Iosif Vaisman

Email: ivaisman@gmu.edu

Database

- **database**: a collection of related structured information about entities
- **file**: a collection of records
- **record**: a set of fields
- **field**: a single characteristic of an entity
- **character**: a symbol used in data field

Database Organization

Database Management System (DBMS)

Four major components of DBMS:
- Data
- Hardware
- Software
- Users
Data Model

- A named logical unit (record type, data item)
- Relationships among logical units

Relationships among logical units

- one to one
- one to many
- many to one

Relational Database Model

<table>
<thead>
<tr>
<th>Protein-code</th>
<th>Protein-name</th>
<th>Length</th>
<th>Species-origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1001</td>
<td>Hemoglobin</td>
<td>145</td>
<td>Bovine</td>
</tr>
<tr>
<td>P1002</td>
<td>Hemoglobin</td>
<td>136</td>
<td>Ovine</td>
</tr>
<tr>
<td>P1003</td>
<td>Eye Lens Protein</td>
<td>234</td>
<td>Human</td>
</tr>
</tbody>
</table>

Object-oriented Database Model

Database administration

- Redundancy eliminated
- Inconsistency avoided
- Data shared
- Standards enforced
- Security applied
- Integrity maintained
- Requirements balanced

Database warehouse

- Operational data
- Data fusion
- Data cleansing
- Metadata

Schadt et al., 2010

G A Komatsoulis et al., 2008

Data integration

Large-scale, complex data sets are shown as a network in which the nodes represent variables of biological interest, such as DNA variation, RNA variation, protein levels, protein states, metabolite levels and disease-associated traits, and the edges between these nodes represent causal relationships between the variables.
InterPro Database

December 1982
680,338 bp
606 seq

December 2012
148,390,863,904 bp
161,140,325 seq

Genome sequencing costs

Sequencing and storage cost

Example of a Genbank entry

Example of a Genbank entry
A simple, five-element descriptor, derived from the Delaunay tessellation of a protein structure in a single point per residue representation, can be assigned to each residue in the protein. The descriptor characterizes main-chain topology and connectivity in the neighborhood of the residue and does not explicitly depend on putative hydrogen bonds or any geometric parameters, including bond lengths, angles, and areas. Rules based on this descriptor can be used for accurate, robust, and computationally efficient secondary structure assignment that correlates well with the existing methods.