to study topological spaces. The basic goal is to find algebraic invariants that classify topological spaces up to homeomorphism (Hatcher 2002). A function $\mathrm{f}\colon \mathrm{X} \to \mathrm{Y}$ between two topological spaces $(X, T_X)$ and $(Y, T_Y)$ is called a **homeomorphism** if $f$ is bijective, continuous, and the inverse function $f^{-1}$ is continuous.

Before we go into examples, let us answer the question: “What is a space?”

# 6 Networks = Graphs + Data

## 6.1 Networks in Biological Systems

**Slide 5-10: Complex Biological Systems: Key Concepts**

The concept of network structures is fascinating, compelling, and powerful and applicable in nearly any domain at any scale.

Network theory can be traced back to **graph theory**, developed by Leonhard Euler in 1736 (see Slide 5-11). However, stimulated by works for example from Barabási et al. (1999), research on complex networks has only recently been applied to biomedical informatics. As an extension of classical graph theory, see for example Diestel (2010), complex network research focuses on the characterization, analysis, modeling and simulation of complex systems involving many elements and connections, examples including the internet, gene regulatory networks, protein–protein networks, social relationships and the Web, and many more. Attention is given not only to try to identify special patterns of connectivity, such as the shortest average path between pairs of nodes (Newman 2003), but also to consider the evolution of connectivity and the growth of networks, an example from biology being the evolution of protein–protein interaction (PPI) networks in different species (Slide 5-11). 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 (Costa et al. 2008).

## 6.2 *Network Theory*

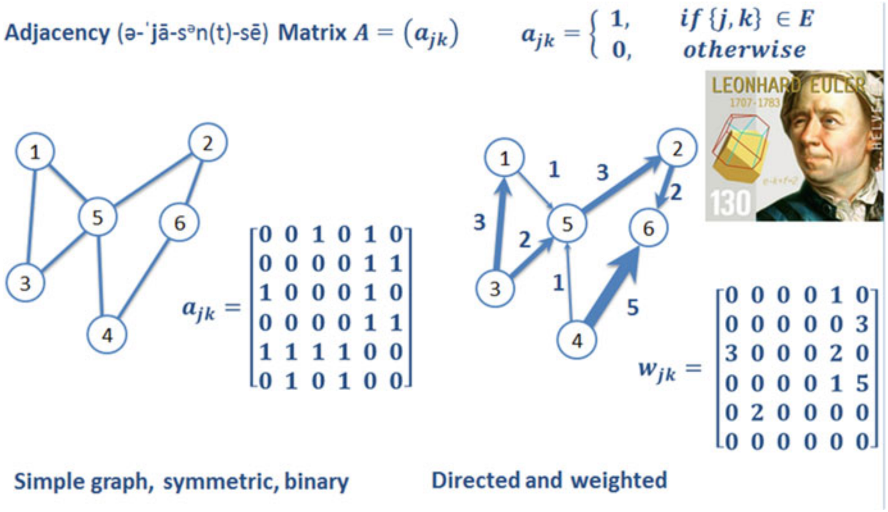
### 6.2.1 Basic Concepts of Networks

**Slide 5-11: Networks on the Example of Bioinformatics**

A graph $\boldsymbol{G}(\boldsymbol{V}, \boldsymbol{E})$ describes a structure which consists of nodes aka vertices $\boldsymbol{V}$, connected by a set of pairs of distinct nodes (links), called edges $\boldsymbol{E}\{\boldsymbol{a}, \boldsymbol{b}\}$ with $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{V}; \boldsymbol{a} \neq \boldsymbol{b}$.

Graphs containing cycles and/or alternative paths are referred to as networks. The vertexes and edges can have a range of properties defined as colors, which also may have quantitative values, referred to as weights. In this slide we see the basic building block symbols of a biological network as used in bioinformatics. The blue dots are serving as network hubs, the red block is a critical node (on a critical link), the white balls are bottle necks, the stars second order hubs etc. (Hodgman et al. 2010).

### 6.2.2 Computational Graph Representation



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

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

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

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

Simple graph, symmetric, binary

Directed and weighted

**Fig. 4** See Slide 5-12

**Slide 5-12: Computational Graph Representation**

In order to represent network data in computers it is not comfortable to use sets; more practical are matrices. The simplest form of a graph representation

(continued)

(continued)

is the so-called adjacency matrix. In this Slide we see an undirected (left) and a directed graph and their respective adjacency matrices. If the graph is undirected, the adjacency matrix is symmetric, i.e., the elements $a_{ij} = a_{ji}$ for any i and j.

Left: a simple undirected binary graph and its mapping in a square adjacency matrix (symmetric), right: a directed and weighted graph (nonsymmetric); there is full correspondence between the network, the graph, and the adjacency matrix.

**Slide 5-13: Example: Tool for Node-Link Visualization**

This Tool is a nice example on the usefulness of adjacency matrices: The InfoVis Toolkit is an interactive graphics toolkit developed by Jean-Daniel Fekete at INRIA (The French National Institute for Computer Science and Control). The toolkit implements nine types of visualization: Scatter Plots, Time Series, Parallel Coordinates, and Matrices for tables; Node-Link diagrams, Icicle trees, and Tree maps for trees; Adjacency Matrices and Node-Link diagrams for graphs. Node-Link visualizations provide several variants (eight for graphs and four for trees). There are also a number of interactive controls and information displays, including dynamic query sliders, fisheye lenses, and excentric labels. Information about the InfoVis toolkit can be found at http://ivtk.sourceforge.net

The InfoVis Toolkit provides interactive components such as range sliders and tailored control panels required to configure the visualizations. These components are integrated into a coherent framework that simplifies the management of rich data structures and the design and extension of visualizations. Supported data structures include tables, trees, and graphs. All visualizations can use fisheye lenses and dynamic labeling (Fekete 2004).

### 6.2.3 Network Metrics

**Slide 5-14: Some Network Metrics (1/2)**

The truly multidisciplinary network science has led to a wide variety of quantitative measurements of their topological characteristics (Costa et al. 2007). The identification between a graph and an adjacency matrix

(continued)

(continued)

makes all the powerful methods of linear algebra, graph theory, and statistical mechanics available to us for investigating specific network characteristics:

**Order** (a in Figure Slide 5-14) = total number of nodes n

**Size** = total number of links:

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

**Clustering Coefficient** (b in Slide 5-14) = 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, i.e., the level of connectedness of the graph. It is calculated as the ratio between the actual number $t_i$ of links connecting the neighborhood (the nodes immediately connected to a chosen node) of a node and the maximum possible number of links in that neighborhood:

$$C_i = \frac{2t_i}{k(k_i - 1)}$$

For the whole network, the clustering coefficient is the arithmetic mean:

$$C = \frac{1}{n} \sum_i C_i$$

**Path length** (c in Slide 5-14) = is the arithmetical mean of all the distances; The characteristic path length of node *i* provides information about how close node *i* is connected to all other nodes in the network and is given by the distance *d(i,j)* between node *i* and all other nodes *j* in the network.

The Path length *l* provides important information about the level of global communication efficiency of a network:

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

Note*: Numerical methods,* e.g.*, the Dijkstra's algorithm (1959), are used to calculate all the possible paths between any two nodes in a network.*

### Slide 5-15: Some Network Metrics (2/2)

**Centrality** (d in Slide 5-15) = the level of "betweenness-centrality" of a node i; it indicates how many of the shortest paths between the nodes of the network pass through node i. A high "betweenness-centrality" indicates that this node is important in interconnecting the nodes of the network, marking a potential hub role (refer to Slide 5-11) of this node in the overall network.

**Nodal degree** (e in Slide 5-15) = number of links connecting *i* to its neighbors. The degree of node *i* is defined as its total number of connections.

$$k_i = \sum_i a_{ij}$$

*The degree probability distribution P(k) describes the p(x) that a node is connected to k other nodes in the network.*

**Modularity** (f in Slide 5-15) = describes the possible formation of communities in the network, indicating how strong groups of nodes form relative isolated subnetworks within the full network (refer also to Slide 5-11).

