Development of Complex Curricula for Molecular Bionics and Infobionics Programs within a consortial* framework**

Consortium leader

PETER PAZMANY CATHOLIC UNIVERSITY

Consortium members

SEMMELWEIS UNIVERSITY, DIALOG CAMPUS PUBLISHER

The Project has been realised with the support of the European Union and has been co-financed by the European Social Fund ***

**Molekuláris bionika és Infobionika Szakok tananyagának komplex fejlesztése konzorciumi keretben

***A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

INTRODUCTION TO BIOINFORMATICS

(BEVEZETÉS A BIOINFORMATIKÁBA)

CHAPTER 2

Knowledge representation and core data-types

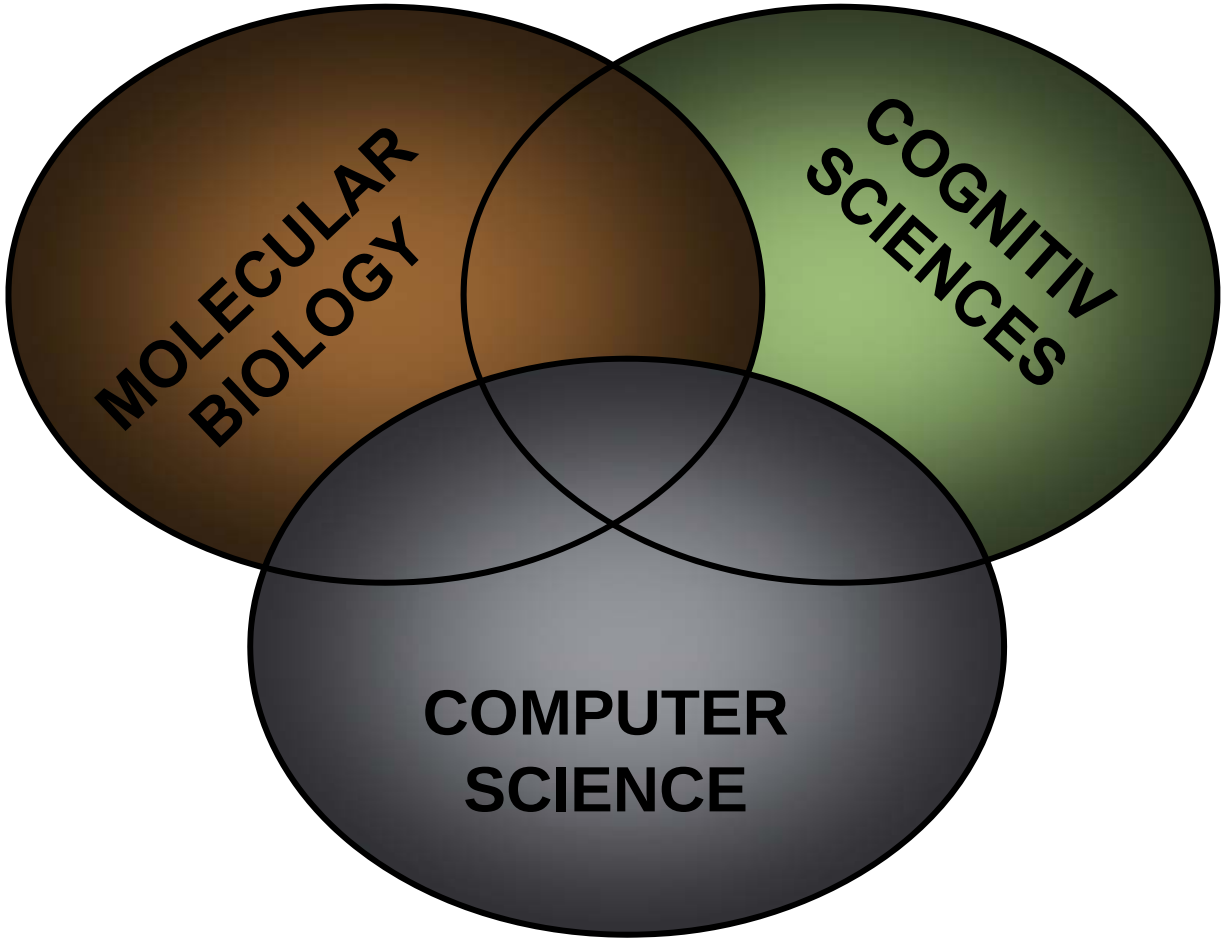
(Ismeretábrázolás és alapvető adattípusok)

Sándor Pongor

What will we speak about?

Core elements. Systems theory of biological knowledge representation. Core data-types: sequences, 3D-structures, networks, texts + database records as a summary.

Bioinformatics is interdisciplinary



What is particular in bioinformatics?

The variety of objects: molecular structures, metabolic pathways, regulatory networks AND their databases

A few methods: analysis and use of similarity;

Complexity of biological knowledge

(and NOT so much the quantity of data...)

The same molecule has many different representations

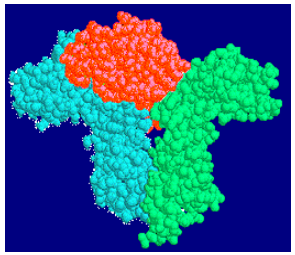
MARTKQTARK
STGGKAPRKQ
LATKAARKSA

Sequences

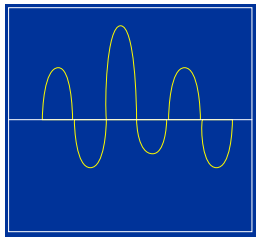
CIPKWNRCGPKMDGVPCCEPYTCTSDYYGNCS

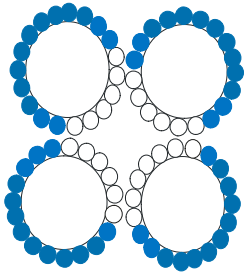
Extended sequences (pl. disulfide topology)

Cartoons of domains or secondary structures

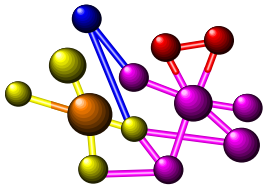


3D structures





Symbolic diagrams (e.g. hydrophobicity plots, helical circle diagrams)



Simplified 3D cartoons

Systems theory, structure and function

Structure and function are concepts of systems theory

Viennese biologist Ludwig von Bertalanffy founded general systems theory to explain commonalities of biological, environmental phenomena.

Ludwig von Bertalanffy

(1901-1972)

It is now used in many fields (social systems, company organization, military).

