Lecture 05

Semi structured, weakly structured data

Graphs, Networks and Homologies

a.holzinger@tugraz.at

Tutor: markus.plass@student.tugraz.at

http://hci-kdd.org/biomedical-informatics-big-data
Schedule

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

- Big data pools
- Complex networks
- Computational graph representation
- Electronic patient record (EPR)
- Homology modeling
- Macroscopic structures
- Medical documentation
- Metabolic network
- Microscopic structures
- Network metrics
- Structural data dimension
- Topological structures
Adjacency matrix = simplest form of computational graph representation, in which 0 or 1 denotes whether or not there is a directed edge from one node to another (in graph theory adjacent nodes in a graph are linked by an edge);

Artifacts = not only a noise disturbance, which is contaminating and influencing the signal (surrogates) but also data which is wrong, however interpreted as to be reliable, consequently may lead to a wrong decision;

Computational graph representation = e.g. by adjacency matrices

Data fusion = data integration techniques that analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion (e.g. combined sensor data in Ambient Assisted Living);

Global Distance Test (GDT) = a measure of similarity between two protein structures with identical amino acid sequences but different tertiary structures. It is most commonly used to compare the results of protein structure prediction to the experimentally determined structure as measured by X-ray crystallography or protein NMRM;

Graph theory = study of mathematical structures to model relations between objects from a certain collection;

Graphs = a hypothetical structure consisting of a series of nodes connected by weighted edges (graphs can be directed/undirected and stoichimetric/non-stoichimetric regarding interaction classes);
Advance Organizer (2/3) H-P

- **Homology** = in mathematics (especially algebraic topology and abstract algebra), it is (ὁμόιος homos = "identical") a certain general procedure to associate a sequence of Abelian groups (i.e. does not depend on their order) or modules with a given mathematical object such as a topological space or a group;

- **Homology modeling** = comparative modeling of protein, refers to constructing an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein (the "template"); in Bioinformatics, homology modeling is a technique that can be used in molecular medicine.

- **In silico** = via computer simulation, in contrast to in vivo (within the living) or in vitro (within the glass);

- **Multi-scale representation** = in a graph, nodes do not have to represent biological objects on the same scale, one node (e.g. a molecule) may have an edge connecting it to a node representing a cell or tissue (the edge indicates that the molecule exerts an effect on the cell/tissue);

- **Network** = graphs containing cycles or alternative paths;

- **Network analysis** = a set of techniques used to characterize relationships among discrete nodes in a graph or a network;

- **Network topology** = the shape or structure of a network;

- **Petri-Net** = a special class of graph, consisting of two general classes or node: place and transition nodes;

- **Predictive modeling** = a set of techniques in which a mathematical model is created or chosen to best predict the probability of an outcome (e.g. regression);

- **P-System** = addresses the slowness of Petri-nets
• **Radius of a graph** = average minimum path length (biological networks are not arranged in a regular or symmetrical pattern);

• **Scale-free Topology** = ensures that there are very short paths between any given pair of nodes, allowing rapid communication between otherwise distant parts of the network (e.g. the Web has such a topology);

• **Semi-structured data** = does not conform with the formal structure of tables/data models assoc. with relational databases, but at least contains tags/markers to separate semantic elements and enforce hierarchies of records and fields within the data; aka schemaless or self-describing structure; the entities belonging to the same class may have different attributes even though they are grouped together;

• **Spatial analysis** = a set of techniques, applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data set;

• **Structural homology** = similar structure but different function;

• **Supervised learning** = machine learning techniques that infer a function or relationship from a set of training data (e.g. classification and support vector machines);

• **Time series analysis** = set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data;

• **Time series forecasting** = use of a model to predict future values of a time series based on known past values of the same or other series (e.g. structural modeling); decomposition of a series into trend, seasonal, and residual components, which can be useful for identifying cyclical patterns in the data;

• **Unstructured data** = complete randomness, noise; (wrongly, text is called unstructured, but there is some structure, too, so text data is a kind of weakly structured data);

• **Vertex degree** = within a topology, the numbers of edges connecting to a node;
Glossary

- ANSI = American National Standards Institute
- CD = cardiac development
- CDA = Clinical Document Architecture
- CHD = congenital heart disease
- CMM = Correlated motif mining
- DPI = Dossier Patient Integre´ = integrated patient record
- E = Edge
- EPR = Electronic Patient Record
- G(V,E) = Graph
- GI = gastrointestinal
- HER = Electronic Health Record
- HL7 = Health Level 7
- KEGG = Kyoto Encyclopedia of Genes and Genomes
- NP = nondeterministic polynomial time
- OWL = Web Ontology Language
- PPI = Protein-Protein Interaction
- SGML = Standard Generalized Markup Language
- TF = Transcription factor
- TG = Target Gene
- V = Vertex
- XML = Extensible Markup Language
Learning Goals ... at the end of the 5th lecture you ...