Two further metrics include:

**Density** = the ratio between m and the maximum possible number of links that a graph may have:

$$\delta = \frac{2m}{n(n-1)}$$

**Path** = a series of consecutive links connecting any two nodes in the network, the distance between two vertices is the length of the shortest path connecting them, the diameter of a graph is the longest distance (the maximum shortest path) existing between any two vertices in the graph:

$$D = \max d_{ij}$$

### Slide 5-16: Some Network Topologies

**Regular network** (a in Slide 5-16) has a local character, characterized by a high clustering-coefficient (c in Slide 5-16) and a high path length (L, Slide

(continued)

(continued)

5-16). It takes a large number of steps to travel from a specific node to a node on the other end of the graph. A special case of a regular network is the:

**Random network**, where *all* connections are distributed randomly across the network; the result is a graph with a random organization (outer right in Slide 5-16). In contrast to the local character of the regular network, a random network has a more global character, with a low C and a much shorter path length L than the regular network. A particular case is the:

**Small-world network** (center of Slide 5-16) which are very robust and combine a high level of local and global efficiency. Watts and Strogatz (1998b) showed that with a low probability p of randomly reconnecting a connection in the regular network, a so-called small-world organization arises. It has both a high C and a low L, combining a high level of local clustering with still a short average travel distance. Many networks in nature are small-world (e.g., Internet, protein networks, social networks, functional and structural brain network), combining a high level of segregation with a high level of global information integration. In addition, such networks can have a heavy tailed connectivity distribution, in contrast to random networks in which the nodes roughly all have the same number of connections.

**Scale-free networks** (B in Slide 5-16) are characterized by a degree probability distribution that follows a power-law function, indicating that on average a node has only a few connections, but with the exception of a small number of nodes that are heavily connected. These nodes are often referred to as hub nodes (see Slide 5-11) and they play a central role in the level of efficiency of the network, as they are responsible for keeping the overall travel distance in the network to a minimum. As these hub nodes play a key role in the organization of the network, scale-free networks tend to be vulnerable to specialized attack on the hub nodes.

**Modular networks** (c in Slide 5-16) show the formation of so-called communities, consisting of a subset of nodes that are mostly connected to their direct neighbors in their community and to a lesser extend to the other nodes in the network. Such networks are characterized by a high level of modularity of the nodes.

### Slide 5-17: Small-World Networks

Taking a connected graph with a high graph diameter and adding a very small number of edges randomly, results in the small world phenomenon: the diameter drops drastically. It is also known as “six degrees of separation”

(continued)

(continued)

since, in the social network of the world, any person turns out to be linked to any other person by roughly six connections (Milgram 1967). The human short-term memory uses small world networks between the neurons.

In the Slide we see the random rewiring procedure for interpolating between a regular ring lattice (rightmost) and a random network (leftmost), without altering the number of vertices or edges in the graph. We start with a ring of $n$ vertices, each connected to its $k$ nearest neighbors by undirected edges. We choose a vertex and the edge that connects it to its nearest neighbor in a clockwise sense. With probability p, we reconnect this edge to a vertex chosen uniformly at random over the entire ring, with duplicate edges forbidden; otherwise we leave the edge in place. We repeat this process by moving clockwise around the ring, considering each vertex in turn until one lap is completed. Then we consider the edges that connect vertices to their second-nearest neighbors clockwise. We randomly rewire each of these edges with probability p, and continue this process, circulating around the ring and proceeding outward to more distant neighbors after each lap, until each edge in the original lattice has been considered once. For intermediate values of p, the graph is a small-world network: highly clustered like a regular graph, yet with small characteristic path length, like a random graph (Watts and Strogatz 1998a).

### 6.2.4 Graphs from Point Cloud Datasets

**Slide 5-18: Graphs from Point Cloud Datasets**

There are many ways to construct a proximity graph representation from a set of data points that are embedded in $\mathbb{R}^d$.

Let us consider a set of data points $\{x_1, \ldots, x_n\} \in \mathbb{R}^d$ .

To each data point we associate a vertex of a proximity graph $\mathcal{G}$ to define a set of vertices $\mathcal{V} = \{v1, v2, \ldots, vn\}$. Determining the edge set $\mathcal{E}$ of the proximity graph $\mathcal{G}$ requires defining the neighbors of each vertex $vi$ according to its embedding $xi$.

Consequently, a **proximity graph** is a graph in which two vertices are connected by an edge *if* the data points associated to the vertices satisfy particular geometric requirements. Such particular geometric requirements are usually based on a metric measuring the distance between two data points. A usual choice of metric is the Euclidean metric. Look at the slide:

(a) is our initial set of points in the plane $\mathbb{R}^2$

(continued)

(continued)

(b) $\varepsilon$-ball graph $vi \sim vj$ *if* $\boldsymbol{xj} \in \mathcal{B}(vi; \varepsilon)$

(c) $k$-nearest-neighbor graph ($k$-NNG): $vi \sim vj$ if the distance between $\boldsymbol{xi}$ and $\boldsymbol{xj}$ is among the $k$-th smallest distances from $\boldsymbol{xi}$ to other data points. The k-NNG is a directed graph since one can have $\boldsymbol{xi}$ among the k-nearest neighbors of $\boldsymbol{xj}$ but not vice versa.

(d) Euclidean Minimum Spanning Tree (EMST) graph is a connected tree subgraph that contains all the vertices and has a minimum sum of edge weights. The weight of the edge between two vertices is the Euclidean distance between the corresponding data points.

(e) Symmetric k-nearest-neighbor graph (Sk-NNG): $vi \sim vj$ if $\boldsymbol{xi}$ is among the k-nearest neighbors of $y$ or vice versa.

(f) Mutual k-nearest-neighbor graph (Mk-NNG): $vi \sim vj$ if $\boldsymbol{xi}$ is among the k-nearest neighbors of $y$ and vice versa. All vertices in a mutual k-NN graph have a degree upper-bounded by k, which is not usually the case with standard k-NN graphs.

(g) Relative Neighborhood Graph (RNG): $vi \sim vj$ if there is no vertex in $B(vi; D(vi, vj)) \cap B(vj; D(vi, vj))$.

(h) Gabriel Graph (GG)

(i) The β-Skeleton Graph (β-SG):

For details please refer to Lézoray and Grady (2012), or to a classical graph theory book, e.g., Harary (1969), Bondy and Murty (1976), Golumbic (2004), Diestel (2010).

## Slide 5-19: Graphs from Images

In this slide we see the examples of:

(a) a real image with the quadtree tessellation,

(b) the region adjacency graph associated to the quadtree partition,

(c) irregular tessellation using image-dependent superpixel Watershed Segmentation (Vincent and Soille 1991)

(d) irregular tessellation using image-dependent SLIC superpixels (Lucchi et al. 2010)

SLIC = Simple Linear Iterative Clustering

### Slide 5-20: Example: Watershed Algorithm

A straightforward implementation of the original Vincent–Soille algorithm is difficult if plateaus occur. Therefore, an alternative approach was proposed by Meijster and Roerdink (1995), in which the image is first transformed to a directed valued graph with distinct neighbor values, called the components graph of f. On this graph the watershed transform can be computed by a simplified version of the Vincent–Soille algorithm, where fifo queues are no longer necessary, since there are no plateaus in the graph (Roerdink and Meijster 2000).

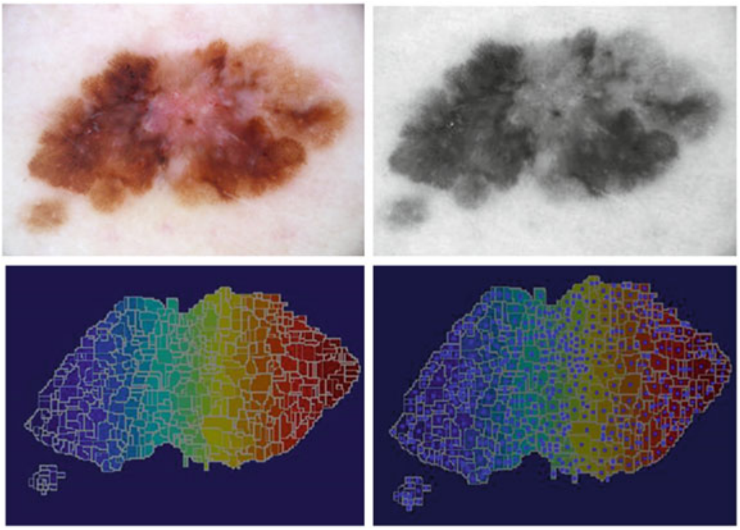


**Fig. 5** See Slide 5-21

### Slide 5-21: From Graphs to Images: Watershed + Centroid

The original natural digital image is first transformed into grey scale, then the Watershed algorithm is applied and then the centroid function calculated, the results are representative point sets in the plane.

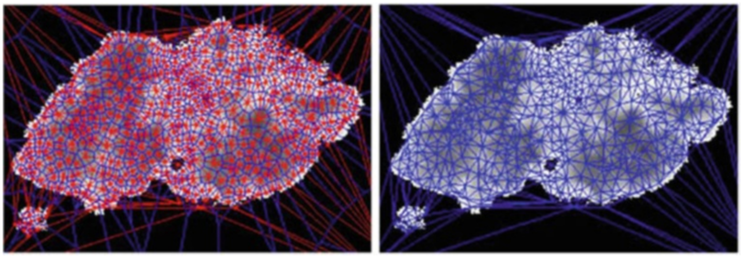


**Fig. 6** See Slide 5-22

### Slide 5-22: Graphs from Images: Voronoi ↔ Delauney

The Delaunay Triangulation (DT): $vi \sim vj$ if there is a closed ball $\mathcal{B}(\cdot;\mathrm{r})$ with $vi$ and $vj$ on its boundary and no other vertex $vk$ contained in it. The dual to the DT is the Voronoi irregular tessellation where each Voronoi cell is defined by the set $\{x \in Rn \mid D(x, vk) \leq D(x, vj)\, for\ all\ vj = vk\}$. In such a graph, $\forall\ vi,\ deg\,(vi) = 3$ (Lézoray and Grady 2012).