Advantage: Qualitative explanations, generalization power, abstraction

Disadvantage: Contains little mathematical or quantitative foundations

What are systems?

Any part of reality that can be ~separated from the environment (by a boundary). A community in an environment.

Consist of interacting parts

Interact with the environment (inputs, outputs)

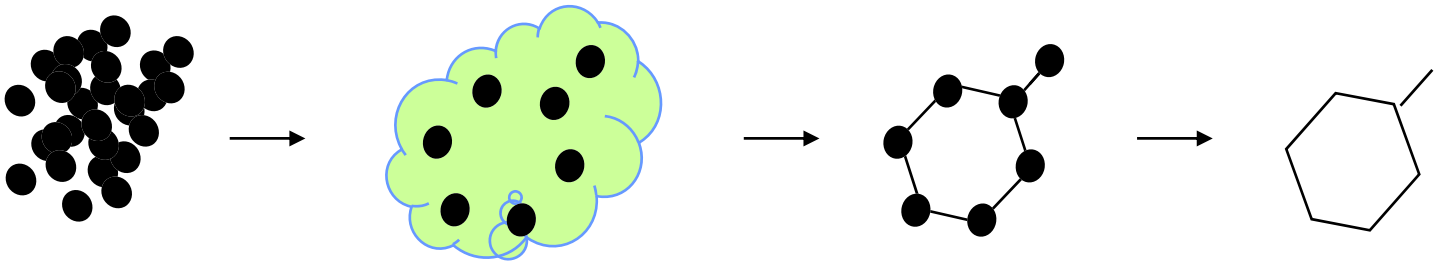
System models are generalizations of reality

Have a structure that is defined by parts and processes

Parts have functional as well as structural relationships between each other.

Systems theory explains the variety of molecular descriptions

Abstract example:



A system of moving particles

Populated positions and a boundary

Structure: Entities and relationships

Form

General definitions for structure and function

Structure is a ~constant spatio-temporal arrangement of elements or properties.

A molecular structure is a subset of this: a constant (spatio-temporal) arrangement of elements (e.g. atoms) and relationships (e.g. bonds)

Substructure: A part of a structure

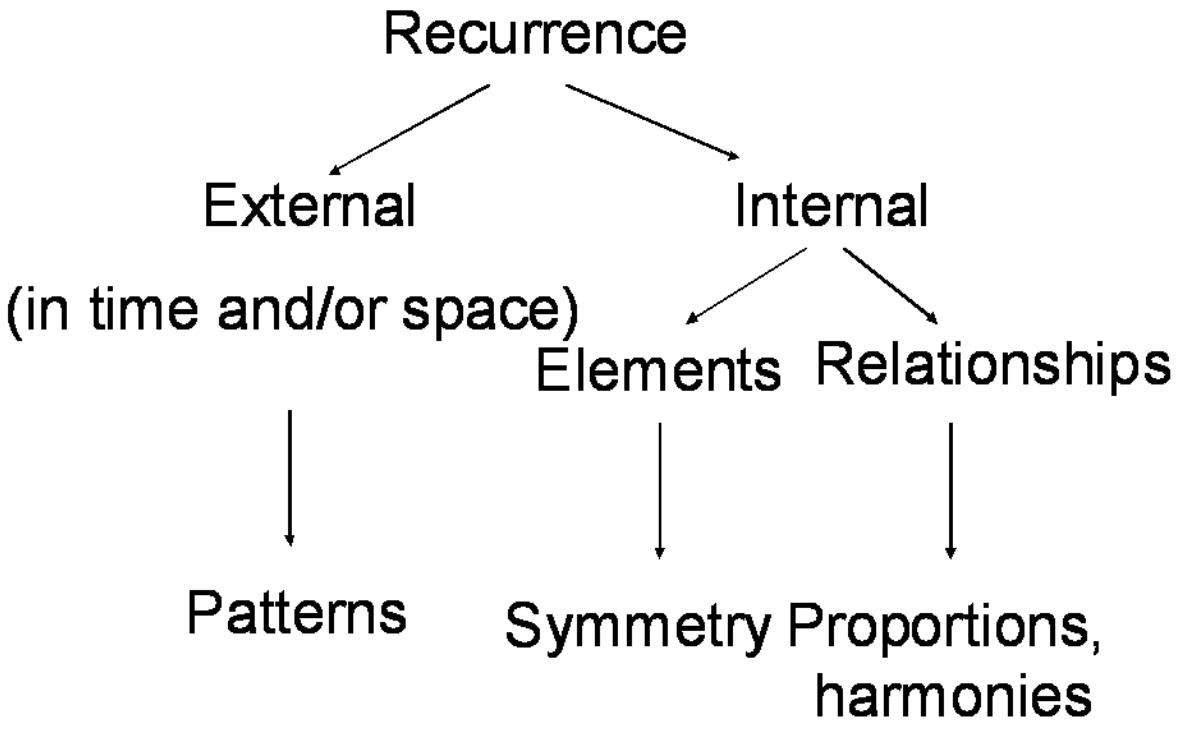
Function is a role played within a system.

A system's function is its role played within a higher system (hierarchical description)

Systems explain various phenomena as repetition (recurrence)

External repetition: same substructures in different systems

Internal repetition: same substructure within the same system



SYSTEM EXAMPLES	Entities	Relationships
a) General examples		
Molecules	Atoms	Atomic interactions (chemical bonds)
Assemblies	Proteins, DNA	Molecular contacts
Metabolic Pathways	Enzymes	Chemical reactions (substrates/products)
Genetic networks	Genes	Co-regulation
b) Examples for proteins		
Protein sequence	Amino acid	Sequential vicinity
Protein structure	Atoms	Chemical bonds
Protein structure (simplified)	Secondary structures	Sequential and topological vicinity
Backbone structure (Fold)	C_α atoms	Peptide bond

Core data-types

A very large number of description can be built from the various entities and relationships. We select a few of them.