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

- Automated Machine Learning algorithms need much training data – focus is on adjusting model parameters without fully understanding the data that the learning algorithm is modeling [1]
- Weakly structured data [4]

Complexity Problem: Time versus Space

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

Slide 5-3: Example: 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 Palace
  - Town/City: LONDON
  - Phone 1: +41 00 000000
  - Email: prince.charles@buckingham.co.uk

Options for this person:
- Admission - Inpatient
- Visit - Outpatient
- Appointments
- Encounters' list
- Medocs
- DRG (composite)
- Diagnostic Results
- Prescriptions
- Notes & Reports
- Immunization
- Measurements
- Birth details
- DB Record's History
- Make PDF document

http://care2x.org
<?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>

Example:

Generic XML template for a med. report

DPI = Dossier
Patient Intégré =
integrated
patient record

Comparison of XML - RDF/OWL in Bioinformatics

Networks = Graphs

STATISTICAL AND MACHINE LEARNING APPROACHES FOR NETWORK ANALYSIS

Edited by
MATTHIAS DEHMER
UMIT – The Health and Life Sciences University. Institute for Bioinformatics and Translational Research. Hall in Tyrol, Austria

SUSHANT C. BASAK
Natural Resources Research Institute
University of Minnesota, Duluth
Duluth, MN, USA

http://www.wired.com/tag/network-science/
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.
$G(V, E)$ Graph

$V$ ... vertex

$E$ ... edge \{a, b\}

a, b \in V; a \neq b

Hodgman, C. T.,
French, A. &
Westhead, D. R.
Adjacency \((ə-'jā-sən(t)-sē)\) Matrix \(A = (a_{jk})\)

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

Simple graph, symmetric, binary

Directed and weighted

Example: Tool for Node-Link Visualization

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

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

Biological Image Analysis Primer, available via
http://www.imagescience.org/meijering/publications/1009/ Erasmus
University Medical Center
**Order** = total number of nodes $n$; **Size** = total number of links $(a)$:

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

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

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

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

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

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

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

**Modularity** (f) = describes the possible formation of communities in the network, indicating how strong groups of nodes form relative isolated sub-networks within the full network (refer also to Slide 5-8).
Slide 5-13: Network Topologies

(a) **Regular**

**Small-world**

**Random**

---

randomness

(b) **Scale-free network**

**Modular network**

Van Heuvel & Hulshoff (2010)
Slide 5-14: Small-World Networks

Increasing randomness


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

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

Slide 5-19 Graphs from Images: Watershed + Centroid


Slide 5-23: Example for a Medical Knowledge Space

# Nodes: 641
# Edges: 1250

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

Nodes
- drugs
- clinical guidelines
- patient conditions (indication, contraindication)
- pharmacological groups
- tables and calculations of medical scores
- algorithms and other medical documents

Edges: 3 crucial types of relations inducing medical relevance between two active substances
- pharmacological groups
- indications
- contra-indications

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

Green nodes (from dark to light):
1. Application (one ore more indications + corresponding dosages)
2. Single indication with additional details (e.g. “VF after 3rd Shock”)
3. Condition (e.g. VF, Ventricular Fibrillation)
Example: The brain is a complex network

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

Slide 5-29: Example: Cell-based therapy

**A. Early phenotypes**

- E1. Abnormal heart tube morphology
- E2. Abnormal looping morphogenesis
- E3. Abnormal sinus venosus
- E4. Abnormal atrioventricular canal morphology

**Intermediate phenotypes**

- I1. Atrial septal defect
- I2. Abnormal endocardial cushion morphology
- I3. Abnormal atrioventricular valve morphology
- I4. Abnormal myocardial trabeculae morphology

**Late phenotypes**

- L1. Abnormal semilunar valve morphology
- L2. Abnormal outflow tract development
- L3. Double outlet right ventricle
- L4. Ventricular septal defect

**Function of clusters**

- BMP/TGFβ signaling
- Transcription regulation
- WNT signaling
- FGF/PDGFR signaling
- Retinoic acid signaling
- Semaphorin signaling
- NOTCH signaling
- ERBB signaling
- Focal adhesion signaling
- Cell cycle regulation
- Other function