### Slide 5-23: Points → Voronoi → Delauney

This animation shows the construction of a Delaunay graph: First the red points on the plane are drawn, then we insert the blue edges and the blue vertices on the Voronoi graph, finally the red edges drawn build the Delaunay graph (Kropatsch et al. 2001).

### Slide 5-24: Example: Graph Entropy Measures

In this Slide we see the evaluated information-theoretic network measures on publication networks. Here from the excellence network of RWTH Aachen University. Those measures can be understood as graph complexity measures which evaluate the structural complexity based on the corresponding concept. A possible useful interpretation of these measures helps to understand the differences in subgraphs of a cluster. For example one could apply community detection algorithms and compare entropy measures of such detected communities. Relating these data to social measures (e.g., balanced score card data) of subcommunities could be used as indicators of collaboration success or lack thereof. The node size shows the node degree, whereas the node color shows the betweenness centrality, and darker color means higher centrality (Holzinger et al. 2013a).

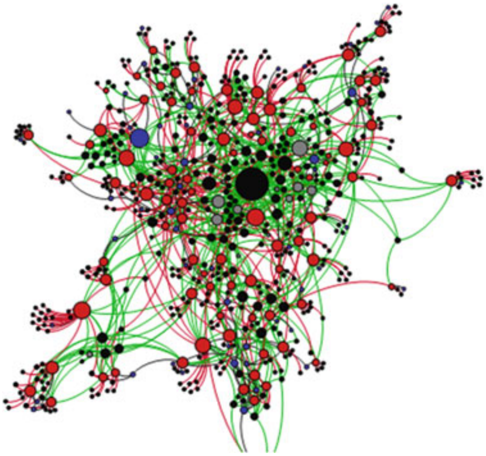


**Fig. 7** See Slide 5-25

### Slide 5-25: Example for a Medical Knowledge Space

A further example shall demonstrate the usefulness of graph theory and network analysis: This graph shows the medical knowledge space of a standard quick reference guide for emergency doctors and paramedics in the German speaking area. It has been subsequently developed, tested in the medical real world and constantly improved for 20 years by Dr. med. Ralf Müller, emergency doctor at Graz-LKH University Hospital and is practically in the pocket of every emergency and family doctor and paramedics in the German speaking area (Holzinger et al. 2013b).

Up to know we know that Graphs and Graph-Theory are powerful tools to map data structures and to find novel connections between single data objects (Strogatz 2001; Dorogovtsev and Mendes 2003). The inferred graphs can be further analyzed by using graph-theoretical and statistical and machine learning techniques (Dehmer et al. 2011). A mapping of the already existing and in the medical practice approved “knowledge space” as a conceptual graph and the subsequent visual and graph-theoretical analysis may provide novel insights on hidden patterns in the data. Another benefit of the graph-based data structure is in the applicability of methods from network topology and network analysis and data mining, e.g., small-world phenomenon (Barabasi and Albert 1999; Kleinberg 2000), and cluster analysis (Koontz et al. 1976; Wittkop et al. 2011).

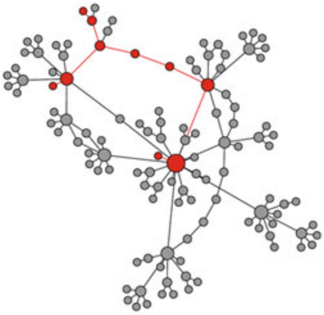
The graph-theoretic data of the graph seen in this slide include:

Number of nodes = 641, number of edges = 1,250, red are agents, black are conditions, blue are pharmacological groups, grey are other documents. The average degree of this graph = 3.888, the average path length = 4.683, the network diameter = 9.

### Slide 5-26: Medical Details of the Graph

The nodes of the sample graph represent: drugs, clinical guidelines, patient conditions (indication, contraindication), pharmacological groups, tables and calculations of medical scores, algorithms, and other medical documents; and the edges represent three crucial types of relations inducing medical relevance between two active substances, i.e., pharmacological groups, indications, and contraindications. The following example will demonstrate the usefulness of this approach.

**Fig. 8** See Slide 5-27



### Slide 5-27: Example for the Shortest Path

This example shows us how convenient we can find which path between two nodes is the shortest as well as the navigation way between these nodes. Computing shortest paths is a fundamental and ubiquitous problem in network analysis. We can for example apply the Dijkstra-algorithm, solves the shortest path problem for a graph with non-negative edge path costs, producing a shortest path tree. This algorithm is often used in routing and as a subroutine in other graph algorithms: For a given node, the algorithm finds the path with lowest cost (i.e., the shortest path) between that node and every other node(Henzinger et al. 1997).

### Slide 5-28: Example for Finding Related Structures

Here we see the relationship between Adrenaline (center black node) and Dobutamine (top left black node), Blue: Pharmacological Group,

(continued)

(continued)

Dark red: Contraindication; Light red: Condition, the Green nodes (from dark to light) are:

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)

## *6.3 Network Examples*

### 6.3.1 The Human Brain as Network

**Slide 5-29: Example: The Brain Is a Complex Network**

Our brain forms one integrative complex network, linking all brain regions and subnetworks together (Van Den Heuvel and Hulshoff Pol 2010). Examining the organization of this network provides insights in how our brain works. Graph theory provides a framework in which the topology of complex networks can be examined, and thus can reveal novelties about both the local and global organization of functional brain networks. In the slide we can see how the modeling of the functional brain by a graph works: edges are the connections between regions that are functionally linked. First, the collection of nodes is defined (A), second the existence of functional connections between the nodes in the network needs to be defined, resulting in a connectivity matrix (B). Finally, the existence of a connection between two points can be defined as whether their level of functional connectivity exceeds a certain predefined threshold (C) (Van Den Heuvel and Hulshoff Pol 2010).

### 6.3.2 Systems Biology and Human Diseases

**Slide 5-30: Representative Examples of Disease Complexes**

Insight into the biology of molecular networks is an important field, as anomalies in these systems underlie a wide spectrum of polygenetic human disorders, ranging from schizophrenia to congenital heart disease (CHD). Understanding the functional architecture of networks that organize the development of organs, see for example Chien et al. (2008), lays the foundation of novel approaches in *regenerative medicine*, since manipulation of

(continued)

(continued)

such systems is necessary for success of *tissue engineering* technologies and *stem cell therapy*.

Lage et al. (2010) developed a framework for gaining new insights into the systems biology of the protein networks driving organ development and related polygenic human disease phenotypes, exemplified with heart development and CHD. In the slide we see examples of four functional networks driving the development of different anatomical structures in the human heart. These four networks are constructed by analyzing the interaction patterns of four different sets of cardiac development (CD): proteins corresponding to the morphological groups “atrial septal defects,” “abnormal atrioventricular valve morphology,” “abnormal myocardial trabeculae morphology,” and “abnormal outflow tract development,” CD proteins from the relevant groups are shown in orange and their interaction partners are shown in grey. Functional modules annotated by literature curation are indicated with a colored background. Centrally in the Figure is a hematoxylin–eosin stained frontal section of the heart from a 37-day human embryo, where tissues affected by the four networks are marked; AS (developing atrial septum), EC (endocardial cushions, which are anatomical precursors to the atrioventricular valves), VT (developing ventricular trabeculae), and OFT (developing outflow tract).

### Slide 5-31: Example: Cell-Based Therapy

In this slide we see an overview of the modular organization of heart development: (A) Protein interaction networks are plotted at the resolution of functional modules. Each module is color coded according to functional assignment as determined by literature curation. The amount of proteins in each module is proportional to the area of its corresponding node. Edges indicate direct (lines) or indirect (dotted lines) interactions between proteins from the relevant modules. (B) Recycling of functional modules during heart development. The bars represent functional modules and recycling is indicated by arrows. The bars follow the color code of (A) and the height of the bars represent the number of proteins in each module, as shown left on the y axis (Lage et al. 2010).

Note: **Phenotype** = an organism’s observable characteristics (traits), e.g., morphology, biochemical/physiological properties, behavior. Phenotypes result from the expression of an organism’s genes as well as the influence of environmental factors and the interactions between them. **Genotype** = inherited instructions within its genetic code.