Biological sequences (character strings built from amino acid alphabet [20 letters] or nucleotides [4 letters]

3D structures (atoms with x,y,z coordinates, chemical bonds)

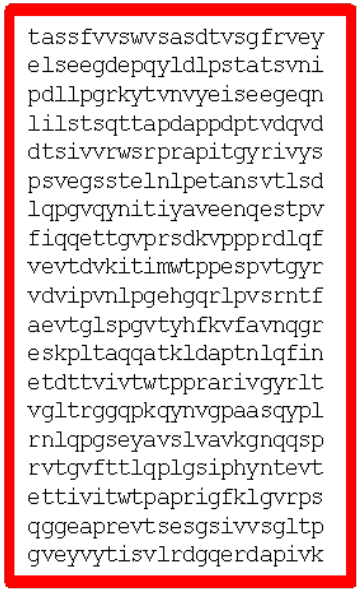
Networks (generalized descriptions, e.g. node can be a gene, edge can be regulatory link)

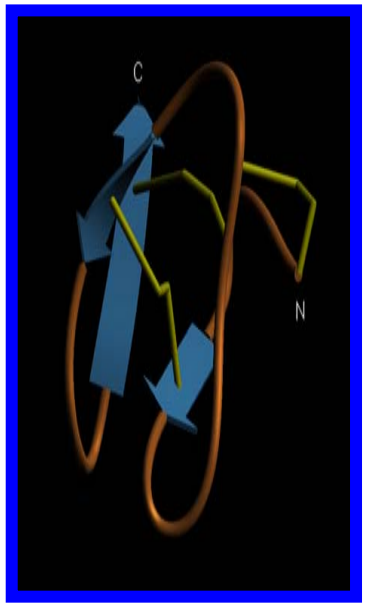
Texts (e.g. PubMed abstracts)

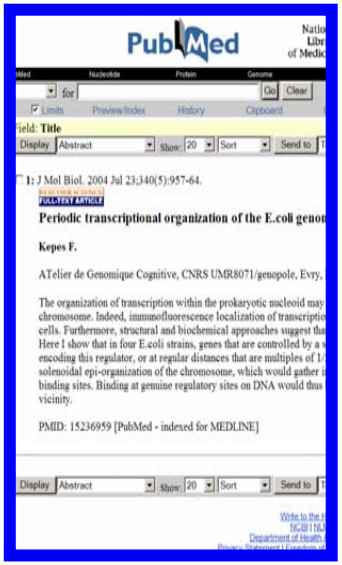
Database records

We discuss them as a standard way to store the core data

Core data-types







SEQUENCES **3-D** **NETWORKS** **TEXT**

BIOLOGICAL SEQUENCES

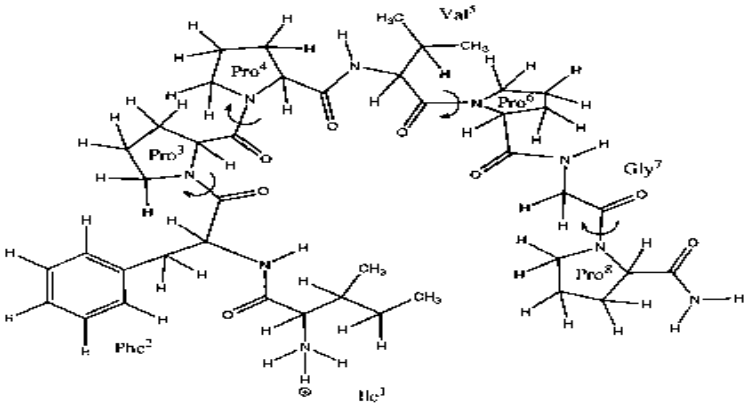
Biological sequences (character strings built from amino acid alphabet [20 letters] or nucleotides [4 letters]

SEQUENCES

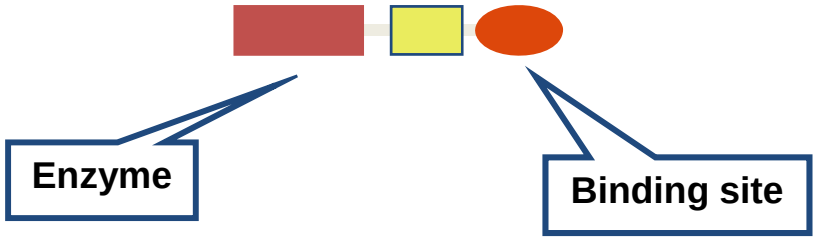
Model: Chemical structure of proteins (far too complicated for large molecules)

Description: Character strings. Characters denote amino acids. (relations – sequential vicinity – are implicit!)

Simplified and/or extended (annotated) forms of visualization



IFPPVPGP



Biological sequences as language

```
qfinetdttvivtwtpprarivg
yrltvgllseegdepqyldlpst
atsvnipdllpgrkytvnvyeis
eegeqnlilstsqttapdappdp
tvdqvddtsivvrwsrprapitg
yrivyspsvegsstelnlpetan
svtlsdlqpgvqynitiyaveen
qestpvfiqqettgvprsdkvpp
prdlqfvevtdvkitimwtppes
pvtgyrvdvipvnlpgehgqrlp
vsrntfaevtglspgvtyhfkvf
avnqgreskpltaqqatkldapt
nlqfinetdttvivtwtpprari
vgyrltvgltrggqpkqynvgpa
asqyplrnlqpgseyavslvavk
gnqqsprvtgvfttlqplgsiph
yntevtettivitwtpaprigfk
lgvrpsqggeaprevtsesgsiv
vsgltpgveyvytisvlrdgger
```

- Sequences are like texts written in an unknown language
- Imperfect analogies to human language and coded messages - we can talk about a “language metaphor”
- Analysis tools (exact and approximate string matching ([=alignment]) were originally developed for texts
- Theory of computer languages (Chomsky) can be applied to biological sequences

3D STRUCTURES

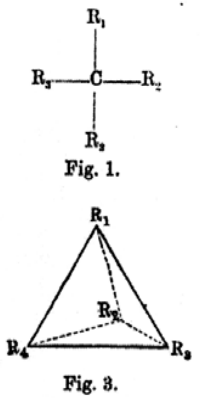
3D structures are atoms with x,y,z coordinates, chemical bonds. For macromolecules we typically simplify them into larger blocks, backbone or surface representations…

Chimie dans l'espace

Dutch chemist (Nobel prize 1902) discovered that some phenomena in chemistry need a 3D description. Before that we had no idea of 3D nature of molecules.

Object metaphore

The analogies with objects (collisions, movements, no overlap in space) is obvious but imperfect. Nevertheless it profoundly influences our thinking about atoms.