**No. of proteins in clusters**

- 10
- 20
- 30
- 40
- 50

**Direct interaction**

**Indirect interaction**

Lage et. al (2010)
Slide 5-31: Three main types of biomedical networks

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

- **Protein-Protein interaction network**

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

Complex networks: the key to systems biology.
*Genetics and Molecular Biology, 31, 3, 591–601.*
Correlated Motif Mining (CMM)

Input: PPI-network $G = (V, E, \lambda)$, $\ell$, $d \in \mathbb{N}$, $d < \ell$

Output: $\{X^*, Y^*\}$ best correlated motif pair found in $G$

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

Boyen et al. (2011)
Matrix contains many sparse elements - In this case it is computationally more efficient to represent the graph as an adjacency list.


http://www.nature.com/msb/journal/v5/n1/fig_tab/msb200940_F6.html
Electronic patient records remain a unexplored, but potentially rich data source for example to discover correlations between diseases.

Example: ὁμολογέω (homologeo)

Homology modeling is a knowledge-based prediction of protein structures.

In homology modeling a protein sequence with an unknown structure (the target) is aligned with one or more protein sequences with known structures (the templates).

The method is based on the principle that homologue proteins have similar structures.

Homology modeling will be extremely important to personalized and molecular medicine in the future.
Slide 5-42: Future Outlook

- **2003**: Genomics
- **2013**: Proteomics
- **2023**: Personalized Medicine

Future Outlook:
- Genomics
- Proteomics
- Personalized Medicine
Thank you!
Sample Questions

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

- http://www.cdisc.org
- http://www.w3.org/Math/
- http://www.sgpp.org/structures.shtml
- http://salilab.org/modeller
- http://www.geneticseducation.nhs.uk
Appendix: clustering network motifs in integrated networks

http://omics.frias.uni-freiburg.de/
Example from Immunology: Structural Homology

http://www.maa.org/cvm/1998/01/tprppoh/article/Pictures/KleinBottle.gif
Medical Documentation – Patient Record
Person registration

New person

- 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
- Blood group: O
- Civil status: Widowed
- Address:
  - Street: Buckingham Palace
  - Town/City: LONDON
  - Phone 1: +41 00 000000
  - Email: prince.charles@buckingham.co.uk
  - Other Hospital Nr.: medical doctor

Options for this person:
- Admission - Inpatient
- Visit - Outpatient
- Appointments
- Encounters' list
- Medocs
- DRG (composite)
- Diagnostic Results
- Prescriptions
- Notes & Reports
- Immunization
- Measurements
- Birth details
- DB Record's History
- Make PDF document

http://care2x.org
... to integrate and analyze these diverse and voluminous data sources to elucidate both normal and disease physiology.

XML is suited for describing semi-structured data including a natural modeling of biological entities, because it allows features as e.g. nesting ...
Example: Comparison of XML and OWL data in bioinformatics