### 6.3.3 Gene Networks

**Slide 5-32: Identifying Networks in Disease Research**

Diseases (e.g., obesity, diabetes, and atherosclerosis) result from multiple genetic and environmental factors, and importantly, interactions between genetic and environmental factors. This slide shows the vast networks of molecular interactions. It can be seen that the gastrointestinal (GI) tract, vasculature, immune system, heart, and brain are all potentially involved in either the onset of diseases such as atherosclerosis or in comorbidities such as myocardial infarction and stroke brought on by such diseases. Further, the risks of comorbidities for diseases such as atherosclerosis are increased by other diseases, such as hypertension, which may, in turn, involve other organs, such as kidney. The role that each organ and tissue type plays in a given disease is largely determined by genetic background and environment, where different perturbations to the genetic background (perturbations corresponding to DNA variations that affect gene function, which, in turn, leads to disease) and/or environment (changes in diet, levels of stress, level of activity, and so on) define the subtypes of disease manifested in any given individual. Although the physiology of diseases such as atherosclerosis is beginning to be better understood, what have not been fully exploited to data are the vast networks of molecular interactions within the cells.

We see clearly in the Slide that there is a diversity of molecular networks functioning in any given tissue, including genomics networks, networks of coding and noncoding RNA, protein interaction networks, protein state networks, signaling networks, and networks of metabolites. Further, these networks are not acting in isolation within each cell, but instead interact with one another to form complex, giant molecular networks within and between cells that drive all activity in the different tissues, as well as signaling between tissues. Variations in DNA and environment lead to changes in these molecular networks, which, in turn, induce complicated physiological processes that can manifest as disease. Despite this vast complexity, the classic approach to elucidating genes that drive disease has focused on single genes or single linearly ordered pathways of genes thought to be associated with disease. This narrow approach is a natural consequence of the limited set of tools that were available for querying biological systems; such tools were not capable of enabling a more holistic approach, resulting in the adoption of a reductionist approach to teasing apart pathways associated with complex disease phenotypes. Although the emerging view that complex biological systems are best modeled as highly modular, fluid systems exhibiting a plasticity that allows them to adapt to a vast array of conditions, the history of science demonstrates that this view, although long the ideal, was never

(continued)

(continued)

within reach, given the unavailability of tools adequate to carrying out this type of research. The explosion of large-scale, high-throughput technologies in the biological sciences over the past 15–20 years has motivated a rapid paradigm shift away from reductionism in favor of a systems-level view of biology (Schadt and Lum 2006).

## *6.4 The Essence: Three Types of Biomedical Networks*

**Slide 5-33: Three Main Types of Biomedical Networks**

In this Slide we see the three main types of biological networks: (1) a transcriptional regulatory network has two components: transcription factor (TF) and target genes (TG), where TF regulates the transcription of TGs; (2) PPI networks: two proteins are connected if there is a docking between them; (3) a metabolic network is constructed considering the reactants, chemical reactions, and enzymes (Costa et al. 2008).

### 6.4.1 Transcriptional Regulatory Networks

**Slide 5-34: Example Transcriptional Regulatory Network**

The extreme complexity of the E. coli transcriptional regulatory network. In this graphical representation, nodes are genes, and edges represent regulatory interactions. The network was reconstructed using data from the RegulonDB. This figure highlights the extreme complexity in regulatory networks. To obtain a deeper understanding of regulatory complexity, scientists must first discover biologically relevant organizational principles to unravel the hidden architecture governing these networks (Salgado et al. 2006).

The complexity of organisms arises rather as a consequence of elaborated regulations of gene expression than from differences in genetic content in terms of the number of genes. The **transcription network** is a critical system that regulates gene expression in a cell. Transcription factors (TFs) respond to changes in the cellular environment, regulating the transcription of target genes (TGs) and connecting functional protein interactions to the genetic information encoded in inherited genomic DNA in order to control the timing and sites of gene expression during biological development. The interactions between TFs and TGs can be represented as a directed graph: The two types

(continued)

(continued)

of nodes (TF and TG) are connected by arcs (see Slide 5-33, arrows) when regulatory interaction occurs between regulators and targets. Transcriptional regulatory networks display interesting properties that can be interpreted in a biological context to better understand the complex behavior of gene regulatory networks. At a local network level, these networks are organized in substructures such as motifs and modules. **Motifs** represent the simplest units of a network architecture required to create specific patterns of inter-regulation between TFs and TGs. Three most common types of motifs can be found in gene regulatory networks:

(1) single input,
(2) multiple input and
(3) feed-forward loop

Target genes belonging to the same single and multiple input motifs tend to be co-expressed, and the level of co-expression is higher when multiple transcription factors are involved.

**Modularity** in the regulatory networks arises from groups of highly connected motifs that are hierarchically organized, in which modules are divided into smaller ones. The evolution of gene regulatory networks mainly occurs through extensive duplication of transcription factors and target genes with inheritance of regulatory interactions from ancestral genes while the evolution of motifs does not show common ancestry but is a result of convergent evolution (Costa et al. 2008).

### 6.4.2 Protein–Protein Networks

**Slide 5-35: Network Representations of Protein Complexes**

The interactions between proteins are essential to keep the molecular systems of living cells working properly. PPI is important for various biological processes such as cell–cell communication, the perception of environmental changes, protein transport and modification. Complex network theory is suitable to study PPI maps because of its universality and integration in representing complex systems. In complex network analysis each protein is represented as a node and the physical interactions between proteins are indicated by the edges in the network.

Many complex networks are naturally divided into communities or modules, where links within modules are much denser than those across modules (e.g., human individuals belonging to the same ethnic groups interact more

(continued)

(continued)

than those from different ethnic groups). Cellular functions are also organized in a highly modular manner, where each module is a discrete object composed of a group of tightly linked components and performs a relatively independent task. It is interesting to ask whether this modularity in cellular function arises from modularity in molecular interaction networks such as the transcriptional regulatory network and PPI network.

The slide shows a hypothetical protein complex (A). Binary PPI are depicted by direct contacts between proteins. Although five proteins (A, B, C, D, and E) are identified through the use of a bait protein (red), only A and D directly bind to the bait. (B) shows the true PPI network topology of the protein complex is shown in. (C) depicts the PPI network topology of the protein complex inferred by the "matrix" model, where all proteins in a complex are assumed to interact with each other. Finally (D) demonstrates the PPI network topology of the protein complex inferred by the "spoke" model, where all proteins in a complex are assumed to interact with the bait; but no other interactions are allowed (Wang and Zhang 2007).

### Slide 5-36: Correlated Motif Mining (CMM)

Correlated motif mining (CMM) is the challenge to find overrepresented pairs of patterns (motifs), in sequences of interacting proteins. Algorithmic solutions for CMM thereby provide a computational method for predicting binding sites for protein interaction. The task is basically to represent motifs X and Y (Fig. 119) to truly represent an overrepresented consensus pattern in the sequences of the proteins in VX, respectively VY, in order to increase the likelihood that they correspond or overlap with a so-called binding site—a site on the surface of the molecule that makes interactions between proteins from VX and VY possible through a molecular lock-and-key mechanism.

We call $\{X, Y\}$ a $(k_x\ k_y\ k_{xy})$-motif pair of a PPI network.

$G = (V, E, \lambda)$ if $|V_x| = k_x$, $|V_y| = k_y$ and $|V_x \cap V_y| = k_{xy}$.

It is called complete if all vertices from $V_x$ are connected with all vertices from $V_y$ (Boyen et al. 2011).

### Slide 5-37: Steepest Ascent Algorithm Applied to CMM

Since the decision problem associated with CMM is in NP,[1] we can efficiently check if a motif pair has higher support than another which makes it possible to tackle CMM as a search problem in the space of all possible (l,d)-motif pairs. If we add the assumption that similar motifs can be expected to get similar support, it has the typical form of a combinatorial optimization problem. In combinatorial optimization, the objective is to find a point in a discrete search space which maximizes a user-provided function f. A number of heuristic algorithms called metaheuristics are known to yield stable results, e.g., the steepest ascent algorithm (Aarts and Lenstra 1997), illustrated as pseudocode in the slide.

## 6.4.3 Metabolic Networks

### Slide 5-38: Metabolic Networks

Metabolism is primarily determined by genes, environment and nutrition. It consists of chemical reactions catalyzed by enzymes to produce essential components such as amino acids, sugars and lipids, and also the energy necessary to synthesize and use them in constructing cellular components. Since the chemical reactions are organized into metabolic pathways, in which one chemical is transformed into another by enzymes and cofactors, such a structure can be naturally modeled as a complex network. In this way, metabolic networks are directed and weighted graphs, whose vertices can be metabolites, reactions and enzymes, and two types of edges that represent mass flow and catalytic reactions. One widely considered catalogue of metabolic pathways available on-line is the Kyoto Encyclopedia of Genes and Genomes (KEGG). In the slide we see a simple metabolic network involving five metabolites M1–M5 and three enzymes E1–E3, of which the latter catalyzes an irreversible reaction (Hodgman et al. 2010).