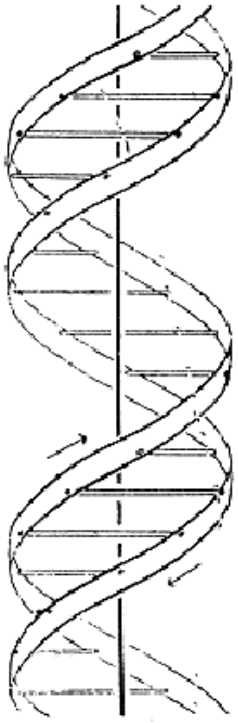


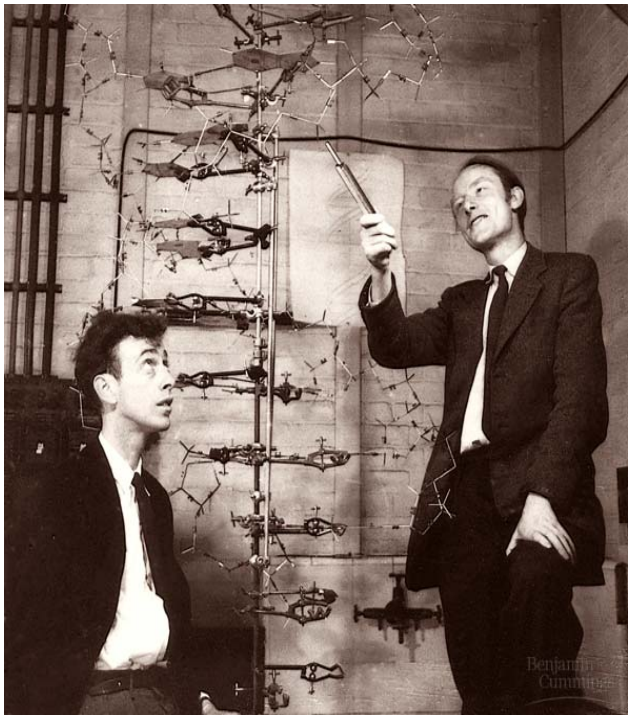
Van t'Hoff 1898

1852-1911

Macromolecules are so complex that only their simplified view make visual sense

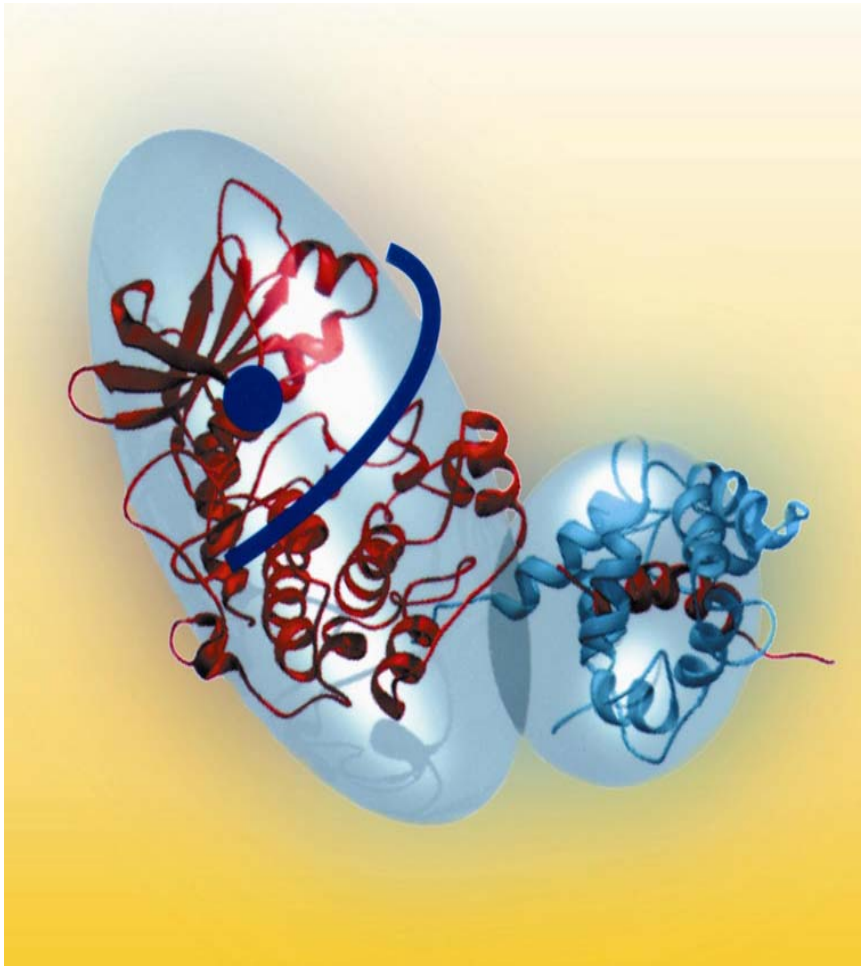
The double spiral was shown in a simplified form already in the first, epoch-making publication.





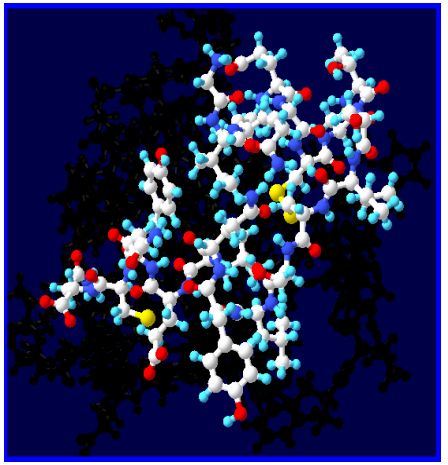
…"This figure is purely diagrammatic. The two ribbons symbolize the the phosphate-sugar chains, and the horizontal rods the pairs of the bases holding the chains together. The vertical line marks the fibre axis" Watson, Crick, 1953

Molecular models today are more an art then science. There are extablished methods of visualization for macromoleculs (backbones, surfaces, color codes etc)



3D structures

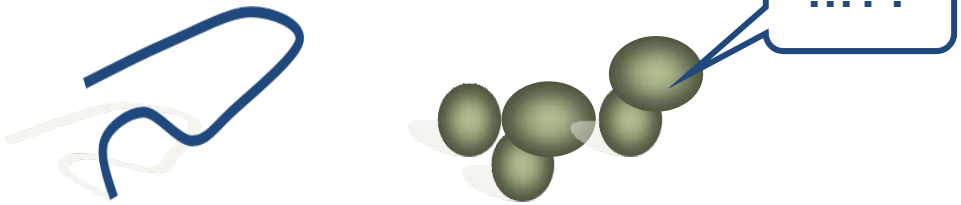
Model: 3D chemical structures



Description: 3D coordinates

$(x_i, y_i, z_i)_n$

Simplified and/or extended (annotated) visualization



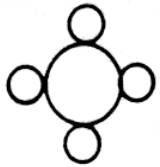
Backbone (main chain) Surface

NETWORKS

are the most generalized entity-relationship models, applicable to any system (e.g. node can be a gene, edge can be regulatory link). Strong analogies with mathematical graphs, week analogies with social systems (“social metaphore”).