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

On time and space of data ...
... to microscopic atomistic structures


<table>
<thead>
<tr>
<th>ATOM</th>
<th>1</th>
<th>N</th>
<th>GLY</th>
<th>A</th>
<th>1</th>
<th>44.842</th>
<th>51.034</th>
<th>101.284</th>
<th>0.01</th>
<th>27.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>2</td>
<td>CA</td>
<td>GLY</td>
<td>A</td>
<td>1</td>
<td>45.640</td>
<td>50.230</td>
<td>100.389</td>
<td>0.01</td>
<td>26.99</td>
</tr>
<tr>
<td>ATOM</td>
<td>3</td>
<td>C</td>
<td>GLY</td>
<td>A</td>
<td>1</td>
<td>46.692</td>
<td>49.648</td>
<td>101.308</td>
<td>0.01</td>
<td>26.80</td>
</tr>
<tr>
<td>ATOM</td>
<td>4</td>
<td>O</td>
<td>GLY</td>
<td>A</td>
<td>1</td>
<td>46.895</td>
<td>50.222</td>
<td>102.381</td>
<td>0.01</td>
<td>26.91</td>
</tr>
<tr>
<td>ATOM</td>
<td>5</td>
<td>N</td>
<td>SER</td>
<td>A</td>
<td>2</td>
<td>47.283</td>
<td>48.516</td>
<td>100.951</td>
<td>1.00</td>
<td>26.26</td>
</tr>
<tr>
<td>ATOM</td>
<td>6</td>
<td>CA</td>
<td>SER</td>
<td>A</td>
<td>2</td>
<td>48.277</td>
<td>47.866</td>
<td>101.761</td>
<td>1.00</td>
<td>26.17</td>
</tr>
<tr>
<td>ATOM</td>
<td>7</td>
<td>C</td>
<td>SER</td>
<td>A</td>
<td>2</td>
<td>49.212</td>
<td>47.031</td>
<td>100.845</td>
<td>1.00</td>
<td>24.21</td>
</tr>
<tr>
<td>ATOM</td>
<td>8</td>
<td>O</td>
<td>SER</td>
<td>A</td>
<td>2</td>
<td>49.060</td>
<td>47.195</td>
<td>99.630</td>
<td>1.00</td>
<td>19.77</td>
</tr>
<tr>
<td>ATOM</td>
<td>9</td>
<td>CB</td>
<td>SER</td>
<td>A</td>
<td>2</td>
<td>47.438</td>
<td>47.091</td>
<td>102.800</td>
<td>1.00</td>
<td>26.31</td>
</tr>
<tr>
<td>ATOM</td>
<td>10</td>
<td>OG</td>
<td>SER</td>
<td>A</td>
<td>2</td>
<td>46.276</td>
<td>46.356</td>
<td>102.404</td>
<td>1.00</td>
<td>27.99</td>
</tr>
<tr>
<td>ATOM</td>
<td>11</td>
<td>N</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>50.147</td>
<td>46.166</td>
<td>101.370</td>
<td>1.00</td>
<td>23.93</td>
</tr>
<tr>
<td>ATOM</td>
<td>12</td>
<td>CA</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>51.129</td>
<td>45.389</td>
<td>100.609</td>
<td>1.00</td>
<td>21.44</td>
</tr>
<tr>
<td>ATOM</td>
<td>13</td>
<td>C</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>50.953</td>
<td>43.905</td>
<td>100.849</td>
<td>1.00</td>
<td>20.32</td>
</tr>
<tr>
<td>ATOM</td>
<td>14</td>
<td>O</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>50.530</td>
<td>43.595</td>
<td>101.950</td>
<td>1.00</td>
<td>22.00</td>
</tr>
<tr>
<td>ATOM</td>
<td>15</td>
<td>CB</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>52.555</td>
<td>45.674</td>
<td>100.990</td>
<td>1.00</td>
<td>19.69</td>
</tr>
<tr>
<td>ATOM</td>
<td>16</td>
<td>CG</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>52.940</td>
<td>47.090</td>
<td>100.611</td>
<td>1.00</td>
<td>21.44</td>
</tr>
<tr>
<td>ATOM</td>
<td>17</td>
<td>ND1</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>53.371</td>
<td>47.470</td>
<td>99.422</td>
<td>1.00</td>
<td>20.87</td>
</tr>
<tr>
<td>ATOM</td>
<td>18</td>
<td>CE2</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>52.956</td>
<td>48.175</td>
<td>101.433</td>
<td>1.00</td>
<td>21.69</td>
</tr>
<tr>
<td>ATOM</td>
<td>19</td>
<td>CE1</td>
<td>HIS</td>
<td>A</td>
<td>3</td>
<td>53.676</td>
<td>48.730</td>
<td>99.476</td>
<td>1.00</td>
<td>20.57</td>
</tr>
</tbody>
</table>
First yeast protein-protein interaction network (2001)

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

The Nature of Space and Time

with a new afterword by the authors

STEPHEN HAWKING and ROGER PENROSE
Let us collect \(n\)-dimensional \(i\) observations: \(x_i = [x_{i1}, \ldots, x_{in}]\)

Example: To predict the folding of a protein

Source: Theoretical and computational Biophysics Group: http://www.ks.uiuc.edu/
Backup Slide: Overview Some Network Metrics

Van Heuvel & Hulshoff (2010)
Catchment basins:

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

Start flooding from local minima, and label ridges wherever differently evolving components meet
FIG. 5

FIG. 6

START

645

COMPUTE VORONOI DIAGRAM OF SAMPLE POINTS
S = \{S_1, S_2, ..., S_n\}

660

COMPUTE THE POLES P OF S

655

COMPUTE THE DELAUNAY TRIANGULATION OF SAMPLE POINTS AND POLES

660

DELETE TRIANGLES IN WHICH ONE OR MORE VERTICES ARE POLES

665

DELETE TRIANGLES WITH NORMAL ANGLES DEVIATING FROM VECTORS TO POLES

670

EXTRACT A TWO-DIMENSIONAL MANIFOLD

END
(a) Initial set of points.

(i) Delaunay Triangulation Graph.
Example: Cell based therapy (1) (Heart transplantation)

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

Chien et al. (2008)
Example: Network Generated by Gene Duplication

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


Wang & Zhang (2007)