[1] NP = nondeterministic polynomial time; in computational complexity theory NP is one of the fundamental complexity classes.

### Slide 5-39: Metabolic Networks are Usually Big . . . Big Data

Such metabolic structures can be very large, as can be seen in this slide. The enzyme-coding genes under TrmB (this is the thermococcus regulator of maltose binding) acts as a repressor for genes encoding glycolytic enzymes and as activator for genes encoding gluconeogenic enzymes control included in the metabolic pathways shown in the slide (13 are unique to archaea and 35 are conserved across species from all three domains of life). Integrated analysis of the metabolic and gene regulatory network architecture reveals various interesting scenarios (Schmid et al. 2009).

### Slide 5-40: Using EPRs to Discover Disease Correlations

Electronic patient records (EPR remain an unexplored, but rich data source for discovering for example correlations between diseases). Roque et al. (2011) describe a general approach for gathering phenotypic descriptions of patients from medical records in a systematic and non-cohort dependent manner: By extracting phenotype information from the “free-text” (=unstructured information) in such records they demonstrated that they can extend the information contained in the structured record data, and use it for producing fine-grained patient stratification and disease co-occurrence statistics. Their approach uses a dictionary based on the International Classification of Disease (ICD-10) ontology and is therefore in principle language independent. As a use case they show how records from a Danish psychiatric hospital lead to the identification of disease correlations, which subsequently can be mapped to systems biology frameworks.

### Slide 5-41: Heatmap of Disease-Disease Correlations (ICD)

Roque et al. (2011) have used text mining to automatically extract clinically relevant terms from 5,543 psychiatric patient records and mapped these to disease codes in the ICD10. They clustered patients together based on the similarity of their profiles. The result is a patient stratification, based on more complete profiles than the primary diagnosis, which is typically used. Figure 124 illustrates the general approach to capture correlations between different disorders. Several clusters of ICD10 codes relating to the same anatomical area or type of disorder can be identified along the diagonal of the heatmap, ranging from trivial correlations (e.g., different arthritis

(continued)

(continued)

disorders), to correlations of cause and effect codes (e.g., stroke and mental/ behavioral disorders), to social and habitual correlations (e.g., drug abuse, liver diseases, and HIV).

## 6.5 *Structural Homologies*

**Slide 5-42: Example: ὁμολογέω (Homologeo)**

Homology (plural: homologies) origins from Greek ὁμολογέω (homologeo) and means "to conform" (in German: übereinstimmen) and has its origins in Biology and Anthropology, where the word is used for a correspondence of structures in two life forms with a common evolutionary origin (Darwin 1859).

In chemistry it is used for the relationship between the elements in the same group of the periodic table, or between organic compounds in a homologous series.

In mathematics homology is a formalism for talking in a quantitative and unambiguous manner about how a space is connected (Edelsbrunner and Harer 2010).

Basically, homology is a concept that is used in many branches of algebra and topology. Historically, the term was first used in a topological sense by Henry Poincaré (1854–1912).

In Bioinformatics, homology modelling is a mature technique that can be used to address many problems in molecular medicine. Homology modelling is one of the most efficient methods to predict protein structures. With the increase in the number of medically relevant protein sequences, resulting from automated sequencing in the laboratory, and in the fraction of all known structural folds, homology modelling will be even more important to personalized and molecular medicine in the future. Homology modelling is a knowledge-based prediction of protein structures. In homology modelling a protein sequence with an unknown structure (the target) is aligned with one or more protein sequences with known structures (the templates).

The method of homology modelling is based on the principle that homologue proteins have similar structures. The prerequisite for successful homology modelling is a detectable similarity between the target sequence and the template sequences (more than 30 %) allowing the construction of a correct alignment. Homology modelling is a knowledge-based structure prediction relying on observed features in known homologous protein structures. By

(continued)

(continued)

exploiting this information from template structures the structural model of the target protein can be constructed (Wiltgen and Tilz 2009).

Two well-known homology modelling programs, which are free for academic research, are

MODELLER (http://salilab.org/modeller) and

SWISSMODEL (http://swissmodel.expasy.org).

The slide shows the comparison of two proteins: The sequences of both proteins are 95 % (53 of 56) identical (only residues 20, 30, and 45 differ), yet the structures are totally different.

### Slide 5-43: Towards Personalized Medicine

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.

## 7 Future Outlook

### Slide 5-: Future Outlook

All these approaches are producing gigantic amounts of highly complex datasets, and the amounts are rising. In particular the amount of unstructured data (or information respectively) rises. Predictive modeling and machine learning are increasingly central to the business models of data-driven businesses (Dhar 2013).

# 8 Exam Questions

## *8.1 Yes/No Decision Questions*

Please check the following sentences and decide whether the sentence is true = YES; or false = NO; for each correct answer you will be awarded 2 credit points.

| | | | |
|---|---|---|---|
| 01 | One of the most exciting and challenging frontiers in neuroscience involves harnessing the power of large-scale genetic, genomic and phenotypic data sets. | ❒ Yes<br>❒ No | 2 total |
| 02 | In the medical domain, many different people work at different times in various locations, therefore standardized data is the basis for accurate communication. | ❒ Yes<br>❒ No | 2 total |
| 03 | XML is often classified as semi-structured, however this is in some way misleading, as the data itself is still structured, but in a flexible rather than a static way. | ❒ Yes<br>❒ No | 2 total |
| 04 | Non-standardized data is an realistic case and is the minority of data but support data quality, data exchange and interoperability in information systems. | ❒ Yes<br>❒ No | 2 total |
| 05 | In order to understand complex biological systems, the three following key concepts need to be considered: emergence, robustness, and standardization. | ❒ Yes<br>❒ No | 2 total |
| 06 | A transcriptional regulatory network has two components: transcription factor (TF) and target genes (TG), where TF regulates the transcription of TGs. | ❒ Yes<br>❒ No | 2 total |
| 07 | The complexity of organisms arises rather as a consequence of elaborated regulations of gene expression than from differences in genetic content in terms of the number of genes. | ❒ Yes<br>❒ No | 2 total |
| 08 | In genetics, a sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance. | ❒ Yes<br>❒ No | 2 total |
| 09 | The decision problem associated with Correlated Motif Mining (CMM) is solvable in P. | ❒ Yes<br>❒ No | 2 total |
| 10 | Our brain forms one integrative complex network, linking all brain regions and sub-networks together. | ❒ Yes<br>❒ | 2 total |

| | | |
|---|---|---|
| Sum of Question Block A (max. 20 points) | | |

## *8.2 Multiple Choice Questions (MCQ)*

The following questions are composed of two parts: the stem, which identifies the question or problem and a set of alternatives which can contain 0, 1, 2, 3, or 4 correct answers, along with a number of distractors that might be plausible—but are incorrect. Please **select the correct answers** by ticking ☒—and do not forget that it can be none. Each question will be awarded 4 points ***only if everything is correct***.