Small molecules – classical graphs

The first network models were the chemical formulas applied in the 19th century. Much of early graph theory was inspired by chemical formulas…



Loschmidt, 1861

Kekulé, 1865

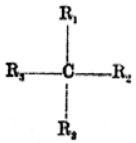


Fig. 1.

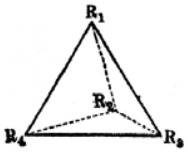
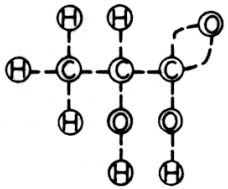
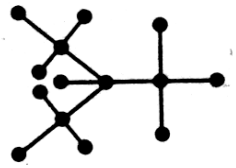


Fig. 3.

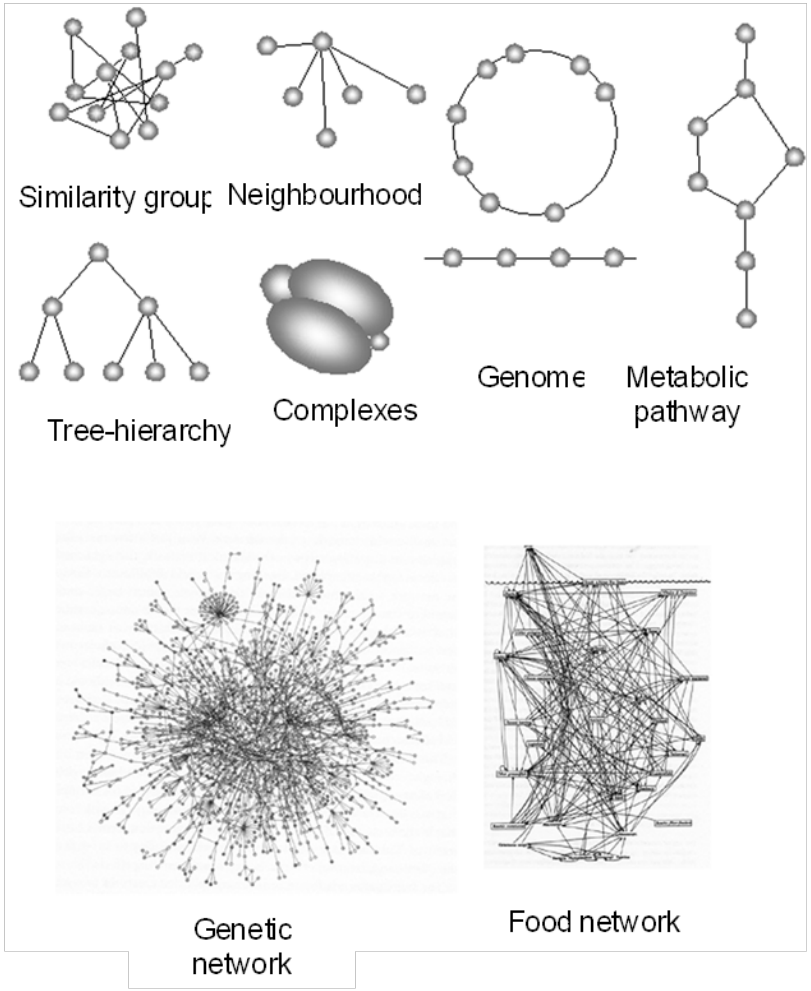
Van ’t Hoff, 1898



Crum Brown, 1861



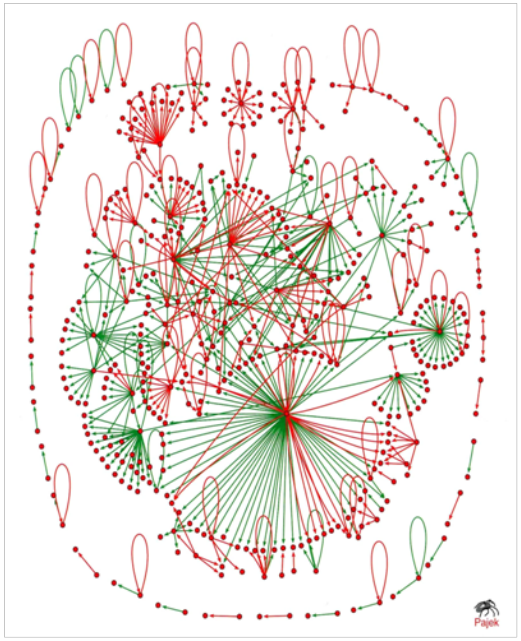
Cayley, 1872



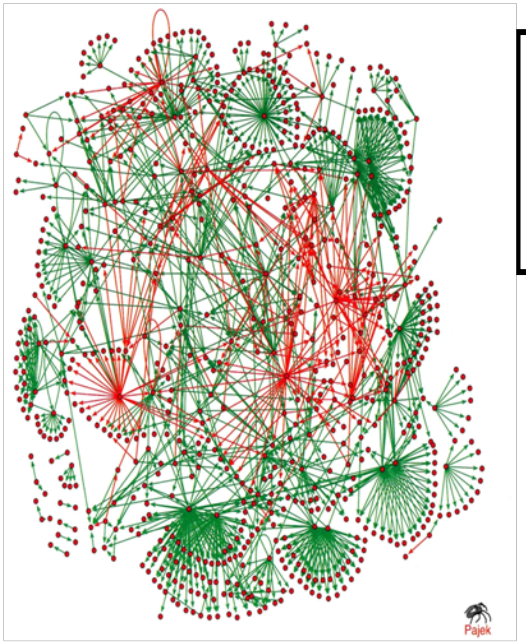
Networks of genomes

Today we employ networks to all biological problems, from the molecular (top left) to the ecological level (bottom right is a food network with species as nodes and predator/prey relations as edges).

The transcription regulatory networks have genes as nodes and up and down regulatory relations as edges.