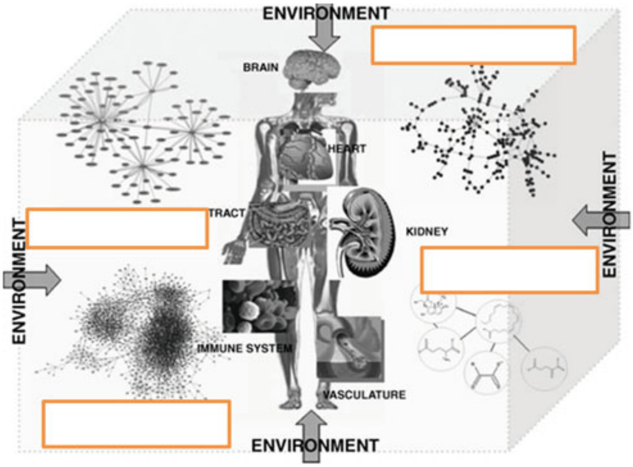
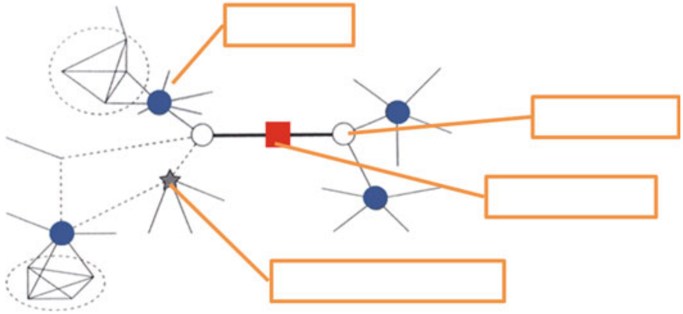
| | | |
|---|---|---|
| 01 | Homology …<br>❐ a) ... In mathematics homology is a formalism for talking in a quantitative and unambiguous manner about how a space is connected.<br>❐ b) ... origins from Greek ὁμολογέω (homologeo) and means "to conform".<br>❐ c) ... is used for a correspondence of structures in two life forms with a common evolutionary origin.<br>❐ d) ... has its origins in Darwinian Biology. | 4 total |
| 02 | The four network representations of protein networks include …<br>❐ a) ... protein complex structure.<br>❐ b) ... true PPI structure.<br>❐ c) ... Spoke model.<br>❐ d) ... Matrix model with bait in the center. | 4 total |
| 03 | Homology modelling …<br>❐ a) … is extremely important for personalized and molecular medicine.<br>❐ b) … is based on the principle that homologue proteins are very different.<br>❐ c) … uses a protein sequence with known structures (targets) to align it with a protein structure with unknown structures (templates).<br>❐ d) … is a knowledge-based prediction of protein structures. | 4 total |
| 04 | Drosophila melanogaster …<br>❐ a) … is an insect which has only some 140 genes.<br>❐ b) … is a very recently found laboratory animal and very important for research in personalized medicine.<br>❐ c) … has been used for many years and is of no more use in genomics.<br>❐ d) … is a model organism and shares many genes with humans. | 4 total |
| 05 | The centrality of a network …<br>❐ a) ... measures the level of "betweeness" of a node (the "importance").<br>❐ b) ... indicates how many of the shortest paths between the nodes of the network pass through node i.<br>❐ c) ... describes the possible formation of communities in the network.<br>❐ d) ... indicates how strong groups of nodes form relative isolated sub-networks within the full network. | 4 total |
| 06 | Scale-free Topology …<br>❐ … ensures that there are short paths between pairs of nodes, allowing rapid communication between otherwise distant parts of the network.<br>❐ … is a set of techniques, applied from statistics, which analyze the topological structure of a network.<br>❐ … is used as a model to predict future values of a topological structure in networks.<br>❐ … is a measure of similarity between two protein structures. | 4 total |

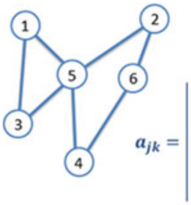
| | | |
|---|---|---|
| 07 | Semi-structured data …<br>❒ a) ... does not conform with the formal structure of tables/data models associated with relational databases.<br>❒ b) ... means randomness, noise and uncertainty.<br>❒ c) ... enforces hierarchies of records and fields within the data.<br>❒ d) ... contains tags/markers to separate semantic elements. | 4 total |
| 08 | Data standardization refers to …<br>❒ a) ... the data content.<br>❒ b) ... the terminologies that are used to represent this data.<br>❒ c) ... how data is exchanged.<br>❒ d) ... how knowledge, e.g. clinical guidelines, protocols, decision support rules, checklists, standard operating procedures are represented in the health information system. | 4 total |
| 09 | Metabolism …<br>❒ a) ... is the study of DNA-sequencing methods and produces a lot of complex data.<br>❒ b) ... is primarily determined by genes, environment and nutrition.<br>❒ c) ... consists of chemical reactions catalyzed by enzymes to produce essential components such as amino acids, sugars and lipids.<br>❒ d) ... is a process within genetics where regulatory metabolic elements have influence on nucleotide sequences. | 4 total |
| 10 | Phenotype …<br>❒ a) ... an organism's observable characteristics (traits).<br>❒ b) ... result from the expression of an organism's genes as well as the influence of environmental factors and the interactions between them.<br>❒ c) ... are inherited instructions within its genetic code.<br>❒ d) ... includes morphology, biochemical/physiological properties, behaviour, etc. | 4 total |

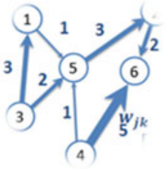
| | | |
|---|---|---|
| **Sum of Question Block B (max. 40 points)** | | |

## 8.3 Free Recall Block

Please follow the instructions below. At each question you will be assigned the credit points indicated if your option is correct (partial points may be given).

| | | |
|---|---|---|
| 01 | Identifying networks in disease research is an important aspect of systems biology, where there is a high diversity of molecular networks within and between cells. Please identify in the following picture the networks and write the name of the network in the appropriate space!<br><br> ENVIRONMENT<br>BRAIN<br>HEART<br>TRACT<br>KIDNEY<br>ENVIRONMENT<br>ENVIRONMENT<br>IMMUNE SYSTEM<br>VASCULATURE<br>ENVIRONMENT | 1 each<br>4 total |
| 02 | A graph $\boldsymbol{G}\ \boldsymbol{V},\boldsymbol{E}$ describes a structure which consists of nodes aka vertices $\boldsymbol{V}$, connected by a set of pairs of distinct nodes (links), called edges $\boldsymbol{E}\ \ \boldsymbol{a},\boldsymbol{b}$ with $\mathbf{a},\mathbf{b}\in\boldsymbol{V};\boldsymbol{a}\neq\boldsymbol{b}$. Please name the symbols in the following network example:<br> | 1 each<br>4 total |

<table>
<tr>
<td>03</td>
<td>In order to represent network data in computers it is not comfortable to use sets; more practical are matrices. The simplest form of a graph representation is the so called adjacency matrix. Please set up the adjacency matrices for the following graphs:<br>



</td>
<td>1 each<br>6 total</td>
</tr>
<tr>
<td>04</td>
<td>In Biomedicine networks of all kind play an extremely important role. Please assign the correct labels to the network metrics below:<br>

| | |
|---|---|
| $\sum_i \sum_j a_{ij}$ | Path length |
| $k_i = \sum_i a_{ij}$ | Network size |
| $C = \frac{1}{n} \sum_i C_i$ | Complete number of nodes |
| $n$ | Clustering coefficient |
| $l = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$ | Nodal degree |

</td>
<td>1 each<br>6 total</td>
</tr>
</table>

| | | |
|---|---|---|
| 05 | Please draw a metabolic network, constructed considering the reactants, chemical reactions and enzymes, consisting of A, B, C, D, E1 and E2 | 1 each<br>6 total |
| 06 | Please provide a simple example of a patient record in XML format: | 1 each<br>4 total |

| | | |
|---|---|---|
| Sum of Question Block C (max. 40 points) | | |

# 9 Answers

## *9.1 Answers to the Yes/No Questions*

Please check the following sentences and decide whether the sentence is true = YES; or false = NO; for each correct answer you will be awarded 2 credit points.

| | | | |
|---|---|---|---|
| 01 | One of the most exciting and challenging frontiers in neuroscience involves harnessing the power of large-scale genetic, genomic and phenotypic data sets. | ☒ Yes<br>❐ No | 2 total |
| 02 | In the medical domain, many different people work at different times in various locations, therefore standardized data is the basis for accurate communication. | ☒ Yes<br>❐ No | 2 total |
| 03 | XML is often classified as semi-structured, however this is in some way misleading, as the data itself is still structured, but in a flexible rather than a static way. | ☒ Yes<br>❐ No | 2 total |
| 04 | Non-standardized data is an realistic case and is the minority of data but support data quality, data exchange and interoperability in information systems. | ❐ Yes<br>☒ No | 2 total |
| 05 | In order to understand complex biological systems, the three following key concepts need to be considered: emergence, robustness, and standardization. | ❐ Yes<br>☒ No | 2 total |
| 06 | A transcriptional regulatory network has two components: transcription factor (TF) and target genes (TG), where TF regulates the transcription of TGs. | ☒ Yes<br>❐ No | 2 total |
| 07 | The complexity of organisms arises rather as a consequence of elaborated regulations of gene expression than from differences in genetic content in terms of the number of genes. | ☒ Yes<br>❐ No | 2 total |
| 08 | In genetics, a sequence motif is a nucleotide or amino-acid sequence pattern that is widespread and has, or is conjectured to have, a biological significance. | ☒ Yes<br>❐ No | 2 total |
| 09 | The decision problem associated with Correlated Motif Mining (CMM) is solvable in P. | ❐ Yes<br>☒ No | 2 total |
| 10 | Our brain forms one integrative complex network, linking all brain regions and sub-networks together. | ☒ Yes<br>❐ No | 2 total |