E. Coli bacterium



Yeast

\+ (up)

\- (down)

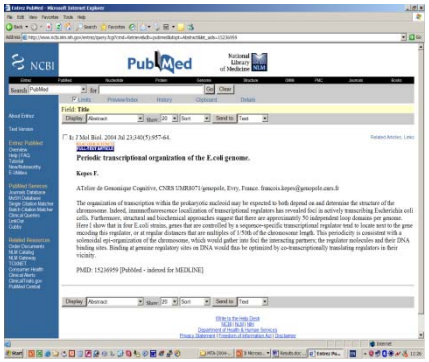
TEXTS (article abstracts in PubMed)

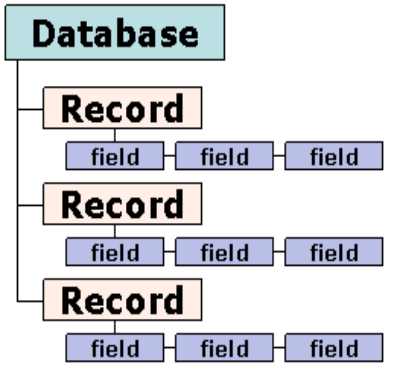
Scientific texts are written in human language. They contain encoded annotations (abbreviated citations, postal addresses etc) and specific language (molecular names, chemical formulas etc). Strong analogies with human semantics.

Scientific texts have a strict or close to strict structure, similar to database records. The meaning of scientific texts is at present not machine-readable. Auxiliary informations (author and journal names, or annotations such as keywords) are machine readable

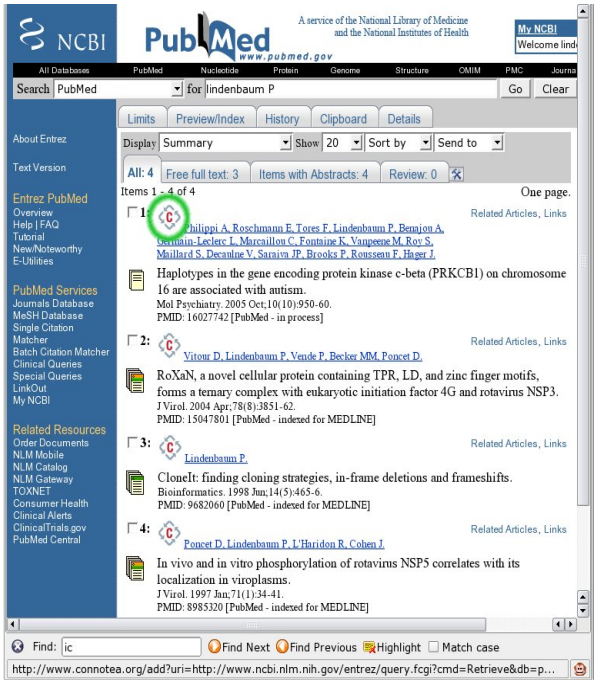
- Keyword collecttions, onthologies, etc.

Model: ?? (none)



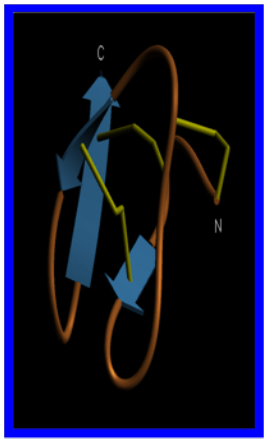


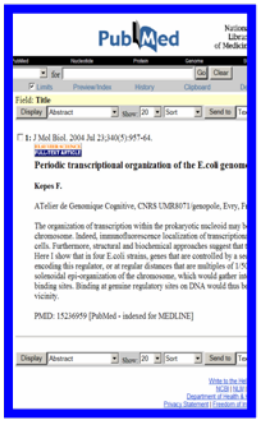
Description: structured files (records, fields), standardized language

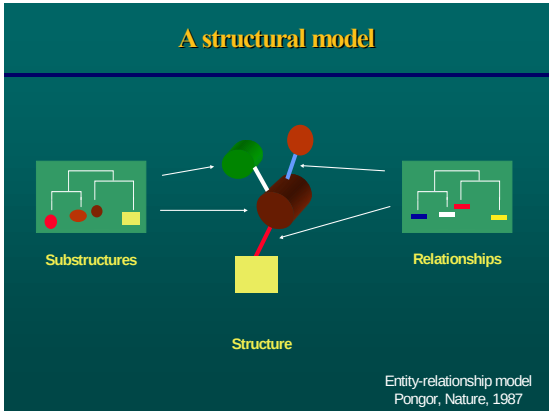


Simplified and/or extended visualization

Introduction to bioinformatics: Core data-types



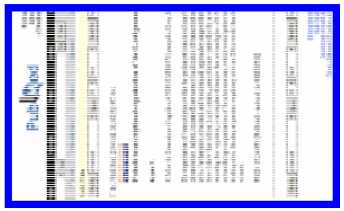




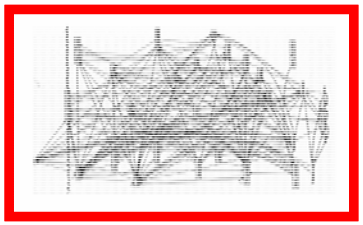
SEQUENCES **3-D** **NETWORKS** **TEXT**

The core data-types are all entity-relationship descriptions. The entities and relationships have to be formally defined, either as concept hierarchies (simplified) or as ontologies that contain descriptions + rules.

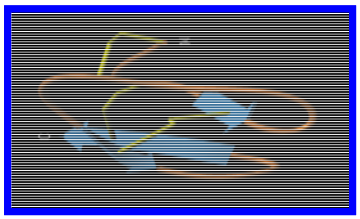
Putting the core-data into database records