| | |
|---|---|
| Sum of Question Block A (max. 20 points) | □□ |

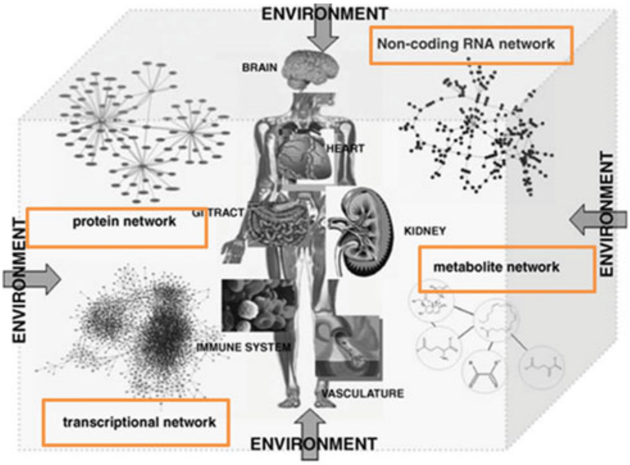
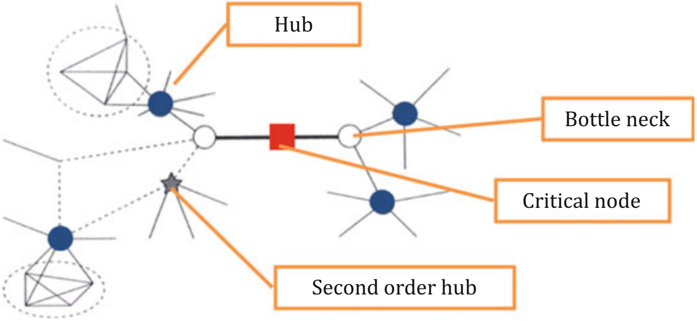
## *9.2 Answers to the Multiple Choice Questions (MCQ)*

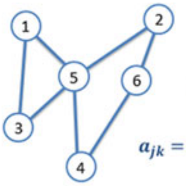
| | | |
|---|---|---|
| 01 | Homology …<br>☒ a) ... In mathematics homology is a formalism for talking in a quantitative and unambiguous manner about how a space is connected.<br>☒ b) ... origins from Greek ὁμολογέω (homologeo) and means "to conform".<br>☒ c) ... is used for a correspondence of structures in two life forms with a common evolutionary origin.<br>☒ d) ... has its origins in Darwinian Biology. | 4 total |
| 02 | The four network representations of protein networks include …<br>☒ a) ... protein complex structure.<br>☒ b) ... true PPI structure.<br>☒ c) ... Spoke model.<br>☒ d) ... Matrix model with bait in the center. | 4 total |
| 03 | Homology modelling …<br>☒ a) … is extremely important for personalized and molecular medicine.<br>❐ b) … is based on the principle that homologue proteins are very different.<br>❐ c) … uses a protein sequence with known structures (targets) to align it with a protein structure with unknown structures (templates).<br>☒ d) … is a knowledge-based prediction of protein structures. | 4 total |
| 04 | Drosophila melanogaster …<br>❐ a) … is an insect which has only some 140 genes.<br>❐ b) … is a very recently found laboratory animal and very important for research in personalized medicine.<br>❐ c) … has been used for many years and is of no more use in genomics.<br>☒ d) ... is a model organism and shares many genes with humans. | 4 total |
| 05 | The centrality of a network …<br>☒ a) ... measures the level of "betweeness" of a node (the "importance").<br>☒ b) ... indicates how many of the shortest paths between the nodes of the network pass through node i.<br>❐ c) ... describes the possible formation of communities in the network.<br>❐ d) ... indicates how strong groups of nodes form relative isolated sub-networks within the full network. | 4 total |
| 06 | Scale-free Topology …<br>☒ … ensures that there are short paths between pairs of nodes, allowing rapid communication between otherwise distant parts of the network.<br>❐ … is a set of techniques, applied from statistics, which analyze the topological structure of a network.<br>❐ … is used as a model to predict future values of a topological structure in networks.<br>❐ … is a measure of similarity between two protein structures. | 4 total |
| 07 | Semi-structured data …<br>☒ a) ... does not conform with the formal structure of tables/data models associated with relational databases.<br>❐ b) ... means randomness, noise and uncertainty.<br>☒ c) ... enforces hierarchies of records and fields within the data.<br>☒ d) ... contains tags/markers to separate semantic elements. | 4 total |

| | | |
|---|---|---|
| 08 | Data standardization refers to …<br>☒ a) ... the data content.<br>☒ b) ... the terminologies that are used to represent this data.<br>☒ c) ... how data is exchanged.<br>☒ d) ... how knowledge, e.g. clinical guidelines, protocols, decision support rules, checklists, standard operating procedures are represented in the health information system. | 4 total |
| 09 | Metabolism …<br>❒ a) ... is the study of DNA-sequencing methods and produces a lot of complex data.<br>☒ b) ... is primarily determined by genes, environment and nutrition.<br>☒ c) ... consists of chemical reactions catalyzed by enzymes to produce essential components such as amino acids, sugars and lipids.<br>❒ d) ... is a process within genetics where regulatory metabolic elements have influence on nucleotide sequences. | 4 total |
| 10 | Phenotype …<br>☒ a) ... an organism's observable characteristics (traits).<br>☒ b) ... result from the expression of an organism's genes as well as the influence of environmental factors and the interactions between them.<br>❒ c) ... are inherited instructions within its genetic code.<br>☒ d) ... includes morphology, biochemical/physiological properties, behaviour, etc. | 4 total |

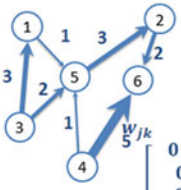
| | | |
|---|---|---|
| **Sum of Question Block B (max. 40 points)** | | |

## 9.3 Answers to the Free Recall Questions

| | | |
|---|---|---|
| 01 | Identifying networks in disease research is an important aspect of systems biology, where there is a high diversity of molecular networks within and between cells. Please identify in the following picture the networks and write the name of the network in the appropriate space!<br><br> ENVIRONMENT<br>Non-coding RNA network<br>BRAIN<br>HEART<br>GI TRACT<br>protein network<br>KIDNEY<br>ENVIRONMENT<br>ENVIRONMENT<br>metabolite network<br>IMMUNE SYSTEM<br>VASCULATURE<br>transcriptional network<br>ENVIRONMENT | 1 each<br>4 total |
| 02 | A graph $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$ describes a structure which consists of nodes aka vertices $\boldsymbol{V}$, connected by a set of pairs of distinct nodes (links), called edges $\boldsymbol{E}\,\{\boldsymbol{a},\boldsymbol{b}\}$with $\mathbf{a},\mathbf{b} \in \boldsymbol{V}; \boldsymbol{a} \neq \boldsymbol{b}$. Please name the symbols in the following network example:<br><br> | 1 each<br>4 total |

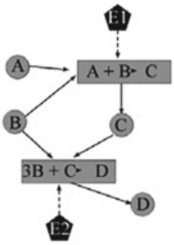
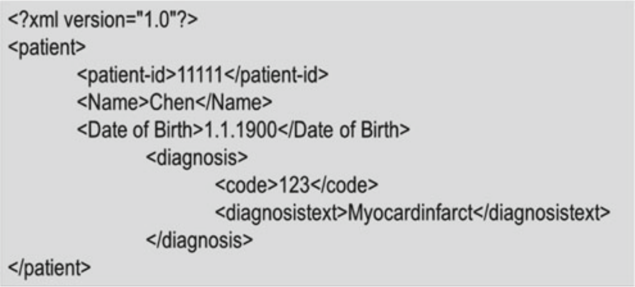
<table>
<tr>
<td>03</td>
<td>In order to represent network data in computers it is not comfortable to use sets; more practical are matrices. The simplest form of a graph representation is the so called adjacency matrix. Please set up the adjacency matrices for the following graphs:<br>


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

Simple graph, symmetric, binary<br>

1<br>2<br>3<br>4<br>5<br>6<br>1<br>3<br>2<br>3<br>2<br>1<br>5

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

Directed and weighted</td>
<td>1 each<br>6 total</td>
</tr>
<tr>
<td>04</td>
<td>In Biomedicine networks of all kind play an extremely important role. Please assign the correct labels to the network metrics below:
<table>
<tr><td>$\sum_i \sum_j a_{ij}$</td><td>Path length</td></tr>
<tr><td>$k_i = \sum_i a_{ij}$</td><td>Network size</td></tr>
<tr><td>$C = \frac{1}{n} \sum_i C_i$</td><td>Complete number of nodes</td></tr>
<tr><td>$n$</td><td>Clustering coefficient</td></tr>
<tr><td>$l = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$</td><td>Nodal degree</td></tr>
</table>
</td>
<td>1 each<br>6 total</td>
</tr>
</table>

| | | |
|---|---|---|
| 05 | Please draw a metabolic network, constructed considering the reactants, chemical reactions and enzymes, consisting of A, B, C, D, E1 and E2<br><br> | 1 each<br>6 total |
| 06 | Please provide a simple example of a patient record in XML format:<br> | 1 each<br>4 total |

```
<?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>
```

| | |
|---|---|
| Sum of Question Block C (max. 40 points) | |

## References

Aarts E, Lenstra J (1997) Local search in combinatorial optimization. Wiley, New York, NY