TEXT

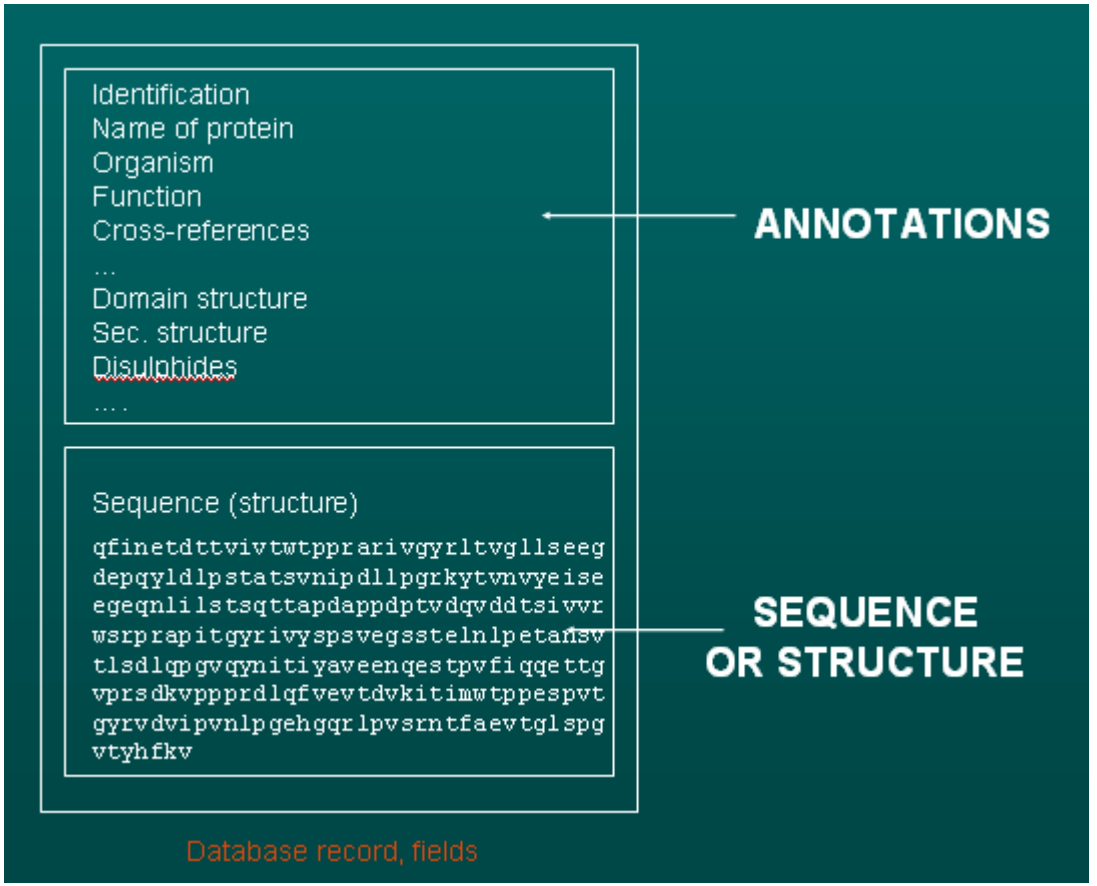


NETWORKS

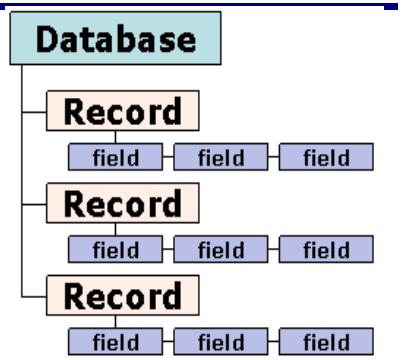


3-D

SEQUENCES



DATABASE RECORD



Biomolecular databases in a nutshell

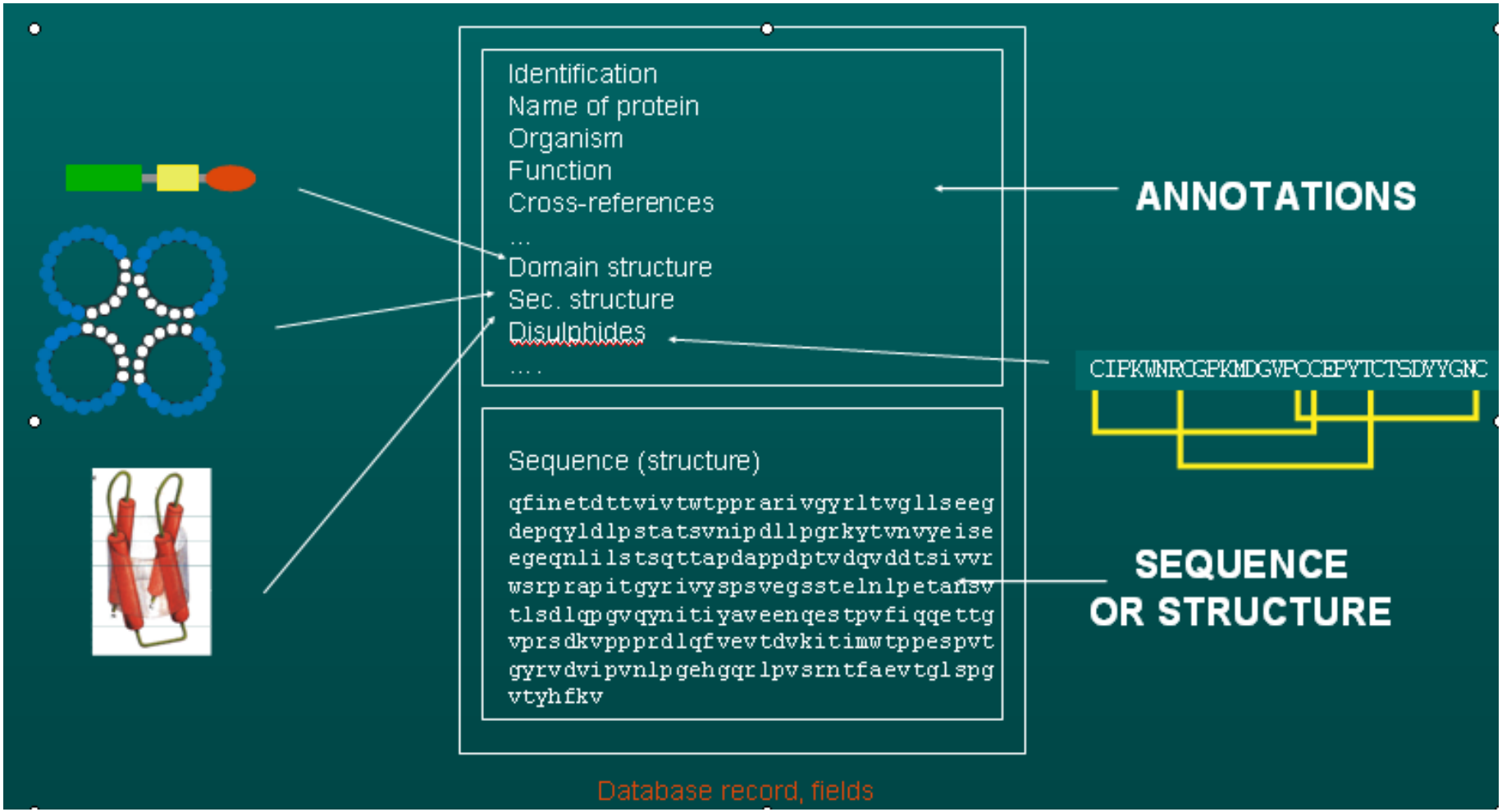
They contain one molecule in a record. Sequence databases are the most developed.