Achard F, Vaysseix G, Barillot E (2001) XML, bioinformatics and data integration. Bioinformatics 17(2):115

Barabási A-L, Albert R, Jeong H (1999) Mean-field theory for scale-free random networks. Phys A Stat Mech Its Appl 272(1–2):173–187

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Bondy JA, Murty USR (1976) Graph theory with applications. Macmillan, London

Boyen P, Van Dyck D, Neven F, Van Ham RCHJ, Van Dijk ADJ (2011) SLIDER: a generic metaheuristic for the discovery of correlated motifs in protein-protein interaction networks. IEEE/ACM Trans Comput Biol Bioinform 8(5):1344–1357

Catchpoole DR, Kennedy P, Skillicorn DB, Simoff S (2010) The curse of dimensionality: a blessing to personalized medicine. J Clin Oncol 28(34):E723–E724

Chien KR, Domian IJ, Parker KK (2008) Cardiogenesis and the complex biology of regenerative cardiovascular medicine. Science 322(5907):1494

Costa LF, Rodrigues FA, Cristino AS (2008) Complex networks: the key to systems biology. Genet Mol Biol 31(3):591–601

Costa LF, Rodrigues FA, Travieso G, Boas PRV (2007) Characterization of complex networks: a survey of measurements. Adv Phys 56(1):167–242

Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London

Dehmer M, Emmert-Streib F, Mehler A (2011) Towards an information theory of complex networks: statistical methods and applications. Birkhäuser, Boston, MA

Dhar V (2013) Data science and prediction. Commun ACM 56(12):64–73

Diestel R (2010) Graph theory, 4th edn. Springer, Berlin

Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: from biological nets to the internet and WWW. Oxford University Press, New York, NY

Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley, New York, NY

Edelsbrunner H, Harer JL (2010) Computational topology: an introduction. American Mathematical Society, Providence, RI

Fekete J-D (2004) The infovis toolkit. Information visualization, INFOVIS 2004. IEEE, Washington, DC, pp 167–174

Forster C, Vossen G (2012) Exploiting XML technologies in medical information systems. In: Proceedings of the 25th Bled eConference eDependability: reliable and trustworthy eStructures, eProcesses, eOperations and eServices for the future, Bled, Slovenia, pp 70–83

Gaal SA (1966) Point set topology, 2nd edn. Academic, New York, NY

Geschwind DH, Konopka G (2009) Neuroscience in the era of functional genomics and systems biology. Nature 461(7266):908–915

Golumbic MC (2004) Algorithmic graph theory and perfect graphs. Elsevier, Amsterdam

Harary F (1969) Graph theory. Addison-Wesley, Reading, MA

Hatcher A (2002) Algebraic topology. Cambridge University Press, Cambridge

Henzinger MR, Klein P, Rao S, Subramanian S (1997) Faster shortest-path algorithms for planar graphs. J Comput Syst Sci 55(1):3–23

Hodgman CT, French A, Westhead DR (2010) Bioinformatics, 2nd edn. Taylor & Francis, New York, NY

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

Holzinger A (2011) Weakly structured data in health-informatics: the challenge for human-computer interaction. In: Baghaei N, Baxter G, Dow L, Kimani S (eds) Proceedings of INTERACT 2011 workshop: promoting and supporting healthy living by design. IFIP, Lisbon, Portugal, pp 5–7

Holzinger A (2012) On knowledge discovery and interactive intelligent visualization of biomedical data: challenges in human–computer interaction & biomedical informatics. In: Helfert M, Fancalanci C, Filipe J (eds) DATA - international conference on data technologies and applications. INSTICC, Rome, pp 5–16

Holzinger A, Ofner B, Stocker C, Valdez AC, Schaar AK, Ziefle M, Dehmer M (2013a) On graph entropy measures for knowledge discovery from publication network data. In: Cuzzocrea A, Kittl C, Simos DE, Weippl E, Xu L (eds) Multidisciplinary research and practice for information systems, vol LNCS 8127, Springer lecture notes in computer science. Springer, Heidelberg, pp 354–362

Holzinger A, Simonic KM, Geier M, Ofner B, Müller R, Heschl S, Prause G (2013) Constraints of list-based knowledge interaction on an android app for emergency medicine. Medicine 2.0 London, Oral Presentation on 23 Sept 2013 (Online). http://www.medicine20congress.com/ocs/index.php/med/med2013/paper/view/1479

Holzinger A, Stocker C, Ofner B, Prohaska G, Brabenetz A, Hofmann-Wellenhof R (2013c) Combining HCI, natural language processing, and knowledge discovery - potential of IBM content analytics as an assistive technology in the biomedical domain, vol LNCS 7947, Springer lecture notes in computer science. Springer, Heidelberg, pp 13–24

Kim PM, Korbel JO, Gerstein MB (2007) Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proc Natl Acad Sci U S A 104 (51):20274–20279

Kleinberg JM (2000) Navigation in a small world. Nature 406(6798):845

Koontz WLG, Narendra PM, Fukunaga K (1976) A graph-theoretic approach to nonparametric cluster analysis. IEEE Trans Comput 100(9):936–944

Kreuzthaler M, Bloice MD, Faulstich L, Simonic KM, Holzinger A (2011) A comparison of different retrieval strategies working on medical free texts. J Univ Comput Sci 17(7):1109–1133

Kropatsch W, Burge M, Glantz R (2001) Graphs in image analysis. In: Kropatsch W, Bischof H (eds) Digital image analysis. Springer, New York, NY, pp 179–197

Lage K, Møllgård K, Greenway S, Wakimoto H, Gorham JM, Workman CT, Bendsen E, Hansen NT, Rigina O, Roque FS (2010) Dissecting spatio-temporal protein networks driving human heart development and related disorders. Mol Syst Biol 6(1):1–9

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. CRC Press, Boca Raton, FL, pp 1–24

Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P (2007) Data integration and genomic medicine. J Biomed Inform 40(1):5–16

Lucchi A, Smith K, Achanta R, Lepetit V, Fua P (2010) A fully automated approach to segmentation of irregularly shaped cellular structures in EM images. Medical image computing and computer-assisted intervention—MICCAI 2010. Springer, Berlin, pp 463–471

Meijster A, Roerdink JB (1995) A proposal for the implementation of a parallel watershed algorithm. Computer analysis of images and patterns. Springer, Berlin, pp 790–795

Milgram S (1967) The small world problem. Psychol Today 2(1):60–67

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256

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. Int J Med Inform 70(2–3):109–115

Roerdink JB, Meijster A (2000) The watershed transform: definitions, algorithms and parallelization strategies. Fundamenta Informaticae 41(1):187–228

Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Søeby K, Bredkjær S, Juul A, Werge T, Jensen LJ, Brunak S (2011) Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 7(8):e1002141

Salgado H, Santos-Zavaleta A, Gama-Castro S, Peralta-Gil M, Peñaloza-Spínola MI, Martínez-Antonio A, Karp PD, Collado-Vides J (2006) The comprehensive updated regulatory network of Escherichia coli K-12. BMC Bioinform 7(1):5

Schadt EE, Lum PY (2006) Reverse engineering gene networks to identify key drivers of complex disease phenotypes. J Lipid Res 47(12):2601–2613

Schmid AK, Reiss DJ, Pan M, Koide T, Baliga NS (2009) A single transcription factor regulates evolutionarily diverse but functionally linked metabolic pathways in response to nutrient availability. Mol Syst Biol 5(1):1–9

Simon HA (1973) The structure of ill structured problems. Artif Intell 4(3–4):181–201

Strogatz SH (2001) Exploring complex networks. Nature 410(6825):268–276

Usdin T, Graham T (1998) XML: not a silver bullet, but a great pipe wrench. ACM Stand View 6 (3):125–132

Van Den Heuvel MP, Hulshoff Pol HE (2010) Exploring the brain network: a review on resting-state fMRI functional connectivity. Eur Neuropsychopharmacol 20(8):519–534

Vincent L, Soille P (1991) Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans Pattern Anal Machine Intell 13(6):583–598

Wang Z, Zhang JZ (2007) In search of the biological significance of modular structures in protein networks. PLoS Comput Biol 3(6):1011–1021

Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. Nature 393 (6684):440–442

Wiltgen M, Tilz GP (2009) Homology modelling: a review about the method on hand of the diabetic antigen GAD 65 structure prediction. Wiener Medizinische Wochenschrift 159 (5):112–125

Wittkop T, Emig D, Truss A, Albrecht M, Böcker S, Baumbach J (2011) Comprehensive cluster analysis with transitivity clustering. Nat Protoc 6(3):285–295