The main part of the record is the structural description which is typically a sequence or a structure.

In addition they contain an annotation part which is a collection of various informations, functional descriptions, crossreferences, and also structural descriptions (info assigned to parts of the structure. So annotation duplicates certain aspects of the molecules.

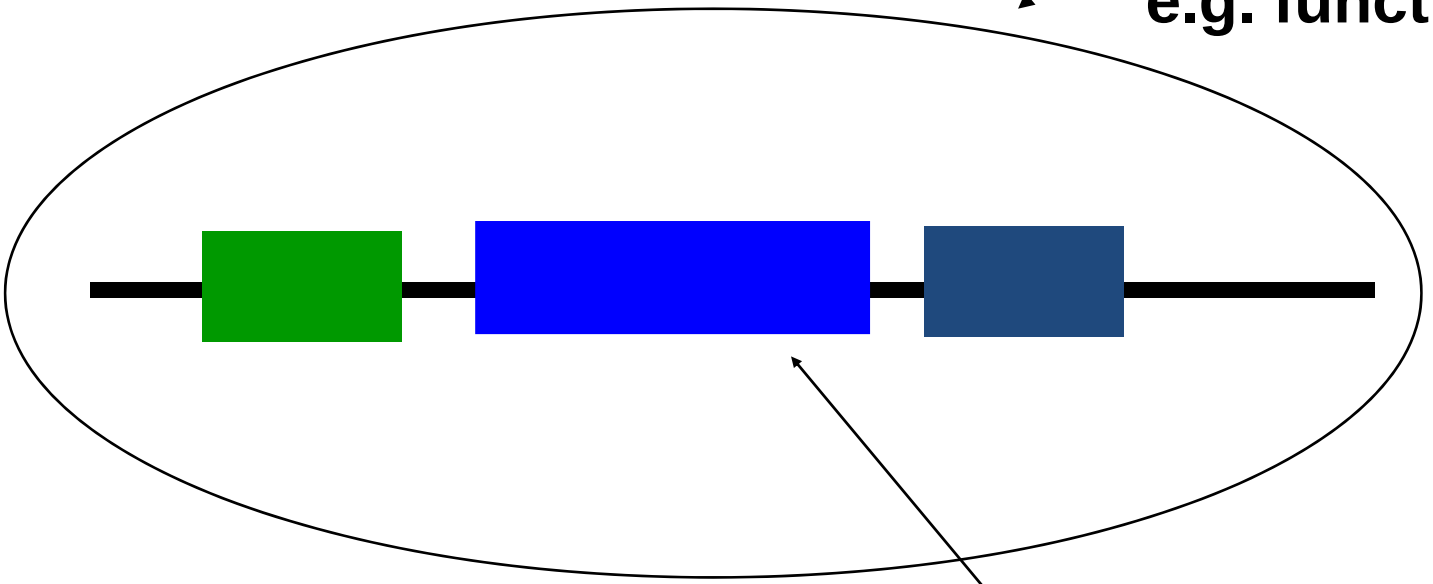
As a result, a sequence database is a complex object that can be handled with dedicated programs (parsers).

Introduction to bioinformatics: Core data-types



Annotation of (sequence) data means assigning global and local descriptors to a molecule

Global descriptors e.g. function



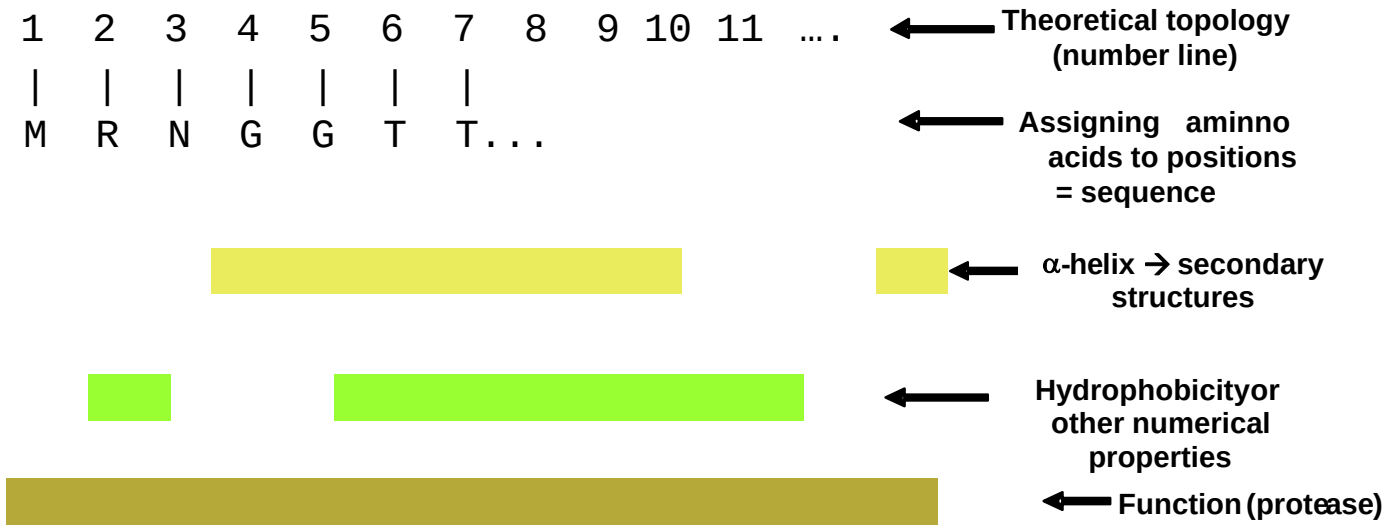
Annotation requires database searching and knowledge of „biology” (chemistry, medicine..)

Local descriptors e.g. binding sites, domains

Generalized annotation

If we take a theoretical topology, the number line, and assign amino acids to it, we obtain sequence.

We can carry on assigning local descriptors or global descriptors and we end up creating a database-record of a structure.This is a database-centric view of a structure.



CORE DATA-TYPES OF BIOINFORMATICS

Molecular structure is a model, an abstract, mental representation that can be described with the tools of systems theory

Concepts of system, structure, function. Structure is an ensemble of elements and relations.

4 core data-types (models): sequence, 3D, network and text

Models are represented by computers with dedicated data-structures, images and/or in a narrative form.

Simplified and extended (annotated) descriptions.

Database records contain a core data-types in machine-readable form and annotations in mostly human-readable forms.